AD_____

Award Number: DAMD17-03-1-0186

TITLE: Design and Clinical Efficacy of a Computer-Aided
         Detection Tool for Masses in Mammograms

PRINCIPAL INVESTIGATOR: Swatee Singh
                          Dr. Joseph Lo

CONTRACTING ORGANIZATION: Duke University Medical Center
                            Durham, NC   27710

REPORT DATE: June 2005

TYPE OF REPORT: Annual Summary

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

DISTRIBUTION STATEMENT: Approved for Public Release;
                        Distribution Unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 01-06-2005	2. REPORT TYPE Annual Summary	3. DATES COVERED *(From - To)* 15 May 2003 – 14 May 2005

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Design and Clinical Efficacy of a Computer-Aided Detection Tool for Masses in Mammograms	
	5b. GRANT NUMBER DAMD17-03-1-0186
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Swatee Singh	
Dr. Joseph Lo	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
E-Mail: swatee.singh@duke.edu	

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Duke University Medical Center Durham, NC 27710	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. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Abstract follows.

15. SUBJECT TERMS
Computer-aided detection, mammography, statistical pattern recognition, classification, feature extraction, FROC

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	UU	108	19b. TELEPHONE NUMBER *(include area code)*

ABSTRACT

Our hypothesis is that a highly sensitive and highly specific CAD scheme, incorporating unique preprocessing techniques and advanced Decision Theory methods, can detect masses and improve the performance of mammographers. To test this hypothesis, we propose to construct a CAD system from two key components: 1) a highly sensitive mass detector, and 2) statistical models designed to reduce false-positives. We feel that it is essential to develop a tool that can identify a high percentage of masses, both spiculated and nonspiculated. It is important for computerized tools to detect as many masses as possible, but not to detect too many regions that are not actual masses. Thus, our program will first concentrate on finding many suspicious regions. Once suspicious regions are identified within the mammogram, we will explore several classification techniques to determine whether the regions are actually masses or some other structure in the breast. The techniques we plan to explore, for both detecting masses and classifying them, include standard, well-known techniques as well as new and novel approaches.

Table of Contents

INTRODUCTION

The most effective early-detection tool for breast cancer currently is screening mammography. To provide a reliable and efficient second-reader to aid mammographers, research has been directed towards developing computer-aided detection (CAD) tools. Although these tools have shown promise in identifying calcifications, detecting masses has proven relatively more difficult.

For this study, we proposed that a highly sensitive and highly specific CAD scheme, incorporating unique preprocessing techniques and advanced Decision Theory methods, could detect masses and improve the performance of mammographers. The proposed CAD system has two key components: 1) a highly sensitive mass detector, and 2) statistical models designed to reduce false-positives.

This pre-doctoral fellowship covers two different students – David Catarious, mentored by Dr. Carey Floyd, and Swatee Singh mentored by Dr. Joseph Lo. It was originally awarded to David Catarious, who graduated in August 2004 from Duke University with his Ph.D. *A Computer-Aided Detection System for Mammographic Masses* (reportable outcome #12) and is now working as a Congressional Science Fellow (outcome #13). Parts of the original aims were concluded as part of that dissertation research. The Army authorized the transfer of the remaining fellowship to Swatee Singh in November of 2004. The progress for the 2003-04 first year of this fellowship is summarized in the report below.

BODY

Task 1: Develop and test unique pre-processing techniques on mammographic regions of interest (ROIs)

In the search for an effective method to both enhance possible masses and reduce the influence of anatomical noise contributed by the background structure, four preprocessing techniques were explored: unsharp masking; local histogram equalization; local region standardization; and combining the previous three techniques with principal component analysis. Among the first three preprocessing techniques, unsharp masking was seen to give best performance. Combining these three techniques was found to have no advantage when compared to using just unsharp masking alone. Hence, it was decided that unsharp masking would be used to compensate for background nonuniformity. This task was concluded on schedule.

Task 2: Develop and test the initial mass detection algorithm on preprocessed ROIs from task 1

After preprocessing, the images were searched for potential masses with a Difference of Gaussians (DOG) filter. Three parameters needed to be specified: the size of the filter template and the two standard deviations of the constituent Gaussians, σ_1 and σ_2. To gather the data required for this task, a total of thirty DOG filters were employed as detection filters over the study database. The study database consisted of 181 CC-view lumisys-scanned mammograms in which over 67,000 potentially suspicious regions are identified. The influence of each of the three parameters of the DOG filter on the following performance measures was studied:

1) <u>Various Properties of the detected regions:</u> It was found that as the size of the filter template increases, the values of the peak response to the DOG filter decrease. Of the 24 features that demonstrate statistically significant differences in mean value across the values of σ_1 and σ_2, only 6 effectively differ: area, major axis length, peak output of the DOG filter, correlation, entropy, and information measure of correlation one. Each of these features increases with both σ_1 and σ_2.

2) <u>The number of true and false positive regions detected:</u>

4

Parameter	Parameter Value	Average Sensitivity (%)	Average FPpl
Template Size (mm^2)	48	96.04	11.75
	56	97.5	11.44
	64	97.5	11.29
σ_1 (mm)	4	97.92	14.34
	5	97.22	10.91
	6	96.18	8.79
	7	94.44	7.26
σ_2 (mm)	5	97.62	17.61
	6	97.92	12.33
	7	96.53	9.28
	8	94.44	7.26

Table 1: The effect of parameters on the malignant sensitivities and false positives per image. For larger filter sizes (values of sigma's), the number of false positives decreases.

As the template size increases from 48 to 64 mm^2, both the initial sensitivity and number of false positives per image (FPpl) slightly improve. However, for σ_1 and σ_2, sensitivity and FPpl demonstrate an inverse relationship. As σ_1 increases, almost half the false positives are eliminated at a relatively low cost of 3.5% reduction in sensitivity; σ_2 eliminates approximately 60% of false positives for the same small decrease in sensitivity.

In order to achieve a compromise among the sensitivity, false positive rate, and the classification performance of the best features, the optimal combination of parameters may be a medium-sized filter template, a small σ_1, and a large σ_2. The medium range of template sizes can achieve slight increases in the false positive rate at no cost in sensitivity. Since increasing σ_2 more aggressively eliminates false positives, a compromise between filter sensitivity and false positive rate could be achieved by mixing a large σ_2 with a small σ_1. Based on this study, it was concluded that the DOG filter employed for our system would be constructed of Gaussians with standard deviations of 4.4 mm and 9 mm (22 and 45 pixels). The size of the DOG filter template was set as a square of side 54 mm (270 by 270 pixels). This approach was also applied to detection of lung nodules in chest radiography, resulting in co-authorship for the student fellow in 3 proceedings papers at SPIE, the primary scientific conference for medical imaging (reportable outcomes #1, 7, and 11). This task was concluded on schedule.

Task 3: Identify potential masses from ROIs

To distinguish suspicious regions from the rest of the image, we employed a multilevel thresholding technique similar to that used by previous researchers (1-3). A set of thresholds was defined based on the gray level histogram of the filtered image and for each level a new image containing suspicious regions was created. To combine these images, the "duration" of the regions was calculated. Although the duration image technique can accurately identify the most suspicious regions in the image, the segmentations of the masses do not reflect the detailed morphology of the mass. The inaccuracy of the method arises mainly because the object borders are determined from the filtered images, not the original images.

As such, in the final version of the CAD system, the duration image technique has been replaced by a segmentation routine using an iterative, gray level, linear segmentation procedure. This modified procedure begins by examining a ROI that has been identified by the CAD system as containing a suspicious region. Unsharp masking is applied to the ROI to compensate for background nonuniformity. The procedure then iterates by estimating the pixels interior and exterior to the object, determining an optimum gray level threshold to separate the

interior and exterior pixels, and constraining the resulting object border. The procedure halts when a stopping criterion has been achieved. A comparison of the developed segmentation algorithm to the previous segmentation procedure is shown in the figure below.
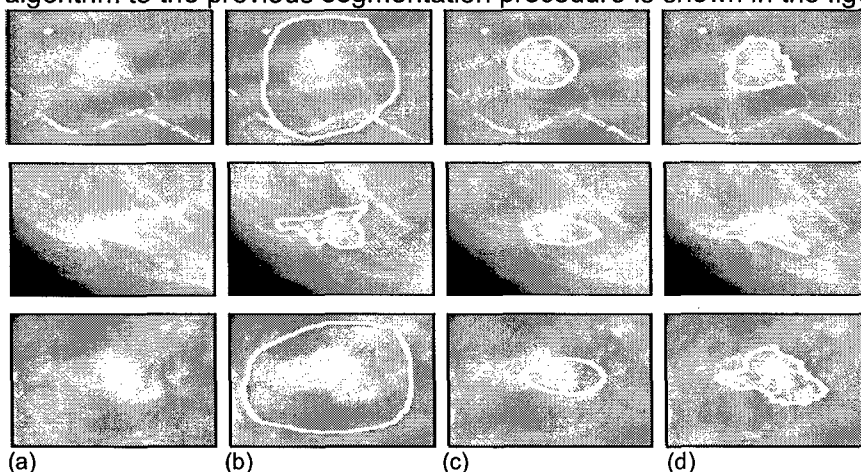

(a)   (b)   (c)   (d)

Fig 1:(a) Three masses (two malignant, one benign) extracted from the original mammographic image, (b) the outline of the mass provided by the DDSM, (c) the segmentation provided by the duration image technique, and (d) the segmentations computed with the new segmentation routine.

The results for this task have been published in SPIE (reportable outcomes #9) and *Medical Physics* (#8 as first author and #11 as co-author), the primary peer-review scientific journal in the field. This task was completed on schedule.

Task 4: Extract shape and textural features from potential masses

For this task, thirteen morphological features were extracted from each suspicious region – area, major axis length, minor axis length, eccentricity, area of the convex hull, equivalent diameter, solidity, extent, and circularity. Also as part of morphological features, the mean, peak, standard deviation, and value at the region's centroid were extracted from the regions' response to the DOG filter. From each suspicious region, six features were extracted that describe the region's boundary. Each of the boundary features is derived from the regions' normalized radial lengths (4): mean, standard deviation, entropy, area ratio, zero crossing count, and range. The remaining fifteen features describe the textural properties of the identified suspicious regions: contrast, average radial gray level change, and the thirteen Haralick et all (5) features. In total we extracted 34 features for each of the approximately 9000 potentially suspicious regions in 1,413 images. This work was the basis for a first-author proceedings paper at SPIE (reportable outcome #6). This aim was accomplished on schedule.

For the development of the classification stage of the CAD system in the next task, we studied numerous rules that would designate which of the suspicious regions correspond to the true positives in the images. Based on this study, we adopted a rule that required the distance between centroids of the true positive and suspicious region to be within 16 mm of each other and an area of overlap between the two regions of 9%.

Task 5.1: Examine linear classification techniques on features extracted to separate masses from non-masses

To reduce the complexity of the predictive models and avoid overtraining problems, we needed a way to reduce the number of features that would need to be computed for determination of the malignancy of a mass. In the stepwise feature selection algorithm implemented for this research, a Fisher's linear discriminant is employed for the internal classifier. After examining different variants of the linear discriminants, we found that the

Fisher's linear discriminant is easily implemented as an iterative process. Hence it was deemed an ideal choice to calculate the threshold values. Several methods to train the discriminant function were explored – such as K-fold cross validation, and resubstitution. For the figure of merit (FOM) four choices were examined – AUC, AUCp, minFPF, and the Mahalanobis distance between mean values of the decision variables in one class compared to the other.
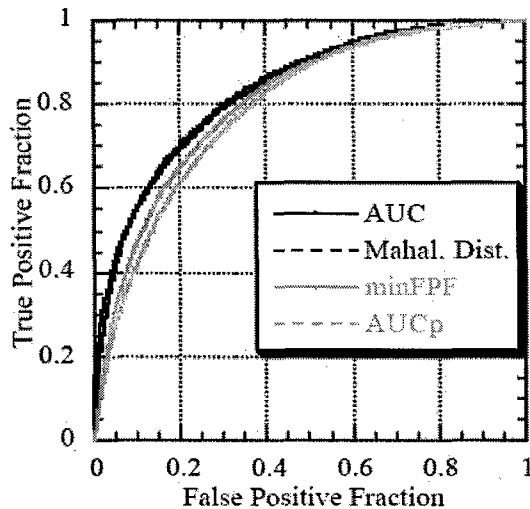

The results for the resubstitution method for the four FOMs are shown in the ROC curves of figure 2. It was found that for all the FOMs, there is no difference between the curves created using the different training algorithms. Also, there was little difference in performance of the two training methods.

Other linear classifiers were explored and their results have been published in 4 co-authored SPIE proceedings papers (reportable outcomes #2, 3, 4, 5).

Fig 2: The average ROC curves over 28 trials for each FOM with training method resubstitution

KEY RESEARCH ACCOMPLISHMENTS

- Optimized parameters of a Difference of Gaussians filter for initial detection of potentially suspicious regions in digitized mammograms, resulting in a final filter that maintains high sensitivity while improving specificity substantially.
- Applied the optimized filter to 1,413 digitized images from the Digital Database for Screening Mammography and identified approximately 9000 potentially suspicious regions.
- Extracted a set of 34 possible cancer descriptors for each of the 9000 potentially suspicious regions, including morphological, boundary, and texture features.
- Examined various linear discriminants and implemented an iterative linear discriminant for our CAD system which merges the extracted features to predict whether the suspicious region contains an actual mass or a false positive.

REPORTABLE OUTCOMES

1. Abbey CK, Eckstein MP, Shimozaki SS, Baydush AH, **Catarious DM**, Jr, Floyd CE, Jr. Human-observer templates for detection of a simulated lesion in mammographic images. In: SPIE Medical Imaging 2002. San Diego, CA, 2002.

2. Baydush AH, **Catarious DM**, Abbey CK, Floyd CE, Jr. Computer Aided Detection of Masses in Mammography using Sub-region Hotelling Observers. Medical Physics 2003; 30:1781-1787.

3. Baydush AH, **Catarious DM**, Floyd CE, Jr. Computer aided detection of masses in mammography using a Laguerre-Gauss channelized Hotelling observer. In: Medical Imaging 2003: Image Perception, 2003.

4. Baydush AH, **Catarious DMJ**. Novel use of the Hotelling Observer for computer-aided diagnosis of solitary pulmonary nodules. In: SPIE Medical Imaging 2001: Image Processing. San Diego, CA, 2001; 1918.

5. Baydush AH, **Catarious DM**, and Floyd CE, Jr. Incorporation of Laguerre-Gauss channelized Hotelling observer into a mammographic mass CAD system. In:

International Workshop on Digital Mammography (2004).

6. **Catarious DM**, Baydush AH, Abbey CK, Floyd CE, Jr. A Mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results. In: Hanson K, ed. Proc. SPIE Int. Soc. Opt. Eng. San Diego, CA, 2003; 111.

7. **Catarious DM**, Jr, Baydush AH, Floyd CE, Jr. Initial development of a computer-aided diagnosis tool for solitary pulmonary nodules. In: SPIE Medical Imaging 2001. San Diego, CA, 2001; 710.

8. **Catarious DM**, Baydush AH, Floyd CE, Jr. Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system. Medical Physics 2004.

9. **Catarious DM**, Jr, Baydush AH, Floyd CE, Jr. Development and application of a segmentation routine in a mammographic mass CAD system. In: SPIE Medical Imaging 2004. San Diego, CA, 2004; 801.

10. Samei E, **Catarious DM**, Jr, Baydush AH, Floyd CE, Jr, Vargas-Voracek R. Bi-plane correlation imaging for improved detection of lung nodules. Import. SPIE Int. Soc. Opt. Eng. San Diego, CA, 2003; 284-297.

11. Tourassi GD, Vargas-Voracek R, **Catarious DM**, Jr. Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. Medical Physics 2003; 30:2123-2130.

12. Ph.D. thesis: **David Catarious**, *A computer-Aided Detection System For Mammographic Masses*, Department of Biomedical Engineering, Duke University, 2004, (available upon request).

13. Job Placement: **David Catarious**, Congressional Science Fellow with The Honorable Edward J. Markey, U.S. House of Representatives.

CONCLUSIONS

We developed the high sensitivity first stage of our CAD system to identify suspicious regions most indicative of malignancy. Pre-processing techniques were used to help identify potential masses from ROIs. A large study was performed to determine optimal DOG parameters to maximize results. We also extracted thirteen textural, morphological, and boundary descriptors of these suspicious regions as mathematical descriptors of the properties of these suspicious regions. Various linear classification techniques were employed to determine the optimum classifier. This project has built the framework for the second stage of the CAD system (to be reported next year) – the high specificity stage that will then reduce the number of false positives per image while maintaining nearly all of actual malignancies.

REFERENCES

1. Carreira MJ, Cabello D, Penedo MG, Mosquera A. Computer-aided diagnoses: Automatic detection of lung nodules. Medical Physics 1998; 25:1998-2006.

2. Giger ML, Doi K, MacMahon H. Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Medical Physics 1988; 15:158-166.

3. Giger ML, Doi K, MacMahon H, Metz CE, Yin F. Pulmonary nodules: Computer-aided detection in digital chest images. RadioGraphics 1990; 10:41-51.

4. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Transactions on Medical Imaging 1993; 12:664-669.

5. Haralick RM. Texture features for image classification. IEEE Trans. Syst. Man Cybern 1973; SMC-3:610-621.

Appendix

Human-observer templates for detection of a simulated lesion in mammographic images

Craig K. Abbey[*]

Dept. of Biomedical Engineering, University of California, Davis, CA

Miguel P. Eckstein and Steven S. Shimozaki

Dept. of Psychology, University of California, Santa Barbara, CA

Alan H. Baydush, David M. Catarious, Carey E. Floyd,

Dept. of Radiology, Duke University, Durham, NC

ABSTRACT

We describe a probit regression approach for maximum-likelihood (ML) estimation of a linear observer template from human-observer data in two-alternative forced-choice experiments. Like a previous approach to ML estimation in this problem [Abbey & Eckstein, *Proc. SPIE*, Vol. 4324, 2001], our approach does not make any assumptions about the distribution of the images. The previous approach utilized a regularizing prior distribution to control the degrees of freedom in the problem. In this work, we constrain the observer template to be represented by a limited number of linear features. Standard methods of probit regression are described for estimating the feature weights, and hence the observer templates.

We have used this probit regression method to estimate human-observer templates for the detection of a small (5mm diameter) round simulated mass embedded in digitized mammograms. Our estimated templates for detecting the mass contain a band of heavily weighted spatial frequencies from 0.08 to 0.3 cycles/mm. We show comparisons between the human-observer template data, and the templates of a number of linear model observers that have been investigated as perceptual models of the human.

Keywords: Visual signal detection, model observer, observer template, classification image, forced-choice detection.

1. INTRODUCTION

The last few years have seen the development of new psychophysical techniques for examining visual strategy in noise-limited detection and discrimination tasks[1-7]. The basic idea behind these techniques is to utilize the images associated with correct responses and those associated with incorrect responses in order to estimate a linear observer template. As such, the estimated templates, also known as "classification images", can be used as a way to understand how observers perform visual tasks and as an alternative to comparisons of performance used to validate perceptual models.

Recently[5], we have described Maximum-Likelihood (ML) and Maximum-a-Posteriori (MAP) procedures for estimating observer templates from real clinical images, as opposed to computer-generated noise textures with a specified Gaussian distribution in two-alternative forced-choice (2AFC) experiments. Because of the large number of free parameters in an observer template (typically the number of parameters is equal to the number of pixels in an image), quadratic priors were used to regularize the template estimates. Here we adopt a somewhat different approach of constraining the template to be represented by a relatively small number of linear features. By using a limited number of features, the problem of finding the ML estimates of the observer template is reduced in degrees of freedom to the problem of finding the ML estimates of the feature weights. We approach this estimation problem through standard methods of probit

[*] Corresponding Author. ckabbey@ucdavis.edu, phone 530-754-9676; fax 530-754-5739; Dept. of Biomedical Engineering, 1 Shields Ave, Davis, CA 95616.

regression[8,9], which implicitly assume a Gaussian-distributed internal noise component in the observer's decision process.

We use the probit-regression approach described here to estimate human-observer templates for the task of detecting a simulated lesion in mammographic image backgrounds. Our image set consists of subregions drawn from a database of digitized mammograms. A total of five subjects have participated in 2AFC detection experiments to obtain psychophysical decision data from human observers. Human-observer templates are estimated using probit regression for a set of features that are defined by radial-frequency bands in Discrete Fourier-Transform (DFT) domain. These templates are compared to the templates of a number of proposed linear model observers. The model observers considered include nonprewhitening matched filter models, a prewhitening matched filter model, and implementations of two difference-of-Gaussian (DOG) channel models.

2. METHODS

2.A. The Linear Observer-Template Model

Here we briefly review the linear-template model for 2AFC detection tasks. A more complete treatment of this model in the context of estimating observer templates is given by Abbey and Eckstein[6,7].

In a 2AFC detection task, an observer is shown two images in each trial asked to identify the image that contains the signal. We will denote an image generically by the vector $\mathbf{g}$. We will refer to the signal present image as $\mathbf{g}^+$, and to the signal-absent image as $\mathbf{g}^-$. The linear template model assumes that the observer performs the 2AFC task by formulating an internal-response variable to each image,

$$\lambda^+ = \mathbf{w}^t \mathbf{g}^+ + \varepsilon^+$$
$$\lambda^- = \mathbf{w}^t \mathbf{g}^- + \varepsilon^-$$

(2.1)

where $\mathbf{w}$ is the observer template – a vector of linear weights – and ε is the observer's internal noise. In a given trial of a 2AFC experiment, the observer correctly identifies the signal-present image if $\lambda^+ > \lambda^-$, gets the trial incorrect otherwise. We define the trial score as 1 if the observer gets the trial correct, and zero otherwise (we assume continuous densities on the responses, and hence an equivocal decision, $\lambda^+ = \lambda^-$, is a zero-probability event).

We can define a variable, o_i, to be the score (the o indicates outcome) of the ith trial. If the observer gets this trial correct, then $o_i = 1$. Otherwise $o_i = 0$. Hence,

$$o_i = \text{step}\left(\lambda_i^+ - \lambda_i^-\right)$$
$$= \text{step}\left(\mathbf{w}^t \Delta \mathbf{g}_i + \Delta \varepsilon_i\right),$$

(2.2)

where $\Delta \mathbf{g}_i = \mathbf{g}_i^+ - \mathbf{g}_i^-$, and $\Delta \varepsilon_i = \varepsilon_i^+ - \varepsilon_i^-$. As defined, o_i is a Bernoulli random variable. If we can assume that $\Delta \varepsilon_i$ is a Gaussian-distributed random variable with zero mean, and a variance of $2\sigma_\varepsilon^2$, then the probability that $o_i = 1$, under the Gaussian assumptions on $\Delta \varepsilon_i$ is

$$p_i = \Phi\left(\frac{\mathbf{w}^t \Delta \mathbf{g}_i}{\sqrt{2}\sigma_\varepsilon}\right),$$

(2.3)

where Φ is the Gaussian cumulative density function. Note that the probability is invariant to a common scaling of $\mathbf{w}$ and σ_ε. Hence for the purposes of this work, we can fix the magnitude of the internal noise component to an arbitrary value of $\sigma_\varepsilon = 1$, yielding

$$p_i = \Phi\left(\frac{\mathbf{w}^t \Delta \mathbf{g}_i}{\sqrt{2}}\right).$$

(2.4)

The binary nature of a trial score and the definition of the score probability in Eqn (2.4) yield a conditional Bernoulli probability distribution for o_i given the observer template $\mathbf{w}$ and the difference image, $\Delta \mathbf{g}_i$, of

$$\Pr\left(o_i|\Delta\mathbf{g}_i,\mathbf{w}\right) = p_i^{o_i}\left(1-p_i\right)^{1-o_i}. \tag{2.5}$$

If the observer makes more than one pass through the data, then the score can be any integer value between 0 and N_{Pass}, the total number of passes through the data. In this case we can consider the score to be distributed with the more general Binomial distribution

$$\Pr\left(o_i|\Delta\mathbf{g}_i,\mathbf{w}\right) = \frac{N_{\text{Pass}}!}{o_i!\left(N_{\text{Pass}}-o_i\right)!}p_i^{o_i}\left(1-p_i\right)^{N_{\text{Pass}}-o_i}. \tag{2.6}$$

This probabilistic link between the observer template and the observed trial score is the basis for estimating the observer template.

The average of the observer scores (divided by the number of passes through the data) is an estimate the proportion of correct responses,

$$\widehat{P}_C = \frac{1}{N_\text{T}}\sum_{i=1}^{N_\text{T}}\frac{o_i}{N_{\text{Pass}}}, \tag{2.7}$$

where N_T is the total number of trials in the experiment. The proportion of correct responses is often used to obtain a detectability measure[10], d_A, which is defined via the inverse of the cumulative normal distribution function as

$$d_A = \sqrt{2}\Phi^{-1}\left(\widehat{P}_C\right). \tag{2.8}$$

2.B. Probit Regression of Trial Scores

The score probability in Eqn (2.4) defines what is known as a probit-link function in the categorical regression literature[8,9]. The link function relates a linear combination of the parameters of interest – the values of the observer template $\mathbf{w}$ in this case – to the mean probability of a binomial random variable via the cumulative Gaussian distribution function.

One potential problem with estimating the observer template from the resulting statistical model is the large number of degrees of freedom in the observer template. Since the observer template has as many elements as there are pixels in $\mathbf{g}$, the number of free parameters can be quite large in an unconstrained model. Generally there will be more parameters in the observer template than there are trials in the 2AFC experiment, which leads to the uncomfortable situation of having more parameters than data points. In previous work[5], we have addressed this problem through the use of a regularizing prior distribution of the observer template data. In this work, we reduce the degrees of freedom by assuming that the observer template can be represented by a relatively small set of known linear feature vectors, $\mathbf{v}_k$, where k runs from 1 to the number of features, N_F. The feature vectors are linearly related to the observer template by the linear equation

$$\mathbf{w} = \mathbf{V}\boldsymbol{\beta}, \tag{2.9}$$

where the columns of the matrix $\mathbf{V}$ are the linear feature vectors ($\mathbf{v}_k$), and $\boldsymbol{\beta}$ is a vector of feature weights with N_F elements. The goal is now to estimate the elements of $\boldsymbol{\beta}$ instead of the entire observer template. The observer template can be synthesized from the feature weights by using Eqn (2.9). The use of feature vectors reduces the degrees of freedom of the problem from the number of pixels in the image to the number of features chosen to represent the observer template. In Section 3 below, this constitutes a reduction from 16,384 free parameters for the entire template to a total of 33 free parameters in the constrained representation.

To use probit regression methods on the feature weights, we must link them to the score probabilities. This can be accomplished by substituting Eqn (2.9) into Eqn (2.4). The resulting expression for the score probability can be written

$$p_i = \Phi\left(\left[\mathbf{X}\boldsymbol{\beta}\right]_i\right) \tag{2.10}$$
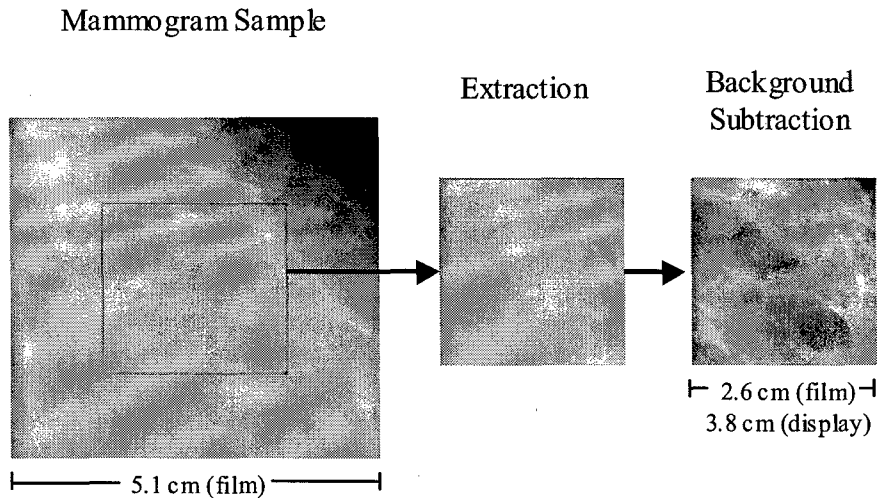
Figure 1. Mammographic sample image preparation. This schematic shows the process of extraction and background subtraction used to create the mammographic images used in the psychophysical studies.

where $\left[\mathbf{X}\boldsymbol{\beta}\right]_i$ is the ith element of the matrix-vector product $\mathbf{X}\boldsymbol{\beta}$, and the matrix $\mathbf{X}$ is defined as

$$\left[\mathbf{X}\right]_{ik} = \frac{\Delta \mathbf{g}_i^t \mathbf{v}_k}{\sqrt{2}}. \qquad (2.11)$$

A standard method of solving for the free parameters in $\boldsymbol{\beta}$ is known as Fisher Scoring or alternatively as Iterative Reweighted Least Squares[9]. We begin this procedure by assembling the score data from the 2AFC trials (with N_{Pass} total passes through the data) into a vector $\mathbf{y}$. The parameter estimation method consists of assuming an initial value of $\boldsymbol{\beta}^0 = \mathbf{0}$, and iterating

$$\boldsymbol{\beta}^{(n+1)} = \boldsymbol{\beta}^{(n)} + \left(\mathbf{X}^t \mathbf{D}^{(n)} \mathbf{X}\right)^{-1} \mathbf{X}^t \left(\mathbf{y} - \mathbf{m}^{(n)}\right), \qquad (2.12)$$

where the vector $\mathbf{m}^{(n)}$ is the predicted mean score value assuming the feature weights,

$$\left[\mathbf{m}^{(n)}\right]_i = N_{\mathrm{Pass}} \Phi\left(\left[\mathbf{X}\boldsymbol{\beta}^{(n)}\right]_i\right),$$

and the diagonal matrix $\mathbf{D}^{(n)}$ is the conditional covariance of score data assuming the feature weights,

$$\left[\mathbf{D}^{(n)}\right]_{ii} = N_{\mathrm{Pass}} \Phi\left(\left[\mathbf{X}\boldsymbol{\beta}^{(n)}\right]_i\right)\left(1 - \Phi\left(\left[\mathbf{X}\boldsymbol{\beta}^{(n)}\right]_i\right)\right).$$

We find this algorithm to converge quickly (typically within 10 iterations) to the maximum-likelihood estimate, $\hat{\boldsymbol{\beta}}$, and the asymptotic error covariance matrix associated with this estimate is given by

$$\mathbf{K}_{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}^t \mathbf{D} \mathbf{X}\right)^{-1}, \qquad (2.13)$$

where $\mathbf{D}$ is $\mathbf{D}^{(n)}$ above, evaluated at $\hat{\boldsymbol{\beta}}$. Because $\mathbf{X}$ is a highly rectangular matrix (its dimensions are N_T by N_F), the matrix inverses necessary for Eqn.s (2.12) and (2.13) are only computed for matrices of size N_F by N_F.

3. RESULTS

3.A. Images for Psychophysical Studies

The images used in the psychophysical studies reported here came from the Digital Database for Screening Mammography, a database of digitized mammograms available at the University of South Florida[11]. The two criteria

Power Spectrum of Mammograms
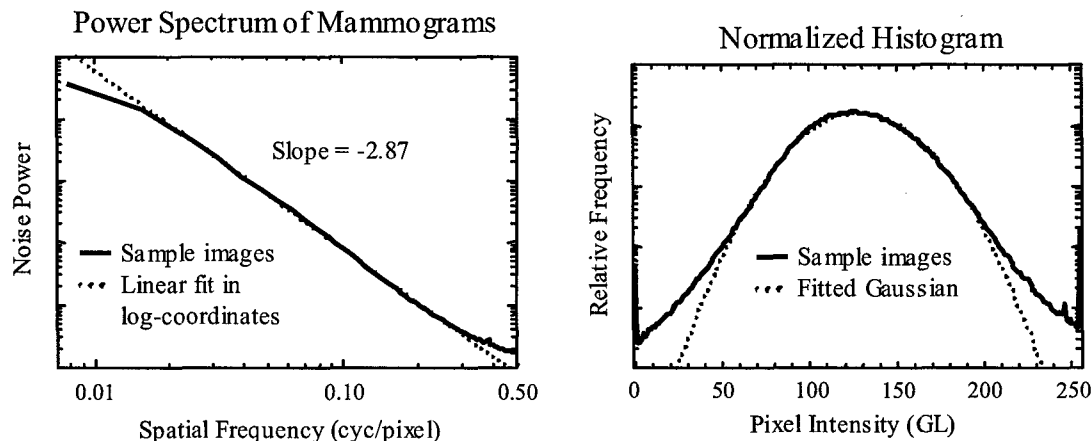
Normalized Histogram

Figure 2. The left plot shows the power spectrum of the mammographic sample images. The slope of the linear fit (in the log-log coordinates of the plot) indicate that the noise power falls off as radial frequency raised to the power –2.87. The right plot shows the normalized histogram of the images with a Gaussian distribution fit to the central portion of the histogram.

used for inclusion in this work were that the patient was classed normal and that the mammographic films were digitized with a Lumisys scanner (Lumisys Inc., Sunnyvale, CA) resulting in a 10-bit digitized image with intensity proportional to log exposure and a pixel size of 0.05mm. Some 656 distinct patches of 1,024 by 1,024, 10-bit data internal to the breast were extracted from cases derived from 82 patients.

Figure 1 shows how these 656 distinct patches were turned into an initial set of 5,904 sample images for use in the psychophysical studies. From each mammogram patch, 9 overlapping 512 by 512 subregions were extracted. The subregions were centered at 1/4, 1/2, and 3/4 of the distance between the image edges in both the vertical and horizontal directions. Because our intent was to study the effect of mammographic structure on detection of a simulated mass, we fit a bilinear function to each subregion and subtracted it from the image to enhance the presence of structure in the images. The parameters of the bilinear function were computed by least-squares fitting to all the pixels in the subregion. After the background had been subtracted, the image was scaled so that the average deviance was 20 gray levels (GL) and the image was added to a pedestal of 128 GL, and finally down-sampled by a factor of 4 for an image size of 128 by 128 with a pixel size of 0.2mm. The scaling and pedestal were chosen so that the images would reside in the middle of the dynamic range on an 8-bit display of a monitor with a linear lookup table. The down-sampling was performed so that the resulting images were of approximately the same size as the film. The 5,904 images were each examined by the author, and 914 of them were removed from the test images. The main criteria for culling the images were that they were too close to the edge of the breast, or there were strong edge artifacts where the film digitizer extended past the edge of the film. The resulting 4,990 images were used for the psychophysical studies.

Figure 2 shows some statistical properties of the mammogram patches. We computed an average noise-power spectrum (NPS) of the images by subtracting the mean image and then windowing the images with a 4th-order Butterworth filter and computing the average of the squared magnitude of the DFT. The radial average of this NPS is plotted with respect to radial frequency on a log-log scale on the left side of Figure 2. The NPS assumes a nearly linear falloff in the log-log plot from 0.02 to 0.3 cycles/pixel (0.1 to 1.5 cycles/mm in the films). The NPS drops by over 3 orders of magnitude in this frequency range. The slope of the log-NPS versus log-frequency line is -2.87. This is very close to the values reported by Burgess[12]. One difference between the NPS plotted here and that reported by Burgess is that the NPS goes below the fitted line at the lowest spatial frequencies. We attribute this to the background subtraction method we used, which will tend to reduce the variability at low spatial frequencies.

The right side of Figure 2 shows the normalized histogram for the entire image set along with a fitted Gaussian distribution (note that the logarithmic y-axis gives the Gaussian distribution its parabolic profile). We see that the Gaussian distribution provides a good fit from approximately 60 to 200 GL. The histogram and the Gaussian fit do not
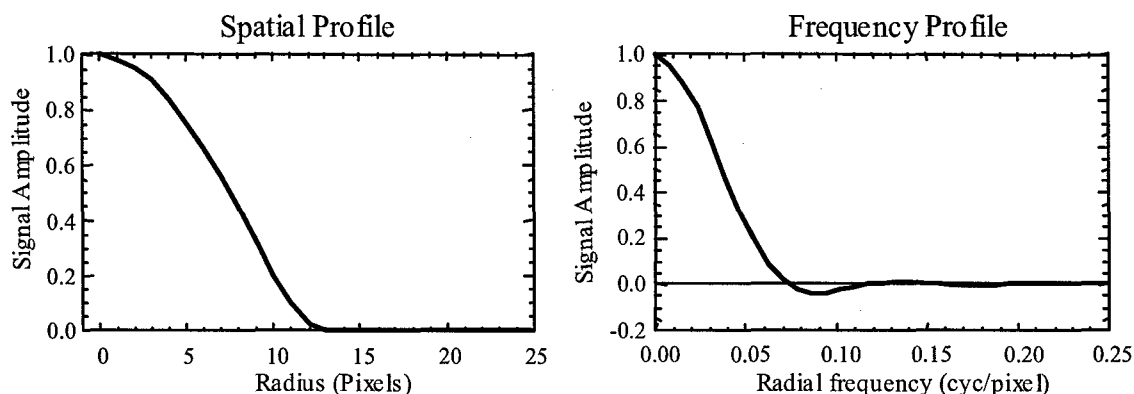
Figure 3. The signal profile used for the psychophysical experiments. On the left the radial profile of the signal is given in terms of distance from the signal center (1 pixel is 0.2mm on a mammographic film). On the right, the signal profile is given as a function of radial frequency.

part until the histogram has fallen off by over two orders of magnitude. The spikes on both ends of the gray-scale range indicate that a small percentage of gray levels were truncated to fit the display range of the monitor.

Figure 3 shows the spatial and spatial-frequency profiles of the signal used to simulate a mass for our experiments. The radial profile of the mass was specified by the function

$$S(r) = A\left(1 - (r/R)^2\right)^{3/2}, \tag{3.1}$$

for pixels whose distance to the signal center, r, was less than the signal radius, $R = 12.5$ pixels (2.5mm). For pixels whose distance from the signal center was more than R, the signal profile was set to zero, yielding a signal diameter of 25.0 pixels (5.0mm). This profile has been used previously by Burgess[13] and others[14] who found that it fit nodule data obtained by Samei et Al.[15] In the experiments, the signal-present images had this signal profile added at a signal amplitude, A.

3.B. Psychometric Study

A total of 5 observers participated in the psychophysical studies. Two of these observers (observers 1 & 2) were authors of this paper (CKA and SSS). The other three observers were naive subjects compensated for participating in the studies. All observers have prior experience as subjects of visual psychophysics experiments. After an initial round of training (50-100 2AFC trials), observers participated in a psychometric study, which evaluated detection performance as a function of signal amplitude.

The signal amplitudes used in the studies ranged from 18 to 50 GL (14% to 39% relative contrast). At each of the signal amplitudes, 200 trials/observer were collected and the proportion of correct responses was computed. The proportion-correct data was converted to detectability according to Eqn (2.8), and plotted in Figure 4 along with linear fits to each observer. We can see in this figure that the observers appear to be reasonably well fit by lines with a y-intercept near or slightly below the origin. The relatively small magnitude of the y-intercept, which is not significantly different from zero for any observer, and the generally good agreement with linear fits suggest that our observers may be well described by the linear model necessary for template estimation. Burgess[16,17,18] has found similar psychometric functions for compact aperiodic signals embedded in noise.

3.C. Observer Template Studies

A second purpose of the psychometric study was to find a reasonable signal amplitude for obtaining data on which to estimate human observer templates. We hoped to achieve a target proportion correct in these experiments between 0.80
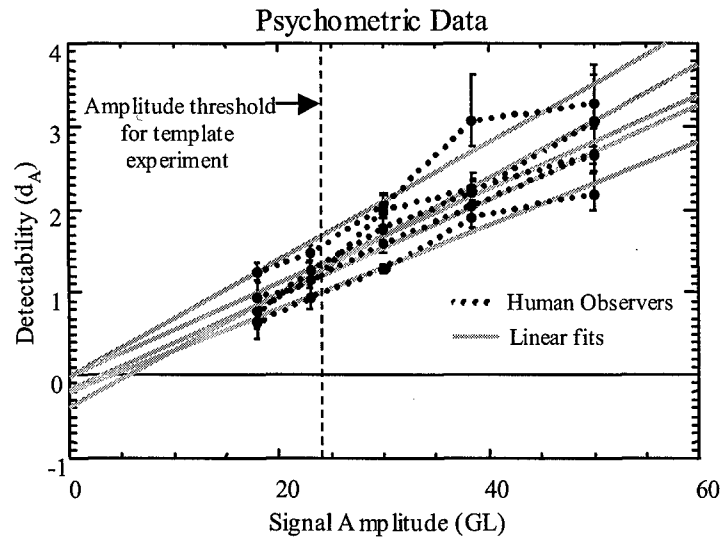
Psychometric Data

Figure 4. Psychometric observer performance data. plotting detectability as a function of signal amplitude along with linear fits for each observer.

Frequency-Band Features

Center freq. = 0.00   Center freq. = 0.10   Center freq. = 0.25
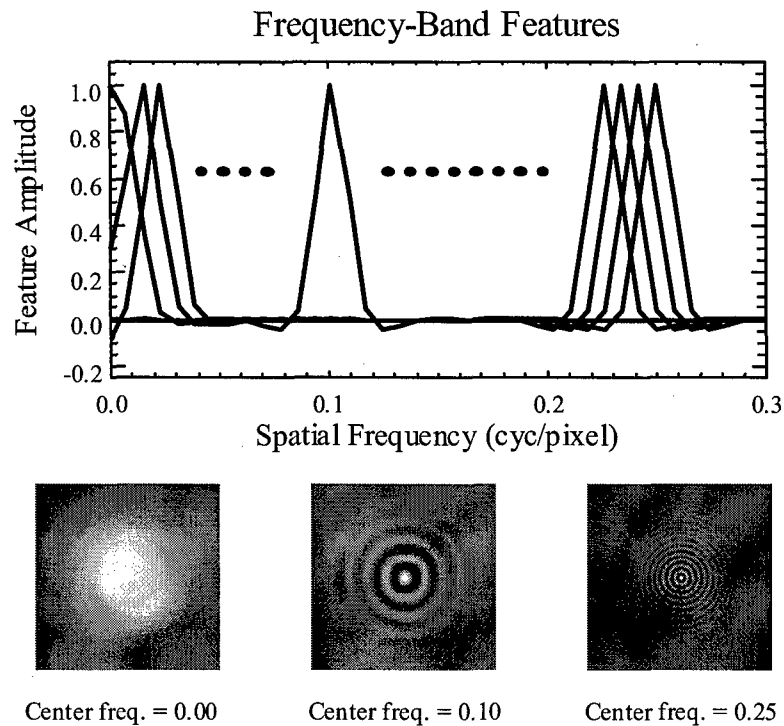
Figure 5. Spatial-frequency band features used to represent the estimated observer templates. The features are defined as radial frequency bands in the DFT domain and then windowed in the spatial domain to reduce ringing artifacts. The plot shows the radial frequency profiles of the features and the images show the spatial appearance of a few selected features.

and 0.85, and based on the psychometric plots in Figure 4, a signal amplitude of 24.0 GL (18.8% contrast) was chosen. With this signal amplitude, observer performance ranged from 0.79 to 0.88 in terms of proportion of correct responses.

After completing the psychometric studies, observers participated in the template experiments. The observer template experiment comprised a total of 2,000 trials, and 3 of the observers (2, 3, and 4) made a second pass through the experiment. For estimating human observer templates, we chose to represent the human observer template by a set of 33 radial frequency bands in the 2D Discrete Fourier Transform domain. These bands ranged from 0.00 to 0.25 cycles per pixel (0.0 to 1.25 cycles/mm in the mammographic films) and extended well beyond the effective spectrum of the signal. Each radial-frequency band had a bandwidth of 0.0078 cycles/pixel before being windowed in the spatial domain to reduce ringing. The spatial window used was a 4th order Butterworth filter with a full-width at half-max of 50 pixels. The radial frequency bands, after this windowing process, are plotted at the top of Figure 5. Images of a few of these frequency-band features can be seen at the bottom of Figure 5. Note that the leftmost of these images (center frequency = 0.0) is also an image of the spatial window.

The images at the top of Figure 6 show the estimated observer templates obtained from the estimation formula given in Eqn (2.12). In addition to an estimated template for each observer, there is one image labeled "All" that consists of a template estimated from the combined observer data. This template treats the score data from the five observers (with three of these having two passes through the data) as if there were only one observer that made eight passes through the data. While it is clearly not valid to ignore observer differences, we find that this composite data is good for visualizing general trends in the individual observer results. The images generally show an area of positive weighting near the signal center, with a pronounced negative fringe starting about 10 pixels from the signal center. This negative surround is fairly narrow relative to the signal size. Radial frequency plots of the observer templates are given at the bottom of Figure 6 with error bars of width +/-1 standard error computed from Eqn (2.13). The plots all show a pronounced band of positive weights from radial frequencies of 0.015 to 0.06 cycles/pixel (0.08 to 0.3 cycles/mm). There also appears to be some lower level oscillations at higher spatial frequencies. This oscillation is particularly well visualized in the composite estimate from all the human observer data.

Figure 7 shows comparisons of the composite human-observer data with various model observers that have been investigated as surrogates for human observers. The model observers are scaled so that their peak values match that of the human-observer plot. In the upper left corner of Figure 7, we see comparisons with two nonprewhitening model observers. The nonprewhitening matched filter (NPW) model observer simply uses the signal profile as the observer template[19,20], and hence the frequency profile of this observer can be found on the right side of Figure 3. We also plot an eye-filtered nonprewhitening (NPWE) observer[21] that consists of modulating the NPW frequency spectrum by a function representing the contrast sensitivity of the human eye. The NPW observer does not capture any of the bandpass character of the human observer data. The NPWE observer does have a bandpass structure, but the band has been shifted to somewhat lower spatial frequencies than the human-observer data would indicate. It may be possible to account for this mismatch by considering a different visual contrast sensitivity function. Both the NPW and NPWE observers are outperformed by at least some of the human observers for this detection task (NPW proportion correct is 0.70; NPWE proportion correct is 0.82).

The upper right corner of Figure 7 shows the comparison with a prewhitened matched filter (PWMF) model observer. The PWMF observer used here consists of modulating the signal frequencies by the computed noise-power spectrum of the images (See Figure 2). The PWMF observer exhibits more low-frequency suppression than the human observer, and oscillates strongly at higher spatial frequencies. The oscillations in the human-observer data appear to be in sync with this model observer, although they are lower in magnitude. With a proportion correct of 0.93, the PWMF significantly outperforms the human observers.

The frequency profiles of both 3-Channel and 10-Channel DOG Channelized-Hotelling model observers[22,14] are plotted on the bottom row of Figure 7. Both channel models have been investigated previously for agreement with human observer data[14], and we refer the reader to the references given for a detailed description of their implementation. Each plot shows the observer model implemented both with, and without internal noise in the channel responses. The 3-channel DOG observers generally fit well at lower spatial frequencies, but diverged from the human-observer templates at frequencies above 0.05 cycles/pixel. The 10-channel DOG observer without internal noise, like the PWMF, more
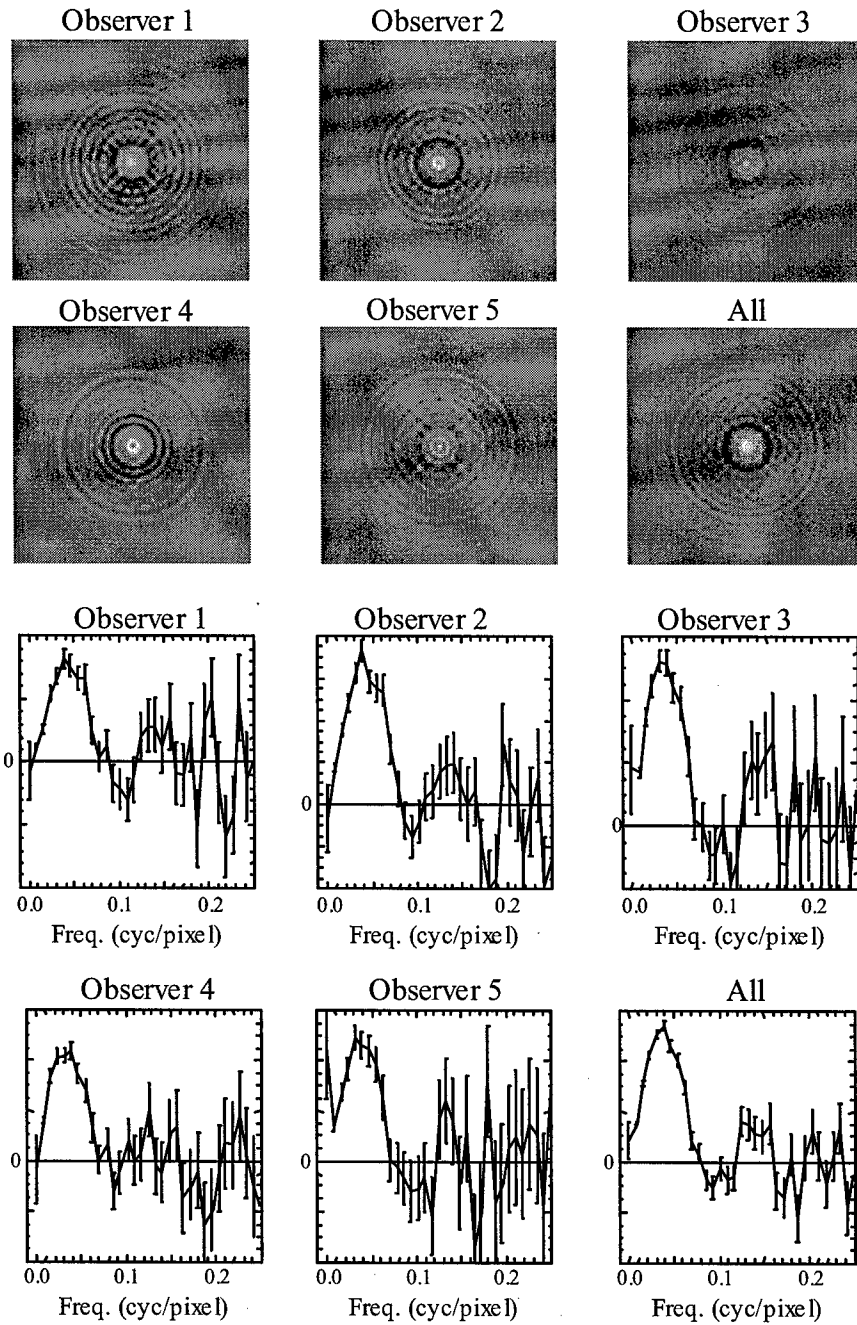
Figure 6. Estimated human observer templates for the lesion detection task. The images at the top show the spatial appearance of the human-observer templates estimated for the 5 observers and the template estimated from a composite dataset from all the observers. The plots show the radial-frequency profiles of the observer templates with error-bars derived from the asymptotic covariance matrix in Eqn (2.13).

strongly suppressed low spatial frequencies than the human observers. However, when internal noise was added to the channel responses, the frequency profile of this observer more closely matched the human observer data at low spatial
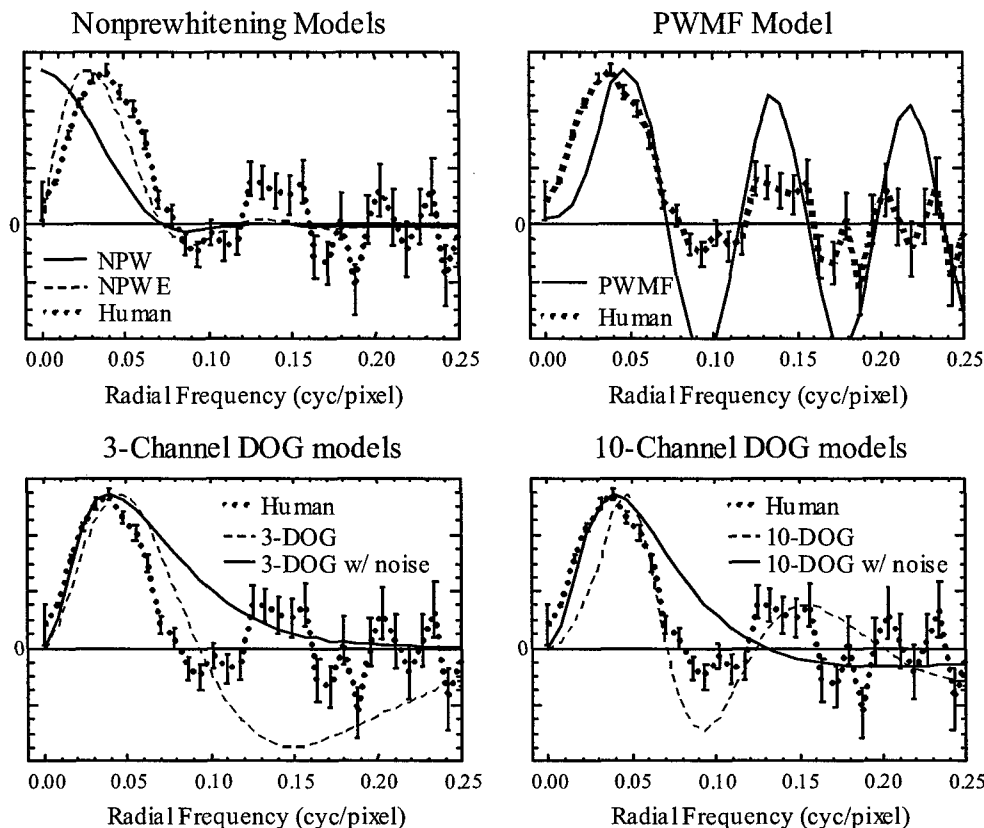
Figure 7. Comparisons of the composite human-observer data and model observer templates. The model observer profiles are normalized so that their peak corresponds to the peak value of the human-observer data.

frequencies. The 10-Channel DOG plots also show how strikingly the inclusion of internal noise in the channel responses can change the weighting scheme of the Channelized-Hotelling observer.

All the models tested here diverged to some degree with the human-observer data at frequencies above 0.05 cycles/pixel, and this raises some concern about their applicability to modeling human observers. However, because both the signal spectrum, and the NPS fall off steeply at higher spatial frequencies, it is not clear how much influence the higher spatial frequencies have on the diagnostic task. On the left side of Figure 8, we plot the NPS – on a linear scale this time – to show the preponderance of noise at low spatial frequencies. The average noise power is substantially reduced by 0.05 cycles/pixel, and an examination of Figure 3 shows that the signal spectrum is also substantially reduced. On the right side of Figure 8, we plot the relative detectability of a PWMF that is constrained to frequencies less than or equal to the x-axis value. The relative detectability is the ratio in detectability between this frequency-constrained PWMF, and the unconstrained PWMF. The plot tells us what percentage of the PWMF detectability is due to the diagnostic information contained in spatial frequencies less than or equal to the x-axis value. Approximately 92% of the relative detectability has been achieved by 0.05 cycles/pixel. This plot tells us that the majority of diagnostic information is contained in the low spatial frequencies. Hence, the Channelized-Hotelling observers are fitting the human observers in the spatial frequencies of greatest diagnostic relevance for this task.

4. CONCLUSION

In this work we have modified a previous approach[5] to maximum-likelihood estimation of observer templates in order to use standard methods of probit regression. Like the earlier approach, this method rests on the assumption of a linear

Relative Noise-Power Spectrum

Relative Detectability

Radial Frequency (cyc/pixel)

Radial Frequency (cyc/pixel)

Figure 8. Plots of relative noise-power spectrum and relative detectability. The noise-power spectrum is equivalent to that plotted in figure 2, but plotted on linear axes and made relative to the DC noise-power. The plot emphasizes the falloff in noise for frequencies above 0.04 cyc/pixel. The relative detectability plot shows the ratio in detectability of a prewhitened matched filter that is only allowed to use frequencies less than or equal to the value of the x-axis. This plot shows that the majority of useful information for performing this task is in the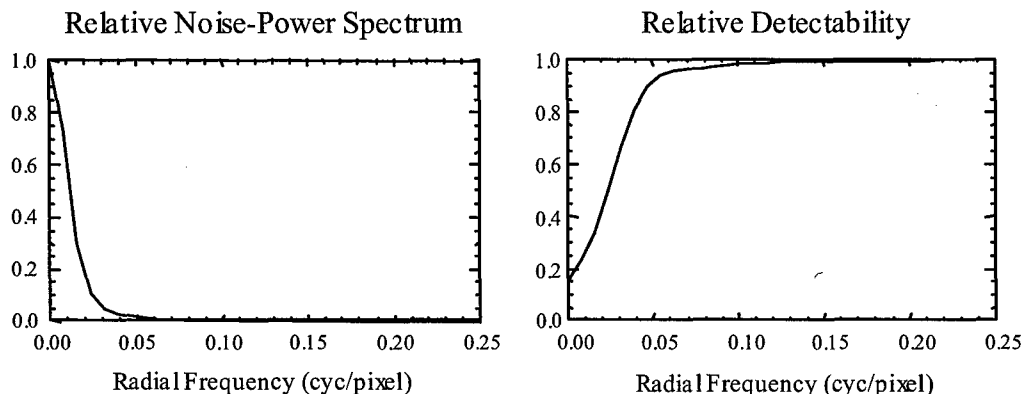 lower spatial frequencies. For example deleting the entire image spectrum for frequencies above 0.06 cycles/pixel results in only a 5% reduction in detectability.

observer and Gaussian-distributed internal noise, and does not make assumptions about the distribution of the images used to perform the task. Hence the method is appropriate for finding linear observer templates in signal-known-exactly tasks involving patient structured backgrounds. The additional assumption necessary for the modification presented here is that the observer template can be described by a linear combination of feature vectors. The feature vectors serve to reduce the degrees of freedom of the estimation problem from the number of image pixels (16,384 in the studies reported here) to the number of feature vectors (33 here). By casting the template estimation problem in terms of probit regression, we can use standard procedures for estimating the feature weights (Fisher Scoring) and the associated error covariance matrix.

We have applied this method to the task of detecting a small low-contrast simulated mass embedded in patient structured backgrounds derived from a set of digitized mammograms. Observer psychometric functions for detecting the mass as a function of lesion contrast show that our observers are reasonably well described by a line with a slightly negative y-intercept, which provides some evidence for linear models. Human observer templates, estimated from one or two passes through 2,000 2AFC trials, show that observers are using a band of spatial frequencies that extend from approximately 0.015 to 0.06 cycles/pixel (0.08 to 0.3 cycles/mm in the film) and peaks between 0.03 and 0.04 cycles/pixel. There is also some evidence of oscillation at higher frequencies.

A number of comparisons are made between a conglomerate of all the human-observer data and various linear model-observer templates suggested as representative of human observers in noise-limited visual tasks. The two nonprewhitening observers we considered, a nonprewhitening matched filter and a nonprewhitening matched filter modulated by a visual contrast-sensitivity function, tended to place too much weight on low spatial frequencies relative to the human observers. Conversely, a prewhitened matched filter model demonstrated a suppression of low spatial frequencies (less than 0.04 cycles/pixel) relative to the human observers, as well as demonstrating relative enhancement of higher spatial frequencies. The fact that the prewhitened matched filter substantially outperforms the human observers indicates that human-observer performance may be limited by an inability to fully suppress low spatial frequencies. This, in turn, suggests that processing the image by filtering low spatial frequencies may improve human-observer performance.

We also compared the human observer data to Channelized-Hotelling observer models derived from two difference-of-Gaussian channel models. These models fit the human observer data at lower spatial frequencies, but both implementations of the 10-channel DOG model diverge from the human observer data at frequencies of 0.05 cycles/pixel and above. This divergence is a concern that we feel should be addressed in future work. However, we have shown that

there is relatively little diagnostic information at frequencies above 0.05 cycles/pixel. Hence we conclude that the two channel models implemented with internal noise in the channel responses, as well as the 3-Channel model without internal noise, are fitting the human observer templates in the diagnostically relevant spatial frequencies for this task.

ACKNOWLEDGMENTS

REFERENCES

1. Beard, B. L., & Ahumada, A. J. (1998). Technique to extract relevant image features for visual tasks. *Proceedings of SPIE, 3299,* 79-85.
2. C.K. Abbey and M.P. Eckstein, "Estimation of human-observer templates for 2 alternative forced choice tasks," *Proc. SPIE* 3663, pp. 284-295 1999.
3. D.C. Edwards, M.A. Kupinski, R.M. Nishikawa, and C.E. Metz, "Estimation of linear observer templates in the presence of multi-peaked gaussian noise through 2AFC experiments" *Proc. SPIE* 3981, 86-96, 2000.
4. C.K. Abbey, M.P. Eckstein, and F.O. Bochud, "Estimates of human observer templates for a simple detection task in correlated noise," *Proc. SPIE* 3981, 70-77 2000.
5. C.K. Abbey and M.P. Eckstein, "Maximum-likelihood and maximum a-posteriori estimates of human-observer templates," *Proc. SPIE* (E.A. Krupinski and D.P. Chakraborty, Ed.s) 4324: 114-122 2001.
6. C.K. Abbey and M.P. Eckstein, "Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments," *Journal of Vision* 2(1):66-78, 2002 (http://www.journalofvision.org/2/1/5/).
7. C.K. Abbey and M.P. Eckstein, "Optimal shifted estimates of human-observer templates in two-alternative forced-choice experiments," *to be published in: IEEE trans. Med. Imag.* May 2002.
8. P. McCullagh and J.A. Nelder, *Generalized Linear Models,* (2nd Edition) Chapman and Hall/CRC, New York, pp. 40-43, 1989.
9. A. Agresti, *Categorical Data Analysis,* John Wiley and Sons, New York, pp. 449-452, 1990.
10. D.M. Green and J.A. Swets, *Signal detection theory and psychophysics,* Wiley, New York, 1966.
11. M. Heath, K.W. Bowyer, D. Kopans et al, "Current status of the Digital Database for Screening Mammography," in *Digital Mammography,* Kluwer Academic Publishers, pp. 457-460 1998.
12. A.E. Burgess, "Mammographic structure: data preparation and spatial statistics analysis," *Proc SPIE* (K.M. Hanson Ed.), 3661:642-653. 1999.
13. A.E. Burgess, X. Li, and C.K. Abbey, "Visual signal detectability with two noise components: anomalous masking effects," *J. Opt. Soc. Am. A,* Vol. 14, No. 9:2420-2442, 1997.
14. C.K. Abbey and H.H. Barrett, "Human and model-observer performance in ramp-spectrum noise: Effects of regularization and object variability." *J. Opt. Soc. Am. A.* 18(3): 473-488. 2001.
15. E. Samei, M. J. Flynn, G. H. Beue, and E. Peterson, "Comparison of observer performance for real and simulated nodules in chest radiography," *Proc SPIE,* (H. L. Kundel, ed.), 2712, 60–70 (1996).
16. A.E. Burgess, and H. Ghandeharian, "Visual signal detection. II. Signal-location identification," *J. Opt. Soc. Am. A,* 1:900-905, 1984.
17. A.E. Burgess, "Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation," *J. Opt. Soc. Am. A,* 2:1498-1507, 1985.
18. A.E. Burgess and B. Colborne, "Visual signal detection. IV. Observer inconsistency," *J. Opt. Soc. Am. A,* 5:617-627, 1988.
19. R.F. Wagner, D.E. Brown, and M.S. Pastel, "Application of information theory to the assessment of computed tomography" *Med. Phys.* 6:83-94, 1979.
20. K.J. Myers, H.H. Barrett, M.C. Borgstrom, D.D. Patton, and G.W. Seeley, "Effect of noise correlation on the detectability of disk signals in medical imaging," *J. Opt. Soc. Am. A,* 2:1752-1759, 1985.
21. A.E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening matched filter model," *J. Opt. Soc. Am. A,* 11:1237-1242, 1994.
22. K. J. Myers and H. H. Barrett, "The addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Am. A* 4:2447–2457, 1987.

Computer aided detection of masses in mammography using subregion Hotelling observers

Alan H. Baydush[a)]
*Department of Radiation Oncology, Physics Division, Duke University Medical Center, Durham,
North Carolina 27710 and Department of Biomedical Engineering, Duke University, Durham, North
Carolina 27710*

David M. Catarious
*Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710 and Digital
Imaging Research Division, Department of Radiology, Duke University Medical Center, Durham,
North Carolina 27710*

Craig K. Abbey
Department of Biomedical Engineering, University of California, Davis, California 95616

Carey E. Floyd
*Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710
and Digital Imaging Research Division, Department of Radiology, Duke University Medical Center,
Durham, North Carolina 27710*

We propose to investigate the use of the subregion Hotelling observer for the basis of a computer aided detection scheme for masses in mammography. A database of 1320 regions of interest (ROIs) was selected from the DDSM database collected by the University of South Florida using the Lumisys scanner cases. The breakdown of the cases was as follows: 656 normal ROIs, 307 benign ROIs, and 357 cancer ROIs. Each ROI was extracted at a size of 1024×1024 pixels and sub-sampled to 128×128 pixels. For the detection task, cancer and benign cases were considered positive and normal was considered negative. All positive cases had the lesion centered in the ROI. We chose to investigate the subregion Hotelling observer as a classifier to detect masses. The Hotelling observer incorporates information about the signal, the background, and the noise correlation for prediction of positive and negative and is the optimal detector when these are known. For our study, 225 subregion Hotelling observers were set up in a 15×15 grid across the center of the ROIs. Each separate observer was designed to "observe," or discriminate, an 8×8 pixel area of the image. A leave one out training and testing methodology was used to generate 225 "features," where each feature is the output of the individual observers. The 225 features derived from separate Hotelling observers were then narrowed down by using forward searching linear discriminants (LDs). The reduced set of features was then analyzed using an additional LD with receiver operating characteristic (ROC) analysis. The 225 Hotelling observer features were searched by the forward searching LD, which selected a subset of 37 features. This subset of 37 features was then analyzed using an additional LD, which gave a ROC area under the curve of $0.9412 +/- 0.006$ and a partial area of 0.6728. Additionally, at 98% sensitivity the overall classifier had a specificity of 55.9% and a positive predictive value of 69.3%. Preliminary results suggest that using subregion Hotelling observers in combination with LDs can provide a strong backbone for a CAD scheme to help radiologists with detection. Such a system could be used in conjunction with CAD systems for false positive reduction. © *2003 American Association of Physicists in Medicine.*
[DOI: 10.1118/1.1582011]

Key words: CAD, mass detection, classification, image processing, breast cancer

INTRODUCTION

Cancer is one of the most devastating and deadly diseases of our time and is the second leading cause of death in the United States.[1] The American Cancer Society estimates that in 2002 alone, breast cancer will be diagnosed in 203 500 women and almost 40 000 women will perish.[2] For women, breast cancer is the most common cause of cancer death. A strong commitment to reducing cancer-related deaths has been put forth by the Department of Health and Human Services. The prime method for detecting breast cancer is through screening mammography.[3] Early detection of suspicious regions helps improve patient outcome and is a key to patient care. We firmly believe that development and application of computer aided detection (CAD) techniques for the automated detection of cancerous breast masses will have a great impact on early detection and hence on overall patient outcome.

Currently, screening mammograms are taken and mammographers examine the images to detect possible abnormalities, some of which are masses. CAD systems have been

researched and are commercially available[4,5] which aid the radiologist in detecting these suspicious regions and thus reduce missed cancers. Most CAD systems can be viewed as two stages. Typically, the first stage uses some type of initial processing, which has high sensitivity and low specificity, to detect a set of potential masses. The second stage consists of classifying these suspicious regions using predictive modeling techniques (neural networks, cluster analysis, etc.) to reject a large number of false positives. With this approach, systems have been developed that effectively detect masses. From the radiologist's point of view, the largest problem with these systems is the false positives per image. It is this second stage of the CAD system which we aim to improve.

Since we wish to help radiologists with the detection task, we have chosen to base our approach on models of the radiologist's vision system. We have pursued this approach previously in chest radiography[6] with great success. We feel that this new approach is innovative and needs to be investigated thoroughly in mammography, as well. We propose that incorporating models from vision science into the classification process will help to reduce the number of false positives while not reducing sensitivity.

For the study presented here, we propose to investigate using subregion Hotelling observers (SRHOs) in conjunction with linear discriminants (LDs) for the automated classification of regions as containing or not containing a mass. This type of classification could be incorporated into a CAD system in the future to aid in false positive reduction.

METHOD

We wish to continue examining incorporating models from the human vision system into the classification stage of CAD systems. Several proposed models[7–10] of the vision system used to predict the performance of visual tasks have utilized an initial linear feature extraction step followed by a reductive feature processing step (usually nonlinear). We have chosen to follow the multilayered general form of these models (a linear mechanism followed by a nonlinear integration of features to perform basic decision tasks). The advantage of this approach is that we are not limited to linear features found in the human vision system, but rather we define these features using locally optimal linear discriminants.

A. Description of three layer classifier

For this study, we will be investigating a SRHO system similar to the one we developed and used for nodule classification in chest radiography.[6] In that study, a three layer system was developed using SRHOs and artificial neural networks (ANNs). The system investigated here will be altered from the previous version by replacing the ANNs used in the system with LDs. The reason for this change is to simplify the overall system and to come up with a single output template that can be used similar to a filter for mass detection. A single filter of this sort can be incorporated via convolution to quickly find regions that "look" like centered ROIs, such as the ones the system presented here is trained on. A flow
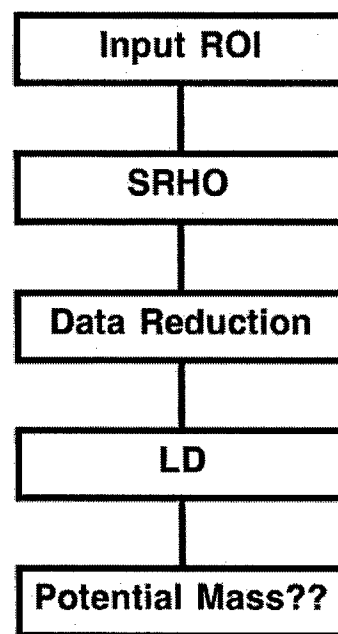
FIG. 1. Flow chart of three layer classifier.

chart of this system is shown in Fig. 1. For this study, our three layer model[7–10] is as follows: Layer 1 models the linear portion of the visual system by using a grid of SRHOs. Layer 2 models the data reduction in the visual process and will be performed by forward searching LD. Layer 3 uses an additional LD to combine the reduced data set and to determine final classification results.

1. Layer 1: Subregion Hotelling observers

The Hotelling observer (HO) is the optimal linear detector for a known signal, known background, and known covariance matrix when statistics are approximately Gaussian.[11] For real medical images, where we do not know the exact signal or background, we use estimates of the signal, the general background, and the covariance matrix to calculate the set of linear weights for the suboptimal observer. These observers are only suboptimal until the sample statistics (average background, signal, covariance matrix) approach the true distribution statistics. If enough samples are used, this approximation should not cause much reduction in performance. The weights or template for the HO are multiplied by the image data and summed to give a test statistic. This test statistic can be used as a decision variable. The test statistic should be larger when the signal is present and smaller when absent. In white noise, the HO reduces to a matched filter. However, in medical images, which have correlated noise, the observer estimates a template to decorrelate the noise.[12] HOs have been shown to be effective in tracking the performance of human observers for detection[13–18] and as a means for measuring image quality.[11,19–22]

Application of the HO to a large region of interest (ROI) is prohibitive, as too many image samples are needed to estimate the covariance matrix.[22] To overcome this difficulty, we have turned to the subregion Hotelling observer (a HO

ROI

128 pixels

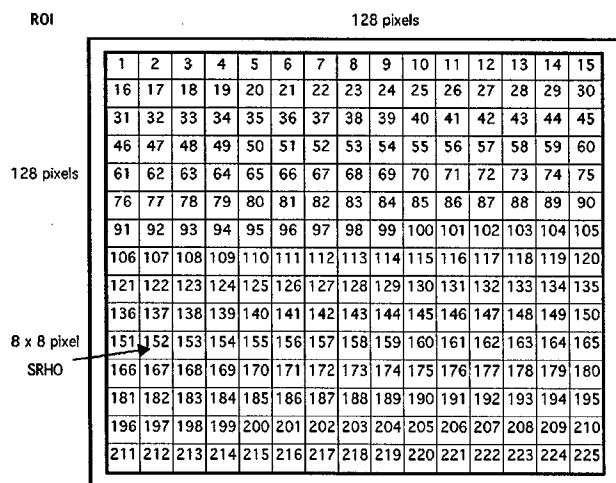| 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135
136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
151	152	153	154	155	156	157	158	159	160	161	162	163	164	165
166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189	190	191	192	193	194	195
196	197	198	199	200	201	202	203	204	205	206	207	208	209	210
211	212	213	214	215	216	217	218	219	220	221	222	223	224	225

128 pixels

8 x 8 pixel SRHO

FIG. 2. The 15×15 grid of SRHO are shown as placed in each ROI. Each of the SRHO covers an area of 8 pixels by 8 pixels. The overall ROI is 128 pixels by 128 pixels, where the center 120×120 of them are covered by the different observers.

for a small subregion). Since it has fewer pixels in the subregion, the SRHO requires significantly fewer samples to properly estimate the necessary covariance matrix. To cover the entire ROI we wish to examine, we tile a matrix of subregions over the full region (Fig. 2). This results in many SRHOs being used to reduce the complexity of the image data down to the number of SRHOs used. The output result of each SRHO is a scalar. N SRHOs tiled over an entire ROI will generate N outputs or "features." These features are then passed on to the second layer for further processing.

2. Layer 2: Forward searching linear discriminant analysis

The result of layer 1 of the classifier is N image features, where each feature is the result of the application of a SRHO to one particular subregion of the full ROI. To further reduce and simplify the algorithm, only certain areas (subregions) or "features" should be selected to be included in the final decision. Regions where misinformation or no useful information is gained can be discarded. To determine which subregions to incorporate, we used a forward searching Fischer's LD, which utilized receiver operating characteristic (ROC) area under the curve (AUC) as the performance criteria. Fischer's LDs are used to optimally divide a two class system into its constituent classes by maximizing the distance between the sample class means relative to the sample variances of the feature set.[23]

A forward searching linear discriminant (FSLD) starts with an empty final set and begins to work by examining the output statistic (AUC) for each of the N image features. The feature that gives the highest output statistic is removed and put into the final set. The forward search continues by taking each of the remaining features one at a time and constructing a LD with the current "final set." Once again, the feature that gives the highest output statistic in conjunction with the previously selected "final set" is included into the new,
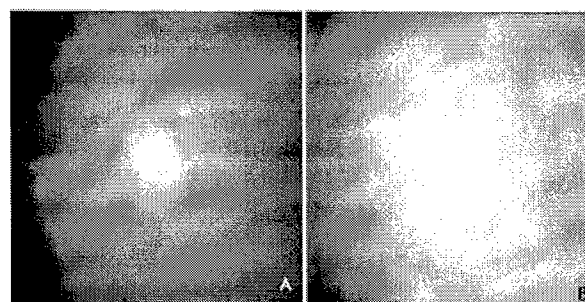

FIG. 3. Average image of the (a) positive and (b) negative ROIs.

larger "final set." This process continues until the output statistic no longer increases with additional features being added. The final chosen feature set is then passed on to the next layer.

3. Layer 3: Combination and classification

The reduced feature data set was then used as the input to an additional LD. A "round robin" or "leave one out" sampling scheme was utilized in order to use all cases for training and testing while still maintaining independence between the training and testing sets. The outputs from this final LD are then used to determine the systems final performance. Again, ROC AUC was used as the output statistic.

B. Image database

A ROI database was generated for this study from cases from the University of South Florida's (USF's) Digital Database for Screening Mammography (DDSM).[24] All of the cases for this study were taken from images that were digitized with a Lumisys scanner at 50 microns. Only images which were normal or contained a mass (either benign or malignant) were used. The DDSM database also contains truth files, which give location and outlines for each mass (benign or malignant), and subtlety ratings. Using the truth files, a database of 1024×1024 pixel ROIs was extracted where each ROI contained a mass abnormality at its center. A number of "normal" tissue ROIs were extracted, as well. The ROIs were extracted at full resolution and then subsampled down to 128×128 pixels (400 micron).

A total of 1320 ROIs were selected. The final breakdown of the cases was 656 normal ROIs, 307 benign ROIs, and 357 cancer ROIs. Since we are interested in a detection task, cancer and benign cases were considered positive and normal cases were considered negative, when calculating performance. Overall, this gives a database of 664 positives and 656 negatives for use in this study. Figure 3 shows images of the numerical average (sum of all cases over number of cases) positive signal (mass present, benign or malignant) and numerical average negative signal (normal tissue only). While the positive average image shows strong radial symmetry and a nicely centered signal, the negative average image is more diffuse and larger. The central portion of the negative image is radially symmetric as well, but there appears to be a small signal from outside the breast in the upper and lower left-hand corners.
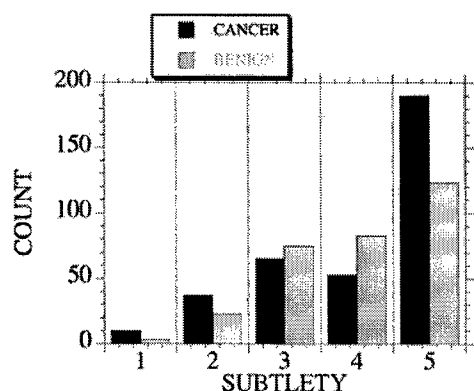
FIG. 4. Histogram of the subtlety rating of the benign and cancerous masses used in the database.

C. Procedure

For the study presented here, 225 separate SRHOs were arranged in a 15×15 grid across the 128×128 pixel ROIs. Each of the SRHOs was designed to "observe" an 8×8 pixel subregion. Therefore, the 8×8 pixel SRHOs covered the center 120×120 pixel region of the ROI (see Fig. 2). A leave-one-out training and testing methodology was used to generate 225 (15×15) features, where each feature is the output of an individual SRHO. Signal and background were

modeled as the average of the positives and negatives (minus the testing case). The overall covariance matrix was formed by combining the positive and negative covariance matrices each weighted by the percentage of total samples which corresponded to that matrix (i.e., number of positive samples/ total number of samples for the positive matrix). All of the test values were then collected as features. The result of this first step was a data reduction from a 128×128 pixels region to 225 values or features per ROI. Each SRHO covered approximately a 3×3 mm area.

For layer 2, these 225 features were used as the inputs to a FSLD. The FSLD was used to select the features that were important and subsequently should be used in the final layer. The forward searching procedure continued until the value of the area under the ROC curve started to decline, at which point the optimal subset of features was chosen. A significance level of 0.05 was used to terminate the selection process.

For layer 3, a LD was applied to only the reduced set of features (derived from layer 2) and additional metrics were calculated using a leave-one-out training and testing methodology. Calculations of ROC area and partial area, as well as statistical comparisons of those metrics, were performed using the ROCKIT program (Charles Metz, University of Chicago).

FIG. 5. Histogram of the AUC values for each of the outputs from (A) the 225 SRHOs and (B) the reduced set of 37 SRHOs.

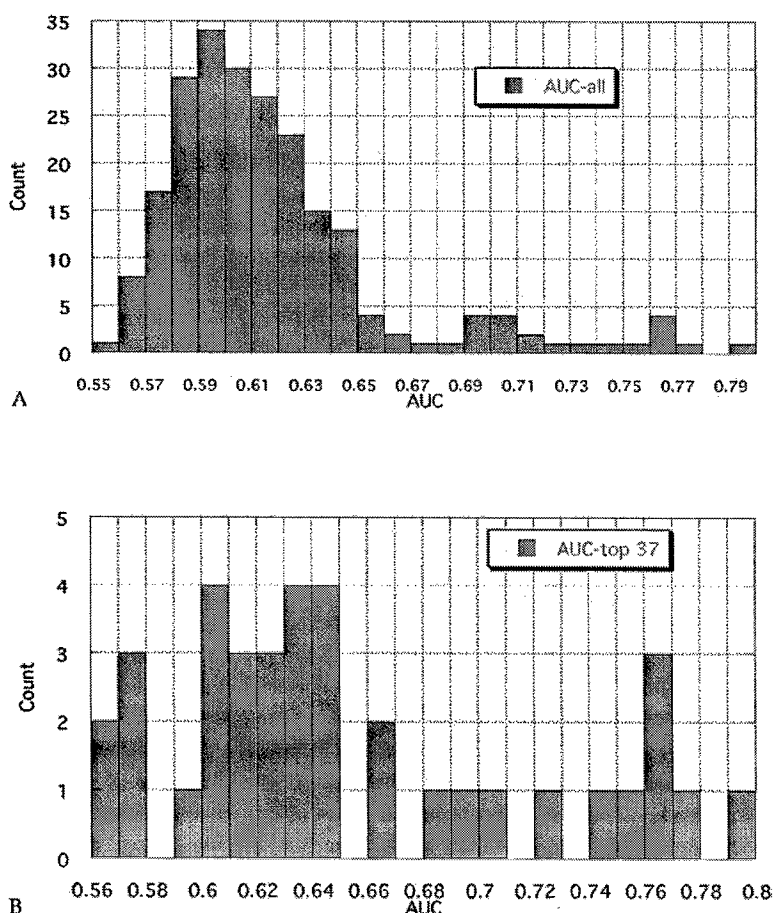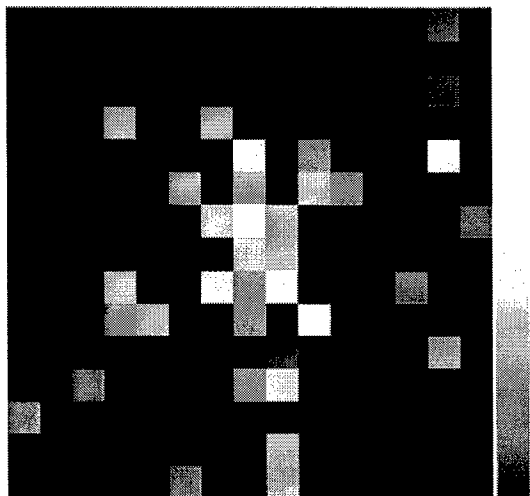FIG. 6. SRHOs selected by the forward searching LD showing the order in which they were selected. The black areas represent regions that were not selected. The included color bar demonstrates the order in which the SRHOs were selected: white first, followed by shades of gray, all the way to black.

FIG. 7. ROC output curves for the fitted and empirical results of the final classification stage. The A_z for the SRHO/LD system is 0.94, with a partial AUC of 0.673.

RESULTS

To demonstrate the varying degree of difficulty of the cases that were used in the database created for this study, a histogram of the subtlety rating of the benign and cancerous masses was generated and is shown in Fig. 4. The subtlety ratings were taken from the data in the DDSM files associated with the images and are based on the assessment of the mammographers who read the case for the database. The histogram shows that most of the cases (both benign and malignant) were of high subtlety, thus validating the complexity of the dataset.

Figure 5(A) shows a histogram of the AUC values for each of the outputs from the 225 SRHOs used in the first layer. The range of individual AUCs for the 225 outputs was from 0.56 to 0.79, with a mean value of 0.62. All of the individual SRHOs performed above chance (0.50). In layer 2, the FSLD was used to reduce the number of features. The FSLD was implemented and proceeded until the AUC objective was maximized using a reduced set of 37 selected features (SRHOs). Figure 5(B) shows a histogram of the individual AUC values for each of the selected SRHOs. The range of individual AUCs for the 37 selected features was from 0.57 to 0.79, with a mean value of 0.65. The reduced set of features was very representative of the full set. The maximum value did not change at all, the minimum value only changed by 0.01, and the mean value only increased by 0.04.

Figure 6 shows a graphical representation of the location and the order in which the SRHOs were selected by the FSLD. The black areas represent regions that were not selected at all. The SRHO shown in white was selected first. As the color bar on the side shows, the order of selection proceeds from white to shades of gray to black. The figure shows a strong preference for selection of regions that were more centrally located over those that were further away.
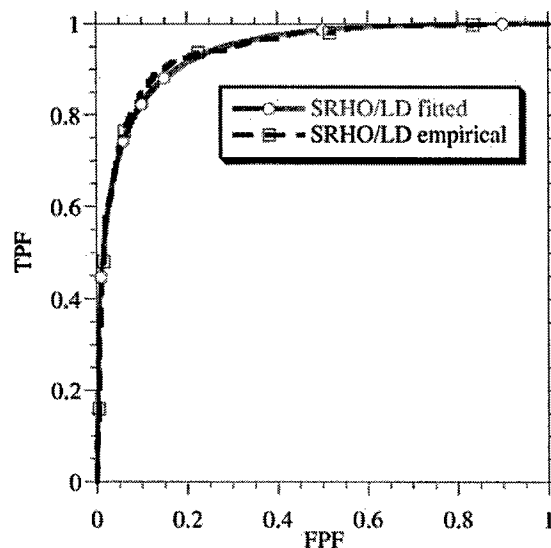
Additionally, only a very weak component of directionality is seen.

This subset of 37 features from layer 2 was then used as input to an additional LD in layer 3. Figure 7 shows the ROC output curves for the fitted and empirical results of this final classification stage. The AUC for the final output and the full SRHO/LD system was $0.94+/-0.006$, which corresponds to a partial AUC (the normalized area above 90% sensitivity on the ROC curve) of $0.673+/-0.028$. Additionally, at 90% sensitivity, the overall classifier had a specificity of 86% and a positive predictive value (PPV) of 86.3%. At 95% sensitivity, the system had a specificity of 69% and a PPV of 75.8%.

DISCUSSION

The purpose of this study was to investigate the use of subregion Hotelling observers in conjunction with linear discriminants for the automated classification of regions containing or not containing a mammographic mass. The exact classification task was to detect the presence or absence of a mass. Both benign and malignant masses were deemed as mass present. It was not the goal of this study to diagnose masses as being either benign or malignant, although similar techniques could be investigated to do so.

For the study presented here, a database of 1320 ROIs was generated from the image cases in the DDSM database. 664 of these cases were positive (benign or malignant mass), while 656 were negative (normal tissue). A histogram of the subtlety ratings from the derived database shows that most of the positive cases were of high subtlety, thus showing the database was a difficult one. A figure of the average positive and average negative signals shows a difference in the two signals profiles.

A three-layer classifier was developed and tested on the above database. The first layer is based on subregion Hotelling observers, the second layer performs data selection and

reduction, and the third layer does the final combination and classification of the remaining features. Figure 5 presented the outputs from the first layer of the system. The AUCs from the 225 SRHOs are seen to range from 0.56 to 0.79. None of the single, small, subregion observers are precise enough to be able to be used on their own; however, when several of them are used together, more information can be obtained and classification improves.

Layer two was used to search the 225 SRHO features and selected a subset of 37 features. It is interesting to note that the FSLD did not just select the best individual features, but chose features which, when combined, gave the best overall final result. Figure 6 demonstrated a strong preference for the selection of regions that were more centrally located and the order of selection fell off as the radius from the center increased. This analysis is consistent with the images of the average positive and negative signals. The average positive has a stronger central profile and is much narrower than the negative profile. As the data reflects, the width of the profiles alone demonstrates that more central region SRHOs should be incorporated and the data reflects this.

Layer three used a LD to classify the regions as mass present or absent by combining the selected subset of 37 features into a final decision. The AUC for this final classification task was $0.94 +/- 0.006$, which corresponds to a partial AUC of $0.673 +/- 0.028$. We calculated the specificity of the system at 95% sensitivity to be 69%. At this threshold, 33 positive cases would be missed, while 453 of the 656 negative regions would be correctly identified as negative. Additionally, at 98% sensitivity (13 missed positives), 345 of the 656 negative regions would be correctly identified. While only missing 2% of the positive cases, this SRHO/LD system could reduce the number of false positives by 53%.

It should be noted, that since the overall observer system was trained on hand selected, centered ROIs, that this is what the system presented here will be best at finding. Differences in mass size, shape, and spiculation may, in fact, reduce system performance, since only one "average" observer is created. Some of the mass differences should be taken into account by the observer by the covariance matrix and this effect should be somewhat reduced. An additional study on training on one type of mass and testing on another would be instructive, but is beyond the scope of this introductory paper. Additionally, the fact that this system was trained on hand selected cases to be the equivalent of "false-positives" is a weakness. This study does, however, give a base line of performance and provide for an instrumental test of the system in the application to mass detection.

This type of highly sensitive classifier could easily be added to available CAD system to improve upon their current performance. The system could be used "as is" or could be retrained with computer selected suspicious masses to determine the overall real effect on false-positive reduction in a CAD setting. Future studies will do just this task.

In conclusion, our preliminary results suggest that using subregion Hotelling observers in conjunction with linear discriminant analysis can provide a successful classification scheme for the detection of masses. Further research will allow this approach to be incorporated into a larger computer aided detection system to aid mammographers with mass detection in the clinic.

[a] Electronic mail: alan.baydush@duke.edu

[1] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 1999 [see comments]," Ca-Cancer J. Clin. **49**, 8–31 (1999).

[2] ACS, "American Cancer Society: Cancer Facts and Figures 2002. Atlanta, Ga: American Cancer Society 2002," 2002.

[3] U. S. D. H. H. S, Healthy People 2010 (Conference Edition, in Two Volumes), Washington, DC, 2000.

[4] 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).

[5] R. A. Castellino, J. Roehrig, and W. Zhang, "Improved computer-aided detection (CAD) algorithms for screening mammography," Radiology **217**, 400 (2000).

[6] A. H. Baydush, D. M. Catarious, Jr., J. Y. Lo, C. K. Abbey, and C. E. Floyd, Jr., "Computerized classification of lung nodules in chest radiographs using sub-region Hotelling observers," Med. Phys. **28**, 2403–2409 (2001).

[7] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, edited by A. B. Watson (MIT, Cambridge, MA, 1993).

[8] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," J. Opt. Soc. Am. A **14**, 2379–2391 (1997).

[9] A. B. Watson, "Detection and recognition of simple spatial forms," in *Physical and Biological Processing of Images*, edited by O. J. S. a. A. J. Seigh (Springer-Verlag, Berlin, 1983), pp. 100–114.

[10] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity quality," in *Digital Images and Human Vision*, edited by A. B. Watson (MIT, Cambridge, MA, 1993).

[11] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," Proc. Natl. Acad. Sci. U.S.A. **90**, 9758–9765 (1993).

[12] M. Eckstein, C. Abbey, and J. Whiting, "Human vs model observers in anatomic backgrounds," Medical Imaging 1998: Image Perception, 1998, pp. 16–26.

[13] R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, "Psychophysical study to test the ability of Th Hotelling trace criterion to predict human-performance," J. Opt. Soc. Am. A Opt. Image Sci. Vis **3**, P126–P126 (1986).

[14] R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, "Hotelling trace criterion and its correlation with human-observer performance," J. Opt. Soc. Am. A Opt. Image Sci. Vis **4**, 945–953 (1987).

[15] H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging," J. Nucl. Med. **39**, 771 (1998).

[16] H. C. Gifford, R. G. Wells, and M. A. King, "A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images," IEEE Trans. Nucl. Sci. **46**, 1032–1037 (1999).

[17] H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging," J. Nucl. Med. **41**, 514–521 (2000).

[18] S. D. Wollenweber, B. M. W. Tsui, D. S. Lalush, E. C. Frey, K. J. LaCroix, and G. T. Gullberg, "Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging," IEEE Trans. Nucl. Sci. **46**, 2098–2103 (1999).

[19] H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao, "Linear discriminants and image quality," Image Vis. Comput. **10**, 451–460 (1992).

[20] H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, "Objective assessment of image quality. 2. Fisher information, Fourier crosstalk, and figures of merit for task-performance," J. Opt. Soc. Am. A Opt. Image Sci. Vis **12**, 834–852 (1995).

[21] W. E. Smith and H. H. Barrett, "Hotelling trace criterion as a figure of merit for the optimization of imaging-systems," J. Opt. Soc. Am. A Opt. Image Sci. Vis **3**, 717–725 (1986).

[22] H. H. Barrett, C. K. Abbey, and B. G. Gallas, "Stabilized estimages of Hotelling-Observer detection performance in patient-structured noise," SPIE Medical Imaging 1998: Image Perception, 1998, pp. 27–43.

[23] M. Nadler and E. P. Smith, *Pattern Recognition Engineering* (Wiley, New York, 1993), p. 588.

[24] M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the digital database for screening mammography," in *Digital Mammography*, edited by N. Karssemeijer, M. Thijssen, and J. Hendriks (Kluwer Academic, New York, 1998), pp. 457–460.

Computer aided detection of masses in mammography using a Laguerre-Gauss channelized Hotelling observer

Alan H. Baydush[a,b], David M. Catarious[b], Carey E. Floyd Jr.[b,c]

[a]Department of Radiation Oncology, Duke University Medical Center, Durham, NC 27710
[b]Department of Biomedical Engineering, Duke University, Durham, NC 27710
[c]Department of Radiology, Duke University Medical Center, Durham, NC 27710

ABSTRACT

We propose to investigate the use of a Laguerre-Gauss Channelized Hotelling Observer (LG-CHO) for the basis of a computer aided detection scheme for masses in mammography.

A database of 1320 regions of interest was selected from the DDSM database collected by the University of South Florida. The breakdown of the cases was: 656 normals, 307 benigns, and 357 cancers. For the detection task, cancer and benign cases were considered positive and normal was considered negative. A 25 channel LG-CHO was designed to best classify regions as containing a mass or not. Application of this LG-CHO to the database gave a ROC area under the curve of 0.936 and a partial area of 0.648. Additionally, at 98% sensitivity the classifier had a specificity of 44.8% and a positive predictive value of 64.2%

Preliminary results suggest that using a LG-CHO can provide a strong backbone for a CAD scheme to help radiologists with detection. These initial results should be able to be incorporated into a larger CAD system for higher performance either as a false positive reduction scheme or as an initial filter used for mass detection.

Keywords: Hotelling observer, computer aided detection, channelized Hotelling observer, mammography, masses.

1. INTRODUCTION

Cancer is a one of the most devastating and deadly disease of our time and is the second leading cause of death in the United States (US).[1] In 1999 alone, over 1.2 million persons in the US were diagnosed with cancer and it was estimated that approximately 563,100 persons would perish.[1] Breast cancer is the most common cause of death in women. A strong commitment to reducing deaths by cancer has been put forth by the Department of Health and Human Services. The prime method for detecting breast cancer is through screening mammography.[2] Early detection of suspicious regions in mammograms is vital to patient outcome and is key to improving patients long term care.

The development and application of image processing techniques for the automated detection of masses will greatly improve early detection. Preliminary results on commercial systems currently available have shown an increase in detection of cancer.[3,4] Once again, this improved early detection is vital to positive patient outcomes.

The long range goal of our group is to build tools which can be incorporated into a full fledged computer aided detection (CAD) system for improving mass detection in mammograms. This CAD system will help radiologists detect breast masses and will increase the chance of early detection of subtle masses. We firmly believe that development and application of CAD techniques for the automated detection of cancerous breast masses will have a great impact on early detection and hence on overall patient outcome.

Most CAD systems can be viewed as a two stage approach. The first stage uses some type of initial linear processing, with high sensitivity and low specificity, to detect a set of potential masses. The second stage consists of classifying these potential masses using predictive modeling techniques to reject a large number of false positives. Using this type of approach, systems have been developed commercially and investigated experimentally from several institutions. These systems have shown great success and hold even more promise in the future. However, some issues still exist with the use of CAD systems in clinical practice. The chief complaint of radiologists and mammographers on CAD systems, such as these, is the number of false positives that the system retains. If too many false positives are reported,

the radiologist not only loses faith in the system, but also loses valuable time. Therefore any efforts into developing new techniques which can be used to reduce the number of false positives should be well received.

Since the system we are trying to model is the radiologist, we have chosen to investigate the incorporation of Hotelling observers into our CAD system. In past research, Hotelling observers have been shown to effectively track human observer performance.[5-10] Specifically, we wish to investigate incorporating Laguerre-Gauss channelized Hotelling observers into the second stage of the CAD system to help improve in false positive reduction. We have pursued this approach previously in chest radiography[11] with great success and wish to extend this experimentation into mammography. Our hypothesis is that incorporating models from vision science into the classification process will help to reduce the number of false positives while not reducing sensitivity.

2. MATERIALS AND METHODS

This section will overview the creation of our region of interest database, background information on Hotelling observers, a description of the channelized Hotelling observer, a description of Laguerre-Gauss channelized Hotelling observers, and finally, the general procedure used in this study.

2.1 Region of Interest Database

To begin our study, we needed to come up with a database of regions of interest (ROIs). This database would be used to train and test our developed observer. Only ROIs and not full size images are needed, as we are envisioning the use of our system to help reduce the number of suspicious regions that remain after initial detection by a CAD system. Because we are focusing of false positive reduction, only a database of potential positive regions which have already been detected is needed.

To create our database, we looked towards the Digital Database for Screening Mammography (DDSM)[12] database collected by the University of South Florida. To further limit the size of the database, we chose to use only images digitized by the Lumisys scanner (digitized at 50 micron). A search was performed on the DDSM database (Lumisys cases) to determine which cases had masses where we could extract a 1k by 1k pixel ROI without going outside the image. It was this subset of cases which we used in this study. The 1k by 1k pixel ROIs were extracted from the viable cases with the mass lesion being centered in the ROI. All of the ROIs were then spatially averaged down to a size of 128 by 128 pixels. This set constituted the set of positive masses. To create a set of normal cases, a similar procedure was followed, except normal DDSM images were used, i.e. images with no abnormality present.

Using the above criteria, a ROI database of 1320 regions was selected. The breakdown of the cases was as follows: 656 normal ROIs, 307 benign ROIs, and 357 cancer ROIs.

Since we are investigating a detection task, cancer and benign cases both constitute being masses and were considered positive. Cases without any abnormalities are normal and were considered negative.

2.2 Hotelling Observers

We wish to continue examining incorporating models from the human vision system into the classification stage of CAD systems. The Hotelling observer (HO) is a mathematical observer which should effectively discriminate between a two class system. The HO incorporates information about the signal, the background, and noise correlation for prediction of positive and negative classes. In white noise, the HO reduces to a matched filter. However, in medical images, which have correlated noise, the observer estimates a template to decorrelate the noise.[13] Additionally, HOs have been shown to be effective in tracking the performance of human observers for detection[5-10] and as a means for measuring image quality.[14-18]

Mathematically, the HO is a set of weights that can be applied to an image to give an output test statistic and this statistic should separate the classes optimally. The weights or template for the HO are defined as:

$$W = [<S+B> - <B>]/K \tag{1}$$

Where S is the signal, B is the background, S+B is the signal in the background, $<>$ represents the mean, and K is the covariance matrix. This covariance matrix should be the weighted mean of the signal and background covariance matrices. To get the output test statistic, L, we multiply these weights by the image data, I, and sum over all the pixels.

$$L = \sum W * I \qquad (2)$$

The test statistic should be larger when the signal is present and smaller when absent. Optimally, this output test statistic will divide the signal present and signal absent cases perfectly, but this rarely happens. To quantify the effectiveness of the observer in properly classifying the cases as signal present or absent, receiver operating characteristic (ROC) analysis is performed. The most common metric examined from ROC analysis is the area under the curve (AUC).

The HO has been shown to be the optimal detector when certain features of the data (signal, background, noise covariance) are known and are approximately Gaussian.[15]

Problematically, however, for real medical images, we do not know the exact signal or background. We therefore have to use estimates of the signal, the general background, and the covariance matrix to calculate the set of linear weights used in the HO. This makes the derived observer a sub-optimal observer. In practice, one uses the average positive signal, the average background signal, and the weighted covariance matrix. Direct application of the HO to a large region of interest (ROI) is prohibitive, as too many image samples are needed to estimate the covariance matrix.[18] For example, if a ROI of size 128 by 128 pixels, the covariance matrix would be of size 16k by 16k elements and would require approximately 5 to 10 fold that number to accurately assess the covariance matrix. This amount of "real„ data is intractable so alternative solutions are necessary.

2.3 Channelized Hotelling Observers
To reduce the number of data samples that are necessary, people have attempted to reduce the dimensionality of the HO. This has been done by applying channelized models to the Hotelling observer to create a channelized Hotelling obsever (CHO).[19] Theoretically, a channel model is used by applying channels to the input data to reduce the dimensionality of the data. Generally speaking, a system of radially symmetric channels is chosen for simplification. Each of the channels would be applied to the data to give a single output, usually by frequency averaging certain expected important frequency bands. These different outputs from each of the channels is then used as the input of a HO, as described above.

This type of CHO reduces the dimensionality of the covariance matrix to the number of channels by the number of channels. For instance, in the case above, for ROIs of size 128 by 128, the covariance matrix is size 16k by 16k. If a 10 channel model is used, the covariance matrix is reduced to 10 by 10. This massive reduction in dimensionality of the covariance matrix allows for the estimation problem to now be tractable with a reasonable sized data set of images.

2.4 Laguerre-Gauss Channelized Hotelling Observer Features
Now that we know we are going to be using a CHO model, the question arises as to what channel basis functions should be used. Barrett et al[20] suggest that since most HOs are smooth, smooth functions should be more favored over non-smooth. Additionally since the objects we are aiming to detect are on average, generally round, a radially symmetric basis should be used. Following Barrett's work, we have also chosen to use a family of functions based on Laguerre-Gauss (LG) channels. LG channels are formed as the product of Laguerre polynomials and Gaussians. Laguerre polynomials are defined as:

$$L_n(x) = \sum_{m=0}^{n} (-1)^n \binom{n}{m} \frac{x^m}{m!}, \text{ where } \binom{n}{m} = \frac{n!}{(n-m)!m!}. \qquad (3)$$

Multiplying these Laguerre polynomials with Gaussians gives LG channels. Each channel is then multiplied by an appropriate channel weight (α_n) determined by applying a HO to the channels, and the sum of all the channels is taken to form the final LG-CHO template, w. In polar coordinate notation, the final template looks like:

$$w(r) = \sum_n \alpha_n \exp\left(\frac{-\pi r^2}{a^2}\right) L_n\left(\frac{2\pi r^2}{a^2}\right) \qquad (4).$$

Here, n is the number of channels. Figure 1(a) shows a 3D representation of a sample LG-CHO template, while figure 1(b) shows a profile through the midline to better show details.
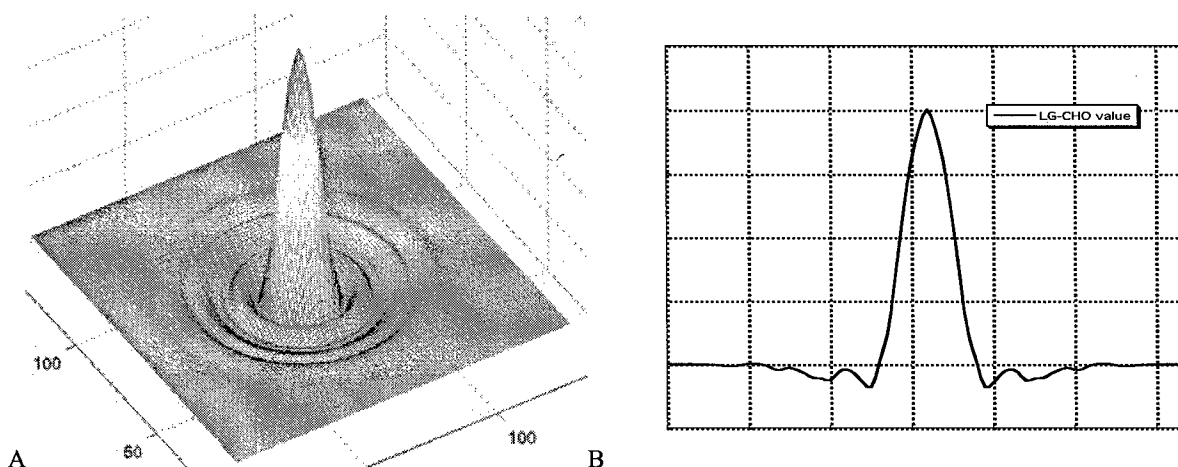
Figure 1: 3D (A) and 2D (B) representation of a 25 channel LG-CHO template. The 2D representation is taken as a slice through the mid-plane of the 3D version.

2.5 Procedure

For the study presented here, a system of Laguerre-Gauss symmetric channels was used for the basis of the CHO. Each of the LG channel templates was applied to each of the members of the ROI database to determine a channel response. Average positive and negative responses and covariance matrices were then calculated. This information was used to determine the weights for the HO. These weights were then applied to the channels and a single LG_CHO template was formed. The response of each ROI to the template was then determined and an output test statistic was calculated. These test statistics were sampled via bootstrap techniques to determine the ROC area under the curve performance metric along with its variance. A variety of channel numbers were empirically tested to maximize the ROC area under the curve.

3. RESULTS

An empirical search methodology was used to determine the optimal number of channels for the ROI database which we used for this study. The maximal ROC area under the curve was determined to occur with 25 channels. Using this number of channels, a LG-CHO template was determined and applied to each of the regions in the ROI database. The responses were collected and analyzed via bootstrap and ROC techniques to determine system performance. The LG-CHO system gave a ROC area under the curve of 0.936 and a partial area under the curve (the normalized area above 90% sensitivity on the ROC curve) of 0.648. Additionally, at 98% sensitivity the overall classifier had a specificity of 45% and a positive predictive value of 64.2%.

Table 1 shows specificities and positive predictive values (PPV) for 90%, 95%, and 98% sensitivity.

Figure 2 shows the ROC output curve for the LG-CHO system.

Sensitivity	Specificity	PPV
90%	82%	83%
95%	73%	78%
98%	45%	64%

Table 1: Specificities and positive predictive values for the LG-CHO at different sensitivities.
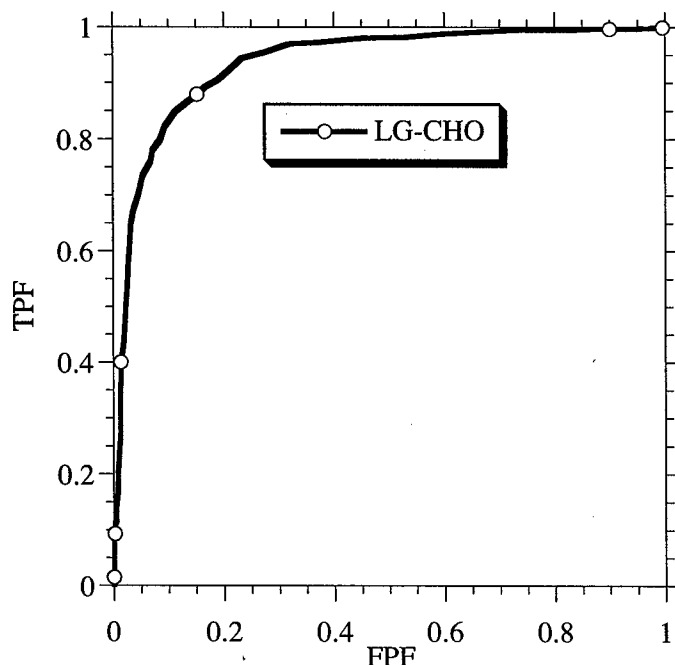
Figure 2: ROC output curve for the overall LG-CHO.

4. DISCUSSION

The goal of this study was to investigate the use of a Laguerre-Gauss channelized Hotelling observer for the automated classification of regions as either containing or not containing a mammographic mass. Our goal was not to determine benign from malignant masses (diagnosis), although similar techniques could be used to perform that study. For the study presented here, a large ROI database was generated from the image cases in the DDSM database. 664 of these cases were positive for having a mass present, while 656 were taken from normal images, so no mass was present.

The LG-CHO system performed quite well for mass detection on our database. The ROC AUC for the classification task was 0.936, which corresponds to a partial AUC of 0.648 +/- 0.028. We calculated the specificity of the system at 95% sensitivity to be 73%. At this threshold setting, 33 positive cases would be missed, while 479 of the 656 negative regions would be correctly identified as negative. Additionally, at 98% sensitivity (13 missed positives), 295 of the 656 negative regions would be correctly identified. This type of highly sensitive classifier could very easily be added to available CAD system to improve upon their current performance.

5. CONCLUSIONS

Preliminary results suggest that using a Laguerre-Gauss channelized Hotelling observer can provide a strong backbone for a CAD scheme to help radiologists with detection. These initial results should be able to be incorporated into a larger CAD system for higher performance either as a false positive reduction scheme or as an initial filter used for mass detection.

ACKNOWLEDGEMENTS

REFERENCES

1. S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 1999 [see comments],,, *Ca: a Cancer Journal for Clinicians* **49**, 8-31, 1, 1999.

2. U. S. D. H. H. S, "Healthy People 2010 (Conference Edition, in Two Volumes),,, Washington, DCWashington, DC, 2000.

3. T. W. Freer, and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center,,, **220**, 781-786, 2001.

4. R. A. Castellino, J. Roehrig, and W. Zhang, "Improved Computer-aided Detection (CAD) Algorithms for Screening Mammography,,, **217(P)**, 400, 2000.

5. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, "Psychophysical Study to Test the Ability of Th Hotelling Trace Criterion to Predict Human-Performance,,, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **3**, P126-P126, 1986.

6. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, "Hotelling Trace Criterion and Its Correlation With Human- Observer Performance,,, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **4**, 945-953, 1987.

7. H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging,,, *J. Nucl. Med.* **39**, 771, 1998.

8. H. C. Gifford, R. G. Wells, and M. A. King, "A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images,,, *IEEE Trans. Nucl. Sci.* **46**, 1032-1037, 1999.

9. H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, "Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging,,, *J. Nucl. Med.* **41**, 514-521, 2000.

10. S. D. Wollenweber, B. M. W. Tsui, D. S. Lalush, E. C. Frey, K. J. LaCroix, and G. T. Gullberg, "Comparison of hotelling observer models and human observers in defect detection from myocardial SPECT imaging,,, *IEEE Trans. Nucl. Sci.* **46**, 2098-2103, 1999.

11. A. H. Baydush, D. M. Catarious, Jr, J. Y. Lo, C. K. Abbey, and C. E. Floyd, Jr, "Computerized classification of lung nodules in chest radiographs using sub-region hotelling observers,,, *Med Phys* **28**, 2403-9, 2001.

12. M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the Digital Database for Screening Mammography,,, Digital Mammography, edited by N. Karssemeijer, M. Thijssen, and J. Hendriks. Kluwer Academic Publishers, 1998, 457-460.

13. M. Eckstein, C. Abbey, and J. Whiting, "Human vs model observers in anatomic backgrounds,,, in Medical Imaging 1998: Image Perception, p.16-26, 1998.

14. H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao, "Linear Discriminants and Image Quality,,, *Image Vis. Comput.* **10**, 451-460, 1992.

15. H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model Observers For Assessment of Image Quality,,, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9758-9765, 1993.

16. H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, "Objective Assessment of Image Quality .2. Fisher Information, Fourier Crosstalk, and Figures of Merit For Task-Performance,,, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **12**, 834-852, 1995.

17. W. E. Smith, and H. H. Barrett, "Hotelling Trace Criterion As a Figure of Merit For the Optimization of Imaging-Systems,,, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **3**, 717-725, 1986.

18. H. H. Barrett, C. K. Abbey, and B. G. Gallas, "Stabilized estimages of Hotelling-Observer detection performance in patient-structured noise,,, in SPIE Medical Imaging 1998: Image Perception, p.27-43, 1998.

19. K. J. Myers, and H. H. Barrett, "Addition of a Channel Mechanism to the Ideal-Observer Model,,, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **4**, 2447-2457, 1987.

20. H. H. Barrett, C. K. Abbey, and B. Gallas, "Stabilized Estimates of Hotelling-Observer Detection Performance in Patient-Structured Noise,,, *Proc. SPIE* **3340**, 27 - 42, 1998.

Novel use of the Hotelling observer for computer aided diagnosis of solitary pulmonary nodules

Alan H. Baydush* and David M. Catarious, Jr.

Department of Radiology, Duke University Medical Center, Durham, NC, USA.
Department of Biomedical Engineering, Duke University, Durham, NC, USA.

ABSTRACT

We propose to investigate a novel use of the Hotelling observer for the task of discrimination of solitary pulmonary nodules from a database of regions that were all deemed suspicious . A database of 239 regions of interest (ROIs) was collected from digitized chest radiographs. Each of these 256x256 pixel ROIs contained a suspicious lesion in the center for which we have a truth file. For our study, 25 separate Hotelling observers were set up in a 5x5 grid across the center of the ROIs. Each separate observer was designed to observe a 15x15 pixel area of the image. Leave-one-out training was used to generate 25 output observer features. These 25 features were then narrowed down using a sequential forward searching linear discriminant analysis. The forward search was continued until the accuracy declined at 13 features and the subset was used as the input layer to an artificial neural network (ANN). This network was trained to minimize mean squared error and the output was the area under the ROC curve. The trained ANN gave an ROC area of .86. In comparison, three radiologists performed at ROC area indexes of .72, .79, and .83.

Keywords: CAD, Lung Nodules, Hotelling Observer, Image Processing

1. INTRODUCTION

Goal: We propose to investigate a novel use of the Hotelling observer for the basis of a computer aided diagnosis (CAD) scheme for the task of discrimination of solitary pulmonary nodules from a database of regions that were all deemed suspicious .

Cancer is one of the most devastating diseases of our time. In 1999, over 1.2 million people in the US were diagnosed with cancer.[1] Lung cancer accounts for about 28 percent of all cancer deaths and estimates show that over 158,000 persons will die from this disease.[1] The prime method for detection of cancer is radiological exams[2], of which, the simplest is the chest x-ray. It has been shown that a radiologist may miss up to 30% of pulmonary nodules in a x-ray image. Since early detection of lung cancer so significantly improves patient outcome, detection of these nodules is very important. The development and use of computer aided detection systems in conjunction with radiologists has been shown to improve detection performance.[3,4]

The goal of the initial study presented here is to begin development of an innovative detection tool for aiding the radiologist in determining if a suspicious region is a pulmonary nodule. This preliminary proposal focuses on investigating the diagnostic accuracy of a combination linear and non-linear classifier to perform the discrimination of pulmonary nodules from suspicious regions.

2. BACKGROUND

Most human sensory processes are understood to work by a linear step followed by a non-linear step for decision tasks. In the case of the visual processing system, the linear step is the receptive fields which process basic visual stimuli and are used to reduce data complexity. This linear step is followed by a non-linear combination of the important data to determine decisions. This multi-layered process is what we have chosen to model and investigate in this study. The process, as we see it, reduces to a 3 layer classification scheme. The first layer models the linear portion of the visual system. We have chosen to use the Hotelling trace observer for this layer. The second layer models the data reduction in the visual process and will be performed using linear discriminant analysis (LDA). The third layer, the non-linear combination of the reduced complexity data, will be performed using an artificial neural network (ANN) for the final classification.

---

* Correspondence: Email: alan.baydush@duke.edu; Phone: (919) 684-2691; Box 2623, Durham, NC 27710.

First, we would like to present some background material on HOs and ANNs before we get into the specifics of this proposal.

2.1. Small region of interest Hotelling observers

The Hotelling trace observer, sometimes just known as the Hotelling observer (HO), is the optimal linear detector for a known signal, known background, and known covariance matrix when statistics are approximately Gaussian. This optimal detector has been shown to be effective in tracking the performance of human observers for detection.[5-10] Many researchers have also used the HO as a means of measuring image quality or as an imaging metric.[11-14] The HO uses information about the signal to be detected, the background, and the image covariance matrix to calculate a set of linear weights. The covariance matrix is a matrix of elements where each element is the covariance between 2 pixels and the diagonal is the variance for each pixel. For real medical images, we do not know these features, so we have to use estimates of the signal to be detected, the general background, and the image covariance matrix to calculate the set of linear weights for the then sub-optimal observer. The weights or template for the HO are defined as

$$W = [<S+B> - <B>]/K, \qquad (1)$$

where $<>$ represents the mean, S is the signal, B is the background, S+B is the signal in the background, and K is the covariance matrix. Multiplying these weights by the image data, I, and summing over all the pixels, p, gives the test statistic,

$$L = \sum W_p * I_p \qquad (2)$$

This test statistic can be used as a decision variable. It will be higher in value when the signal is present and lower when it is absent. In white noise, the HO is a matched filter; however, in correlated noise, such as in medical images, this observer estimates a template that decorrelates the noise.[15]

Application of the HO to a large region of interest (ROI) is prohibitive, as too many image samples would be needed to properly estimate the covariance matrix. For instance, in the database we have developed, the 256 by 256 pixel ROIs would require a covariance matrix of size 65,536 (256x256) by 65,536 (256x256) elements. Collecting a database of real images large enough to obtain a stable estimate of a covariance matrix of this size would prove to be overly difficult.

To combat this size difficulty, many researchers have investigated using a channelized HO model, where radially symmetric vision channels are used to reduce the dimensionality of the problem. Initially we tried this approach, only to find that it did not work well for lung nodule detection. We felt that this failure was due to neither the normal anatomy nor the nodule signal in the lungs to be radially symmetric. Deciding to relax this radial symmetry constraint caused us to re-think the pixel-wise HO.

We then decided to use many small region of interest Hotelling observers (SRHO), because a small region observer would require significantly less samples to properly estimate the necessary covariance matrix. Our proposal was to tile a small matrix of small observers over the full region of interest we wished to examine. This will result in many SRHO being used to reduce the complexity of the image data; however, each small observer will be observing a portion of the full resolution image. We chose not to sub-sample the image data as we felt that the HO would be able to model and incorporate the image texture into the covariance matrix. By doing this, we hoped to maintain the sensitivity to the high frequency content of the image. These small observers will be sensitive to changes in high frequency noise power spectra and structured noise, including anatomy. The result of applying these many SRHO would be a matrix of outputs or features, one list of features for each SRHO used.

These output features, the output of the small individual Hotelling classifiers, will then be examined by analysis (LDA, neural network) to further reduce the dimensionality of the problem. The final reduced set of features/classifiers will then be combined using a non-linear ANN to determine the final decision as to if a region should be classified as a pulmonary nodule or not. In essence, the adoption of a multi-layered approach allows not having to lose the high frequency content, which we feel plays an important role in nodule classification.

2.2. Artificial neural networks

The methods of developing the artificial neural network models which we will use have been described in the previous studies from our lab and will only be summarized here. The multi-layer ANNs use a three layer (one hidden layer), feed-forward, error-backpropagation ANNs. When a perceptron is used, no hidden layer is incorporated into the network. Each ANN is presented with the input findings for each case and the corresponding known truth outcome. The ANN merges all the findings nonlinearly to generate a single output value between zero and one corresponding to its prediction of the

likelihood of a nodule being present for that case. The ANN is trained and learns iteratively under this supervised training process in order to improve its performance.

A "round robin" or "leave-one-out" sampling scheme is utilized in order to use all cases for training and testing while still maintaining independence between the training and testing sets. Network training can be halted when the ROC area index, A_z, is maximized over the testing cases. Our custom ANN software was written in the C language and runs on Sun Ultra 60 workstations (Sun Microsystems Inc., Mountain View, Ca.). Initial training requires up to several minutes for each new combination of parameters, but a finalized ANN can evaluate each new case within a fraction of a second.

As stated previously, for each case the model produces as its prediction a number between zero and one. To use the ANN as a diagnostic aide, one could select a certain threshold value, such that those cases with output values below the threshold would be considered probably not being a nodule. The remainder of cases with values exceeding the threshold would be considered a pulmonary nodule. The sensitivity is the number of correctly classified nodules divided by the number of all actual nodules; the specificity is the number of correctly diagnosed negative lesions out of all actual negative lesions. Varying this threshold value results in a trade off between sensitivity and specificity and will generate an ROC curve for analysis.

3. METHODS

Here is a summary of the methods for this study:

3.1. Image database

We have previously[16] collected a database of 239 ROIs for nodule classification and detection studies. Each ROI is 256 pixels by 256 pixels. For the purposes of this database, a nodule was defined as any lesion that represented a tumor or granuloma (calcified or noncalcified). All of the original images were taken between 1991 and 1996. A truth file was prepared by two board certified radiologists for the digitized 2048 pixel by 2048 pixel images based on the PA radiograph, CT results when applicable, the full radiology report, and the pathology report when applicable. Overall, the database consists of 94 negative pulmonary nodule ROIs and 145 positive ROIs. Please note that for this database, all of the negative regions were deemed suspicious for a nodule upon initial examination by the radiologist, which makes this a very difficult database.

In addition to having this database, 3 radiologists have performed a ROC study over all of the images by selecting a probability of the region being a nodule for 237 of the regions in this database (2 regions were used for training). Analysis of the radiologists ROC ratings yielded areas which ranged from .72 to .83. This level of radiologist performance for area under the ROC curve corresponds well to other studies of lung nodule databases for sets of cases which were deemed to be at a level of complexity of very subtle (.753) to subtle (.876).[17]

3.2. SRHO

For our study, 25 separate Hotelling observers were set up in a 5x5 grid across the center of the full size ROIs. The Hotelling observers were set up in a matrix and numbered as shown in Figure 1. Each separate observer was designed to observe or discriminate a 15 x 15 pixel area of the image, thus the 25 sub regions cover the 75 x 75 pixel center of the ROI. A leave-one-out training and testing methodology was used to generate 25 features, where each feature is the output of the individual observers. Signal and background were modeled as the average of the positives and negatives, respectively, and the covariance matrix was calculated over the images to be trained on.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Figure 1. Feature numbering matrix for the 5x5 grid of SRHO covering the center 75x75 pixels of each ROI.

3.3 LDA and ANN

These 25 features derived from separate Hotelling observers were then narrowed down by using a sequential forward searching linear discriminant analysis (LDA) where percent correct (total number of correct identifications over total number) was used as the performance metric. The forward search was continued until the accuracy started to decline and then the chosen subset was used as the input layer to a three layer artificial neural network (ANN). This network was trained to minimize mean squared error and the output was the area under the curve given by receiver operating characteristic (ROC) analysis. Once again, a leave-one-out methodology was incorporated into the training and testing of the ANN. We should note that our ANNs were not trained with optimal training weights nor optimal training iterations, so our result presented here could improve.

4. RESULTS

The 25 Hotelling observer features, as laid out in Fig. 1, were searched by the LDA. The output of the LDA is shown in table 1, where each region, in order of importance cumulatively is shown. For each region, the independent accuracy is shown as well as the cumulative accuracy, based on using that region and the previous regions. A maximal percent correct of 76.6% was reached using 13 of the 25 features. This subset of 13 features was then used as the input layer into the ANN, which when trained gave out a ROC area of .86.

Order Selected	Region	Independent	Cumulative
1	5	0.6402	0.6402
2	10	0.5774	0.682
3	13	0.6192	0.6946
4	17	0.5481	0.7029
5	8	0.5732	0.6987
6	9	0.5941	0.7029
7	22	0.6318	0.7113
8	19	0.6318	0.7197
9	6	0.6234	0.7322
10	12	0.6109	0.749
11	15	0.5732	0.749
12	7	0.5983	0.7531
13	20	0.6025	0.7657
14	24	0.6234	0.7615
15	1	0.523	0.7573
16	23	0.6276	0.7448
17	3	0.5774	0.7531
18	2	0.59	0.7364
19	4	0.5732	0.728
20	16	0.5774	0.728
21	25	0.4895	0.7197
22	11	0.6234	0.7155
23	18	0.5565	0.7071
24	14	0.5607	0.6904
25	21	0.6192	0.6946

Table 1. Table showing independent and cumulative accuracy (percent correct) for each of the 25 features as the LDA searched through the set. A maximum is reached at 76.6% at 13 features selected.

5. DISCUSSION

This work represents an initial study into using small regions of an image as the input to Hotelling observers to obtain image features. These features were then reduced using LDA. The reduced set was then fed into an ANN to perform the task of CAD for the ROI database we have collected. 25 features (15x15 sub regions) were calculated using the SRHO technique. LDA was used to determine when accuracy started to decline by adding more features. A subset of 13 features gave the highest percent correct. These 13 features were then used as the input layer to an ANN. The trained ANN gave an ROC area index of .86. For comparison, the three radiologists who had performed this same ROC study on these ROIs had areas of .72, .79, and .83.

Preliminary results suggest that using sub region Hotelling observers in combination with ANNs can provide a strong backbone for a CAD scheme to help radiologists with diagnostic decisions. Our initial results already compare well to radiologists performance for the classification of suspicious regions for pulmonary nodules.

6. CONCLUSIONS

The immediate benefit of this proposal is to develop the ground work for a highly accurate computer-aided diagnosis system for pulmonary nodule classification which would be using a very different approach then what has been used historically in the field. This ground work should yield enough preliminary results and validation to support continuing this project on a larger scale and building such a system to assist the radiologist.

REFERENCES

1. S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, Cancer statistics, 1999 [see comments], *Ca: a Cancer Journal for Clinicians* **49**, 8-31, 1, 1999.
2. K. Shaffer, Role of radiology for imaging and biopsy of solitary pulmonary nodules, *Chest* **116**, 519S-522S, 1999.
3. T. Kobayashi, X. W. Xu, H. MacMahon, C. E. Metz, and K. Doi, Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs, *Radiology* **199**, 843-8, 1996(96220553).
4. M. L. Giger, K. Doi, H. MacMahon, C. E. Metz, and F. Yin, Pulmonary nodules: Computer-aided detection in digital chest images, *RadioGraphics* **10**, 41-51, 1990.
5. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, Psychophysical Study to Test the Ability of Th Hotelling Trace Criterion to Predict Human-Performance, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **3**, P126-P126, 1986.
6. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, Hotelling Trace Criterion and Its Correlation With Human-Observer Performance, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **4**, 945-953, 1987.
7. H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging, *J. Nucl. Med.* **39**, 771, 1998.
8. H. C. Gifford, R. G. Wells, and M. A. King, A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images, *IEEE Trans. Nucl. Sci.* **46**, 1032-1037, 1999.
9. H. C. Gifford, M. A. King, D. J. de Vries, and E. J. Soares, Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging, *J. Nucl. Med.* **41**, 514-521, 2000.
10. S. D. Wollenweber, B. M. W. Tsui, D. S. Lalush, E. C. Frey, K. J. LaCroix, and G. T. Gullberg, Comparison of hotelling observer models and human observers in defect detection from myocardial SPECT imaging, *IEEE Trans. Nucl. Sci.* **46**, 2098-2103, 1999.
11. H. H. Barrett, T. Gooley, K. Girodias, J. Rolland, T. White, and J. Yao, Linear Discriminants and Image Quality, *Image Vis. Comput.* **10**, 451-460, 1992.
12. H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, Model Observers For Assessment of Image Quality, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9758-9765, 1993.
13. H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, Objective Assessment of Image Quality .2. Fisher Information, Fourier Crosstalk, and Figures of Merit For Task-Performance, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **12**, 834-852, 1995.
14. W. E. Smith, and H. H. Barrett, Hotelling Trace Criterion As a Figure of Merit For the Optimization of Imaging-Systems, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **3**, 717-725, 1986.
15. M. Eckstein, C. Abbey, and J. Whiting, Human vs model observers in anatomic backgrounds, in Medical Imaging 1998: Image Perception, 1998.
16. J. A. Drayer, N. Vittitoe, R. Vargas-Voracek, A. H. Baydush, C. E. Floyd, Jr, and C. E. Ravin, Characteristics of regions suspicious for pulmonary nodules on chest radiographs, *Acad Radiol* **5**, 613-9, 1998.

17. J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, *Am. J. Roentgenol.* **174**, 71-74, 2000.

Incorporation of a Laguerre-Gauss channelized Hotelling observer into a mammographic mass CAD system

Alan H. Baydush[a], David M. Catarious[b], Carey E. Floyd Jr.[b,c]

[a]Department of Radiation Oncology, Duke University Medical Center, Durham, NC 27710

[b]Department of Biomedical Engineering, Duke University, Durham, NC 27710

[c]Department of Radiology, Duke University Medical Center, Durham, NC 27710

ABSTRACT

Previously, we have developed and tested a Laguerre-Gauss channelized Hotelling observer (LG-CHO) for mass detection. This previous work optimized and used the LG-CHO on a database of regions of interest (ROIs) that had been selected from a mammographic image database derived from the DDSM. Positive images contained masses (malignant and benign) and negative cases contained normal tissue. Additionally, we have also re-optimize the LG-CHO on the data from the initial detection stage of a CAD system we are developing, thus incorporating computer selected false positives into the training set. For the study presented here, we will incorporate the optimized observer results as an additional feature to be used in a false positive reduction stage in a CAD system we are developing. The resultant performance of this re-optimized system will be compared with the previous performance. Results are expected to show increased ability of the system to properly classify CAD suspicious regions as positive or negative.

Keywords: Hotelling observer, computer aided detection, channelized Hotelling observer, mammography, masses.

1. Introduction

Early detection of suspicious regions in mammograms is vital to patient. The development and use of Computer Aided Detection (CAD) systems has shown an increase in detection of cancer (Castellino, Roehrig et al. 2000; Freer and Ulissey 2001). The long range goal of our group is to build tools which can be incorporated into CAD systems for improving the detection of suspicious masses in mammograms.

Our CAD system (Catarious, Baydush et al. 2004), as developed so far, consists of a six stage approach. The first stage is initial filtration with a difference of Gaussian (DOG) filter. This filter has been empirically determined and is applied using normalized cross correlation. The second stage is suspicious region localization, where a thresholding technique has been applied and regions are not allowed to grow into one another. The third stage is suspicious region segmentation, which uses an iterative, linear classifier to determine inside and outside pixels. The fourth stage is feature extraction, followed by feature selection. The last stage is classification and false positive reduction. Here, we discuss the development and incorporation of a Laguerre-Gauss channelized Hotelling observer (LG-CHO) as an additional feature to be used in stages four through six of our developing CAD system in hopes that it can be used to help further reduce false positives and improve the performance of the overall system.

2. Materials and Methods

2.1 Hotelling Observers

The Hotelling observer (HO) is a mathematical construct, which should discriminate a two class system. The HO incorporates information about the signal, the background, and noise

correlation for prediction of class. In correlated noise, the observer estimates a template to decorrelate the noise (Eckstein, Abbey et al. 1998) which improves its effectiveness. HOs have been shown to track the performance of human observers for detection tasks (Fiete, Barrett et al. 1986; Fiete, Barrett et al. 1987; Gifford, King et al. 1998; Gifford, Wells et al. 1999; Wollenweber, Tsui et al. 1999; Gifford, King et al. 2000).

Mathematically, the HO is a set of weights that can be applied to an image to give an output test statistic. This test statistic should separate the classes optimally. The weights for the HO are:

$$W=[<S+B> - <B>]/K \tag{1}$$

Where S is the signal, B is the background, S+B is the signal in the background, $<>$ represents the mean, and K is the covariance matrix. To get the output test statistic (L) we take the dot product of the weights and the image data (I). The test statistic should divide the signal present and signal absent cases perfectly, but this rarely happens in realistic cases. The HO has been shown to be the optimal detector when certain features of the data (signal, background, noise covariance) are known and are approximately Gaussian (Barrett, Yao et al. 1993). Problematically, we do not know the exact signal or background for medical images. We therefore use estimates and these estimates reduce the performance of the HO. Additionally, direct application of the HO to a large region of interest (ROI) is prohibitive, as too many image samples are needed to estimate the covariance matrix (Barrett, Abbey et al. 1998).

2.2 Channelized Hotelling Observers

Channelized Hotelling observers (CHO) (Myers and Barrett 1987) are created by applying some type of channels to the input data to reduce the dimensionality. Generally speaking, a system of

radially symmetric channels is chosen for simplification. Each of the channels is applied to the data to give a single output. These different channel outputs are then used as the input of a HO, as described above. This type of CHO reduces the dimensionality of the covariance matrix to the number of channels by the number of channels. This massive reduction in dimensionality of the covariance matrix allows for the estimation problem to now be tractable with a reasonable sized data set of images.

2.3 Laguerre-Gauss Channelized Hotelling Observer

For the study presented here, we have followed Barrett's work and have chosen to use a family of functions based on Laguerre-Gauss (LG) channels. LG channels are formed as the product of Laguerre polynomials and Gaussians. Laguerre polynomials are defined as:

$$L_n(x) = \sum_{m=0}^{n} (-1)^n \binom{n}{m} \frac{x^m}{m!}, \text{ where } \binom{n}{m} = \frac{n!}{(n-m)! \, m!}. \tag{2}$$

Multiplying these Laguerre polynomials with Gaussians gives LG channels. Each channel is then multiplied by an appropriate channel weight (α_n) determined by applying a HO to the channels, and the sum of all the channels is taken to form the final LG-CHO template. In polar coordinate notation, the final template, w, looks like:

$$w(r) = \sum_n \alpha_n \exp\left(\frac{-\pi r^2}{a^2}\right) L_n\left(\frac{2\pi r^2}{a^2}\right) \tag{3}.$$

Here, n is the number of channels. Figure 1 shows a 3D representation of a sample LG-CHO template.
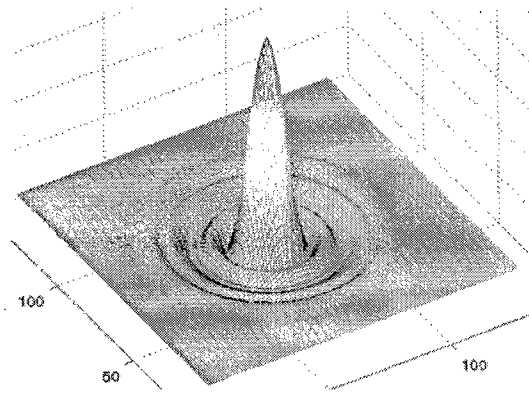
Figure 1: 3D plot representation of a sample 25 channel LG-CHO template.

2.4 Image Database

The mammograms that were used for this study were extracted from the University of South Florida's Digital Database for Screening Mammography (DDSM) (Heath, Bowyer et al. 1998). 183 images from 169 patients were pulled from the DDSM. Specifically, 83 images contained 50 benign and 50 malignant masses and 100 "normal" images contained no abnormalities. The images were chosen from the set scanned with a Lumisys scanner at a resolution of 50 microns per pixel at a bit depth of 12, but were resampled to 200 micron per pixel. Even though the images in the study database were randomly selected, the distribution of mass descriptors closely matched that of the entire collection of masses.

Since we are investigating a detection task, a positive detection is considered if either a cancer or a benign mass is correctly identified. Cases without any abnormalities are normal and were considered negative. Results are shown for detection, as presented above, and for classification, which is only the detection of malignant masses.

2.5 Procedure

In a previous study (Baydush, Catarious et al. 2003), we used a hand selected region of interest (ROI) database to train and test the LG-CHO. While the receiver operating characteristic (ROC) results for that study were promising, we realized we needed to train the observer on computer selected false positives and test the observer in a full CAD system. For the study presented here, an image database, described above, was used as input to our CAD system and the CAD system was used to perform stages one through three, as detailed above. At this point, ROIs of all the suspicious regions were extracted based on their centroid location, determined from the segmentation output. The initial sensitivity of the system was ~98% of the malignant masses with approximately 9.76 false positives per image (FPpI). These ROIs were used to train and test the LG-CHO observers. The response of each ROI to the template was then determined and an output test statistic was calculated. These test statistics were analyzed and a variety of channel numbers and channel parameters were empirically tested to maximize the ROC area under the curve. The LG-CHO with the best overall area under the curve was chosen to be used.

This LG-CHO was then applied to the entire image for each image in the database. Four features were calculated from the output of the normalized cross correlation within each suspicious region. The mean, standard deviation, peak value, and the value at the centroid were calculated for each suspicious region. These four new features were included into the set of features that were already measured by our system. Stages four through six of the system were then completed both with and without the incorporation of the four LG-CHO based features. FROC results were calculated.

3. Results

The LG-CHO which used 40 channels and an a value of 65 was shown to give the highest area under the curve results of 0.7625 with the training ROIs. This LG-CHO was used to generate four features which were included in the feature selection stage of the CAD system. Before the inclusion of these four features, the CAD system had selected the following features: average Haralick correlation, normalized radial length (NRL) spread, NRL change, and average Haralick sum average. These features had the following ROC areas: .83, .81, .80, and .76 respectively. With the inclusion of the LG-CHO features, the system chose the exact same features as previously; however, the centroid value from the LG-CHO was chosen fourth and the average Haralick sum average was chosen last. This feature had a ROC area of .74. Figure 2 shows the FROC results of the CAD system both with and without the LG-CHO being incorporated.

FROC results of CAD system with and without LG-CHO
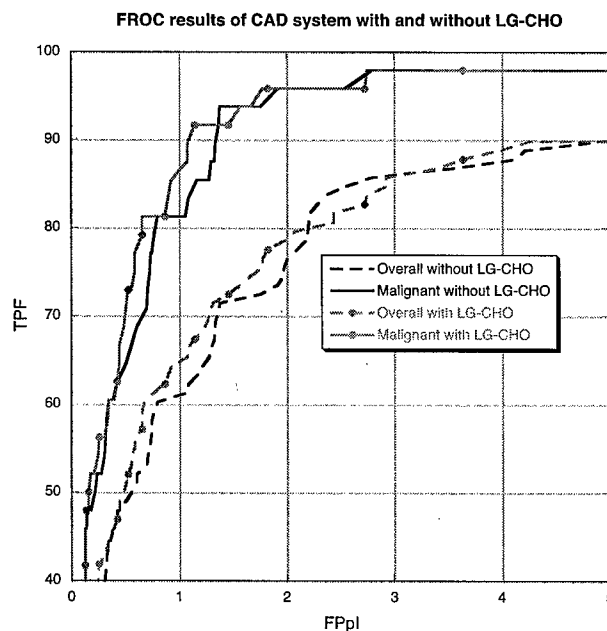
Figure 2: FROC results of the CAD system both with and without the LG-CHO being incorporated. Results for classification (malignant versus not) and detection (benign and malignant versus normal) are shown.

4. Discussion

The goal of this study was to investigate the incorporation of a LG-CHO into a CAD system. The CAD system did automatically select one of the LG-CHO features as being important and the inclusion of this feature did improve both overall performance (detection) and malignant (classification) performance, especially in high sensitivity regions of the classification FROC curve. Of interest is that the system chose exactly the same features both with and without the inclusion of the LG-CHO except the LG-CHO feature was chosen as well. These results show that observer templates can be used to improve CAD results. In the future, more advanced channelized observer models should be investigated.

References

Barrett, H. H., C. K. Abbey, et al. (1998). Stabilized estimages of Hotelling-Observer detection performance in patient-structured noise. SPIE Medical Imaging 1998: Image Perception.

Barrett, H. H., J. Yao, et al. (1993). "Model Observers For Assessment of Image Quality." Proceedings of the National Academy of Sciences of the United States of America **90**(21): 9758-9765.

Baydush, A. H., D. M. Catarious, et al. (2003). Computer aided detection of masses in mammography using a Laguerre-Gauss channelized Hotelling observer. Medical Imaging 2003: Image Perception.

Castellino, R. A., J. Roehrig, et al. (2000). "Improved Computer-aided Detection (CAD) Algorithms for Screening Mammography." Radiology 217(P): 400.

Catarious, D. M., A. H. Baydush, et al. (2004). "Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system." Medical Physics 31: 1512-1520.

Eckstein, M., C. Abbey, et al. (1998). Human vs model observers in anatomic backgrounds. Medical Imaging 1998: Image Perception.

Fiete, R. D., H. H. Barrett, et al. (1986). "Psychophysical Study to Test the Ability of Th Hotelling Trace Criterion to Predict Human-Performance." Journal of the Optical Society of America a-Optics Image Science and Vision 3(13): P126-P126.

Fiete, R. D., H. H. Barrett, et al. (1987). "Hotelling Trace Criterion and Its Correlation With Human- Observer Performance." Journal of the Optical Society of America a-Optics Image Science and Vision 4(5): 945-953.

Freer, T. W. and M. J. Ulissey (2001). "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center." Radiology 220: 781-786.

Gifford, H. C., M. A. King, et al. (1998). "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging." Journal of Nuclear Medicine 39(5): 771.

Gifford, H. C., M. A. King, et al. (2000). "Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging." Journal of Nuclear Medicine 41(3): 514-521.

Gifford, H. C., R. G. Wells, et al. (1999). "A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images." IEEE Transactions On Nuclear Science **46**(4): 1032-1037.

Heath, M., K. W. Bowyer, et al. (1998). Current status of the Digital Database for Screening Mammography. Digital Mammography. N. Karssemeijer, M. Thijssen and J. Hendriks, Kluwer Academic Publishers: 457-460.

Myers, K. J. and H. H. Barrett (1987). "Addition of a Channel Mechanism to the Ideal-Observer Model." Journal of the Optical Society of America a-Optics Image Science and Vision **4**(12): 2447-2457.

Wollenweber, S. D., B. M. W. Tsui, et al. (1999). "Comparison of hotelling observer models and human observers in defect detection from myocardial SPECT imaging." IEEE Transactions On Nuclear Science **46**(6): 2098-2103.

A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results

David M. Catarious, Jr.[*a], Alan H. Baydush[b,a], Craig K. Abbey[d], Carey E. Floyd, Jr.[c,a]

[a]Dept. of Biomedical Engineering, Duke University, Durham, NC 27710
[b]Dept. of Radiation Oncology, Duke University Medical Center, Durham, NC 27710
[c]Dept. of Radiology, Duke University Medical Center, Durham, NC 27710
[d]Dept. of Biomedical Engineering, University of CA, Davis, Davis, CA 95616

ABSTRACT

In this paper, we present preliminary results from a highly sensitive and specific CAD system for mammographic masses. For false positive reduction, the system incorporated features derived from shape, fractal, and channelized Hotelling observer (CHO) measurements. The database for this study consisted of 80 craniocaudal mammograms randomly extracted from USF's digital database for screening mammography. The database contained 49 mass findings (24 malignant, 25 benign). To detect initial mass candidates, a difference of Gaussians (DOG) filter was applied through normalized cross correlation. Suspicious regions were localized in the filtered images via multi-level thresholding. Features extracted from the regions included shape, fractal dimension, and the output from a Laguerre-Gauss (LG) CHO. Influential features were identified via feature selection techniques. The regions were classified with a linear classifier using leave-one-out training/testing. The DOG filter achieved a sensitivity of 88% (23/24 malignant, 20/25 benign). Using the selected features, the false positives per image dropped from ~20 to ~5 with no loss in sensitivity. This preliminary investigation of combining multi-level thresholded DOG-filtered images with shape, fractal, and LG-CHO features shows great promise as a mass detector. Future work will include the addition of more texture and mass-boundary descriptive features as well as further exploration of the LG-CHO.

Keywords: computer aided detection, mammography, masses, channelized Hotelling observers, fractal dimension

1. INTRODUCTION

For women in the United States, breast cancer is the second-most deadly type of cancer[1]. The American Cancer Society (ACS) estimates that in 2002, breast cancer will be diagnosed in 203,500 women and will kill almost 40,000 women[1]. Survival rates are significantly higher when the cancer is detected at an early stage[2-4]. The 5-year survival rate for patients with localized breast cancer is 96%. Patients with distant metastases see their 5-year survival rate drop to 21%[1]. Thus, detecting breast cancer at an early stage is critical to patient care.

The most common and effective early-detection tool currently available to clinicians is screening mammography. In fact, half of the cancers detected in screening mammography are impalpable[5]. Studies have shown that mammography is the only screening program proven to reduce mortality[5]. Mammography is also inexpensive and widely available.

Unfortunately, screening mammography has some drawbacks. Mammography is very difficult because there is no normal appearance of the breast that can be memorized; every breast is uniquely individual[6]. In addition, in the United States, mammography's low positive predictive value (PPV) (15% to 30%[5, 7]) means a high proportion of women who are subject to biopsies have benign breast disease. The low PPV of mammography increases patient anxiety, discomfort, and cost of care. It also contributes to reduced patient participation.

---

[*] david.catarious@duke.edu; phone 919.668.2539; fax 919.684.3934; DUMC Box 2623, DUMC, Durham, NC 27710

To aid mammographer's in identifying mammographic abnormalities, much research has been directed towards developing computer-aided detection (CAD) systems. These systems are meant to serve as second-readers to provide mammographer's with a second opinion. Studies have demonstrated these systems to have a beneficial effect on mammographers' sensitivity while not being detrimental to their specificity[8].

We have created a preliminary CAD system designed to detect mammographic masses. The proposed CAD system will consist of the components given in Figure 1. The system input will be a mammographic image. Using a pattern template and a pattern matching procedure, the CAD system will highlight areas of the image that are suspicious of being masses. From the highlighted image, specific regions of high suspicion will be identified and localized. These regions will then be described by a specific set of features. Specifically, we have investigated the combination of morphological, fractal, and channelized Hotelling observer (CHO) features. Using these descriptors, each region will be identified as being a mass or nonmass via classification and false positive reduction. The output of the system will be an image with highly suspicious regions identified. The system performance will be judged via free-response receiver operating characteristic (FROC) analysis.

2. MATERIALS AND METHODS

2.1 Database of Mammograms
The database of cases employed in this study was extracted from the Digital Database for Screening Mammography (DDSM) provided by the University of South Florida[9]. The DDSM contains 2,620 cases compiled by three institutions. Three scanners, at three different resolutions, were employed to digitize the mammographic films. For this study, we chose to use cases scanned by the Lumisys scanner at fifty microns-per-pixel.

From the Lumisys-scanned images, we randomly selected eighty images. Of these eighty images, forty images contained forty-nine masses (twenty-five malignant and twenty-four benign). The remaining forty images contained no mass findings. Although the DDSM contains both craniocaudal (CC) and mediolateral oblique (MLO) view mammograms, we chose to examine only images taken from the CC view.

At a resolution of fifty microns, the image size for the mammograms varied but averaged roughly 6,000 by 4,000 pixels. To obtain images of a uniform size; the maximum number or rows and columns was computed and each image was padded with the appropriate number of zeros. The images were then spatially averaged down to a size of 1,508 by 1,064 pixels (a resolution of 200 microns-per-pixel).

From the information contained in the DDSM, we extracted outlines of the masses. These outlines defined our ground truth.

2.2 Overview of CAD System
An overview of the developed CAD system is given in Figure 1. The initial input is a CC view mammogram. First, the image is filtered to enhance possible mass locations (A). From the filtered image, a multi-level gray level thresholding procedure defines specific suspicious regions (B). From these suspicious regions, features are extracted (C). A subset of these features is then selected (D) and used to classify the suspicious regions and reduce the number of false positives (E).
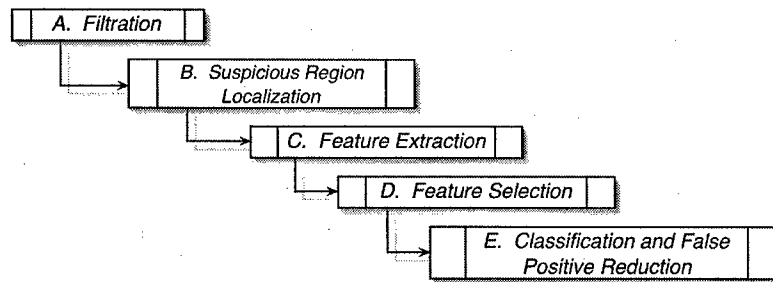
Figure 1. Overview of the CAD system.

A. Filtration

To identify the locations of abnormalities in mammograms, previous researchers have employed a variety of methods [10-17]. We chose to identify suspicious masses through filtering the images with a mass-like filter template. This approach requires the selection of an appropriate mass template and a template matching procedure.

Since masses are typically round and can have varying degrees of border-sharpness, a Gaussian filter is a natural choice to model the masses. The Gaussian chosen must have a relatively small width[1]. Alternatively, Gaussian filters may also be used as averaging filters. To achieve an overall smoothing effect, the Gaussians chosen for averaging should have a relatively broad extent. Because of the versatility of the Gaussian filter, a difference of Gaussians (DOG) filter is an effective mass template. Past research has demonstrated the usefulness of DOG filters for similar tasks[10-12].

A DOG filter is created by subtracting a rotationally-symmetric, two-dimensional Gaussian with width parameter σ_1 from a rotationally-symmetric, two-dimensional Gaussian with width parameter σ_2, where $\sigma_1 > \sigma_2$. Subtracting the Gaussians results in a filter that has a narrow, positive peak in the center surrounded by negative lobes that gradually increase back to zero.

The DOG filter has three parameters: the widths of its constituent Gaussians and the template window size. Note that the template window size does not affect the performance of the template window unless it truncates the Gaussians. It does, however, affect computation time and thus should be kept as small as possible. We selected width parameters of 90 and 45, respectively. The window size for the DOG filter template was 120 pixels.

To employ the DOG filter template to locate suspicious masses, we implemented normalized cross correlation (NCC)[18-20]. Although cross correlation is familiar and computationally efficient, it is amplitude dependent. Since the density in mammograms can widely vary, cross-correlation is of limited usefulness for this task. Alternatively, NCC is invariant to varying background scale. NCC computes the correlation between the mass template and the underlying mammographic image. Areas of the image that follow the same profile will return a value of one; areas that are exact opposites of the template will return negative one.

Although NCC cannot be entirely implemented in the frequency domain, a fast implementation is available through the use of running sum matrices[21]. The only additional parameter for NCC is the size of the windowing operator. In this case, the window size was selected to be equivalent to the size of the DOG filter template.

B. Suspicious Region Identification

The areas in the filtered image that best match the filter template will contain the brightest grayscale values. To distinguish these areas from the rest of the image, previous researchers[19, 22, 23] have employed a gray level thresholding technique. By selecting the pixels with values above certain thresholds, the most suspicious regions will be identified. Determining thresholds based on percentiles of the gray scale histogram provides a general procedure that can be performed on an image-by-image basis.

---

[1] Note that we will refer to a Gaussian's "width,, instead of its variance. This is because these Gaussians are being used as static filter templates, not as distributions of random variables.

Our method of suspicious region identification is also based on thresholding the histogram of the filtered images. Initially, a set of threshold values is determined by selecting the increasing percentages of the gray levels. In our current implementation, we select the pixels from the top 1% to the top 21% in steps of 2%, for a total of eleven thresholds.

The result from the thresholding process is a set of images, each containing suspicious regions. To reduce the number of suspicious regions, we combine the regions on each level into one image, called the duration image. The duration image consists of regions that have not merged with a neighboring region as the threshold levels progressed. If a region did merge with another region, it was extracted at the level before it merged. The number of thresholds that a region existed as an independent entity is denoted as the region's duration. The regions remaining in the duration image were then passed onto the feature extraction stage.

C. Feature Extraction
For this initial system, we extracted forty-seven features derived from each region's duration, morphology, fractal dimension, and response to a Laguerre-Gauss CHO (LG CHO).

C.1 Morphological Features
The morphological features extracted were area, eccentricity, and convex area minus area. Convex area is defined as the area of a region's convex hull (basically a rubber band surrounding the region).

C.2 Fractal Dimension Features
Because of its ability to measure the roughness of an object, fractal dimension has been adopted as a textural feature. There have been many methods proposed to measure fractal dimension[24]. In this research, we adopted the covering-blanket method (CBM)[25, 26]. The CBM method takes advantage of the fact that certain measurements of fractal objects follow Richardson's Power Law: $M(\varepsilon) = K\varepsilon^{d-D}$, where ε is the scale value, $M(\varepsilon)$ is the value of some measured property at scale ε (such as surface area), K is a constant of proportionality, d is the topological dimension, and D is the fractal dimension.

Thus, to measure the fractal dimension, we must measure a property of the image over several scales (that is, using windows of several sizes). In this case, we measured surface area. If the surface is truly fractal in nature, plotting the surface area vs. scale on a log-log plot should result in a straight line with slope d-D. The slope and intercept will be estimated using regression. In this implementation, the overall slope and intercept is not measured. Instead, the *local* slope and intercept are measured. That is, only three points are considered at once to determine the slope and intercept. This is because real objects rarely exhibit a true linear behavior. The slope and y-intercept over nine scales were extracted and used as features. As additional features, we also measured the derivative and standard deviations of the fractal dimension and y-intercepts over all scales. This resulted in a total of forty fractal features.

C.3 Laguerre-Gauss Channelized Hotelling Observer Features
Our final set of features was collected from each region's response to a LG CHO. The LG CHO is a mathematical observer model designed to process different frequencies present in an image. Ideally, a Hotelling observer (HO)[27-29] would be employed that could observe an entire region of interest. However, to create a HO large enough to observe a meaningful region would require an inordinate amount of sample images[30].

One method to reduce the dimensionality problem is to use linear functions of the pixels instead of operating on the pixels directly. In the literature, these functions are known as *channels*. By using channels, the dimensionality of the problem can be reduced to equal the number of channels. In practice, the number of channels selected is much less than the number of pixels, making the problem tractable. After obtaining a region's response to each channel, a HO can be created. HOs designed for channel outputs are called *channelized HOs* (CHOs).

The next issue is to decide what to use for the channels. Barrett *et al*[30] state that since most HOs are smooth, smooth functions should be favored. Also, since the task is to locate masses (which are usually round), the channels should be rotationally symmetric. Following these constraints, Barrett *et al*[30] suggest exploring a family of functions known as Laguerre-Gauss (LG) functions. The parameters for the LG CHO employed in this study were determined empirically by Baydush et al[31].

By filtering the mammograms with the LG CHO filter template, we extracted three features for each suspicious region: the mean, peak, and standard deviation of the LG CHO output.

D. Feature Selection

Although we have measured a set of features to describe each suspicious region, not all of these features will prove useful in discrimination. Thus, we will select a subset of features, A, that best separates the masses from the nonmasses. To select an effect feature subset, we implemented forward searching stepwise feature selection (FS-SWFS)[32]. In FS-SWFS, A begins as the empty set and is constructed by sequentially adding features that maximize a performance criterion, θ. FS-SWFS also deletes features from A if their removal improves θ. The selection process halts when a subset of a particular size has been found or when θ does not improve for a given number of iterations. For this research, we judged performance via a linear classifier and the set θ equal to the resulting area under the receiver operating characteristic (ROC) curve.

E. Classification and False Positive Reduction

Once the suspicious regions have been subject to feature extraction and the subset of features has been selected, the regions are classified as masses or nonmasses. Previous researchers have employed a number of methods to perform region classification, including linear discriminant analysis (LDA)[33, 34], artificial neural networks[35-37], and rule-based methods[10, 38]. For our system, we chose to employ LDA via Fisher's linear discriminant[32].

Fisher's linear discriminant is defined as $a \square cS^{-1} |\vec{m}_1 - \vec{m}_2$, where $\vec{m}_i$ is the sample mean vector for class i and S is the sample covariance matrix for the features. c is an arbitrary constant. This value of a is known as Fisher's linear discriminant. Note that when $c=1$, Fisher's linear discriminant is equivalent the Hotelling observer computed with sample, instead of population, statistics. Also, if we assume the features are multivariate normally distributed with equal covariance matrices, and we employ sample statistics, Fisher's linear discriminant is a Bayes' classifier.

To train and test the linear classifier, we implemented a round-robin training and testing procedure.

3. RESULTS

Before any classification and false positive reduction, the duration images contained an average of ~20 false positives per image (FPpI). The initial sensitivity was 88% (43/49).

The stepwise feature selection chose the features in Table 1. Table 1 indicates each feature's individual ROC performance as well as the cumulative ROC performance, where ROC performance is given as the area under the ROC curve (AUC). The final classifier was constructed using these features.

Features in the Order Chosen	Individual AUC	Cumulative AUC
Peak of the LG CHO output	0.90	0.90
Area	0.87	0.93
Duration	0.81	0.94
Convex Area - Area	0.77	0.95
Mean of the LG CHO output	0.80	0.95
Standard Deviation of the Fractal Dimension, scale 8	0.75	0.95
Standard Deviation of the y-Intercept, Scale 8	0.76	0.95
Fractal Dimension, Scale 2	0.69	0.95

Table 1: Table of the features chosen by stepwise feature selection. The left column specifies the feature, the center column provides the feature's individual ROC performance, and the right column indicates the cumulative ROC performance.

The overall system performance is given in the FROC curve in Figure 2. Also given is the system's performance when only considering the best feature, the peak output of the LG CHO. Note that since no masses were missed in the FPpI range from ~20 to ~5 for the final system, this portion of the FROC curve was truncated.

False Positives per Image
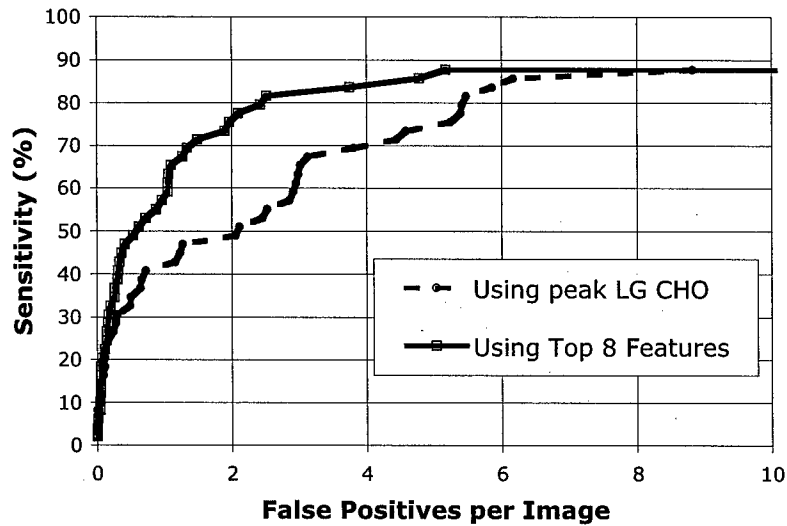
Figure 2: FROC Curve describing system performance. The solid line describes the system performance when eight features are utilized by the classifier. The dotted line describes the performance when only the peak output of the LG CHO is considered by the classifier.

An example of a mammogram processed with this CAD system is given in Figure 3. In the left panel is a CC view mammogram that contains a malignant mass. The right panel displays the suspicious regions remaining after the classification stage.
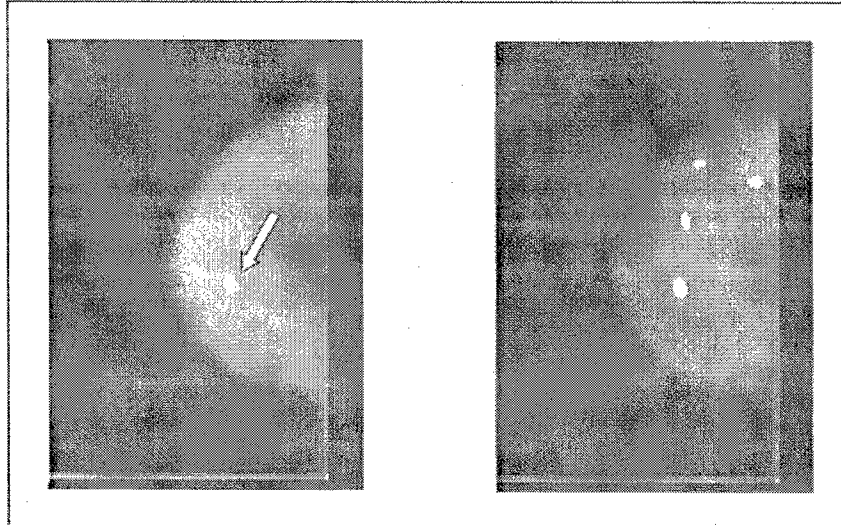
Figure 3: The original CC-view mammogram (left) contains a malignant mass. The resulting output of the CAD program detected the mass as well as three false positives.

4. CONCLUSIONS

We have developed an initial CAD system for detecting mammographic masses. Our system utilized features based on suspicious regions' morphology, fractal dimension, and response to a LG CHO. The system is able to achieve ~88% sensitivity with ~5 FPpI. It is also able to maintain greater than 80% sensitivity until ~2.5 FPpI. As can be seen in Table

1, the most influential feature was the peak output of the LG CHO. Figure 2 compares the system performance when using the best eight features to the performance when using just the peak output of the LG CHO. While the system performance is better when eight features are considered, the system performs remarkably well when just the peak LG CHO output is employed. At ~5 FPpI, the simplified system is still able to achieve ~75% sensitivity. Thus, as demonstrated in previous studies, the LG CHO is very effective at distinguishing mammographic masses from background structures[31]. In future work, we will continue to explore the capabilities of the LG CHO.

It was also informative to examine the individual performances of the features selected by the FS-SWFS. Although mist did not perform extremely well as single features, they were collectively able to increase by ROC AUC from 0.90 to 0.95. Therefore, these features must exhibit some degree of independence and capture different information about the suspicious regions.

The performance of the fractal features was somewhat disappointing. As judged by ROC AUC, the fractal features only incrementally increased system performance. They also exhibited low ROC AUCs when considered alone. In fact, two out of the three morphological features were selected before any fractal features.

Although this system exhibits a high sensitivity at a moderate level of FPpI, there is still more work to be performed. In the future, this system will be extended by adding more images to the database, adding a step to further reduce the influence of noisy backgrounds, incorporating a wider range of morphological features, extending the set of textural measures, incorporating a finer region segmentation, and exploring additional DOG filters and other mass identifying filters.

ACKNOWLEDGEMENTS

REFERENCES

1.   ACS, "American Cancer Society: Cancer Facts and Figures 2002. Atlanta, Ga: American Cancer Society 2002.," (2002).

2.   E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," *Radiology* 191, 241-244 (1994).

3.   I. Anttinen, M. Pamilo, M. Soiva, and M. Roiha, "Double reading of mammography screening films: one radiologist or two?," *Clinical Radiology* 48, 414-421 (1993).

4.   W. R. Hendee, C. Beam, and E. Hendrick, "Proposition: all mammograms should be double-read," *Medical Physics* 26, 115-118 (1999).

5.   A. K. Tucker and Y. Y. Ng, *Textbook of Mammography,* 2 ed. (Churchill Livingstone, London, 2001), p. 334.

6.   M. J. Homer, *Mammographic Interpretation: A Practical Approach,* 1 ed. (McGraw-Hill, Inc., New York, 1991), p. 248.

7.   D. B. Kopans, "The positive predictive value of mammography," *AJR. American Journal of Roentgenology* 158, 521-526 (1992).

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.   M. Heath, K. W. Bowyer, and D. Kopans, "Current status of the Digital Database for Screening Mammography," in *Digital Mammography*, N. Karssemeijer, M. Thijssen, and J. Hendriks, eds. (Kluwer Academic Publishers, 1998), pp. 457-460.

10.   B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," *Academic Radiology* 2(11), 959-966 (1995).

11.   B. Zheng, Y. H. Chang, and D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms," *Academic Radiology* 3(10), 806-814 (1996).

12.   W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," *IEEE Trans Med Imaging* 16(6), 811-819 (1997).

13.   H. Kobatake, M. Murakami, H. Takeo, and S. Nawano, "Computerized Detection of Malignant Tumors on Digital Mammograms," *IEEE Transactions on Medical Imaging* 18(5), 369-378 (1999).

14.   Y.-H. Chang, W. F. Good, J. H. Sumkin, B. Zheng, and D. Gur, "Computerized Localization of Breast Lesions from Two Views: An Experimental Comparison of Two Methods," *Investigative Radiology* 34(9), 585-588 (1999).

15.   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," *Medical Physics* 29(2), 238-247 (2002).

16.   N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Med Phys* 26(8), 1642-1654 (1999).

17.   N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Medical Physics* 23(10), 1685-1696 (1996).

18.   R. C. Gonzalez and R. E. Woods, *Digital Image Processing,* Third ed. (Addison-Wesley, New York, 1993), p. 716.

19.   M. J. Carreira, D. Cabello, M. G. Penedo, and A. Mosquera, "Computer-aided diagnoses: Automatic detection of lung nodules," *Med Phys* 25(10), 1998-2006 (1998).

20.   M. G. Penedo, M. J. Carreira, A. Mosquera, and D. Cabello, "Computer-aided diagnosis: a neural-network-based approach to lung nodule detection," *IEEE Trans Med Imaging* 17(6), 872-880 (1998).

21.   J. P. Lewis, "Fast Normalized Cross-Correlation", retrieved www.idiom.com/~zilla/Work/nvisionInterface/nip.html.

22.   M. L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields," *Medical Physics* 15, 158-166 (1988).

23.   M. L. Giger, K. Doi, H. MacMahon, C. E. Metz, and F. Yin, "Pulmonary nodules: Computer-aided detection in digital chest images," *RadioGraphics* 10, 41-51 (1990).

24.   H. O. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science* (Springer-Verlag, New York, New York, 1992).

25.   J. L. Solka, C. E. Priebe, and G. W. Rogers, "An initial assessment of discriminant surface complexity for power law features," *Simulation* 58(5), 311-318 (1992).

26.   S. Peleg, J. Naor, R. Hartley, and D. Avnir, "Multiple Resolution Texture Analysis and Classification," *IEEE PAMI* 6(4), 518-523 (1984).

27.   A. H. Baydush, D. M. Catarious, and J. Y. Lo, "Computer Aided Detection of Masses in Mammography using Sub-region Hotelling Observers," presented at *Medical Imaging Perception Society*, Warrenton, VA, Warrenton, VA, 2001.

28.   A. H. Baydush, D. M. Catarious, Jr, J. Y. Lo, C. K. Abbey, and C. E. Floyd, Jr, "Computerized classification of lung nodules in chest radiographs using sub-region hotelling observers," *Medical Physics* 28, 2403-2409 (2001).

29.   A. H. Baydush and D. M. J. Catarious, "Novel use of the Hotelling Observer for computer-aided diagnosis of solitary pulmonary nodules," *Proceedings of the SPIE* 4322, 1918 (2001).

30.   H. H. Barrett, C. K. Abbey, and B. Gallas, "Stabilized estimates of Hotelling-observer detection performance in patient-structured noise," presented at *SPIE*, 3340, 27-43, 1998.

31.   A. H. Baydush, J. Catarious, D M, and C. E. Floyd, "Computer aided detection of masses in mammography using a Laguerre-Gauss channelized Hotelling observer," presented at *SPIE Medical Imaging*, Sand Diego, CA, Sand Diego, CA, 2003.

32.   M. Nadler and E. P. Smith, *Pattern Recognition Engineering* (John Wiley and Sons, New York, New York, 1993), p. 588.

33.   H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space," *Physics in Medicine and Biology* 40(5), 857-876 (1995).

34.   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. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology* 212(3), 817-827 (1999).

35.   Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology* 187, 81-87 (1993).

36.   G. te Brake and N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms," *Medical Physics* 28(2), 259-266 (2001).

37.   B. Zheng, Y. Chang, W. F. Good, and D. Gur, "Performance gain computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering," *Medical Physics* 28(11), 2302-2308 (2001).

38.   L. Li, Y. Zheng, L. Zheng, and R. A. Clark, "False-positive reduction in CAD mass detection using a comptetive classification strategy," *Medical Physics* 28(2), 250-258 (2001).

Initial development of a computer-aided diagnosis tool for solitary pulmonary nodules

David M. Catarious, Jr.[*a], Alan H. Baydush[b,a], Carey E. Floyd, Jr.[b,a]

[a]Department of Biomedical Engineering, Duke University, Durham, NC
[b]Department of Radiology, Duke University Medical Center, Durham, NC

ABSTRACT

This paper describes the development of a computer-aided diagnosis (CAD) tool for solitary pulmonary nodules. This CAD tool is built upon physically meaningful features that were selected because of their relevance to shape and texture. These features included a modified version of the Hotelling statistic (HS), a channelized HS, three measures of fractal properties, two measures of spicularity, and three manually measured shape features. These features were measured from a difficult database consisting of 237 regions of interest (ROIs) extracted from digitized chest radiographs. The center of each 256x256 pixel ROI contained a suspicious lesion which was sent to follow-up by a radiologist and whose nature was later clinically determined. Linear discriminant analysis (LDA) was used to search the feature space via sequential forward search using percentage correct as the performance metric. An optimized feature subset, selected for the highest accuracy, was then fed into a three layer artificial neural network (ANN). The ANN s performance was assessed by receiver operating characteristic (ROC) analysis. A leave-one-out testing/training methodology was employed for the ROC analysis. The performance of this system is competitive with that of three radiologists on the same database.

Keywords: computer-aided diagnosis, artificial neural networks, linear discriminant analysis, pulmonary nodule classification, ROC analysis, feature extraction

1. INTRODUCTION

In the year 2000, it is estimated that cancers of the lung and bronchus will account for 31% of the cancer deaths in men and 25% of the cancer deaths in women[1]. For both genders, lung cancer is the leading cause of death among all cancers[1]. Early detection is key to a patient surviving lung cancer with survival rates being 3 to 4 times higher in patients whose cancers were discovered early compared to those discovered late[1]. Solitary pulmonary nodules are the first sign of cancer found in 20-30% of lung cancer cases and thus are extremely important to detect in a chest radiograph. Since up to 20% of suspected nodules turn out to be other entities, it is important to be able to detect and correctly identify lung lesions as nodules or non-nodules[2].

It has been shown in the literature that the use of a CAD tool can aid a radiologist in the detection and diagnosis of pulmonary nodules[3-9]. While some CAD systems have relied on combining radiologist s observations with image data, our goal in this study was to develop a CAD system that can aid a radiologist in discriminating between lung nodules and normal lung lesions based on image data alone. This is desirable because the subjectivity of the radiologist s measurements will not affect the performance of the system.

2. MATERIALS AND METHODS

2.1. Image database

The region of interest (ROI) image database consisted of 237 256x256 pixel regions of interest (ROIs) extracted from digitized chest radiographs. Each ROI contained a centered nodule that was sent to follow-up (i.e., fluoroscopy or CT) by a radiologist and whose nature was later clinically determined. The ROIs were extracted by hand using image display

---

* Correspondence: Email: dmc2@duke.edu; Phone: (919) 668-2539; Radiology Dept/Digital Imaging, Box 2623 DUMC, Durham, NC 27710

software. The three radiologists who examined this database achieved ROC areas of .72 (.03), .79 (.03), and .83 (.03). For further information on this database, see Drayer et al[10]. Note that the images which contained a lung nodule will be referred to as positive images while those that contained no nodule will be referred to as negative images. Examples of a positive image and a negative image are given in Figure 1.
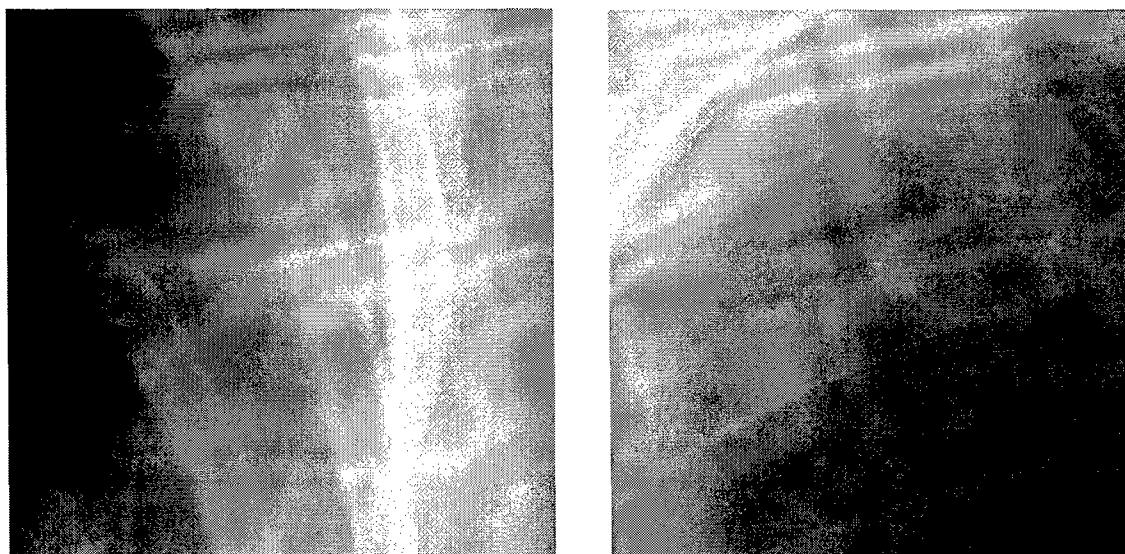

Figure 1. Sample ROIs from the image database. The image on the left is positive (contains a nodule) while the image on the right is negative (does not contain a nodule).

2.2. Features and feature extraction

Ten features were selected for use in this CAD system. They included a modified Hotelling statistic (HS), a channelized Hotelling statistic (CHS), three measures of fractal properties, two spiculation measures, lesion radius, lesion circularity, and lesion compactness. The first seven of these features were algorithmically computed while the latter three were measured by hand. They were selected based on their relevance to the shape and texture of the lesions being discriminated.

The Hotelling statistic (HS) is a measure that has been proven to be effective in the detection of signals in correlated noise[11-16]. It can be derived from the Hotelling trace criterion (HTC), which has been found to correlate highly with human performance on observer tests[17]. In fact, signal detection theory tells us that the HS is the optimal detector in the case where the signal (lung nodule) and the statistical properties of the noise (background) are known. The HS is computed as:

$$HS = x\Sigma^{-1}s^{T},$$

(1)

where x is the image to be classified (stored as a 1xN vector), Σ^{-1} is the inverse covariance matrix of the images (an NxN matrix), and s^{T} is the transpose of the known signal (also stored as a 1xN vector). Thus, the HS is a scalar descriptor of the image. Note that Eq.(1) is equivalent to the log-likelihood function derived in signal detection theory. The HS also assumes that the pixels of the image should be approximately normally distributed (at least locally) over the entire image population.

In this setting of classifying a lesion as lung nodule (positive) or not a nodule (negative) in real chest radiographs, neither the true signal nor the statistics of the background noise are completely known. Thus, they both must be estimated. Estimating the signal is a simple matter of averaging all of the positive cases, averaging all of the negative cases, and subtracting the two average images. This will provide a decent approximation of a positive image though this type of estimation is highly dependent upon the size of the database. The average images of the positives and negatives are shown in Figure 2 while their center profile is shown in Figure 3. It can been seen in these figures that the average images and central

profiles are very similar for the positives and negatives, which helps to explain why pulmonary nodule detection is such a difficult task.
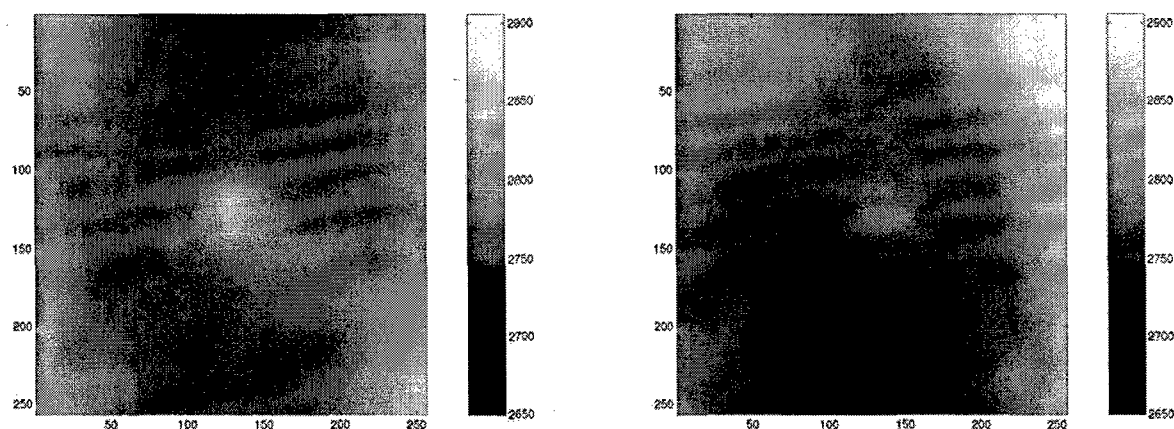
Figure 2. The image on the left is the average positive image while the image on the right is the average negative image. Notice how each image contains a perturbation in the center. Both images are displayed over a range of gray-values from 2650 to 2905.
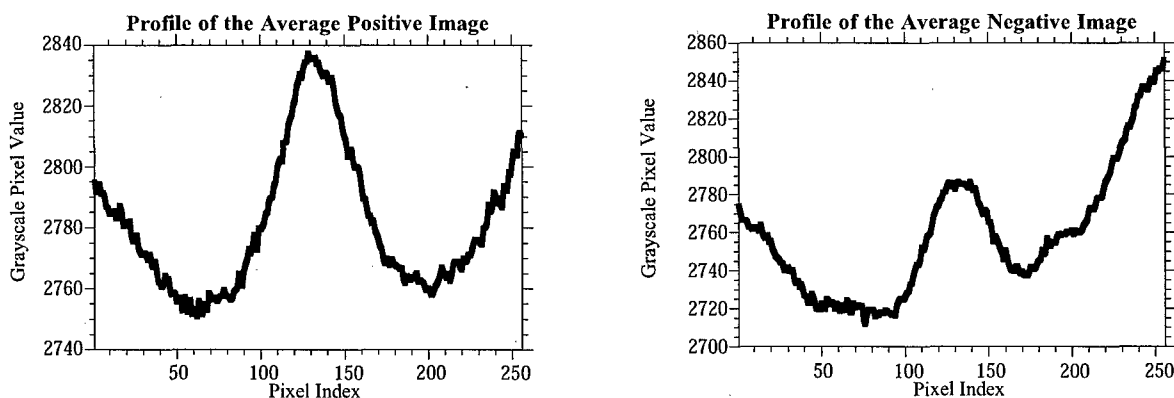
Figure 3. Plots of the profiles of the average images of the positive cases and the negative cases. They are profiles of the images in Figure 2 going horizontally across the center. Note the similarity in the central region and general trend of the pixel values.

Obtaining an estimate of the covariance matrix is more problematic. Since the images being used in this study are 256x256 pixels, at least 256^2 images would be needed to compute the covariance matrix. It is recommended that 3 to 10 times this number be used so as to have an accurate estimate of the covariance matrix[15]. Collecting this many chest radiographs would be an intractable task. To avoid the need for this many images, the 256x256 pixel images were subsampled into 8x8 pixel images by segmenting the images into 64 subregions and then averaging the pixels within each subregion. The covariance matrix of this new image set is only 64x64 pixels and thus can be estimated from the database. Note that the estimated signal will now only be 8x8 in size as well.

Another way to reduce the dimensionality of the covariance matrix is to use the channelized Hotelling statistic (CHS). In the channelized version, the images are filtered by a set of frequency-selective channels which are meant to simulate the human visual system[14,15]. In this study, ten channels were chosen to extract data from the images. The channels were defined by Laguerre-Gauss functions, which are a class of radially-symmetric curves[16]. By selecting ten channels, the

images are effectively described by a length ten vector, reducing the covariance matrix to 10x10 in size and thus making the computation of the covariance matrix possible. The CHS is then computed as:

$$CHS = fS^{-1}h^{T},$$

(2)

where f is the 1x10 feature vector, S^{-1} is the 10x10 inverse covariance matrix of the channel-filtered images, and h^{T} is the transpose of the 1x10 average signal feature vector.

Another feature that describes the texture of an object in an image is the fractal dimension. In a general sense, the fractal dimension of a surface is a measure of its roughness. Fractal dimension provides a means to quantify how irregular, jagged, crinkly, curvy, and space-filling a surface is. Note that unlike Euclidean dimensions, the fractal dimension of an object is not an integer since it describes how much space an object fills between Euclidean dimensions.

In the literature, there are many definitions and many different ways to compute the fractal dimension of an object. In this study, we adopted the method put forth by Peli[18] and Peleg et al[19] which was derived from one of Mandelbrot s methods to measure the coastline of Britain in The Fractal Geometry of Nature [20]. This method, called the covering-blanket method, basically works by defining an upper and lower bound on the image and iteratively raising and lowering the surfaces in a window of increasing size via erosion and dilation. The size of the window at each iteration is denoted by ε. At each iteration of this process, a measure of the surface area is calculated. It is known that the surface area of a fractal follows a power law behavior with respect to its fractal dimension. This power law behavior is described explicitly by Richardson s Power Law [21]:

$$M(\varepsilon) = K\varepsilon^{d-D},$$

(3)

where $M(\varepsilon)$ is some measured property at scale ε, K is a constant, d is the Euclidean dimension, and D is the fractal dimension. In this situation, $M(\varepsilon)$ is the surface area at scale ε. Thus, by taking the natural logarithm of both sides of Eq.(3), it can be seen that the fractal dimension represents the slope of a line. This fractal dimension is determined by finding the regression line through the data points and calculating its slope. Along with the slope, the y-intercept of the regression line can easily be determined and used as an additional fractal feature.

For the images in this study, three fractal properties were selected: the overall fractal dimension of the entire image, the overall y-intercept, and the fractal dimension of the center portion of the ROI (where the center portion refers to the central region when the image is divided into nine equal subregions).

Spicularity is an important property to describe because of its prevalence in lung nodules[2]. To measure the degree of spicularity of the lesions, the convex hulls of the masses were computed. The convex hull is defined as the convex polygon of least area that completely covers an object [22]. The convex hull can be easily pictured as the outline of a rubber band wrapped around an object. Two measures involving the convex hull were used in this study: the area of the convex hull and the ratio of the mass area to the convex hull area.

Besides the algorithmically determined features described above, three hand-measured features were also used in this study: lesion radius, circularity, and compactness. Although it may be undesirable to use non-machine calculated features, these features were included because they were readily available and can be algorithmically calculated.

The lesion radius was determined as the radius of the circle that had the same area as the lesion. Circularity was calculated as the ratio of the number of pixels inside a circle of the same area to the number of pixels inside the outline of the lesion. Compactness was calculated as $N^{2}/(4\pi)S$, where N is the number of pixels in the lesion s perimeter and S is the number of pixels in the region.

2.3. Rank ordering of features

It is well known that redundant or linearly dependant features can degrade the performance of an artificial neural network (ANN). To identify which features aided and/or limited the discriminatory ability of this CAD system, the entire feature space was searched via a sequential forward search using linear discriminant analysis (LDA). During the sequential forward search using LDA, percent correct was used as the performance metric. During the first stage, each feature was examined individually by the LDA and the resulting percent correct was determined. At the end of the first stage, the feature which garnered the highest accuracy was noted and removed from the feature set. During the second stage, each remaining feature was paired with the best feature and again examined by the LDA. At the end of the second stage, the feature which provided the best accuracy, when paired with the best feature, was removed from the feature set. This process continued until a rank ordering of the features had been determined.

This was an essential step because optimizing an artificial neural network can be a time consuming process due to the number of free parameters involved. For example, to exhaustively search the feature space by looking at every combination of features would require 1,023 comparisons of feature subsets, not to mention optimizing the momentum rates, learning rate, and number of hidden nodes. Since LDA has no free parameters to adjust, it is much easier to determine the rank ordering of the features using the LDA than an ANN. By identifying the most and least important features before implementing the ANN, the optimization process could proceed in an organized and efficient manner. Note that it is also desirable to keep the number of inputs low (especially when using a limited training set) since networks with a large number of connections, in comparison to the number of training cases, tend to lose their generalization capabilities.

2.4. Classification

The final classification was performed via the use of a multi-layer ANN trained by backpropagation. ANNs have become a very popular method to perform the classification of lung nodules in CAD systems[3-6,8,9]. Initially, the network was given all ten features as inputs. To optimize the network s performance and reduce the complexity of the network, one feature at a time was removed from the network, in the order prescribed by the LDA, until the best performance was found. The activation function used was the logistic sigmoid function. As is usually the case, the number of hidden nodes was determined experimentally. The input data was normalized to be between 0 and 1. The network was trained so that 0 represented a negative and 1 represented a positive. The neural network was trained via a leave-one-out training/testing methodology. Mean-squared error (MSE) was used as the minimization criteria for training. Receiver operating characteristic (ROC) analysis was performed on the output of the ANN so that the results from the ANN could be compared to those of the radiologists.

3. RESULTS

The ten features being studied for use in the CAD system are listed in Table 1 along with the ROC areas that each feature provides when used on its own. As seen in Table 1, the ROC areas range from very poor to reasonably well (.533 to .706).

Table 1. ROC Areas for each of the features when analyzed individually.

Feature	ROC Area (A_Z)
Modified Hotelling Statistic	.706
Channelized Hotelling Statistic	.660
Overall Fractal Dimension	.535
Overall Y-Intercept from Fractal Regression Line	.550
Center Fractal Dimension	.533
Convex Hull Area	.603
(Mass Area)/(Convex Hull Area)	.615
Lesion Radius	.592
Lesion Compactness	.575
Lesion Circularity	.611

After searching the entire set via the sequential forward searching LDA, it was determined that the features follow the subsequent order in terms of contributing to the classification ability: modified HS, lesion radius, center fractal dimension, channelized HS, convex hull area, (mass area)/(convex hull area), overall fractal dimension, overall y-intercept from the fractal regression line, lesion compactness, and lesion circularity. The accuracy of the LDA, with the features added in the prescribed order as well as individually, is given in Table 2.

Table 2. The order and accuracy of the features determined by the LDA.

Order of Features Selected by LDA	Individual Percent Correct	Cumulative Percent Correct
Modified Hotelling Statistic	.66	.66
Lesion Radius	.54	.70
Center Fractal Dimension	.58	.69
Channelized Hotelling Statistic	.62	.70
Convex Hull Area	.54	.72
(Mass Area)/(Convex Hull Area)	.57	.72
Overall Fractal Dimension	.54	.72
Overall Y-Intercept from Fractal Regression Line	.62	.71
Lesion Compactness	.54	.73
Lesion Circularity	.60	.73

After the classification order of the features was determined by the LDA, the neural network optimization was performed. First, all ten features were used as inputs to the network and the ROC area was calculated. Next, the least valuable feature (determined form the LDA analysis) was removed and the remaining nine features were used as inputs to the network. This process continued all the way down to just using one input to the network. After examining the results for the different number of inputs, it was determined that the optimum performance was achieved when the top seven features were selected as inputs to the network. Thus, the final network had seven input nodes, three hidden nodes and one output node. The set of features used was the first seven features in Table 2, ranging from the modified Hotelling statistic to the overall fractal dimension. This combination of features produced a network that achieved an ROC area of .78. The resulting ROC curve is given in Figure 4.

ROC Curves from Radiologists and ANN

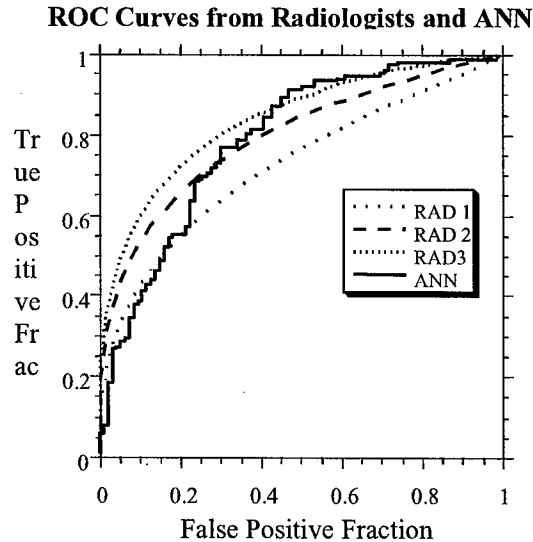

Figure 4. ROC (Receiver Operating Characteristic) curve generated from the ANN. The curve has an area of .78. The x-axis represents the False Positive Fraction (1-Specificity) while the y-axis represents the True Positive Fraction (Sensitivity). The ideal CAD system would achieve an A_z of 1, denoting that it is 100% sensitive (no false negatives) and 100% specific (no false positives).

4. CONCLUSIONS

As can be seen from Tables 1 and 2, the only feature to exhibit any reasonable classification ability when used alone was the modified Hotelling statistic. For the most part, the rest of the features performed poorly. Thus, it is interesting that combining features that individually had almost no discriminatory ability (like the fractal features) would produce a CAD tool that performed at a level competitive with that of three radiologists. This suggests that the reason lung nodule detection is such a difficult task is because so many features must be taken into account and no one feature can be used to make a confident decision.

The fact that the modified Hotelling statistic performed so well is also interesting to note. Since the modified Hotelling statistic performed its analysis on images which had been subsampled and averaged, it made its decisions based on the low frequency content of the image. This is surprising since most of the fine details of the nodule have effectively been eliminated in each of the subregions. Therefore, it seems that this CAD system may concentrate first on low-frequency content of the image and then fine tune its decisions based on more high-frequency details.

Although this CAD system performed admirably on this data set, more research needs to be performed before its true performance can be assessed. The largest limitation of this analysis is the size of the database used and thus more images need to be collected.

Overall, the performance of this CAD system is high enough to merit further research. It also suggest that a system based solely on texture and shape measures could be a viable CAD tool.

ACKNOWLEDGMENTS

REFERENCES

1. R. T. Greenlee, T. Murray, S. Bolden, and P. A. Wingo, Cancer statistics, 2000, **50**, 7-33, 2000.
2. J. Erasmus, J. Connolly, H. McAdams, and V. Roggli, Solitary Pulmonary Nodules: Part I. Morphologic Evaluation for Differentiation of Benign and Malignant Lesions, *RadioGraphics* **20**, 43 - 58, 2000.
3. K. Nakamura, H. Yoshida, R. Engelmann, H. MacMahon, S. Katsuragawa, T. Ishida, A. Ashizawa, and K. Doi, Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks, *Radiology* **214**, 823-830, 2000.
4. H. MacMahon, R. Engelmann, F. M. Behlen, K. R. Hoffmann, T. Ishida, C. Roe, C. E. Metz, and K. Doi, Computer-aided diagnosis of pulmonary nodules: Results of a large-scale observer test, *Radiology* **213**, 723-726, 1999.
5. X. W. Xu, K. Doi, T. Kobayashi, H. MacMahon, and M. L. Giger, Development of an improved CAD scheme for automated detection of lung nodules in digital chest images, *Med Phys* **24**, 1395-403, 1997(97449551).
6. S.-C. B. Lo, M. T. Freedman, J.-S. Lin, and S. K. Mun, Automatic Lung Nodule Detection Using Profile Matching and Back-Propagation Neural Network Techniques, *J. Dig. Img.* **6**, 48 - 54, 1993.
7. H. Yoshimura, M. L. Giger, K. Doi, H. MacMahon, and S. Montner, Computerized scheme for the detection of pulmonary nodules: A nonlinear filtering technique, *Invest Radiol* **27**, 124-129, 1992.
8. M. G. Penedo, M. J. Carreira, A. Mosquera, and D. Cabello, Computer-aided diagnosis: A neural-network-based approach to lung nodule detection, *IEEE Trans. Med. Imaging* **17**, 872-880, 1998.
9. Y. C. Wu, K. Doi, and M. L. Giger, Detection of lung nodules in digital chest rad iographs using artificial neural networks: a pilot study, *J Digit Imaging* **8**, 88-94, 1995.
10. J. A. Drayer, N. Vittitoe, R. Vargas-Voracek, A. H. Baydush, C. E. Floyd, Jr, and C. E. Ravin, Characteristics of regions suspicious for pulmonary nodules on chest radiographs, *Acad Radiol* **5**, 613-9, 1998.

11. R. D. Fiete, and H. H. Barrett, Using the Hotelling Trace Criterion For Feature Enhancement in Image-Processing, *Opt. Lett.* **12**, 643-645, 1987.
12. M. P. Eckstein, C. K. Abbey, and F. O. Bochud, Visual signal detection in structured backgrounds. IV. Figures of merit for model performance in multiple-alternative forced- choice detection tasks with correlated responses, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **17**, 206-217, 2000.
13. W. E. Smith, and H. H. Barrett, Hotelling Trace Criterion As a Figure of Merit For the Optimization of Imaging-Systems, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **3**, 717-725, 1986.
14. T. K. Narayan, and G. T. Herman, Prediction of human observer performance by numerical observers: an experimental study, *J, Opt. Soc. Am. A* **16**, 679 - 693, 199.
15. M. Eckstein, C. Abbey, and J. Whiting, Human vs model observers in anatomic backgrounds, in Medical Imaging 1998: Image Perception, 1998.
16. H. H. Barrett, C. K. Abbey, and B. Gallas, Stabilized Estimates of Hotelling-Observer Detection Performance in Patient-Structured Noise, *Proc. SPIE* **3340**, 27 - 42, 1998.
17. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, Hotelling Trace Criterion and Its Correlation With Human- Observer Performance, *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.* **4**, 945-953, 1987.
18. T. Peli, Multiscale fractal theory and object characterization, *J. Opt. Soc. Am. A* **7**, 1101 - 1112, 1990.
19. S. Peleg, J. Naor, R. Hartley, and D. Avnir, Multiple Re solution Texture Analysis and Classification, *Trans Pat Anal Mach Intel* PAMI-**6**, 518 - 523, 1984.
20. B. Mandelbrot, <u>The Fractal Geometry of Nature</u>, W H Freeman, New York, 1983.
21. J. L. Solka, C. E. Priebe, and G. W. Rogers, An initial assessment of discriminant surface complexity for power law features, *Simulation* **58**, 311 - 318, 1992.
22. U. Manber, <u>Introduction to Algorithms</u>, Addison-Wesley, New York, 1989, 478.

Incorporation of an iterative, linear segmentation routine into a mammographic mass CAD system

David M. Catarious, Jr.[a)]
Department of Biomedical Engineering, Duke University and Digital Imaging Research Division,
Department of Radiology, Duke Medical Center, Durham, North Carolina 27710

Alan H. Baydush
Department of Radiation Oncology, Physics Division, Duke University Medical Center
and Digital Imaging Research Division, Department of Radiology, Duke University Medical Center

Carey E. Floyd, Jr.
Digital Imaging Research Division, Department of Radiology, Duke University Medical Center and
Department of Biomedical Engineering, Duke University

In previous research, we have developed a computer-aided detection (CAD) system designed to detect masses in mammograms. The previous version of our system employed a simple but imprecise method to localize the masses. In this research, we present a more robust segmentation routine for use with mammographic masses. Our hypothesis is that by more accurately describing the morphology of the masses, we can improve the CAD system's ability to distinguish masses from other mammographic structures. To test this hypothesis, we incorporated the new segmentation routine into our CAD system and examined the change in performance. The developed iterative, linear segmentation routine is a gray level-based procedure. Using the identified regions from the previous CAD system as the initial seeds, the new segmentation algorithm refines the suspicious mass borders by making estimates of the interior and exterior pixels. These estimates are then passed to a linear discriminant, which determines the optimal threshold between the interior and exterior pixels. After applying the threshold and identifying the object's outline, two constraints on the border are applied to reduce the influence of background noise. After the border is constrained, the process repeats until a stopping criterion is reached. The segmentation routine was tested on a study database of 183 mammographic images extracted from the Digital Database for Screening Mammography. Eighty-three of the images contained 50 malignant and 50 benign masses; 100 images contained no masses. The previously developed CAD system was used to locate a set of suspicious regions of interest (ROIs) within the images. To assess the performance of the segmentation algorithm, a set of 20 features was measured from the suspicious regions before and after the application of the developed segmentation routine. Receiver operating characteristic (ROC) analysis was employed on the ROIs to examine the discriminatory capabilities of each individual feature before and after the segmentation routine. A statistically significant performance increase was found in many of the individual features, particularly those describing the mass borders. To examine how the incorporation of the segmentation routine affected the performance of the overall CAD system, free-response ROC (FROC) analysis was employed. When considering only malignant masses, the FROC performance of the system with the segmentation routine appeared better than the previous system. When detecting 90% of the malignant masses, the previous system achieved 4.9 false positives per image (FPpI) compared to the post-segmentation system's 4.2 FPpI. At 80% sensitivity, the respective FPpI were 3.5 and 1.6. © *2004 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1738960]

Key words: mammographic mass segmentation, computer-aided detection (CAD), mammography, image processing, linear discriminant

I. INTRODUCTION

Breast cancer is the second-most deadly type of cancer for women in the United States.[1] The American Cancer Society estimates that in 2003, invasive breast cancer will be diagnosed in 211 300 women and will kill almost 40 000 women.[1] Survival rates are significantly higher when the cancer is detected at an early stage.[2–4] The 5-year survival rate for patients with localized breast cancer is 97%, while patients with distant metastases have a 5-year survival rate of 23%.[1] It is clear that detecting breast cancer at an early stage is critical to patient care.

The most common and effective early-detection tool currently available to clinicians is screening mammography. To aid mammographers in reading mammograms, research has been directed towards developing computer-aided detection (CAD) and computer-aided diagnosis tools. CAD algorithms

may operate differently than mammographers and thus may have the ability to add a unique viewpoint. Mammograms are read more accurately when read by more than one mammographer;[2,5,6] unfortunately, having multiple mammographers read the same case is neither time- nor cost-efficient. CAD systems have been demonstrated to be able to serve as reliable, accurate, and efficient second-readers to aid mammographers.[5,7,8]

One of the key components to most CAD systems is the segmentation of regions that are potential masses. Segmentation algorithms are designed to accurately identify the border of a particular object, such as a mass or calcification in a mammogram. Since the border of a mass may be indicative of its pathology, describing the mass border can have an impact on the diagnostic performance of the CAD system. In addition, the accuracy of both morphological and textural measurements of a mass is influenced by the correct identification of the mass border. If a segmentation procedure does not perform well, the features used to describe the suspicious region may not be accurate, causing the CAD system to perform at a suboptimal level.

In the past, CAD researchers have implemented several different segmentation schemes. For example, Huo et al. employed gray level region growing, which successively adds neighboring pixels to a region if they meet a specified criterion.[9,10] Petrick et al. developed a method known as density-weighted contrast enhancement which combined adaptive filtering and edge detection.[11] te Brake et al. explored discrete dynamic contour models, which segment objects by balancing internal and external energy functions.[12]

In this paper, we present a simple and efficient procedure to segment potential masses based on an iterative, gray level-based linear discrimination. We examine the capability of the segmentation routine as applied to mammographic masses. We also incorporate the segmentation procedure into a mammographic mass CAD system and examine its effect on overall system performance.

II. MATERIALS AND METHODS

II.A. Overview

We present a new segmentation routine for use with mammographic masses. Briefly, the proposed segmentation method constructs outlines of mammographic structures by employing linear decision models to differentiate the structure's interior and exterior pixels. After applying a decision threshold to estimate the object's border, two border constraints are applied to decrease the influence of background noise on the result. This procedure iterates until a stopping criterion is achieved.

The performance of the segmentation routine is judged by (1) its influence on morphological and textural features measured from CAD-identified suspicious regions and (2) the change in the FROC performance achieved by incorporating the segmentation routine into the CAD system. The CAD system is reviewed in Sec. II B. After discussing the system's components, we detail the implementation of the segmentation procedure in Sec. II C. Section II D discusses the mam-
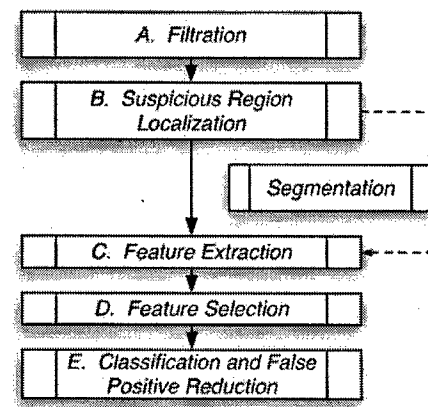
FIG. 1. Flowchart showing the components of the previously developed mass CAD system. The dashed lines show the placement of the new segmentation algorithm.

mographic images that were selected for this study. The evaluation procedure is provided in Sec. II E.

II.B. Previous CAD system

The CAD system developed previously (Fig. 1) is a multi-stage algorithm consisting of (A) filtration, (B) suspicious region localization, (C) feature extraction, (D) feature selection, and (E) classification/false-positive reduction. The filtration (A) is performed with a difference of Gaussians (DOG) filter implemented via normalized cross correlation. The suspicious region localization (B) is based on a progressive gray level thresholding procedure. The features extracted in (C) include both morphological and textural measurements and are selected (D) via a stepwise procedure. Finally, the classification and false-positive reduction (E) were performed with Fisher's linear discriminant. Each portion of the algorithm is discussed below. This system follows the same basic structure that was described in Catarious et al.[13]

1. Filtration

To identify potential masses, we employed a DOG filter. Past research[14-16] has demonstrated the usefulness of DOG filters for similar tasks because they perform both mass detection and background suppression in one step. To employ the DOG filter to search for potential masses, we used normalized cross correlation (NCC).[17,18]

2. Suspicious region localization

The areas in the filtered image that best match the filter template will contain the highest NCC output values. To distinguish suspicious regions from the rest of the image, we employed a multi-level thresholding technique similar to that used by previous researchers.[17,19,20] We define a set of thresholds based on the gray level histogram of the filtered image. At each threshold level, a new image containing suspicious regions is created.

To combine these images into one, we calculated the *duration* of the regions. The duration of a region is defined as

the number of thresholds that the region exists as an *independent* entity (i.e., number of sequential thresholds for which it grows without merging with a neighboring region). As the threshold percentage level gets higher (thresholds get lower), regions grow and merge with one another. To prevent this merging from occurring, the regions are extracted from the thresholded images at the end of their duration (i.e., just before merging) and are combined into one binary image, called the *duration image*.

3. Feature extraction

A total of 20 features, both morphological and textural, were measured for each suspicious region in the duration images. The morphological features measured were area, eccentricity, major and minor axis length, area of the convex hull, equivalent diameter, solidity (area/area of the convex hull), extent (area/area of the bounding box), circularity, and seven features derived from the normalized radial length (NRL):[11] NRL mean, standard deviation, entropy, area ratio, zero crossing count, spread, and change. Details on the first five of the NRL features can be found in Petrick *et al.*[11] and Kilday *et al.*,[21] while details of the latter two can be found in Catarious *et al.*[13] The texture features included the mean, peak, and average output of the DOG filter within each suspicious region as well as contrast.

4. Feature selection

Once the features have been extracted from the images, a reduced set of features are identified.[22,23] The goal of feature selection is to identify a subset of features, denoted A, that improves the discrimination of the regions. Feature selection can reduce computation time, eliminate redundant/linear dependent features, eliminate noisy features, and simplify the classification process.

We implemented a version of stepwise feature selection (SWFS).[24] In SWFS, features are alternately added and deleted from A. A feature is added to A if its inclusion results in higher classification accuracy (as judged by the empirical area under the ROC curve, denoted AUC). Similarly, a feature is removed from A if classification performance improves with the deletion of a previously included feature. The process of adding and deleting features from A halts when the performance criterion stops improving or a certain number of features is achieved.

Although SWFS is not guaranteed to provide the optimal subset of features, performing an exhaustive search with 20 features is computationally intractable.

5. Classification and false-positive reduction

Once the reduced set of features has been selected a discrimination function is employed to make the overall classification decision. Some of the more popular classifiers are based on linear discriminant analysis,[22,25] artificial neural networks,[12,26,27] and rule-based methods.[14,28] Each has shown success in both detection and diagnostic settings.

To separate the masses from other mammographic structures, we implemented a linear classifier, or a linear discrimi-

nant function. Specifically, we implemented Fisher's linear discriminant,[24] which, given a set of multidimensional data from two classes, projects the data onto the line that maximally separates the means of the two classes while minimizing the variance within each class.

II.C. Mass segmentation

The proposed segmentation routine has been developed because, although the duration image technique (Sec. II B2) can accurately identify the most suspicious regions in the image, the segmentations of the masses do not reflect the detailed morphology of the mass. The inaccuracy of the method arises mainly because the object borders are determined from the filtered images, not the original images. Since the DOG filter is designed to look for round masses, the filtered versions of the original images contain round blobs. Even masses that are not round are replaced with semi-round blobs. Details about the mass border, such as fine spiculations, are lost in this procedure. Thus, the features that relate to the mass border are affected.

To recover these features, we have developed an iterative, gray level, linear segmentation procedure. The procedure begins by examining a region of interest (ROI) that is identified by the CAD system as containing a suspicious region. Unsharp masking is applied to the ROI to compensate for background nonuniformity. The procedure then iterates by estimating the pixels interior and exterior to the object, determining an optimum gray level threshold to separate the interior and exterior pixels, and constraining the resulting object border. The procedure halts when a stopping criterion has been achieved.

The input to the algorithm is a ROI containing a suspicious region [Fig. 2(a)]. For each suspicious region in the duration image, the initial seed point is selected as the pixel with the highest gray value within 3 mm (15 pixels) of the centroid of the region. Around the seed point, a square, 42.6 mm (213 pixel) ROI, centered at the seed point, is extracted from the unsharp masked image [Fig. 2(b)].

For the initial iteration, the border of the object is selected to be a circle of radius 16 mm that surrounds the center of the ROI [Fig. 2(b)]. All pixels inside the circle are considered interior, while all pixels outside the circle are considered exterior. To refine the estimate of the object's border, a threshold to separate the object's interior and exterior pixels is computed via Fisher's linear discriminant:

$$t = \overline{x}^T S^{-1}(\overline{x}_{int} - \overline{x}_{ext}) - \frac{1}{2}(\overline{x}_{int} - \overline{x}_{ext})^T S^{-1}(\overline{x}_{int} + \overline{x}_{ext}),$$

where the scalar t is the threshold, $\overline{x}$ is the vector of pixel values, $\overline{x}_{int}$ and $\overline{x}_{ext}$ are the sample means of the values of the interior and exterior pixels as defined in the previous segmentation, and S is the sample covariance matrix. In this instance, gray level value is the only feature used to discriminate between the interior and exterior pixels. Thus, each of the vectors in the discriminant function reduces to a scalar. The covariance matrix simplifies to the pooled variance of the gray levels of the interior and exterior pixels. Fisher's

(a)   (b)   (c)   (d)
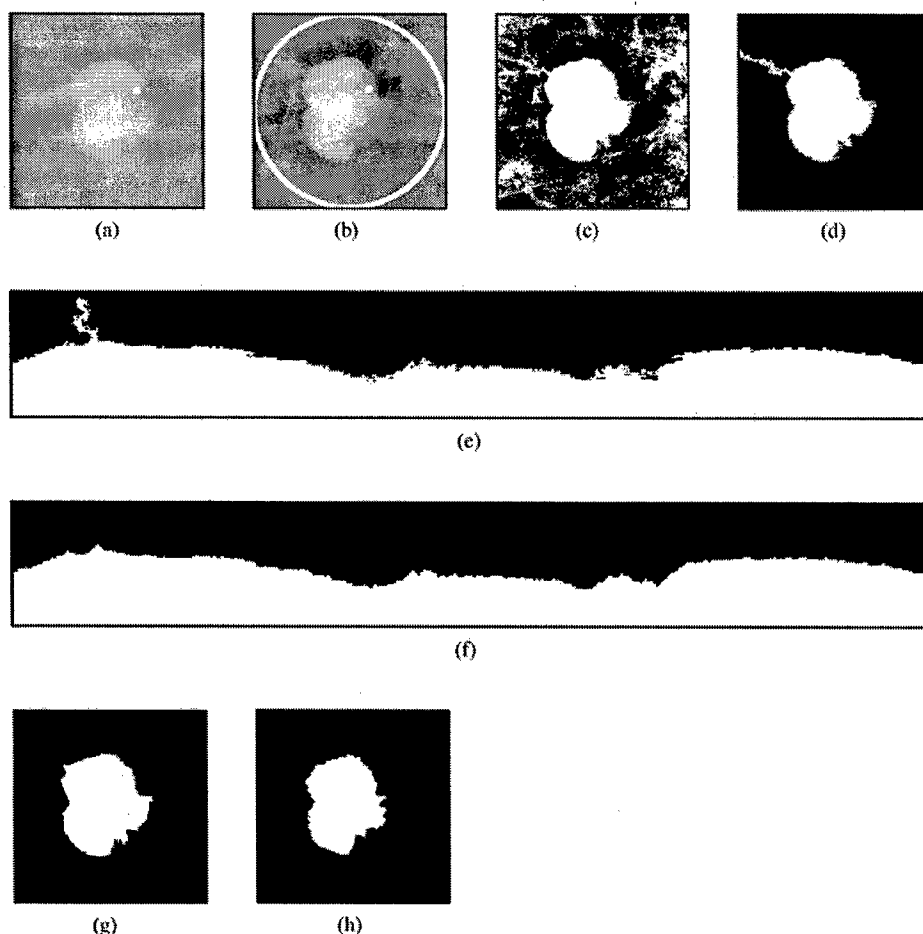

(e)

(f)

(g)   (h)

FIG. 2. The progression of the new segmentation method. (a) A 42.6 mm by 42.6 mm (213 pixels) ROI containing a malignant mass. (b) The ROI in (a) after applying the unsharp masking to reduce the influence of a nonuniform background. The circle around the mass represents the initial segmentation. (c) The ROI after the first optimal threshold has been applied. (d) The center connected component of the thresholded ROI in (b). (e) The $[r, \theta]$ matrix computed from the central connected component in (b). (f) The $[r, \theta]$ matrix after the constraints have been applied. (g) The segmentation after the algorithm's first iteration. (h) The final segmentation (after four iterations).

linear discriminant projects the data onto the line that best separates the class means relative to the variance. The threshold t is the midpoint on the projected line.

The resulting threshold t is then applied to the ROI [Fig. 2(c)]. The connected region that contains the center of the thresholded ROI is selected as the candidate region segmentation [Fig. 2(d)]. Selecting only the center region eliminates any neighboring, but unconnected, structures that were above the threshold t.

At this stage, it is possible that background structures are identified as being part of the interior. Two constraints are applied to the new object outline: (1) the interior pixels on each ray emanating from the center must have gaps of no more than d pixels, and (2) the pixels along the object's border must be within a specified distance of their immediate neighbors.

Before applying the constraints, the binary ROI containing the segmentation estimate is transformed into polar coordinates. The center of the ROI serves as the origin and rays of length r are extracted at each angle, θ. The rays, r, have a length of 80 pixels. The result of the transformation is a matrix with dimensions $[r, \theta]$, where ones and zeros represent the segmentation's interior and exterior pixels, respectively. An example of the resulting $[r, \theta]$ matrix is given in Fig. 2(e).

After the matrix is fully constructed, the first constraint is applied. The algorithm searches in the r dimension for interior pixels separated by more than d pixels. In vectors where this occurs, the "border pixel" is selected to be the last interior pixel before the gap. This constraint helps to eliminate random structures that may cross through the region from being included in the segmentation. One example of such a spurious structure appears at the 10 o'clock position in Fig. 2(c) and again in Fig. 2(d). As seen in Fig. 2(f), the structure is eliminated by this constraint. Note that some gap between neighbors is allowed in case the suspicious mass does not have interior pixels uniformly above the chosen threshold t.

The second constraint controls the roughness of the segmented border. Since large distances between neighboring border pixels may be caused by the presence of noisy background structures, the border is adjusted to limit the distances between each border pixel in the $[r, \theta]$ matrix. Beginning at the first r-vector in the matrix, the border is traversed. The traversal proceeds as follows: Let $[r_{-1}, \theta_{-1}]$ and $[r_0, \theta_0]$ represent the previous border pixel and the current border pixel being examined. If the city-block distance between $[r_{-1}, \theta_{-1}]$ and $[r_0, \theta_0]$ is less than a specified distance n, $[r_0, \theta_0]$ is accepted as a border pixel. If the distance is greater than n, $[r_0, \theta_0]$ is adjusted to be n pixels from $[r_{-1}, \theta_{-1}]$. Although large well-defined spiculations will still be included after applying this constraint, fine spiculations and other border subtleties may be excluded from the

segmentation. An example of the border remaining after applying these constraints is shown in Fig. 2(f).

Once both constraints have been applied, the current iteration is complete [Fig. 2(g)]. If the stopping criterion is met, the procedure is completed. The change in segmentations between the current and previous segmentations is used as a stopping criterion. If the stopping criterion is not met, another iteration is performed. The result of the current iteration is used as the initial segmentation for the next iteration. Since there is no previous segmentation on the first iteration, this algorithm will iterate at least twice. An example of a final segmentation is given in Fig. 2(h).

Since the determination of the gray level threshold is a major portion of this algorithm, it should be noted that other threshold selection techniques have been developed. For example, Otsu[29] developed a method that selects the threshold value that maximizes the between-class variance (and thus minimizes the interclass variance) of the gray level histogram. Using the refinement suggested by Reddi et al.,[30] this threshold can be calculated in an efficient manner. However, unlike Fisher's linear discriminant, their method is unsupervised and thus is not suitable for an iterative framework. Since the proposed routine improves after each iteration, the threshold selection technique must be adaptable to iterative implementation. Making no assumptions about the underlying distributions, Fisher's linear discriminant determines the line that best separate the gray level means of the interior and exterior pixels. Since it is easily implemented as an iterative process, Fisher's linear discriminant is an ideal choice to calculate the threshold values.

II.D. Database of mammograms

The mammograms for this study were extracted from the University of South Florida's Digital Database for Screening Mammography (DDSM).[31] With each mammographic image, the DDSM contains information about any lesions it contains, including BI-RADS™[32] assessment, subtlety, patient age, and breast density. Also included with each DDSM image is a chain code that defines the lesion boundary that was indicated by a radiologist. Using this information, a "truth" image was created for each image in the study database. Since the system only examines masses, only mass locations were stored in the truth images for this study.

A "study database" of 183 mammographic images from 169 patients was collected from the DDSM. The study database consists of 83 images containing 50 benign and 50 malignant masses, and 100 "normal" images containing no abnormalities. All images were originally scanned with a Lumisys scanner at a resolution of 50 microns per pixel at a bit depth of 12.[31] Although the images in the study database were randomly selected, the distribution of mass descriptors closely matched that of the entire collection of masses scanned with the Lumisys scanner. No masses having a shape of "architectural distortion" as the primary finding were included in the study database. The average diameter of the masses in the study database was 17 mm.

After the cases were selected, they were resized via zero padding to be a uniform size. They were then subsampled by a factor of 4, resulting in a pixel resolution of 200 microns per pixel, which is in a range consistent with that of other researchers.[14,33,34]

II.E. Procedure

To perform this study, we created two CAD systems (systems A and B). System A is the previously developed system, while system B includes the new segmentation routine inserted after the suspicious region localization stage. Thus, the first two stages of each system are exactly the same. To begin our examination, we ran the study database through the first two stages of the CAD system, that is, the filtration and suspicious region localization stages. The systems employed a DOG filter constructed of Gaussians with symmetric widths of 18 and 9 mm (90 and 45 pixels). The size of the DOG filter template and the NCC templates was 24 mm (120 pixels). The duration images were created with seven thresholds of 1%–13% in steps of 2%.

At this point, the regions identified by system A were processed by the feature extraction stage. In system B, the suspicious regions were processed first by the new segmentation stage and then by the feature extraction stage. For the segmentation routine, several combinations of d and n were explored. The parameters that provided the best results and were used in the final version were 3 and 2, respectively. The stopping criterion employed was that the object boundary ceased changing. The weighting on the unsharp masking operation was 0.9.

To determine the effect of the segmentation algorithm, a ROC study was performed to compare the discriminatory power of the 20 individual features extracted from each of region before and after the new segmentation algorithm was applied. Since the shapes of the regions were different after they were segmented, a few suspicious regions no longer corresponded to a mass and vice versa. Thus, a paired t-test could not be employed to judge the performance differences. In order to take advantage of the regions that were paired, partially paired t-tests were performed using the ROCKIT software package (Charles Metz, University of Chicago, Chicago, IL).

After the feature extraction stage, each system progressed through the feature selection and classification stages. The stopping criterion adopted was when the empirically measured AUC did not increase more than 0.005. During the feature selection process, the entire study database was used for training. In the classification stage, the systems were trained and tested using a round-robin sampling. The overall systems' performances were examined using both ROC analysis and FROC curves. The ROC analysis was performed on the ROIs to determine if there was a statistical difference in the final performance of the systems at a significance level of 0.05. When using ROC analysis to examine system performance, the system input was the same set of ROIs for systems A and B. However, since the CAD system also performs the detection task, it is useful to examine the results via FROC curves. Using the FROC curve, the
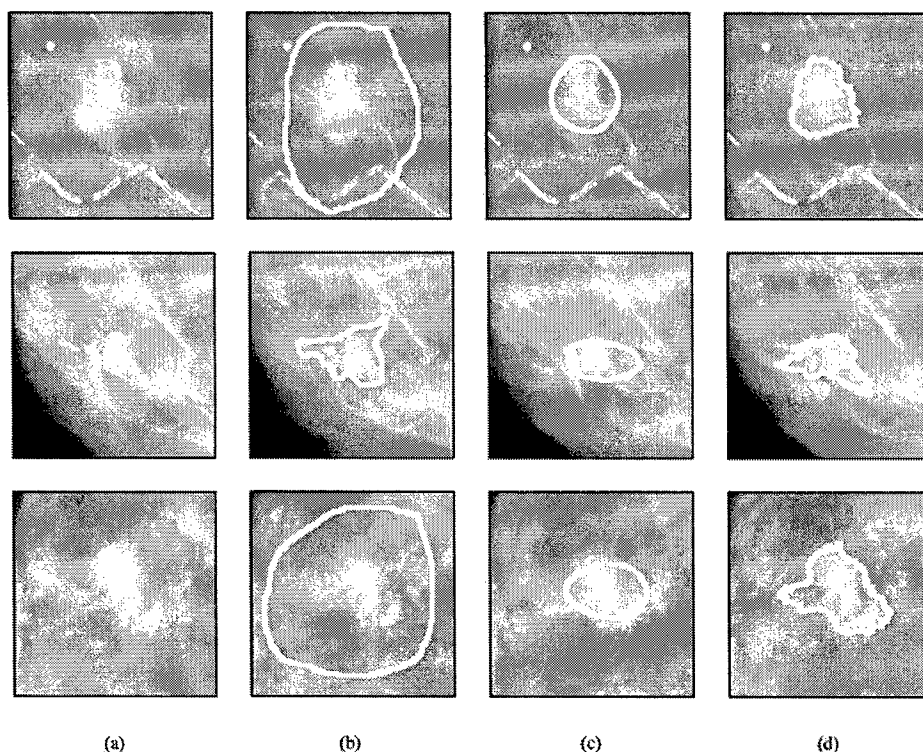
FIG. 3. 40 mm by 40 mm ROIs showing two malignant masses and one benign (bottom) mass. (a) The detected mass from the unprocessed mammographic image. (b) The mass outline provided by the DDSM. (c) The segmentations provided by the duration image technique. (d) The masses segmented with the segmentation routine.

(a)   (b)   (c)   (d)

sensitivity can be compared against false positives per image (FPpI) instead of false positive fraction. Along with the final system performance, the performance on only the malignant masses was also examined.

A suspicious region was deemed to be a true positive if it met the following two criteria: (1) the region intersected with the true positive region, outlined in the truth images, and (2) the centroid of the suspicious region was no more than 16 mm (80 pixels) from the centroid of the region in the truth file.

The computations for this study were performed on a machine with dual 1.8 GHz AMD (Advanced Micro Devices, Inc., Sunnyvale, CA) processors. The CAD and segmentation systems were programmed in MATLAB® (The Math Works, Inc., Natick, MA).

III. RESULTS

Before the Feature Extraction stage (and after the suspicious region localization stage in Fig. 1), systems A and B were able to detect 98% (49/50) of the malignant masses and 88% (44/50) of the benign masses, for an overall detection performance of 93%. The total number of false positive regions identified in the 183-image study database was approximately 3600, for an average of 19.7 FPpI.

Some examples of segmented objects can be seen in Fig. 3. Shown in the figure are (a) three masses (two malignant, one benign) extracted from the original mammographic image, (b) the outline of the mass provided by the DDSM, (c) the segmentation provided by the duration image technique, and (d) the segmentations computed with the new segmentation routine. It is clear that the segmentations provided by the

iterative algorithm in system B have much more structure than the segmentations provided by system A's duration images. As seen in Fig. 3, the region boundaries in the truth files were not necessarily intended to follow the mass morphology and so were not compared to our segmentation results. The segmentation routine in system B required an average of 7.8 iterations for each suspicious region with an average of 1.7 s per iteration.

Table I shows the AUC achieved by each feature extracted from the regions in systems A and B. Two AUCs are provided for each feature, one for each system. The order in which the features were selected by the SWFS algorithm as well as the cumulative AUCs as the features were added are given in the columns next to the AUC values. There were statistically significant performance changes for 11 of the 20 features, shown in the p-values in the right hand column of Table I. Both systems selected four features. Of the four features chosen, the systems had only one in common, the peak output of the DOG filter. Additionally, system A selected NRL mean, equivalent diameter, and NRL change, while system B selected minor axis length, NRL spread, and solidity.

System A achieved an overall AUC of 0.91 while system B achieved an overall AUC of 0.91. There is no statistical difference between the overall performances of the systems (p-value of 0.82). When considering the performance on just the malignant masses, system A achieved an AUC of 0.92 while system B achieved an AUC of 0.93. Once again, no statistical difference between the systems was found (p-value of 0.41).

For both systems, Fig. 4(a) shows FROC curves describ-

TABLE I. The individual AUCs for each of the 20 features measured for system A and system B. The order in which the features were selected by the stepwise feature selection (SWFS) and the cumulative AUC is also provided. The final column provides the p-values of the change in performances from system A to system B.

Features	System A AUC	System A Order selected (cumulative AUC)	System B AUC	System B Order selected (cumulative AUC)	p-value
Area	0.79		0.81		0.31
Eccentricity	0.63		0.75		<0.0001
Major axis length	0.73		0.68		0.0032
Minor axis length	0.82		0.83	1 (0.83)	0.17
Area of the convex hull	0.79		0.79		0.085
Equivalent diameter	0.79	3 (0.91)	0.81		0.31
Solidity	0.64		0.71	4 (0.91)	0.0009
Extent	0.52		0.69		<0.0001
Mean DOG output	0.77		0.69		0.087
Peak DOG output	0.83	1 (0.83)	0.82	3 (0.91)	0.71
Std. dev. of DOG output	0.81		0.79		0.0096
Circularity	0.67		0.82		<0.0001
Contrast	0.72		0.73		0.06
NRL mean	0.64	2 (0.90)	0.82		<0.0001
NRL std. dev.	0.64		0.72		0.0046
NRL entropy	0.62		0.64		0.20
NRL area ratio	0.64		0.80		<0.0001
NRL zero crossing	0.59		0.70		0.0012
NRL spread	0.65		0.80	2 (0.87)	<0.0001
NRL change	0.73	4 (0.91)	0.76		0.11

ing both the overall performance and the performance on only the malignant masses. As can be seen, both systems perform better on malignant masses than on overall masses. Figure 4(b) shows a partial view of the FROC curve in Fig. 4(a), showing only the area above 60% sensitivity and less than 6 FPpI. In the range from ~1 to 6 FPpI, the overall performances of systems A and B cross and overlap in several places. For malignant masses, system A has a performance advantage from 5.8 to 9 FPpI, a range of 94% to 96% sensitivity. However, system B outperforms system A below 5.8 FPpI, achieving 1.6 FPpI at 80% sensitivity on the malignant masses compared to system A's 3.5 FPpI.

IV. DISCUSSION

From examining Table I, it can be seen that the segmentation routine made a statistically significant difference in the discriminatory ability of 11 of the features (9 increased with segmentation while 2 decreased). Intuitively, since the shape of masses is distinctive, we would expect more accurate segmentations to increase the AUCs of the individual features describing the border. The fact that the segmentation routine captures important information in the details of the border can be seen in the significant performance increases of five of the seven NRL features: NRL area ratio, NRL mean, NRL spread, NRL standard deviation, and NRL zero crossing. The remaining two NRL features did not change with statistical significance (p-values greater than 0.11).

Additionally, the improved accuracy in the description of the masses' overall shapes is evident from the improvement of circularity, extent, eccentricity, and solidity. Circularity made one of the more dramatic increases in performance,

rising from 0.67 to 0.82 with a p-value of <0.0001, making it one of the better performing features in system B. This increase is not surprising because, after the segmentation, the majority of nonmass objects should be less circular than they were previously. Overall, the increase in the effectiveness of many of the morphological features validates the segmentation algorithm.

Only 2 of the 20 features significantly decreased in ROC performance: major axis length and standard deviation of the DOG filter output. We feel that the major axis length was more effective presegmentation because many of the nonmass objects in the duration image were long and thin. After being segmented, the long, thin objects become more constrained in size, making it more difficult to make a classification based on the major axis length.

The effectiveness of the standard deviation of the DOG output decreased due to the increased accuracy of the object borders. Since the segmentation procedure groups pixels with similar gray values, the standard deviation of the DOG output does not vary greatly between mass and nonmass objects.

Due to the limited size of the study database, all of the data used to select the features was also examined in the classification stage. Thus, some bias is present in our results. In an effort to minimize this bias, the mammograms in the study database were chosen to be representative (in terms of BI-RADS™ descriptors) of the entire Lumisys-scanned set of mammograms. This issue will be resolved in the future by gathering a larger study database and separating the cases and for training and testing.

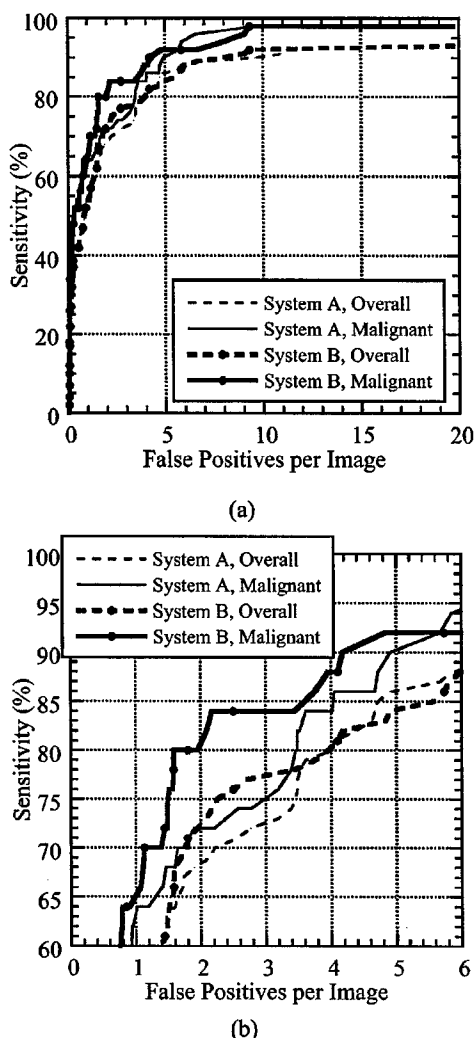Unfortunately, we were not able to observe a statistical

FIG. 4. (a) FROC curves comparing the performances of systems A (without circles) and B (with circles). Both overall performance (dashed lines) and malignant performance (solid lines) are shown. (b) A magnified view of the FROC curve in (a).

difference between the total ROC performances of the systems. However, we still find value in the segmentation routine because, as mentioned above, the segmentation algorithm improved the efficacy of several of the features, particularly the border-describing features. Although the ROC analysis could not provide statistical significance to the difference in overall system performance, the FROC curves indicate how the incorporation of the segmentation routine positively affected the CAD system's performance in some key ranges. Although the curves for the overall performance crossed a few times in the range from 1 to 10 FPpI, the performance of system B on the malignant masses clearly exceeded that of system A from ~0.9 to 5.8 FPpI, corresponding to a range from 60% to 94% sensitivity. Thus, the segmentation routine was able to capture the distinct border characteristics of malignant masses and more easily distinguish them from other structures.

Since it can be argued that detecting malignant masses is more important than detecting benign masses, we are encour-

aged that our system performs better on the malignant masses in the lower range of FPpI. When detecting 90% of the malignant masses, system A achieves 4.9 FPpI compared to system B's 4.2 FPpI, a decrease of 14%. At 80%, system A and system B's respective FPpI are 3.5 and 1.6, a decrease of 54%. Preserving a high level of sensitivity as false positives are reduced is key to the success of a CAD system; it has been demonstrated that, at a constant system sensitivity, reducing the system's FPpI increases a mammographer's performance.[35]

The segmentation routine, however, did not successfully segment each region. For less than 1% of the suspicious regions, the resulting segmentation was a single pixel. Only one of the single pixel segmentations corresponded to a benign mass. In that instance, the mass was larger than the segmentation window. Since the interior of the mass was relatively uniform, no structure was present to be segmented. To deal with masses larger than the window size, a future improvement will be to adaptively set the window size. In each of the remaining single pixel segmentations, the suspicious regions were in a flat gray level region. They were identified as suspicious regions mainly because they were neighbors of other structures identified by the DOG filter. Thus, since no structure was present, a single pixel is an acceptable segmentation.

The ROIs shown in Fig. 3 demonstrate the qualitative effectiveness of this new segmentation procedure. The segmentations presented seem to closely follow the border of the masses in each case. Given that the average segmentations only took 7.8 iterations, we feel it should be incorporated into our CAD system.

V. CONCLUSION

In this study, we integrated a new algorithm to segment suspicious regions in a mammographic mass CAD system. The proposed segmentation algorithm is an iterative procedure that utilizes a linear discriminant function to separate an object's interior pixels from its exterior pixels. The algorithm requires only two parameters: d, the maximum distance between neighboring pixels on each ray, and n, the allowable distance between neighboring border pixels. The inclusion of the two constraints on the boundaries helps to exclude spurious background structures. On average, the procedure completes in only 7.8 iterations. The procedure is based upon established statistical techniques and is straightforward to implement.

Unfortunately, the accuracy of the segmentation routine could not be assessed against the radiologist-drawn boundaries included in the DDSM database. In many cases, the provided outlines were generous and went beyond the borders present on even the most well defined masses.

However, the increased accuracy of the individual mass features validates the segmentation routine's performance. It was found that the segmentation routine affected the performance of individual features in a predictable and intuitive manner; most of the features describing the mass border increased with statistical significance. As seen in Fig. 4, the

segmentation routine greatly aided the performance of the CAD system on malignant masses at a critical region of the FROC curve, where sensitivity is greater than 80% and FPpI are less than 5. The addition of the segmentation made the largest difference in the system's efficacy on malignant masses.

The borders of mammographic masses have been shown to be important in discriminating them from other structures. Since the discrimination ability of most of the features used to measure a mass' border increased significantly and the system's performance on malignant masses was improved, we feel that the introduced segmentation routine is an appropriate and necessary addition to our CAD system.

ACKNOWLEDGMENTS

[a]Electronic mail: david.catarious@duke.edu

[1] ACS, "American Cancer Society: Cancer Facts and Figures 2003. Atlanta, Ga: American Cancer Society 2003" (2003).

[2] E. L. Thurfjell, K. A. Lernevall, and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology 191, 241–244 (1994).

[3] I. Anttinen, M. Pamilo, M. Soiva, and M. Roiha, "Double reading of mammography screening films: one radiologist or two?" Clin. Radiol. 48, 414–421 (1993).

[4] W. R. Hendee, C. Beam, and E. Hendrick, "Proposition: all mammograms should be double-read," Med. Phys. 26, 115–118 (1999).

[5] 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).

[6] C. E. Metz and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images," Med. Decis Making 12, 60–75 (1992).

[7] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa, and Y. Jiang, "Computer-aided diagnosis in radiology: potential and pitfalls," Eur. J. Radiol. 31, 97–109 (1999) (review).

[8] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology 219, 192–202 (2001).

[9] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 3rd ed. (Addison–Wesley, New York, 1993), p. 716.

[10] Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," Med. Phys. 22, 1569–1579 (1995).

[11] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," Med. Phys. 23, 1685–1696 (1996).

[12] G. te Brake and N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms," Med. Phys. 28, 259–266 (2001).

[13] D. M. Catarious, Jr., A. H. Baydush, C. K. Abbey, and C. E. Floyd, Jr., "A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results," in SPIE Medical Imaging 2003, San Diego, CA (San Diego, CA, 2003), p. 1927.

[14] B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," Acad. Radiol. 2, 959–966 (1995).

[15] B. Zheng, Y. H. Chang, and D. Gur, "Adaptive computer-aided diagnosis scheme of digitized mammograms," Acad. Radiol. 3, 806–814 (1996).

[16] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Trans. Med. Imaging 16, 811–819 (1997) (98192265).

[17] M. J. Carreira, D. Cabello, M. G. Penedo, and A. Mosquera, "Computer-aided diagnoses: Automatic detection of lung nodules," Med. Phys. 25, 1998–2006 (1998).

[18] G. M. te Brake and N. Karssemeijer, "Single and multiscale detection of masses in digital mammograms," IEEE Trans. Med. Imaging 18, 628–639 (1999).

[19] M. L. Giger, K. Doi, and H. MacMahon, "Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields," Med. Phys. 15, 158–166 (1988).

[20] M. L. Giger, K. Doi, H. MacMahon, C. E. Metz, and F. Yin, "Pulmonary nodules: Computer-aided detection in digital chest images," Radiographics 10, 41–51 (1990).

[21] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging 12, 664–669 (1993).

[22] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space," Phys. Med. Biol. 40, 857–876 (1995).

[23] B. Sahiner, H. P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue," Med. Phys. 23, 1671–1684 (1996).

[24] M. Nadler and E. P. Smith, Pattern Recognition Engineering (Wiley, New York, 1993), p. 588.

[25] 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. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," Radiology 212, 817–827 (1999).

[26] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," Radiology 187, 81–87 (1993).

[27] B. Zheng, Y. Chang, W. F. Good, and D. Gur, "Performance gain in computer-assisted detection schemes by averaging scores generated from artificial neural networks with adaptive filtering," Med. Phys. 28, 2302–2308 (2001).

[28] L. Li, Y. Zheng, L. Zheng, and R. A. Clark, "False-positive reduction in CAD mass detection using a comptetive classification strategy," Med. Phys. 28, 250–258 (2001).

[29] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern. 9, 62–66 (1979).

[30] S. S. Reddi, S. F. Rudin, and H. R. Keshavan, "An optimal multiple threshold scheme for image segmentation," IEEE Trans. Syst. Man Cybern. 14, 661–665 (1984).

[31] M. Heath, K. W. Bowyer, and D. Kopans, "Current Status of the Digital Database for Screening Mammography," in Digital Mammography, edited by J. Hendriks (Kluwer Academic, Dordrecht, 1998), pp. 457–460.

[32] BI-RADS, American College of Radiology Breast Imaging—Reporting and Data System (BI-RADS), 3rd ed. (American College of Radiology, 1998).

[33] Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," Med. Phys. 28, 455–461 (2001).

[34] 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," Acad. Radiol. 6, 22–33 (1999).

[35] B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-copy mammographic radings with different computer-assisted detection cuing environments: Preliminary findings," Radiology 221, 633–640 (2001).

Development and application of a segmentation routine in a mammographic mass CAD system

David M. Catarious, Jr.[a], Alan H. Baydush[b,a], and Carey E. Floyd, Jr.[c,a]

[a]Department of Biomedical Engineering, Duke University,
Box 2623 DUMC, Durham, NC, USA;
[b]Department of Radiation Oncology, Duke University Medical Center,
Box 2623 DUMC, Durham, NC, USA;
[c]Department of Radiology, Duke University Medical Center,
Box 2623 DUMC, Durham, NC, USA

ABSTRACT

The purpose of this paper is to present a new segmentation routine developed for mammographic masses. We previously developed a computer-aided detection (CAD) system for mammographic masses that employed a simple but imprecise segmentation procedure. To improve the systems performance, an iterative, linear segmentation routine was developed. The routine begins by employing a linear discriminant function to determine the optimal threshold between estimates of an objects interior and exterior pixels. After applying the threshold and identifying the objects outline, two constraints are applied to minimize the influence of extraneous background structures. Each iteration further refines the outline until the stopping criterion is reached. The segmentation algorithm was tested on a database of 181 mammographic images that contained forty-nine malignant and fifty benign masses. A set of suspicious regions of interest (ROIs) was found using the previous CAD system. Twenty features were measured from the regions before and after applying the new segmentation routine. The difference in the features discriminatory ability was examined via receiver operating characteristic (ROC) analysis. A significant performance difference was observed in many features, particularly those describing the object border. Free-response ROC (FROC) curves were utilized to examine how the overall CAD system performance changed with the inclusion of the segmentation routine. The FROC performance appeared to be improved, especially for malignant masses. When detecting 90% of the malignant masses, the previous system achieved 4.4 false positives per image (FPpI) compared to the post-segmentation systems 3.7 FPpI. At 85%, the respective FPpI are 4.1 and 2.1.

Keywords: segmentation, computer-aided detection (CAD), mammographic masses, ROC analysis

1. INTRODUCTION

Over the past few years, we have been developing a computer-aided detection (CAD) system designed to detect masses in mammograms.[1,2] An important component of any CAD system is the ability to identify and accurately outline suspicious regions. Since the shape of a mass is highly indicative of its pathology, capturing the description of mass borders is paramount to the success of a mass CAD system. To achieve accurate segmentations, other researchers have employed several methods, including region growing, active contour segmentation, and threshold-based procedures.[3-6]

In this work, we develop a new method to segment masses as well as other mammographic structures. The quality of the segmentation routine is explored by examining its effect on the ability of morphological and textural descriptors to separate masses from non-masses. We also examine the impact that incorporating the proposed routine has on the overall performance of our existing CAD system.

Send correspondence to D.M.C.: E-mail: david.catarious@duke.edu, Telephone: 1 919 668 2539

2. METHODS AND MATERIALS

2.1. Overview

We present a newly developed segmentation routine for use with mammographic masses. Briefly, the segmentation routine estimates the borders of objects by iterative implementation of a linear decision model combined with two constraints to eliminate extraneous background influence.

The performance of the segmentation routine is judged by (1) its influence on morphological and textural features measured from CAD-identified suspicious regions (using receiver operating characteristic (ROC) analysis) and (2) the change in the free-response ROC (FROC) performance after incorporating the segmentation routine into the CAD system. We begin by discussing the database of mammograms used in this study in section 2.2. Sections 2.3 and 2.4 provide a brief review and updates to the CAD system and the extracted features. In section 2.5 we provide details of the implementation of the segmentation procedure. The study procedure is provided in 2.6.

2.2. Database of mammograms

The mammograms for this study were selected from the University of South Florida's Digital Database for Screening Mammography (DDSM).[9] Along with each image, the DDSM provides specific information on each lesion contained within the image. Using this information, a study database of 181 mammographic images from 169 patients was collected from the DDSM. The study database consists of 81 images containing 50 benign and 49 malignant masses, and 100 normal images containing no abnormalities. All images were originally scanned with a 12-bit Lumisys scanner at a resolution of 50 microns per pixel. Although the images in the study database were randomly selected, the distribution of mass descriptors closely matched that of the entire collection of masses scanned with the Lumisys scanner. Although the DDSM describes some masses as having a shape of architectural distortion, no masses with architectural distortion as the primary finding were included in the study database. The average diameter of the masses in the study database was 17 mm.

After the cases were selected, each image was subsampled by a factor of four, resulting in a pixel resolution of 200 microns per pixel, a range which is consistent with that of other researchers.[10–12]

2.3. Previously developed CAD system

Since the CAD system used in this research has been previously presented,[1] we will give only a brief overview and discuss only the portions which have changed. The CAD system consists of five components: filtration, suspicious region localization, feature extraction, feature selection, and classification and false-positive reduction.

In the filtration stage, the mammograms are filtered with a difference of Gaussians (DOG) filter using normalized cross correlation (NCC),[3, 7, 8] as described by the following equation:

$$\gamma(s,t) = \frac{\sum_x \sum_y [f(x,y) - \bar{f}(x,y)][w(x-s,y-t) - \bar{w}]}{\{\sum_x \sum_y [f(x,y) - \bar{f}(x,y)]^2 \sum_x \sum_y [w(x-s,y-t) - \bar{w}]^2\}^{\frac{1}{2}}} \tag{1}$$

where γ is the filtered image, s and t index the position of the filter template w within the image f, x and y index the pixels interior to both f and w, $\bar{w}$ is the average value of the template, and $\bar{f}(x,y)$ is the average value of the portion of the image coincident with the filter template. The denominator in Eq (1) serves to normalize the filter response between -1 and 1.

In the previous implementation, the NCC operation was implemented exactly as specified in Eq (1). However, because the gray values at the skin boundary drop rapidly, $\bar{f}(x,y)$ changes quickly until the filter template is completely inside the breast. The rapid change in $\bar{f}(x,y)$ due to the dark region surrounding the breast causes the filter response to be suppressed along the skin boundary, making it difficult to detect fine structures. To correct for this response, we adjusted Eq (1) to examine only the pixels interior to the breast (as defined by our breast outline). This adjustment causes the $\bar{w}$ term to be a constant value when the template is completely interior to the breast but to vary when the template coincides with the skin boundary.

In addition to making corrections for the skin boundary, we also adjusted the image to correct for the edge next the the chest wall. Since the current version of our CAD system examines only craniocaudal view images,

at the chest wall, the pixel values drop off to zero. Since this hard edge causes filtering effects, to keep the filter response steady, we mirrored the region adjacent to the chest wall over the chest wall boundary.

By examining the filtered images with a gray level thresholding technique, regions suspicious of being masses are localized. From these regions, twenty morphological and textural features are extracted. Using a stepwise feature selection routine, the most effective features are selected and used for classification purposes. Previously, the feature selection algorithm selected features until the performance criterion, area under the ROC curve (AUC), did not increase. This implementation resulted in the selection of a majority of the features. Currently, the feature selection algorithm will only add a feature when its incorporation results in a statistically significant increase in AUC. The statistical significance is judged by comparing the results of 1,500 bootstrap samples with the percentile method.

Once the final subset of features is selected, the CAD system employs a linear classifier to designate regions as being masses or non-masses.

2.4. Extracted Features

For this study, sixteen morphological and four textural features were extracted for each suspicious region.

The measured morphological features included area, eccentricity, major and minor axis length, area of the convex hull, equivalent diameter, solidity ($\frac{\text{area}}{\text{area of the convex hull}}$), extent ($\frac{\text{area}}{\text{area of the bounding box}}$), and circularity. In addition, seven features were measured that are derived from the normalized radial length (NRL)[5]: NRL mean, standard deviation, entropy, area ratio, zero crossing count, spread, and change. Details on the first five of the NRL features can be found in Petrick $et\ al$[5] and Kilday $et\ al$,[13] while details of the latter two can be found in Catarious $et\ al$.[1] The main purpose of the NRL features is to examine the roughness of the border of an object.

The four textural measurements examined are contrast and three features measured from the output of the DOG filter in each region: the mean, peak, and standard deviation.

2.5. Developed segmentation routine

The developed segmentation routine is an iterative thresholding procedure. Briefly, the procedure begins with an initial estimate of the border of the selected region. Then, using Fisher's linear discriminant,[14] a threshold is computed to separate the region's interior pixels from the background pixels. The resulting outline is processed by two constraints designed to minimize the influence of noisy background structures. Once the constraints are applied, the procedure repeats until there is no change in the computed outline.

To begin, a region detected in the suspicious region localization stage is selected. The center of the region is selected to be the pixel with the largest gray value within 3 mm (15 pixels) of the centroid of the region. A 42.6 mm by 42.6 mm (213 pixels by 213 pixels) region is then extracted around the chosen center (Figure 1(a)). To enhance the detected structure, the region is subjected to unsharp masking (see Figure 1(b)).

Using the unsharp masked region of interest (ROI), the initial border is estimated by a circle of radius 16 mm (80 pixels) surrounding the center of the region, as shown in Figure 1(b). Using the pixels interior and exterior to the circle, Fisher's linear discriminant is used to compute the threshold the best separates the two regions:

$$t = \bar{x}^{\mathrm{T}}\mathbf{S}^{-1}(\bar{x}_{int} - \bar{x}_{ext}) - \frac{1}{2}(\bar{x}_{int} - \bar{x}_{ext})^{\mathrm{T}}\mathbf{S}^{-1}(\bar{x}_{int} + \bar{x}_{ext}), \tag{2}$$

where the scalar t is the threshold, $\bar{x}$ is the vector of pixel values, $\bar{x}_{int}$ and $\bar{x}_{ext}$ are the sample means of the values of the interior and exterior pixels as defined in the previous segmentation, and $\mathbf{S}$ is the sample covariance matrix. Since gray level value is the only feature used to discriminate between the interior and exterior pixels, each vector in Eq (2) reduces to a scalar. The covariance matrix simplifies to the pooled variance of the gray levels of the interior and exterior pixels. Fishers linear discriminant provides the optimal separation of the two classes of sample data because it projects the data onto the line that best separates the class means relative to the variance; the threshold t is the midpoint on the projected line. Figure 1(c) shows the region in Figure 1(a) after thresholding.

Since there will be pixels above the threshold that are not part of the object of interest, only the center connected region is preserved, as in Figure 1(d).
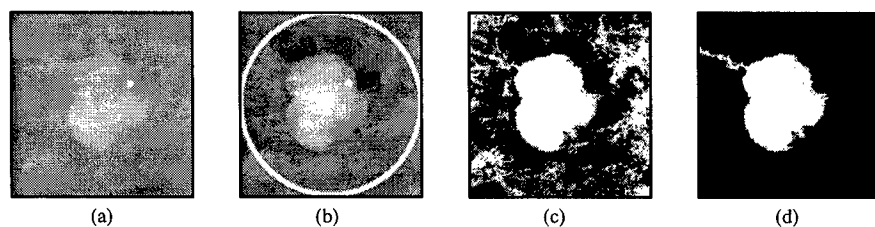
Figure 1. The first steps in the segmentation procedure. (a) The 42.6 mm by 42.6 mm (213 pixels by 213 pixels) ROI containing the object to be segmented. (b) The unsharp masked ROI from (a) with a 16 mm (80 pixel) radius circle representing the initial estimate of the object's outline. (c) The ROI after thresholding. (d) The center-connected region from (c).

Since it is possible that background structures are still included in the segmentation, two constraints are applied to minimize their influence. The first constraint examines each ray emanating from the center of the region and eliminates any pixels that are greater than d mm from the previous interior pixel. This constraint is designed to remove background structures that may cross through the region and be very close to the structure of interest. The second constraint forces the border pixels to be within n mm of their immediate neighbors. This constraint will prevent the segmentation from following a random structure that managed to pass the first constraint.

To facilitate the implementation of the constraints, the ROI is transformed into polar coordinates. The first constraint is applied by examining each ray in the r direction independently. Beginning at the pixel at the last row of the $[r, \theta]$ matrix, the distance between interior pixels is calculated. As soon as a gap of d mm or more is encountered, the remaining pixels on the ray are marked as exterior to the region. The effect of this constraint can be seen in Figure 2.

The second constraint is applied by traversing the $[r, \theta]$ matrix and examining each pair of neighboring border pixels (where a border pixel is the last interior pixel along each ray in the r direction). For any given pair, $([r_0, \theta_0], [r_1, \theta_1])$, the city-block distance between them is computed. If the distance is greater than n mm, $[r_1, \theta_1]$ is adjusted to be n mm from $[r_0, \theta_0]$; otherwise, $[r_1, \theta_1]$ is accepted as the border pixel. This procedure of pairwise comparisons continues until all border pixels meet the constraint.

(a)

(b)
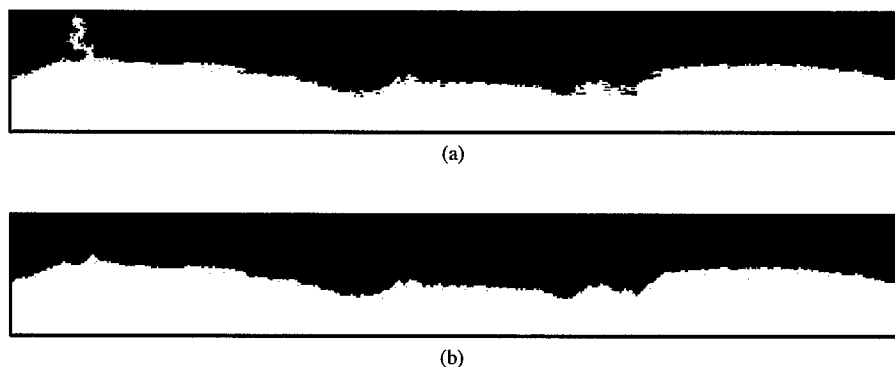
Figure 2. (a) The polar-transformed $[r, \theta]$ matrix of the ROI in Figure 1(d) before application of the constraints. (b) The $[r, \theta]$ matrix after application of the constraints. Note the removal of the spurious structure at the 10 o'clock position in Figure 1(d).

Once the constraints have been enforced, the region is transformed back into spatial coordinates (Figure 3). The resulting region constitutes the input to the next iteration of the segmentation algorithm. The procedure halts once there in no change in the border between iterations.
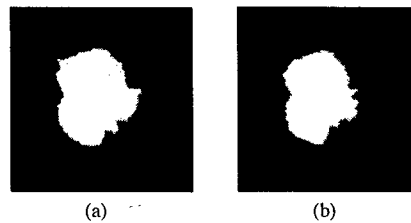
Figure 3. (a) The resulting segmentation after the first iteration. (b) The final segmentation (after the fourth iteration).

2.6. Procedure

For this study, we compared the performance of the CAD system before and after the segmentation routine was inserted, denoted as the pre-seg and post-seg systems. Both systems employed a DOG filter constructed of symmetric two-dimensional Gaussians with widths of 18 and 9 mm (90 and 45 pixels). The size of both the DOG filter and NCC templates was 24 mm (120 pixels). The regions localized by the previous system were identified by increasing the gray level threshold from 2% through 13% in steps of 2%.

After exploring several combinations of values, the parameters d and n in the segmentation routine were set to .6 mm (3 pixels) and .4 mm (2 pixels), respectively. 0.9 was set as the weighting factor on the unsharp masking operation.

ROC analysis was employed to compare the discriminatory power of the twenty individual features extracted from each of region before and after the new segmentation algorithm was applied. Bootstrap sampling was employed to determine the statistical significance between the empirical AUCs achieved by each feature. A feature was deemed to have changed significantly if the change in AUC was significant at the 0.05 level.

As mentioned earlier, the feature selection stage also employed bootstrap sampling to determine whether or not the addition or deletion of a feature from the model resulted in a statistically significant performance difference. The significance levels for both adding and deleting a feature were set to 0.05. The AUCs used in the feature selection stage were computed using the entire dataset for training and testing.

Round-robin training and testing was employed in the classification stage. Since the CAD system also performs the detection task, the performances of the pre-seg and post-seg systems were examined via FROC curves. In addition to looking at the performance of the system on masses vs. non-masses, we also examined how the system performed on just malignant masses.

The computations for this study were performed on a machine with dual 1.8 GHz AMD (Advanced Micro Devices, Inc., Sunnyvale, CA) processors. The CAD and segmentation systems were programmed in MATLAB release 13 (The MathWorks, Inc., Natick, MA).

3. RESULTS

Some sample results of the segmentation algorithm are shown in Figure 4. Shown in the figure are (a) 3 masses (2 malignant, 1 benign) extracted from the original mammographic image, (b) the segmentation provided by the previous system, and (c) the segmentations computed with the new segmentation routine. By comparing columns (b) and (c) of Figure 4, it can be seen that the new segmentation algorithm produces segmentations with finer detail than that of the previous system. The newly developed segmentation routine required an average of 7.8 iterations for each suspicious region with an average of 0.37 seconds per iteration.

Table 1 shows the AUC values and the p-values of the differences of each of the twenty features in the pre-seg and post-seg systems. The AUCs of seven features increased significantly: eccentricity, solidity, extent, circularity, NRL mean, NRL ratio, and NRL spread; only one feature, the mean output of the DOG filter, had a significant decrease in AUC.

Although each system chose three features, the specific features selected by the stepwise feature selection algorithm differed from the pre-seg and post-seg systems. The pre-seg system chose one textural feature and two
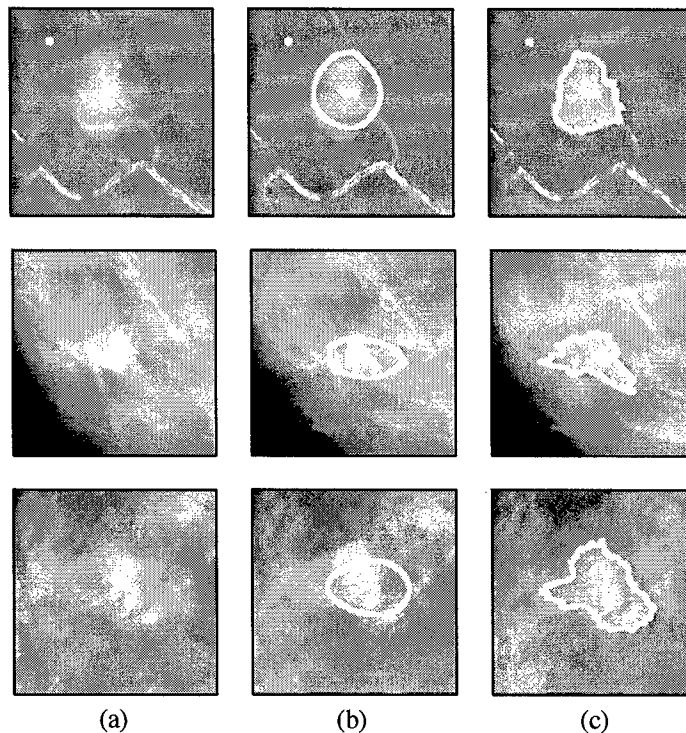
Figure 4. Two malignant masses (top two) and one benign mass segmented with the segmentation routine. (a) The original mass. (b) The segmentation used in the previous version of the system. (c) The new segmentation.

morphological features: the peak output of the DOG filter, the NRL mean, and the NRL spread. The post-seg system chose three morphological features: minimum axis length, solidity, and NRL spread.

The FROC performances of the pre-seg and post-seg systems are shown Figure 5. The overall performances of the systems (i.e., the performance in separating masses from non-masses) are about equal; there is very little difference across the entire range of the FROC curve. However, in the range from 1.5 false positives per image (FPpI) to 4 FPpI, the segmentation algorithm clearly made an improvement in the performance of the system on malignant masses. For example, at 90% sensitivity, the pre-seg system marked 4.4 FPpI compared to the post-seg system's 3.7 FPpI. At 85%, the difference was even greater: 4.1 FPpI for the pre-seg system compared to 2.1 FPpI for the post-seg system.

4. DISCUSSION

From examining Table 1, it can be seen that a total of eight of the features experienced a statistically significant change in performance, with seven increasing and one decreasing. Since the segmentation algorithm has provided sharper outlines of the regions, the increase in performance of the features describing the border agrees well with intuition. The fact that the segmentation routine captures important information in the details of the border can be seen in the significant performance increases of three of the seven NRL features: NRL mean, NRL area ratio, and NRL spread. The remaining NRL features increased in performance but without statistical significance.

Additionally, the improved accuracy in the description of the masses overall shapes is evident from the improvement of circularity, extent, eccentricity, and solidity. For example, circularity made one of the more dramatic increases in performance, rising from 0.67 to 0.79 with a p-value of < 0.01, making it one of the better performing features after incorporating the segmentation algorithm. This increase is not surprising because, after the segmentation, the majority of non-mass objects should be less circular than they were previously. Overall, the increase in the effectiveness of many of the morphological features validates the segmentation algorithm.

Table 1. The features, AUCs before and after incorporation of the segmentation routine, and the statistical significance of their change. The standard deviations and p-values were computed using the bootstrap sampling technique and the percentile method.

Feature	Pre-seg System AUC	Post-seg System AUC	p-value of Difference
Area	0.79 ± 0.02	0.82 ± 0.02	0.13
Eccentricity	0.64 ± 0.03	0.71 ± 0.02	0.02
Major Axis Length	0.72 ± 0.02	0.70 ± 0.02	0.26
Minor Axis Length	0.82 ± 0.02	0.84 ± 0.02	0.23
Area of Convex Hull	0.79 ± 0.02	0.80 ± 0.02	0.32
Equivalent Diameter	0.79 ± 0.02	0.83 ± 0.02	0.13
Solidity	0.67 ± 0.03	0.77 ± 0.02	< 0.02
Extent	0.52 ± 0.02	0.76 ± 0.02	0.05
Mean DOG filter Output	0.76 ± 0.02	0.70 ± 0.02	0.05
Peak DOG filter Output	0.83 ± 0.02	0.81 ± 0.02	0.25
Std dev DOG filter Output	0.81 ± 0.03	0.76 ± 0.02	0.09
Circularity	0.67 ± 0.03	0.79 ± 0.02	< 0.01
Contrast	0.73 ± 0.03	0.77 ± 0.02	0.06
NRL Mean	0.65 ± 0.03	0.80 ± 0.02	< 0.01
NRL Std Dev	0.64 ± 0.03	0.71 ± 0.03	0.06
NRL Entropy	0.61 ± 0.03	0.63 ± 0.03	0.32
NRL Area Ratio	0.64 ± 0.03	0.76 ± 0.03	< 0.01
NRL Zero Crossing	0.60 ± 0.02	0.64 ± 0.03	0.10
NRL Spread	0.66 ± 0.03	0.78 ± 0.02	< 0.01
NRL Change	0.74 ± 0.03	0.79 ± 0.02	0.11

Only one of the twenty features significantly decreased in ROC performance: mean output of the DOG filter output. Another DOG filter-extracted feature, the standard deviation of the filter output, almost decreased significantly with a p-value of 0.06. We feel the effectiveness of these two feature decreased due to the increased accuracy of the object borders. Since the segmentation procedure groups pixels with similar gray values, it would be expected that the mean and standard deviation of the DOG output would not vary as much as they did in the previous segmentations, particularly in non-mass objects. Without the variation present, it is not surprising that it was not as capable of separating masses from non-masses.

The FROC curve in Figure 5 indicates how the incorporation of the segmentation routine positively affected the CAD systems performance, particularly in the key range above 80% sensitivity and below 4 FPpI. Although the curves for the overall performance were close over the entire range of the curve, the performance of the post-seg system on the malignant masses clearly exceeded that of 1.5 FPpI to 4 FPpI, corresponding to a range from 75% to 90% sensitivity.

Since detecting malignant masses is more important than detecting benign masses, we are encouraged that our system performs better on the malignant masses in the lower range of FPpI. When detecting 90% of the malignant masses, the pre-seg system achieves 4.4 FPpI compared to the post-seg system's 3.7 FPpI, a decrease of 16%. At 85%, the pre-seg and post-seg systems respective FPpI are 4.1 and 2.1, a decrease of 49%. At 80% sensitivity, the FPpI's for the pre- and post-seg systems are 3.3 and 1.8, respectively, a decrease of 45%. Since it has been demonstrated that, at a constant system sensitivity, reducing the systems FPpI increases a mammographers performance,[15] preserving a high level of sensitivity as false positives are reduced is key to the success of a CAD system.
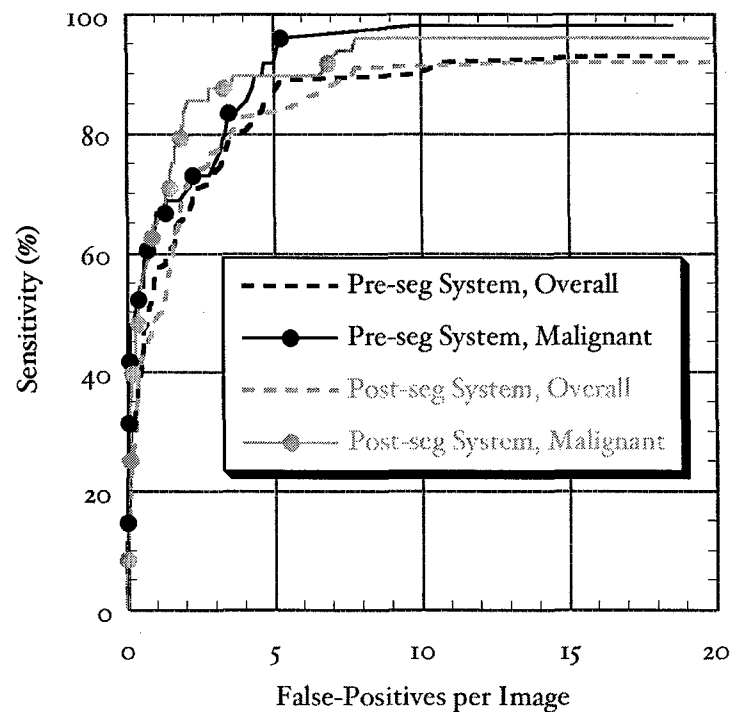
Figure 5. FROC curve comparing the system performance before (black) and after (gray) incorporation of the segmentation algorithm. The overall performance (i.e., masses vs. non-masses) is shown by the dotted lines; performance on only malignant masses (i.e., malignant masses vs. non-masses) is shown by the solid lines.

5. CONCLUSION

In this study, a new algorithm to segment suspicious regions was incorporated into a mammographic mass CAD system. The proposed segmentation algorithm is an iterative procedure that utilizes a linear discriminant function to separate an objects interior pixels from its exterior pixels and requires only two parameters: d, the maximum distance between neighboring pixels on each ray, and n, the allowable distance between neighboring border pixels. The inclusion of the two constraints on the boundaries helps to minimize the influence of spurious background structures. On average, the procedure completes in only 7.8 iterations.

The performance and importance of the segmentation routine was validated by its impact on the accuracy of the individual mass features validates the segmentation routines performance. It was found that the segmentation routine affected the performance of individual features in a predictable and intuitive manner; many of the features describing the mass border increased with statistical significance. As seen in Figure 5, the segmentation routine greatly aided the performance of the CAD system on malignant masses at a critical region of the FROC curve, where sensitivity is greater than 80% and FPpI are less than 4. The addition of the segmentation made the largest difference in the systems efficacy on malignant masses.

Because the borders of masses hold much of the information regarding their pathology, it is critical to accurately measure this information. From the impact of the segmentation routine on the performance of the individual features as well as the FROC performance of our system, we feel that the introduced segmentation routine is a critical addition to our CAD system.

6. ACKNOWLEDGEMENTS

REFERENCES

1. D. M. Catarious, Jr, A. H. Baydush, C. K. Abbey, and C. E. Floyd, Jr, "A mammographic mass CAD system incorporating features from shape, fractal, and channelized Hotelling observer measurements: preliminary results," in "Proc. of the SPIE", *Medical Imaging 2003: Image Processing*, Milan Sonka, J. Michael Fitzpatrick; Eds., Vol. 5032, p. 111-119, 2003;

2. A. H. Baydush, D. M. Catarious, Jr., and C. E. Floyd, Jr., "Computer-aided detection of masses in mammography using a Laguerre-Gauss channelized hotelling observer," in "Proc. of the SPIE", *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*, Dev P. Chakraborty, Elizabeth A. Krupinski; Eds., Vol. 5034, p. 71-76, 2003.

3. R. C. Gonzalez, and R. E. Woods, *Digital Image Processing*, 716, Third ed., Addison-Wesley, New York, 1993.

4. Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," *Med Phys* **22**, 1569-79, 1995.

5. N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification," *Med Phys* **23**, 1685-96, 1996.

6. G. te Brake, and N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms," *Med Phys* **28**, 259-266, 2001.

7. M. J. Carreira, D. Cabello, M. G. Penedo, and A. Mosquera, "Computer-aided diagnoses: Automatic detection of lung nodules," *Med Phys* **25**, 1998-2006, 1998.

8. G. M. te Brake, and N. Karssemeijer, "Single and multiscale detection of masses in digital mammograms," *IEEE Trans Med Imaging* **18**, 628-39, 1999.

9. M. Heath K. Bowyer, D. Kopans, R. Moore and P. Kegelmeyer Jr., "The Digital Database for Screening Mammography", in *The Proceedings of the 5th International Workshop on Digital Mammography*, Medical Physics Publishing, Toronto, 2000.

10. B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis," *Acad Radiol* **2**, 959-66, 1995.

11. Y. H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment," *Med Phys* **28**, 455-461, 2001.

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," *Acad Radiol* **6**, 22-33, 1999.

13. J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans Med Imaging* **12**, 664-669, 1993.

14. M. Nadler, and E. P. Smith, *Pattern Recognition Engineering*, 588, John Wiley and Sons, New York, New York, 1993.

15. B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings," *Radiology* **221**, 633-640, 2001.

Bi-plane correlation imaging for improved detection of lung nodules

Ehsan Samei[1,2,3], David M. Catarious, Jr.[1,2], Alan H. Baydush[1,2,4]
Carey E. Floyd, Jr.[1,2], Rene Vargas-Voracek[1]

[1] Department of Radiology, [2] Department of Biomedical Engineering
[3] Department of Physics, [4] Department of Radiation Oncology
Duke University, Durham, NC 27710

ABSTRACT

Bi-plane correlation imaging (BCI) is a new imaging approach that utilizes angular information from a bi-plane digital acquisition in conjunction with computer assisted detection (CAD) to reduce the degrading influence of anatomical noise in the detection of subtle lesions in planar images. An anthropomorphic chest phantom, supplemented with added nodule phantoms (5-13 mm at the image plane), was imaged from different posterior projections within a ±12° range by moving the x-ray tube vertically and horizontally with respect to the detector. Each image was analyzed using a basic front-end single-view CAD algorithm. The correlation of the suspect lesions from the PA view with those from each of the oblique views was examined using *a priori* knowledge of the acquisition geometry. The correlated suspect lesions were registered as positive. Using an optimum −3° vertical geometry and processing parameters, BCI resulted in 62.5% sensitivity, 1.5 FP/image, and 0.885 PPV. The corresponding values from the observer experiment were 56% sensitivity, 10.8 FP/image, and 0.45 PPV, respectively. Compared to single-view CAD results, the BCI reduced sensitivity by 20%. However, the corresponding reduction in FPs was notably higher (94%) leading to 140% improvement in the PPV. Changes in processing parameters could result in higher PPV and lower FP/image at the expense of lower sensitivity. Similar findings were indicated for small (5-9 mm) and large (9-13 mm) nodules, but the relative improvement was significantly higher for smaller nodules. (The research was supported by a grant from the NIH, R21CA91806.)

Keywords: Chest radiography, digital radiography, stereoscopy, lung nodules, lung cancer, computer aided detection (CAD)

1. INTRODUCTION

Lung cancer is a leading cause of death in the US, surpassing the mortality associated with breast, prostate, colon, and cervical cancers combined (ACS 2002). In its early stages, lung cancer is often discovered in the form of solitary lung nodules when a chest radiograph of a patient is taken for another purpose. Many studies suggest that the probability of localized disease, and thus patient survival, is inversely proportional to the size of a nodule at the time of diagnosis (Padilla 1997, Mori 1989). Therefore, any improvement in the poor prognosis of lung cancer relies on improving the early detection of associated lung nodules when they are still small, and thus the probability of the spread of the disease is still low. In chest radiographs, small cancerous nodules are difficult to detect. Even very experienced radiologists often miss subtle lung nodules that can be detected if the image is viewed retrospectively after the disease is confirmed (Heelan 1984). In spite of much technological advancement in chest radiography in the last four decades, there has been little or no improvement in the detection of small lung nodules (Revesz 1977, Muhm 1983, Heelan 1984, Gavelli 1998).

There are three main factors limiting the detection of subtle lung nodules and early diagnosis of lung cancer: nodule contrast to noise ratio (CNR), perceptual errors, and anatomical noise. The detection of lung nodules can be influenced by their low CNR. There have been significant advancements in radiologic technology, including the development of digital radiographic systems, that have led to significant improvements in the resolution, noise, and latitude characteristics of thoracic images leading to improved CNR of lung lesions. Perceptual errors, at both visual and cognitive levels, are the second obstacle contributing to the low detection rate of subtle lung nodules (Kundel 1978, Kundel 1975, Carmody 1980). Computer assisted diagnosis (CAD) algorithms have also been developed as a method to provide a complete search of the image data and thus minimize perceptual errors in the detection of lung nodules in chest radiographs (Giger 1988). The third and perhaps the most significant obstacle with detrimental effects on the detection of lung nodules in chest

radiographs is anatomical noise, the normal thoracic structures surrounding and overlaying a lesion masking its appearance (Burgess 1997, Samei 2000, Revesz 1974, Neitzel 1998).

Several promising methods have been developed to reduce the influence of anatomical noise in thoracic images. Two such techniques that aim to improve lung nodule detection by minimizing the appearance of ribs and other overlaying thoracic structures are dual-energy imaging (Stewart 1990, Kido 1995) and digital tomosynthesis (Zwicker 1997, Dobbins 1998). The former technique, with only two systems commercially available, is currently under clinical evaluation, while the latter awaits further development and clinical implementation. Computed tomography is probably the optimal modality for minimizing anatomical noise in chest imaging as it eliminates overlays of anatomical structures associated with projection imaging. There has been recent excitement over the use of low-dose CT for lung cancer screening (Henschke 1999). However, at the present, utilization of CT as a wide-spread screening method for the detection of subtle lung nodules is controversial because of associated economic (cost and technology availability), patient care (e.g., over-treatment), and epidemiological (e.g., patient dose) issues.

This study proposes a new image acquisition and processing approach, bi-plane correlation imaging (BCI), for improving the detection of subtle lung nodules. In this approach, two digital images of the thorax are acquired within a short time interval from two slightly different posterior projections (Fig. 1). The image data are then incorporated into an enhanced CAD algorithm in which nodules are detected by examining the geometrical correlation of the detected signals in the two views. The underlying hypothesis of the proposed approach is that the anatomical noise associated with normal anatomical features in the thorax is the main factor limiting the detection of subtle lung lesions. Angular information is used to minimize this limiting influence by identifying and positively reinforcing the nodule signals, which remain relatively constant against a variation in the background structure. This approach does not promise to completely eliminate anatomical noise (as CT does), but aims to cost-effectively reduce its influence without an increase in the patient dose. Using correlation of signals between two views to identify "true" signals, CAD can be utilized at high sensitivity levels, lowering the detection thresholds, without an undesirable increase in the number of false positives. The hybrid approach of utilizing angular information in conjunction with digital acquisition and CAD addresses all three major obstacles to the detection of subtle lung nodules discussed above. The angular information reduces the effects of anatomical noise, the high signal-to-noise ratio of digital acquisition assures sufficient nodule contrast, and CAD incorporates a complete search. This paper reports on a study aimed to explore the feasibility of BCI for improved early detection of subtle lung nodules via a phantom experiment.
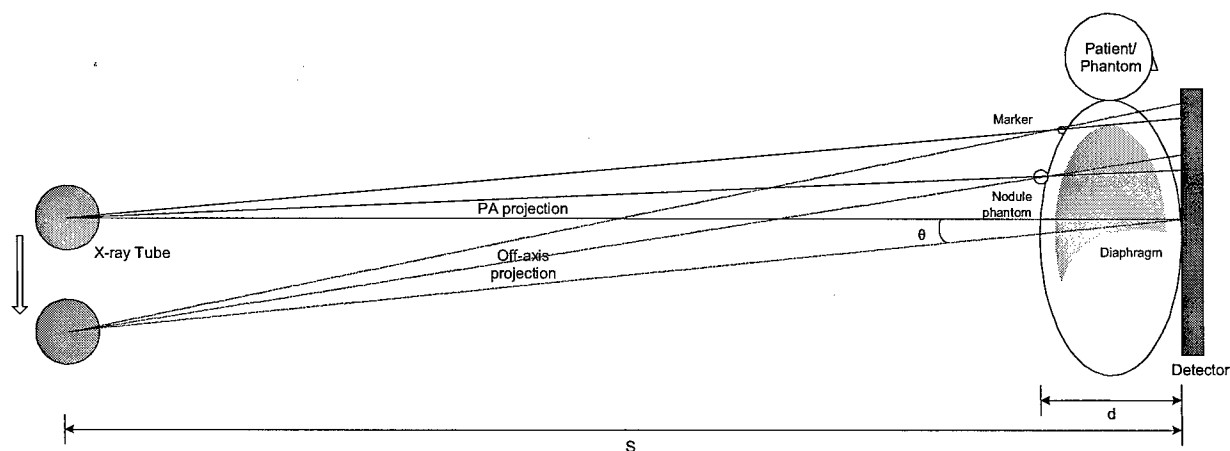

Fig. 1: The schematic geometry for the acquisition of BCI bi-plane image pairs at 0 (PA) and –6 degrees

2. MATERIALS AND METHODS

1. Image acquisition

This study was performed based on images acquired from an anthropomorphic chest phantom (RSD, Inc., Long Beach, CA). The phantom was superimposed with 16 additional nodule phantoms of 8 different sizes made of Teflon emulating the appearance of subtle tissue-equivalent lesions in chest radiographs (4-11 mm in diameter) with a physical density within a 0.95-1.1 g/cm^3 range (Samei 1997). Table 1 lists the diameter, thickness, and contrast characteristics of the nodule phantoms. As the nodule phantoms were placed on the back of the chest phantom in PA acquisition geometry, they were magnified by about 20% to 5-13 mm in diameter. Four small fiducial markers were also added at four corners of the phantom for verifying the acquisition geometry. The supplemented phantom was imaged using a conventional PA geometry with a flat-panel digital radiographic system (GE, XQ/i). The exact locations of the nodules were recorded via an additional image with the locations of the nodules marked with fiducial markers (Fig. 2).

In addition to the PA view, by vertically adjusting the x-ray tube, the supplemented phantom was imaged using seven additional projection orientations at $-6°$, $-4°$, $-3°$, $+3°$, $+4°$, $+6°$, and $+12°$ (Fig. 1). The x-ray tube was moved between these angular positions using a precise programmable tube mover (Fig. 3). The above acquisitions were repeated with the 16 nodule phantoms placed in a different arrangement configuration to allow superimposition of a given nodule against various local anatomical backgrounds. A conventional kVp (120) and standard photo-timed exposure (mAs = 5) were used for the acquisitions. The acquisitions were repeated with the supplemented phantom rotated 90 degrees to assess the utility of BCI with horizontal (i.e. lateral) displacement of the x-ray tube in the two projections. The images were corrected for offset and gain non-uniformities without any additional image processing. The total of 32 projection images (8 projections x 2 nodule configurations x 2 orientations) were stored electronically.

Table 1 illustrates the realization of the nodule phantoms in one of the PA radiographs. As evident in the illustration, the lesions were extremely subtle and most of them were below the size considered the threshold of detectability, 10 cm, on chest radiographs, 10 mm (Kundel 1981).

All the acquired oblique images were paired with a corresponding 0 degree/ PA image to be used for determining the optimum acquisition geometry as described below. In the acquired image set, the relative angular separation of any oblique view from the corresponding PA view was verified by correlating the coordinates of the four fixed fiducial markers placed at the corners of the image area. The results showed excellent geometrical accuracy, with sub-mm precision for geometrical correlation of anatomical features. For each image, a truth file was also generated from the known location of the nodule phantoms to be used for evaluating the performance of the CAD algorithm described below.

Table 1: The characteristics of the nodule phantoms. The diameters were 20% higher in the imaging plane because of magnification. The estimated peak contrasts were determined from the maximum thickness of the phantoms, an assumed scatter-to-primary ratio of 0.68, and an effective linear attenuation coefficient of 0.045126 mm^{-1}as defined in Samei et al. (Samei 1997) and estimated for a 0.5 mm thick CsI detector using the xSpect x-ray simulation routine.
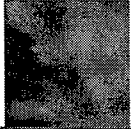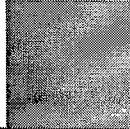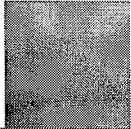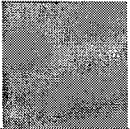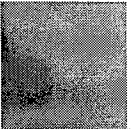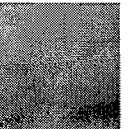
Diameter (mm)	3.9	5.5	5.0	7.8	7.0	9.3	8.4	11.0
Maximum thickness (mm)	1.52	2.12	2.36	3.00	3.35	3.56	3.95	4.27
Estimated peak contrast, dE/E (%)	4.1	5.7	6.3	8.1	9.0	9.6	10.6	11.5
Emulated peak physical density (g/cm^3)	0.95	0.95	1.1	0.95	1.1	0.95	1.1	0.95
Nodule appearance								

Fig. 2: A PA chest image with the location of the added the nodules marked with fiducial markers
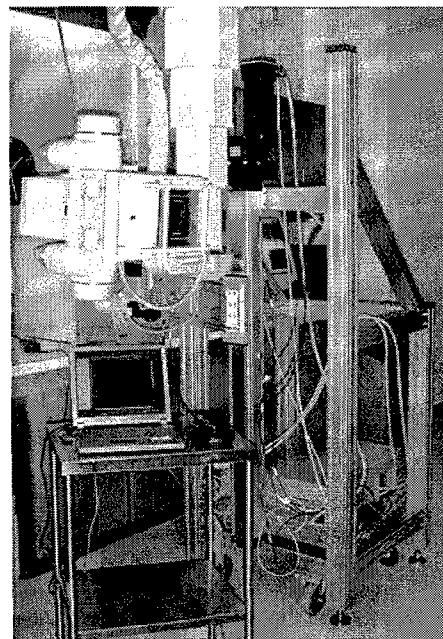
Fig. 3: The tube mover used to acquire bi-plane data.

2. Single-view CAD

A single-view CAD algorithm has been under development at our research laboratory. The acquired phantom images were processed by the algorithm and the results were used as input to the BCI scheme described in the next section.

The CAD algorithm consisted of four stages: A) image preprocessing, B) filtration, C) suspicious region localization, D) feature extraction, and D) classification/false positive reduction. At the image preprocessing stage, the images were first inverted and then converted to a logarithmic scale. The lung fields were segmented via hand-drawn outlines, a segmentation step expected to be automated in the next version of the algorithm. The image preprocessing stage also included a background regularization process via unsharp-masking (USM) or local histogram equalization (LHE) (Gonzalez 1993). USM suppresses low-frequency background information while emphasizing the high-frequency content of an image. Subtracting a low-pass filtered version of the original image enhances the high-frequency content and corrects for non-uniformities in the background, therefore potentially raising the detectability of the nodules. Alternatively, LHE is able to accentuate local details while preserving the overall, or low-frequency, structure of the image, leading to increased local contrast. The USM processing was applied according to $I(x,y) = A.O(x,y)-L[O(x,y)]$, where $I(x,y)$ is the new image, $O(x,y)$ is the original image, A is a scalar in [0, 1], and L is a low-pass rectangular average filtering operator. In the LHE process, each pixel in the original image was transformed into a new pixel by $I(x,y) = A(x,y).[O(x,y)-m(x,y)]+m(x,y)$, where $A(x,y) = kM/\sigma(x,y)$, M represents the global mean of the image, k is a scalar in [0, 1], and $\sigma(x,y)$ and $m(x,y)$ are the local standard deviation and local mean of pixels in a kernel/window around (x, y). In this study, the kernel sizes for this operator were varied between 28 and 52 mm for a fixed $A = k = 0.5$.

After pre-processing, the images were filtered for enhancing nodule-like features within the images. Since it has been demonstrated that lung nodules generally follow a Gaussian-like profile (Samei 1997), a Difference Of Gaussian (DOG) filter was used (Zheng 1995). Two DOG filters were utilized with two different combinations of the standard deviation widths of the two defining Gaussian components, 8/4 or 8/2 mm. These particular combinations were selected based on an iterative empirical approach for best performance. The kernel size of the DOG and the kernel size of the preprocessor were always chosen to be equal. The DOG filter was applied using the normalized cross-correlation (NCC) method (Gonzalez 1993, Carreira 1998, Penedo 1998). Unlike conventional cross-correlation, NCC is amplitude independent, and thus

suitable to the widely varying background of chest images. The output of the filtration was an image with values ranging between −1 and +1, with the extremes corresponding to the perfect mismatch or match of the original image to the targeted DOG profile, respectively.

The filtered images were further processed to identify suspicious nodule locations via a multi-level thresholding procedure. In this procedure, regions were identified at eleven gray level thresholds. As the threshold levels progress, some of the regions would grow and merge with their neighbors. The final set of suspicious regions was determined by extracting the suspicious regions at the threshold level before they merged with another region.

Finally, from each of the suspicious regions, twelve features were extracted. These features included area, eccentricity, major axis length, minor axis length, convex area (the area of the convex hull), equivalent diameter (diameter of the circle that has an area equal to that of the region), orientation, the filled region (the area of the region including internal voids), Euler number (the number of objects in the region minus the number of holes in those objects), solidity (the area of the region divided by area of the convex hull), and extent (the area of the region divided by area of the bounding box - smallest enclosing rectangle). To classify the suspicious regions as being nodules or nonnodules, a multistep linear classifier was employed. Specifically, each pair of extracted features was combined via Fisher's linear discriminant (Nadler 1993), and classification decisions were made. The thresholds on the classification outputs were empirically determined so as to minimize the number of true positives eliminated. Once each pair of features was compared and classification decisions were made, all decisions were logically "ANDed" to make the final classification decision. Round-robin training and testing was employed in the classification procedure.

The ground truth was specified by binary images that indicated the location and sizes of the true lesions in the images. A nodule was counted as being "hit" when any part of the suspicious region fell within 5 mm of the centroid of the true lesion.
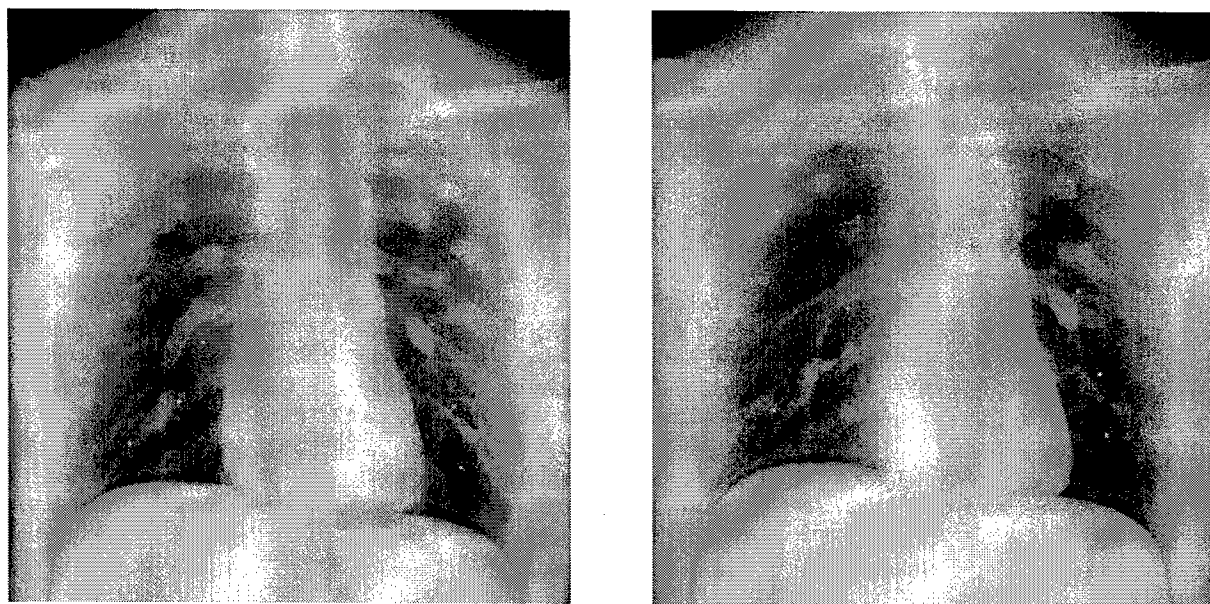
Fig. 4: A PA (left) and −3° oblique (right) radiograph of the chest phantom. The fiducial markers mark the center of true lesions, while the bright areas are suspect lesions identified by single-view. In single-view CAD, a TP is registered if any area of the CAD "island" is within a 5 mm radius of the true lesion. In the BCI scheme, a TP is registered when a TP in the PA view coincides/correlates with a suspect lesion in the oblique view based on the known angular separation of the two views. If a FP in the PA view correlates with a suspect lesion (true or false) in the other view, the suspect lesion is considered a FP in the BCI scheme.

3. BCI detection scheme

The single-view CAD algorithm described above was supplemented with a bi-plane correlation routine. In BCI, the geometrical correlation of the detected signals in the two views of a bi-plane image pair data is examined in order to detect subtle lung nodules with a high-sensitivity while minimizing the number of false-positives by applying a geometrical correlation rule. In the routine, the PA image was used as the reference image. For each suspected region in the PA image, the known angular separation between the PA and an oblique image was used to locate the possible location where the geometrically-shifted image of the candidate nodule might expected to appear in the oblique view depending on nodule's location within the thoracic cavity (Fig. 4).

To take into account the shift in the location of the suspect regions due to overlapping thoracic structure, a margin parameter defined a degree of tolerance from the perfect geometrical correlation between the two views in the horizontal and vertical directions. The resultant rectangular mask had a width equal to twice the margin size, and a length equal to maximum possible displacement based on angular separation plus the margin size. If a suspect region was identified within the mask, the original suspect region in the PA view was scored as positive. If more than one suspect region was found within the mask, only one of them was counted. A true-positive was indicated when a correlated suspect region pair corresponded to a true-positive in the PA view. Otherwise the correlated pair was registered as false-positive.

Additional correlation rules were also applied based on the closeness in the area and the eccentricity of suspect lesion pairs calculated from an area correlation index or an eccentricity correlation index defined as $2|A_{PA}-A_{obl}|/(A_{PA}+A_{obl})$ or $2|X_{PA}-X_{obl}|/(X_{PA}+X_{obl})$ where A and X are the area and eccentricity, respectively. A pair of suspect lesions was registered as FP if their associated indices fell outside of specific thresholds.

Table 2: Performance of the single-view CAD as a function of processing parameters averaged over all the acquired images.

DOG filter size	Pre-processing method	Preprocessing kernel size, mm	Sensitivity (%)	FP	PPV
8/4 mm	None	28	76.76	100.4	0.109
		36	74.22	56.0	0.175
		44	72.46	32.1	0.266
	LHE	28	84.57	20.3	0.400
		36	80.08	23.7	0.351
		44	77.15	23.4	0.345
	USM	28	84.18	64.6	0.172
		36	78.71	38.0	0.249
		44	75.78	20.2	0.375
8/2 mm	None	28	85.55	117.9	0.104
		36	82.03	80.4	0.140
		44	81.64	56.4	0.188
	LHE	28	90.82	48.6	0.230
		36	88.48	40.2	0.261
		44	84.96	36.7	0.271
	USM	28	90.62	85.2	0.145
		36	86.13	69.1	0.166
		44	83.98	52.1	0.205

4. BCI evaluation

The acquired phantom images were processed using the CAD and BCI processing schemes described above. Each oblique view projection image was paired with its corresponding PA radiograph. The results were analyzed to find the best processing and acquisition parameters for optimum performance. The independent parameters were the following:

BCI PC
□ 0.9-1
□ 0.8-0.9
▨ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

PCr
□ 0.9-1
□ 0.8-0.9
▨ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

BCI FP
□ 27-30
□ 24-27
□ 21-24
▨ 18-21
▨ 15-18
▨ 12-15
■ 9-12
■ 6-9
■ 3-6
■ 0-3

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

FPr
□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

BCI PPV
□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

PPVr
□ 1.85-2
□ 1.7-1.85
□ 1.55-1.7
▨ 1.4-1.55
▨ 1.25-1.4
▨ 1.1-1.25
■ 0.95-1.1
▨ 0.8-0.95
▨ 0.65-0.8
□ 0.5-0.65

Angle

LHE 20   LHE 28   LHE 36   LHE 44   LHE 52
Preprocessing window size

Fig. 5: Variation in percent correct (sensitivity), false positive rate, and PPV of BCI (left column) and of BCI compared to single-view CAD (BCI/CAD ratio, right column) as a function of vertical displacement angle and pre-processing kernel size (LHE pre-processing, 8/4 mm DOG) (no area or eccentricity correlation rule).

1. Angular separation (-6 to 12 degrees)
2. Displacement orientation (horizontal and vertical)
3. CAD pre-processing method (USM, LHE, and none)
4. CAD pre-processing and DOG kernel size (28, 36, 44, and 52 mm)
5. Standard deviations of the DOG filter (8/4 mm and 8/2 mm)
6. Correlation margin size (2-10 mm)
7. Area correlation index (0-2)
8. Eccentricity correlation index (0-2)

The optimization/evaluation of these parameters was performed based on six figures of merit:

1. Percent correct (PC) when using BCI as the ratio of true positives to true positives plus false negatives
2. False positives per image (FP) when using BCI
3. Positive predictive value (PPV) when using BCI
4. The relative improvement of the PC when using BCI relative to single-view CAD (PCr)
5. The relative improvement of the FP when using BCI relative to single-view CAD (FPr)
6. The relative improvement of the PPV when using BCI relative to single-view CAD (PPVr)

In order to identify the optimum set of processing parameters, the figure of merit

results were initially scanned for optima across the ranges of influencing processing variables, identifying initial parameter values that yield the best results. Fixing the variables at those specific values, the figures of merit were then examined at each acquisition angle across each single variable. The values of the fixed parameters were then iteratively changed until the optimum parameter combinations were found.

In addition to the computer analysis of the images, the 16 vertical displacement images were read by an experienced chest radiologist who was asked to identify any suspect extremely subtle lesions in the images. The performance of the radiologist was used as a benchmark for the subtlety of the lesions. The observer's PC, FP, and PPV, averaged over all 16 images, were compared to the corresponding figures from the BCI and from the single-view CAD to assess the relative merit of BCI and the advantage of utilizing image data from a second view.

3. RESULTS

The single-view CAD showed notable variability as a function of processing parameter. The sensitivity, false positive, and positive predictive figures, averaged over all images, are listed in Table 2. There was a general tradeoff between sensitivity and false positives, as parameters leading to higher sensitivity also generated a larger number of false positives. The best combination of parameters (8/4

BCI PC
0.9-1
0.8-0.9
0.7-0.8
0.6-0.7
0.5-0.6
0.4-0.5
0.3-0.4
0.2-0.3
0.1-0.2
0-0.1

Preprocessing/ DOG method

PCr
0.9-1
0.8-0.9
0.7-0.8
0.6-0.7
0.5-0.6
0.4-0.5
0.3-0.4
0.2-0.3
0.1-0.2
0-0.1

Preprocessing/ DOG method

BCI FP
27-30
24-27
21-24
18-21
15-18
12-15
9-12
6-9
3-6
0-3

Preprocessing/ DOG method

FPr
0.9-1.0
0.8-0.9
0.7-0.8
0.6-0.7
0.5-0.6
0.4-0.5
0.3-0.4
0.2-0.3
0.1-0.2
0.0-0.1

Preprocessing/ DOG method

BCI PPV
0.9-1
0.8-0.9
0.7-0.8
0.6-0.7
0.5-0.6
0.4-0.5
0.3-0.4
0.2-0.3
0.1-0.2
0-0.1

Preprocessing/ DOG mehod

PPVr
1.9-2.0
1.7-1.9
1.6-1.7
1.4-1.6
1.3-1.4
1.1-1.3
1.0-1.1
0.8-1.0
0.7-0.8
0.5-0.7
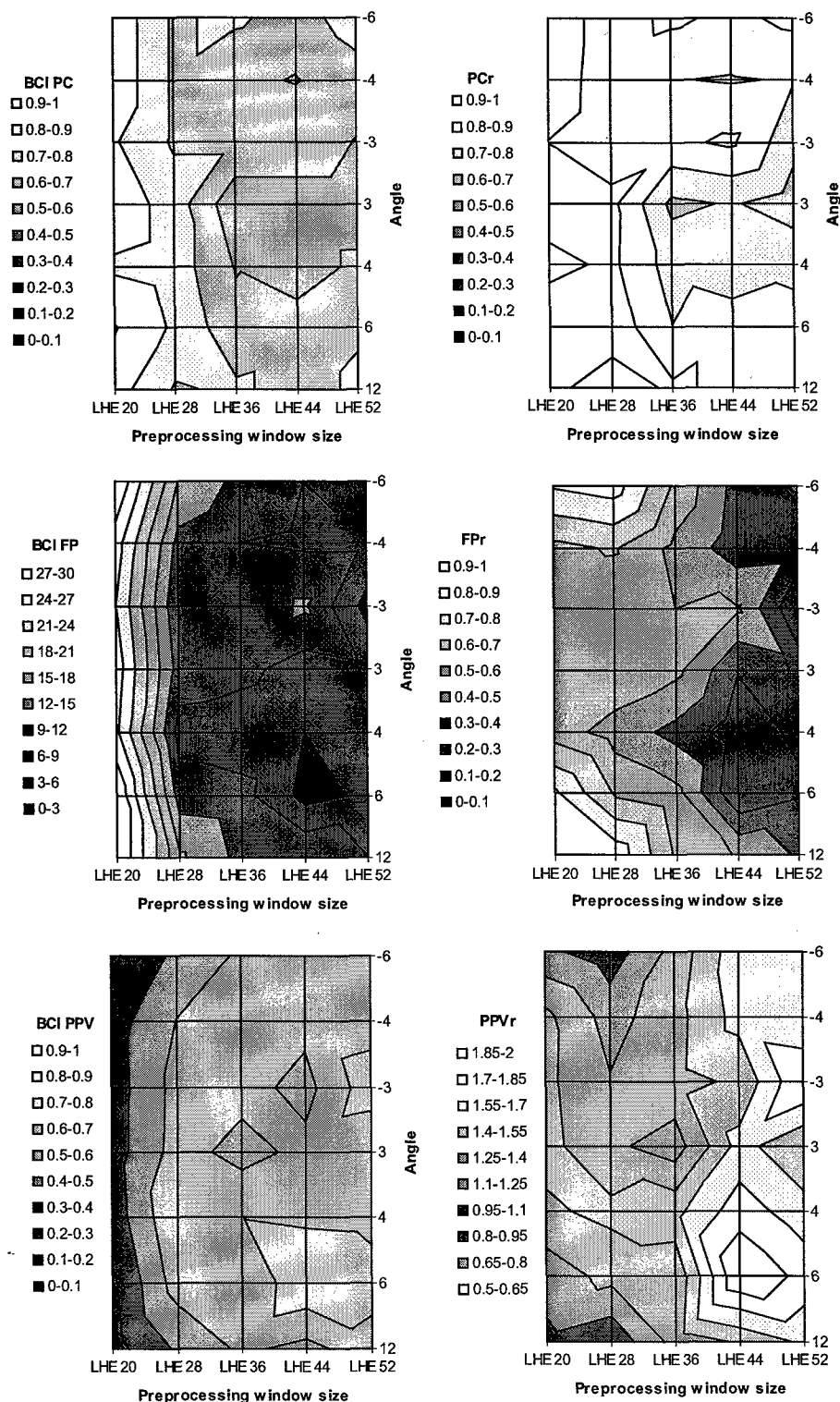
Preprocessing/ DOG method

Fig. 6: Variation in percent correct (sensitivity), false positive rate, and PPV of BCI (left column) and of BCI compared to single-view CAD (BCI/CAD ratio, right column) as a function of vertical displacement angle and pre-processing method (fixed 44 mm kernel size) and DOG filter size (8/2 and 8/4 mm) (no area or eccentricity correlation rule).
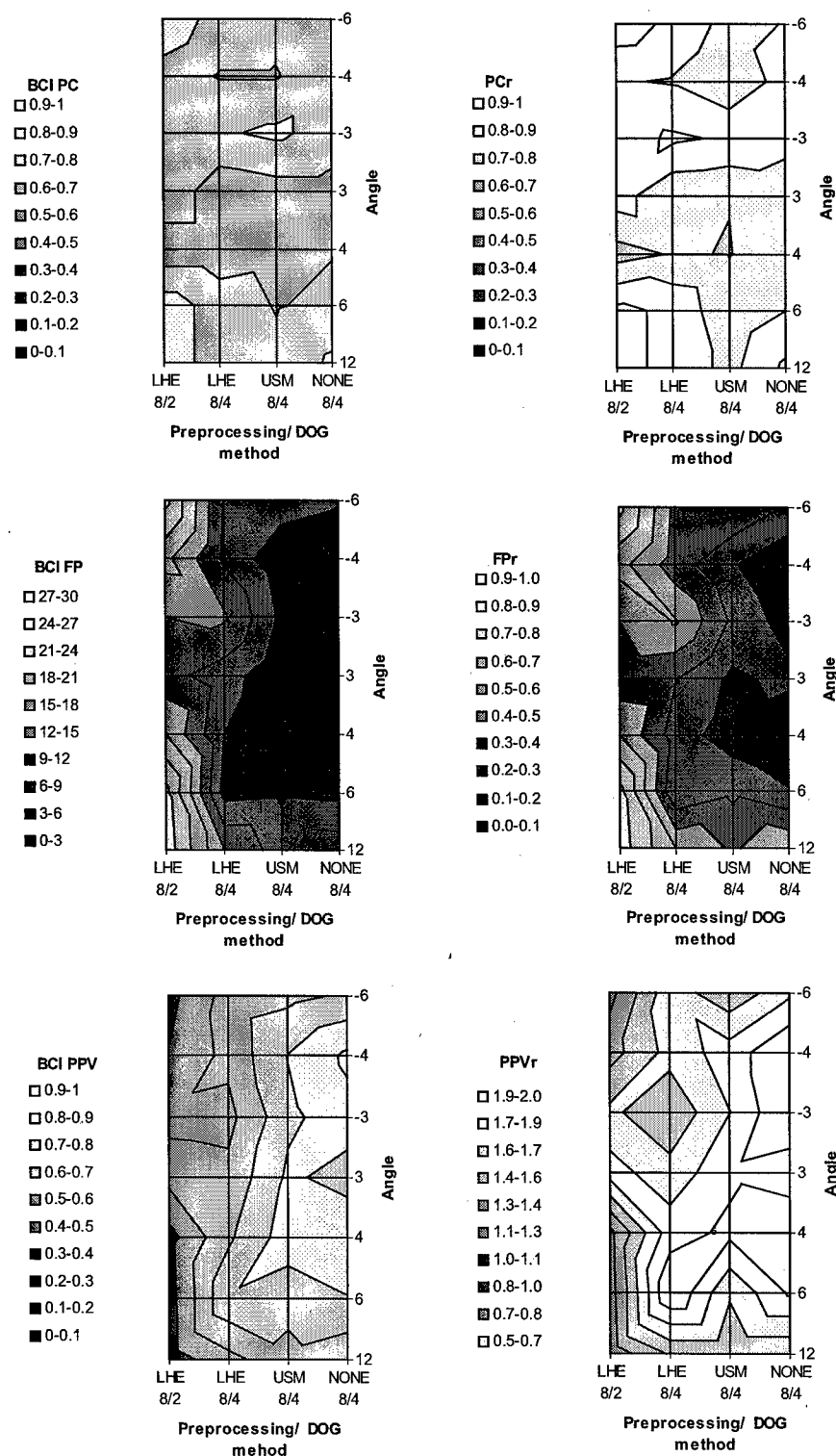
Fig. 7: Variation in percent correct (sensitivity), false positive rate, and PPV of BCI (left column) and of BCI compared to single-view CAD (BCI/CAD ratio, right column) as a function of vertical displacement angle and correlation area margin in mm (no pre-processing, 44 mm kernel size, 8/4 DOG, no area or eccentricity correlation rule). Optimal parameters are marked with two circles.

mm DOG, LHE, 28 mm) provided 85% sensitivity at 20 false-positives per image for a PPV of 0.4. Clearly this level of performance is insufficient for a conventional clinical CAD implementation. This performance, however, exceeds what is expected from a single-view CAD algorithm without more sophisticated false-positive reduction strategies, given the subtly of the lesions under consideration. Furthermore, operating at a high sensitivity/FP region, the BCI scheme was specifically designed to reduce the number of false-positives via its geometrical correlation rule. Nevertheless, research is in progress to further improve the performance of our single-view CAD with additional false-positive reduction methods.

Even though some of the processing parameters combinations exhibited better performance in single-view CAD processes, it was unclear which ones might yield optimum BCI performance. Thus all the combinations of CAD processing parameters were considered for the BCI processing. In terms of pre-processing, Fig. 5 illustrates the dependency of the BCI performance on the kernel size (for LHE pre-processing) at all the examined oblique acquisition angles. The results clearly indicate that the BCI scheme reduces the sensitivity and FP of the single-view CAD. However, the corresponding reduction in FPs is notably higher leading to an improvement in PPV, regardless of the kernel size and the acquisition angle. The results further suggest that a kernel size of 44 mm provides better overall performance in

BCI PC
- 0.9-1
- 0.8-0.9
- 0.7-0.8
- 0.6-0.7
- 0.5-0.6
- 0.4-0.5
- 0.3-0.4
- 0.2-0.3
- 0.1-0.2
- 0-0.1

Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

PCr
- 0.9-1
- 0.8-0.9
- 0.7-0.8
- 0.6-0.7
- 0.5-0.6
- 0.4-0.5
- 0.3-0.4
- 0.2-0.3
- 0.1-0.2
- 0-0.1

Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

BCI FP
- 9-10
- 8-9
- 7-8
- 6-7
- 5-6
- 4-5
- 3-4
- 2-3
- 1-2
- 0-1

Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

FPr
- 0.9-1
- 0.8-0.9
- 0.7-0.8
- 0.6-0.7
- 0.5-0.6
- 0.4-0.5
- 0.3-0.4
- 0.2-0.3
- 0.1-0.2
- 0-0.1

Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

BCI PPV
- 0.9-1
- 0.8-0.9
- 0.7-0.8
- 0.6-0.7
- 0.5-0.6
- 0.4-0.5
- 0.3-0.4
- 0.2-0.3
- 0.1-0.2
- 0-0.1

Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

PPVr
- 1.85-2
- 1.7-1.85
- 1.55-1.7
- 1.4-1.55
- 1.25-1.4
- 1.1-1.25
- 0.95-1.1
- 0.8-0.95
- 0.65-0.8
- 0.5-0.65

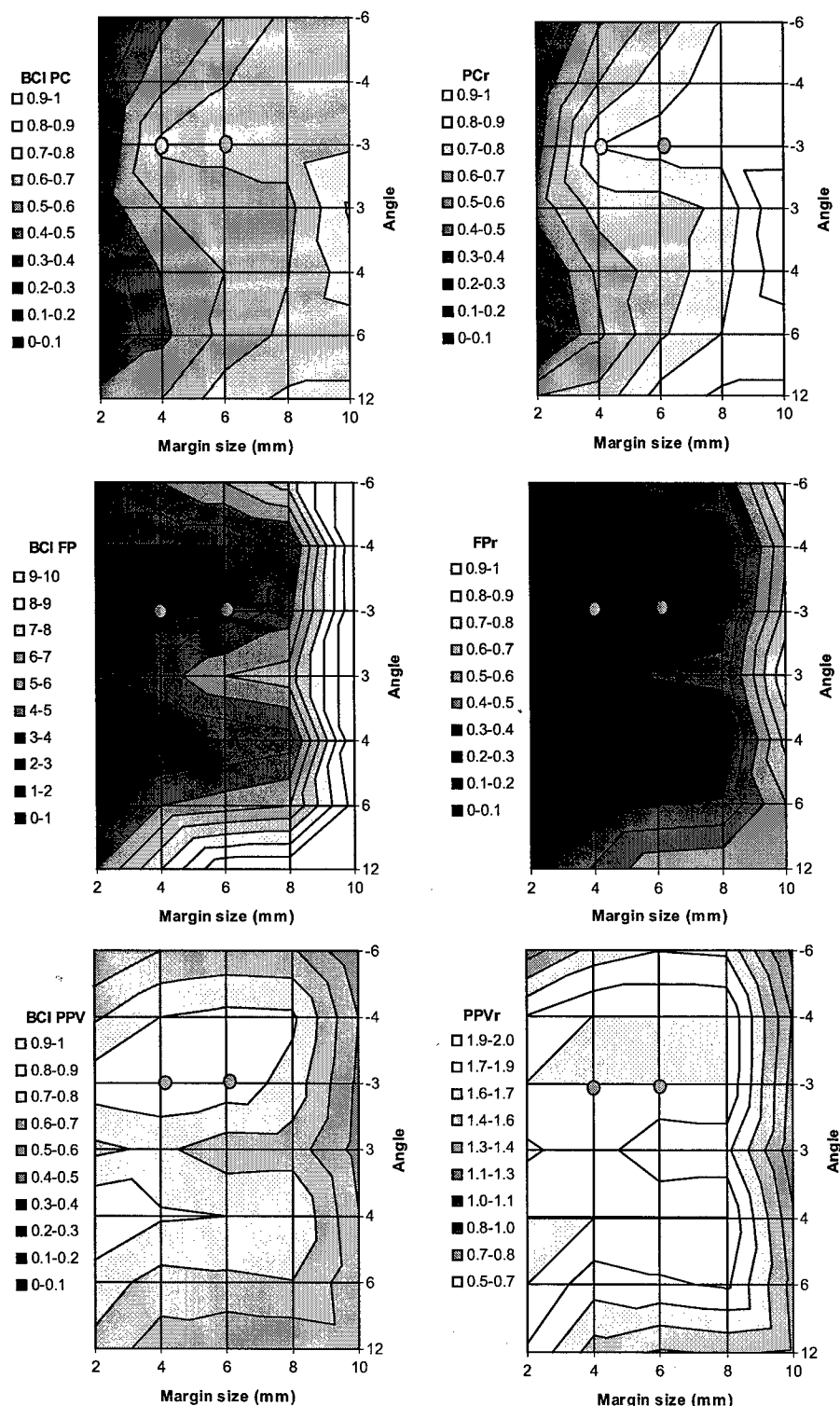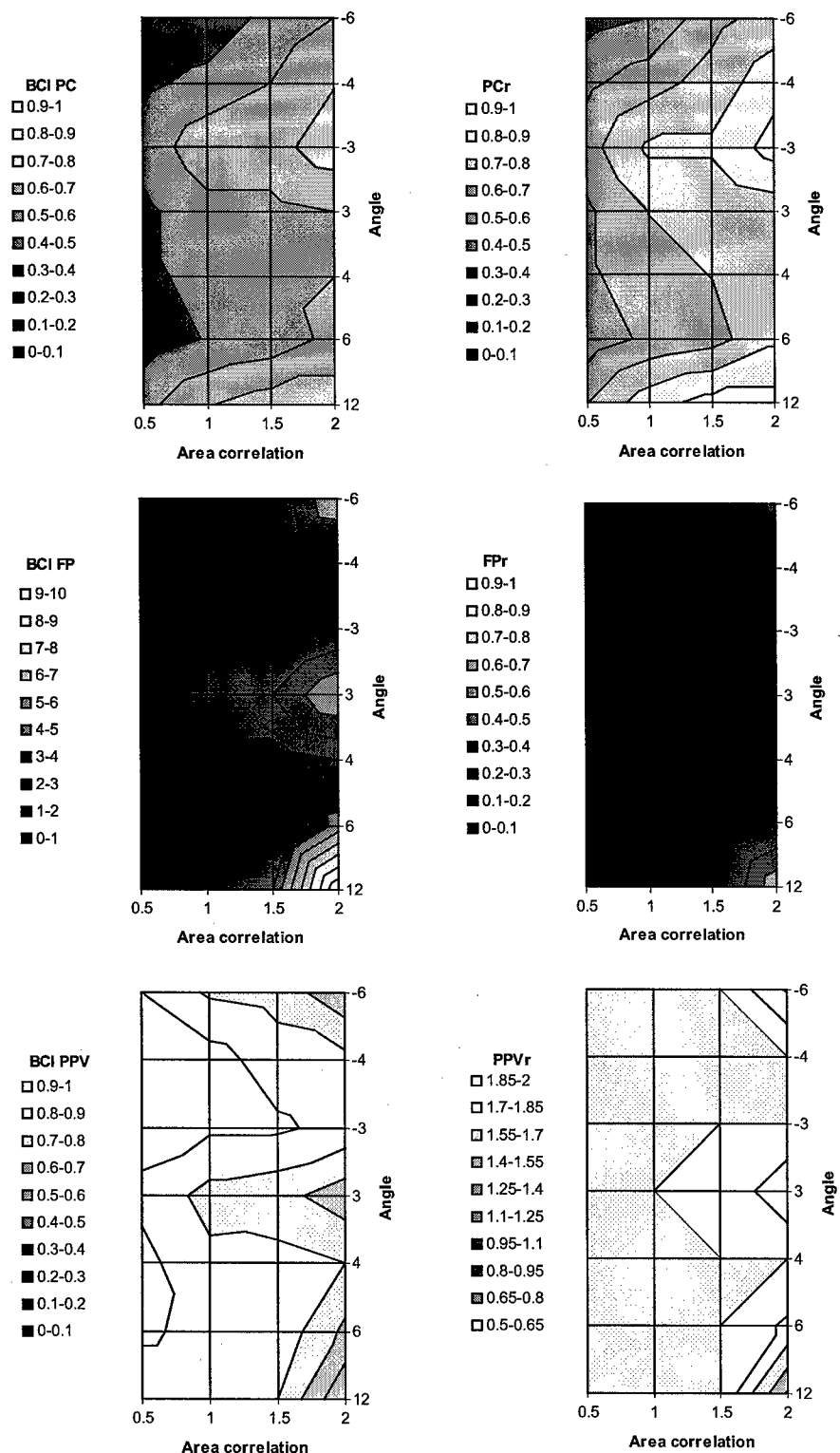Angle: -6, -4, -3, 3, 4, 6, 12
Area correlation: 0.5, 1, 1.5, 2

Fig. 8: Variation in percent correct (sensitivity), false positive rate, and PPV of BCI (left column) and of BCI compared to single-view CAD (BCI/CAD ratio, right column) as a function of vertical displacement angle and area correlation index (no pre-processing, 44 mm kernel size, 8/4 mm DOG, 6 mm correlation margin).

terms of the PPV and PPV improvement.

Fig. 6 similarly illustrates the BCI performance as a function of DOG filter size and pre-processing method for a fixed 44 mm kernel size. The results indicate that 8/4 mm DOG filter size provides better performance, a finding consistent with the single-view CAD results of Table 2. However, no pre-processing of the images proves to provide a better performance for the BCI method, as opposed to the LHE pre-processing, which was found optimum for single-view CAD (Table 2).

Basing the follow-up analysis on no pre-processing method, Fig. 7 illustrates the impact of the correlation margin size on performance. As the margin size increases from a restrictive 2 mm value, both sensitivity and the number of FPs increases. Aiming to maintain a sensitivity higher than 60%, optimum performance, in terms of both the BCI performance alone, and the relative improvement with respect to single-view CAD, is exhibited at a margin size of 4-6 mm at −3° acquisition angle. Beyond 4-6 mm margin size, the number of false positives increases at a faster rate leading to a reduction in the PPV. The optimal regions are marked in the figure.

Basing the follow-up analysis on 6 mm correlation margin size, Fig. 8 illustrates the impact of imposing additional area correlation rule on the results. The results suggest that addition area correlation reduces both sensitivity and FPs with similar rate without causing any improvement in

BCI PC

□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Margin size (mm)

PCr

□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Margin size (mm)

BCI FP

□ 9-10
□ 8-9
□ 7-8
▨ 6-7
▨ 5-6
▨ 4-5
■ 3-4
■ 2-3
■ 1-2
■ 0-1

Margin size (mm)

FPr

□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Margin size (mm)

BCI PPV

□ 0.9-1
□ 0.8-0.9
□ 0.7-0.8
▨ 0.6-0.7
▨ 0.5-0.6
▨ 0.4-0.5
■ 0.3-0.4
■ 0.2-0.3
■ 0.1-0.2
■ 0-0.1

Margin size (mm)

PPVr

□ 1.85-2
□ 1.7-1.85
□ 1.55-1.7
□ 1.4-1.55
▨ 1.25-1.4
▨ 1.1-1.25
■ 0.95-1.1
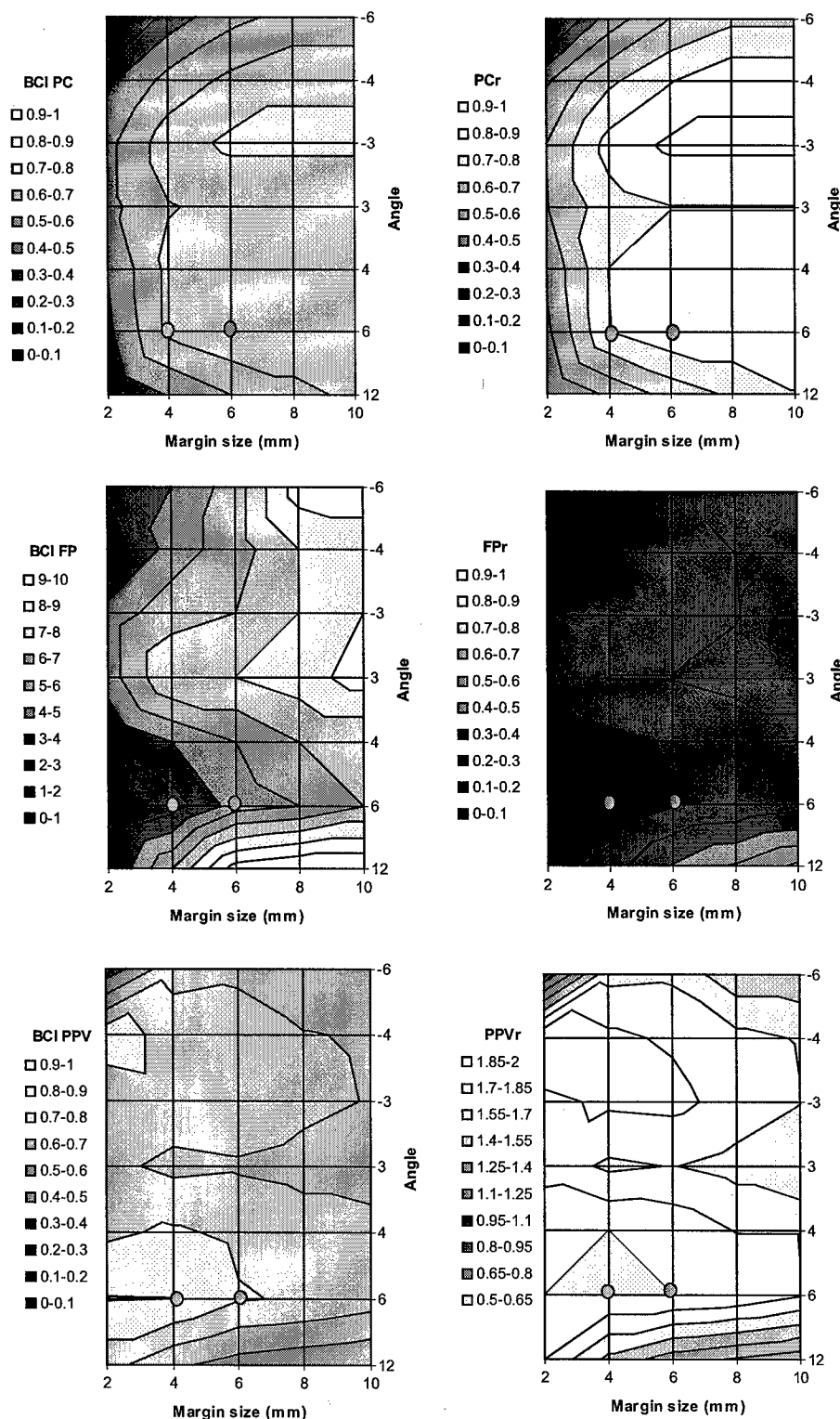▨ 0.8-0.95
▨ 0.65-0.8
□ 0.5-0.65

Margin size (mm)

Fig. 9: Variation in percent correct (sensitivity), false positive rate, and PPV of BCI (left column) and of BCI compared to single-view CAD (BCI/CAD ratio, right column) as a function of horizontal displacement angle and correlation area margin in mm (USM pre-processing, 44 mm kernel size, 8/4 DOG, no area or eccentricity correlation rule). Optimal parameters are marked with two circles.

the PPV. Similar results were obtained when applying an addition eccentricity rule.

The results for horizontal displacement exhibited similar dependencies, except that the best performance was provided by USM processing (with 44 mm kernel size and 8/4 DOG). Fig. 9 illustrates the horizontal displacement results for this parameter combination as a function of correlation margin size and displacement angle. Again, aiming to maintain a sensitivity higher than 60%, optimum performance, in terms of both the BCI performance alone, and the relative improvement with respect to single-view CAD, was exhibited for a margin size of 4-6 mm at +6° acquisition angle. Similar to vertical displacement results, the horizontal results did not show any improvement in the PPV with the additional area or eccentricity correlation rules.

Taking into account all the dependencies described above, the optimum oblique view for the acquisition angle appeared to be at around -3° vertical displacement or +6° horizontal displacement, 8/4 mm DOG, 4-6 mm margin size, and no area or eccentricity correlations. Table 3 provides the actual figures of merit for the optimum performance of the BCI method. The summary results show that vertical displacement yields better results. Furthermore, breaking the lesions into small (4.7-8.7 mm) and large (8.7-13.3 mm) sizes indicate that the percent improvement compared to single-view CAD is higher for smaller lesions.

Using the optimum geometry (0/–3° vertical displacement) and processing parameters, BCI results in 62.5% sensitivity, 1.5 FP/image, and 0.885 PPV. The corresponding values from the observer experiment were 56%, 10.8, and 0.45, respectively. Compared to single-view CAD results, the BCI reduced sensitivity by 20%. However, the corresponding reduction in FPs was notably higher (94%) leading to 140% improvement in the PPV. Adjustment in the processing parameters could yield higher sensitivity at the expense of higher FPs.

4. DISCUSSION

Currently, no radiologic screening program exists for early detection of lung cancer. Early detection presently relies on chest radiography examinations performed on asymptomatic patients for other diagnostic purposes. There have been studies that favor using chest radiography as a screening tool (Shimizu 1992, Brett 1969). However, the usefulness of such a screening program for lung cancer has been questioned based upon its ineffectiveness either in diagnosis (Gurney 1995) or in changing the mortality rate once the cancer is diagnosed (Fontana 1991). If lung nodules could be reliably detected in earlier stages of lung cancer, a screening program could be justified. Apart from low-dose CT, which is currently under investigation, current radiologic technology is unable to either visually or computationally (i.e., using CAD) image/detect lung nodules at a sufficiently early stage without generating a large number of false positives. The percentage of false-positive diagnoses unfavorably affects the predictive value of a screening program, especially when the prevalence of the disease is low (Kundel 1981). CT can surpass many of the limitation of chest radiography in imaging lung cancer (Henschke 1999). However, its utilization as a screening method raises economical (cost and technology availability), patient care (e.g., over-treatment), and epidemiological (e.g., patient dose) issues.

BCI is a new imaging technique that has not been investigated in the past. The phantom-based findings reported in this paper, the first public report of the technique, suggest significant potential of BCI to surpass the fundamental anatomical noise limitations of chest radiographic imaging and chest CAD and to improve the early detection of subtle lung nodules. A more sensitive and specific diagnostic approach for smaller lesions (4-11 mm diameter unmagnified size in this study), BCI has the potential to change the current state of practice, perhaps leading to a preventive lung cancer screening program for high-risk populations. The cost associated with the technology is minimal, and thus it can be implemented cost-effectively at doses comparable to chest radiography.

These findings of this study require further important validations. An open issue is the sensitivity of the BCI performance to the initial performance level of the single-view CAD algorithm. We intend to test the BCI scheme using different CAD algorithms, algorithms with more aggressive FP reduction strategies, and an iteratively combined dual-view CAD scheme. In terms of acquisition, plans are underway to assess the sensitivity of the BCI performance to exposure, potentially reducing the total exposure to that of a single PA chest exam. Finally, the performance of the technique should be measured on human subjects with confirmed lung nodules, with additional strategies to minimize possible motion artifacts.

Table 3: The optimal performance of the BCI for lesions within various size ranges for vertical and horizontal displacement of the x-ray tube. The vertical displacement images were processed with no preprocessing, 44 mm kernel size, and 8/4 mm DOG. The horizontal displacement images were processed with USM pre-processing, 44 mm kernel size, and 8/4 mm DOG.

	0/-3° Vertical Displacement						0/+6° Horizontal Displacement					
	5-13 mm		5-9 mm		9-13 mm		5-13 mm		5-9 mm		9-13 mm	
Margin	4 mm	6 mm	4 mm	6 mm	4 mm	6 mm	4 mm	6 mm	4 mm	6 mm	4 mm	6 mm
BCI PC	62.5%	65.6%	62.5%	68.8%	62.5%	62.5%	62.5%	68.8%	68.8%	81.3%	56.3%	56.3%
BCI FP	1.5	2.0	1.5	2.0	1.5	2.0	2.5	4.5	2.5	4.5	2.5	4.5
BCI PPV	0.885	0.867	0.833	0.818	0.786	0.750	0.802	0.708	0.691	0.591	0.646	0.500
PCr	0.801	0.840	0.929	1.000	0.708	0.708	0.798	0.878	0.786	0.929	0.817	0.817
FPr	0.058	0.077	0.058	0.077	0.058	0.077	0.108	0.191	0.108	0.191	0.108	0.191
PPVr	2.404	2.350	4.322	4.250	3.149	2.959	2.319	2.038	3.014	2.565	3.459	2.630

ACKNOWLEDGEMENTS

REFERENCES

1. American Cancer Society. Cancer facts and figures - 2002, Atlanta, GA, 2002.
2. G.Z. Brett. Earlier diagnosis and survival in lung cancer. *British Medical Journal*, 4:260-262, 1969.
3. A.E. Burgess, X. Li and C.K. Abbey. Nodule detection in two component noise: toward patient structure. Volume 3036 of *SPIE Medical Imaging*, pages 2-13, 1997.
4. D.P. Carmody, C.F. Nodine, and H.L. Kundel. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, 9:339-344, 1980.
5. M.J. Carreira, D. Cabello, et al. Computer-aided diagnoses: Automatic detection of lung nodules, Med Phys 25(10): 1998-2006, 1998.
6. J.T. Dobbins, R.L. Webber, and S.M Hames. Thomosynthesis for improved pulmonary nodule detection. *Radiology*, 209(P):280, 1998.
7. R.S. Fontana, D.R. Sanderson, L.B. Woolner, W.F. Taylor, W.E. Miller, J.R. Muhm, P.E. Barnatz, W.S. Payne, P.C. Pairolero, and E.J. Bergstralh. Screening for lung cancer: a critique of the Mayo lung project. *Cancer*, 67:1155-1164, 1991.
8. G. Gavelli and E. Giampalma. Sensitivity and specificity of chest x-ray screening for lung cancer. Proceedings of the International Conference on Prevention and early Diagnosis of Lung Cancer, pages 103-108, 1998.
9. M.S. Giger, K. Doi, and H. MacMahon. Image feature analysis and computer-aided diagnosis in digital radiography, 3. Automated detection of nodules in peripheral lung fields. *Medical Physics*, 15(2):158-166, 1988.
10. R.C. Gonzalez, R.E. Woods, Digital Image Processing. New York, Addison-Wesley. 3rd ed., 1993.
11. J.W. Gurney. Why chest radiography became routine. *Radiology*, 195:245-246, 1995.
12. C.I. Henschke, D.I. McCauley, D.F. Yankelevitz, D.P. Naidich, G. McGuiness, et al., Early lung cancer action project: overall design and findings from baseline screening, Lancet 354: 99-105, 1999.
13. S. Kido, J. Ikezoe, H. Naito, J. Arisawa, S. Tamura, T. Kozuka, W. Ito and H. Kato. Clinical evaluation of pulmonary nodules with single-exposure dual-energy subtraction chest radiography with an iterative noise-reduction algorithm. *Radiology*, 194(2):407-412, 1995.
14. H.L. Kundel. Predictive value and threshold detectability of lung tumors. *Radiology*, 139:25-29. 1981.
15. H.L. Kundel, C.F. Nodine and D. Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13:175-181, 1978.
16. H.L. Kundel and C.F. Nodine. Interpreting chest radiographs without visual search. *Radiology*, 116:527-532, 1975.
17. J. Padilla, V. Calvo, J.C. Penalver, G.Sales, and A. Morcillo. Surgical results and prognostic factors in early non-small cell lung cancer. *Annals of Thoracic Surgery*, 63(3):324-326, 1997.
18. R.T. Heelan, B.J. Flehinger, M.R. Melamed, M.B. Zaman and W.B. Perchick. Non-small-cell lung cancer: results of the New York screening program. *Radiology*, 151:289-293, 1984.
19. K. Mori, N. Yanase, M. Kaneko, R. Ono, and S. Ikeda. Diagnosis of peripheral lung cancer in cases of tumors 2 cm or less in size. *Chest*, 95:304-308, 1989.
20. J.R. Muhm, W.E. Miller, R.S. Fontana, D.R. Sanderson, and M.A. Uhlenhopp. Lung cancer detected during a screening program using four-month chest radiographs. *Radiology*, 148:561-565, 1983.
21. M. Nadler, E.P. Smith, Pattern Recognition Engineering. New York, New York, John Wiley and Sons, 1993.
22. U. Neitzel, T. Pralow, C. Schaefer-Prokop and M. Prokop. Influence of scatter reduction on lesion signal-to-noise ratio and lesion detection in digital chest radiography. volume 3336 of *SPIE Medical Imaging*, 1998.
23. M.G. Penedo, M.J. Carreira, et al. Computer-aided diagnosis: a neural-network-based approach to lung nodule detection, IEEE Trans Med Imaging 17(6): 872-80, 1998.
24. G. Revesz and H.L. Kundel. Psychophysical studies of detection errors in chest radiology. *Radiology*, 123:559-562, 1977.
25. G. Revesz, H.L. Kundel and M.A. Graber. The influence of structured noise on the detection of radiologic abnormalities. *Investigative Radiology*, 9:479-486, 1974.

26. E. Samei, W.R. Eyler, and L.F. Baron. Effects of Anatomical Structure on Signal Detection. *Handbook of Medical Imaging, Vol. 1 Physics and Psychophysics*. Eds. J. Beutel, R. Van Metter, and H. Kundel. SPIE Press, Bellingham, WA, 2000.

27. E. Samei, M.J. Flynn, W. Eyler, Simulation of subtle lung nodules in projection chest radiography, *Radiology* 202(1): 117-124, 1997.

28. N. Shimizu, A. Ando, S. Teramoto, Y. Moritani, and K. Nishii. Outcome of patients with lung cancer detected via mass screening as compared to those presented with symptoms. *J Surgical Oncology*, 50(1):7-11, 1992.

29. B.K. Stewart and H.K. Huang. Single-exposure dual-energy computed radiography. *Method Physics*, 17(5):866-875, 1990.

30. B. Zheng, Y.H. Chang, et al. Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis." Academic Radiology 2(11): 959-966, 1995.

31. R.D. Zwicker and N.A. Atari, Transverse tomosynthesis on a digital simulator. *Medical Physics,* 24(6):867-71, 1997.

Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information

Georgia D. Tourassi[a] and Rene Vargas-Voracek
Department of Radiology, Duke University Medical Center, Durham, North Carolina 27710

David M. Catarious, Jr.
Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

Carey E. Floyd, Jr.
Department of Radiology, Duke University Medical Center, Durham, North Carolina 27710 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27710

The purpose of this study was to develop a knowledge-based scheme for the detection of masses on digitized screening mammograms. The computer-assisted detection (CAD) scheme utilizes a knowledge databank of mammographic regions of interest (ROIs) with known ground truth. Each ROI in the databank serves as a template. The CAD system follows a template matching approach with mutual information as the similarity metric to determine if a query mammographic ROI depicts a true mass. Based on their information content, all similar ROIs in the databank are retrieved and rank-ordered. Then, a decision index is calculated based on the query's best matches. The decision index effectively combines the similarity indices and ground truth of the best-matched templates into a prediction regarding the presence of a mass in the query mammographic ROI. The system was developed and evaluated using a database of 1465 ROIs extracted from the Digital Database for Screening Mammography. There were 809 ROIs with confirmed masses (455 malignant and 354 benign) and 656 normal ROIs. CAD performance was assessed using a leave-one-out sampling scheme and Receiver Operating Characteristics analysis. Depending on the formulation of the decision index, CAD performance as high as $A_z = 0.87 \pm 0.01$ was achieved. The CAD detection rate was consistent for both malignant and benign masses. In addition, the impact of certain implementation parameters on the detection accuracy and speed of the proposed CAD scheme was studied in more detail. © *2003 American Association of Physicists in Medicine.*
[DOI: 10.1118/1.1589494]

Key words: mammography, computer-assisted detection (CAD), knowledge-based, mutual information, receiver-operating characteristic (ROC)

I. INTRODUCTION

Breast cancer is one of the most devastating and deadly diseases for women.[1] While there are many exciting new techniques on the horizon, for the time being mammography remains the screening test in the battle against breast cancer. Patients with early-detected malignancies have a significantly lower mortality rate.[2,3] Unfortunately, it is reported that up to 30% of breast lesions go undetected in screening mammograms[4–6] and up to 2/3 of those lesions are visible in retrospect.[7] Breast masses comprise a significant portion of missed cancers.[4,5] The clinical significance of early diagnosis and the difficulty of the diagnostic task have generated a tremendous interest in developing computer-assisted detection (CAD) schemes for mammographic interpretation. Several studies have demonstrated that CAD technology has a positive impact on early breast cancer detection.[5,7,8] However, there are still unresolved issues related to the clinical role of CAD in mammography. For example, the CAD detection accuracy is reportedly lower for masses than for calcifications.[9,10] Since high sensitivity is essential in screening mammography, CAD is often compromised by a higher false-positive rate in the detection of breast masses. The impact of false-positive CAD cues on the recall rate of mammograms is under investigation.[5,8] Generally, it is assumed that the radiologists will be able to discard easily most of the false-positive cues. However, a recent study has challenged this belief.[11] The study also showed that low-performing CAD tools degrade radiologists' performance in noncued areas. Therefore, it is recommended that a cueing CAD tool should be used by an experienced interpreter to effectively process all cues.[10] However, the medical and legal implications of dismissing CAD cues are currently unknown.[10] Consequently, CAD research efforts in mammography are ongoing.

Thus far, the overwhelming majority of CAD techniques follow a two-step approach (e.g., Refs. 9, 12–23). Initially, traditional image processing is performed to identify suspicious mammographic regions. Subsequently, morphological and/or textural features are automatically extracted from these regions. The features are merged with linear classifiers or artificial intelligence techniques to further refine the detection and often the diagnosis (benign versus malignant) of potential abnormalities. The suspicious mammographic re-

gions detected by the CAD system serve as cues to the radiologists. Commercially available products are designed to operate as black boxes that provide diagnostic cues but not comprehensible decision models. In addition, some researchers have developed mathematical models to describe the statistical nature of mammograms.[24,25] Such models could be potentially extended to perform as CAD tools.

The purpose of this study is to develop a knowledge-based (KB) CAD scheme for the detection of breast masses in digitized mammograms. Generally, knowledge-based CAD (KB-CAD) systems aim to provide evidence-based decision support using a knowledge databank. Much like a physician relates a present case to those seen in the past, a KB system relates a new case to similar cases stored in its knowledge databank. Based on the similar cases, a diagnosis is assigned to the new case by analogy or by copying the answer if the match is close enough. The main benefits of using KB-CAD systems are the following: (1) KB-CAD systems take full advantage of growing data libraries without further re-training of the CAD system and, (2) they can be interactive allowing physicians to formulate their own questions and get interpretable answers.

The computational demands of maintaining, indexing, and querying a large knowledge databank have limited the application of these tools in mammography. Furthermore, defining similarity between two images is nontrivial. There is not a single similarity metric that is known to produce the best results in all applications. Common practice is to select diagnostically important features and feature-based distance metrics to determine similarity. Case-based reasoning (CBR) is a typical example of a KB system and it has been successfully applied for mammographic diagnosis based on radiologist-extracted BIRADS findings.[26,27] In addition, Chang *et al.* showed the feasibility of using a KB-CAD system for the detection of mammographic masses.[28] Their system employed a feature-based similarity metric that required segmentation of the suspected masses.

In contrast, our proposed KB-CAD scheme follows an image retrieval approach that is not feature-based but uses template matching with a global similarity metric. Template matching requires comparison of a given image with a template image. Each mammographic case stored in the knowledge databank serves as a template. Given a query mammographic region, the KB-CAD scheme retrieves similar cases from its knowledge databank. The focus of this study is to investigate mutual information (MI) as a potential similarity metric for knowledge-based detection of masses in screening mammograms.

MI is a fundamental concept in information theory.[29] It is defined in terms of two objects (i.e., images) and it measures how much one object can explain the other. Thus, MI captures the similarity or the amount of relevant information between two objects.[29] In medical imaging, MI has been a very effective similarity metric for image registration tasks.[30] The basic idea is that when two images are properly aligned, their MI is maximal. Our study aims to evaluate if MI can serve as a similarity metric in a template-matching scheme for the detection of mammographic masses. We hypothesize

that if two mammographic regions depict similar structures, they should contain relevant diagnostic information for each other. Therefore, by measuring their MI we can potentially quantify their diagnostic similarity. Furthermore, the MI is calculated directly from the images without any preprocessing. By using MI as a global similarity metric, we avoid issues related to image segmentation, feature extraction, and feature selection that are typically associated with feature-based similarity metrics or feature-based CAD schemes.

II. MATERIALS AND METHODS

A. The image database

The CAD system was developed and evaluated using the Digital Database for Screening Mammography (DDSM) that was collected at the University of South Florida under the DOD Breast Cancer Research Program Grant No. DAMD17-94-J-4015.[31] DDSM is intended as a benchmark database for CAD tools on screening mammograms. The database includes normal, cancer, and benign cases. A DDSM mammogram is considered normal if no further evaluation was required and the patient had a normal screening exam at least four years later. A cancer case is a screening mammogram with at least one biopsy-proven malignancy. A benign case is a screening mammogram with a suspicious finding that was determined to be benign by pathology or additional imaging.

DDSM includes three volumes, each containing mammograms digitized with a different digitizer (LUMISYS, HOW-TEK, and DBA). Each DDSM screening exam consists of two images for each breast (standard craniocaudal and mediolateral oblique views). Our study focused on the DDSM mammograms digitized using the LUMISYS scanner. Initially, these mammograms were downloaded and archived. From those, all mammograms with annotated masses were selected. Specifically, all malignant masses present in the sets "cancer_02," "cancer_05," "cancer_09," and "cancer_15" were identified. Similarly, all benign masses present in the sets "benign_01," "benign_04," "benign_06," "benign_13," and "benign_14" were identified. There were 260 studies with malignant masses and 146 studies with benign masses. Some masses were visible in one mammographic view only.

The DDSM includes information describing the location of the masses. 512×512 pixel regions of interest (ROIs) centered on the known location of each annotated mass were extracted. In addition, 512×512 pixel ROIs depicting normal tissue were also extracted. The normal ROIs were extracted from the sets "normal_09" and "normal_10." The two sets included 82 patients with normal screening mammograms. Two 512×512 pixel ROIs were randomly chosen from each view per breast. Thus, eight ROIs were extracted from each DDSM patient with a normal screening exam. There were 1465 ROIs in total; 455 ROIs depicting a biopsy-proven malignant mass, 354 ROIs with a benign mass, and the remaining 656 ROIs were normal. To facilitate detailed analysis according to the difficulty level of the detection task, all extracted ROIs were furthered indexed according to the den-
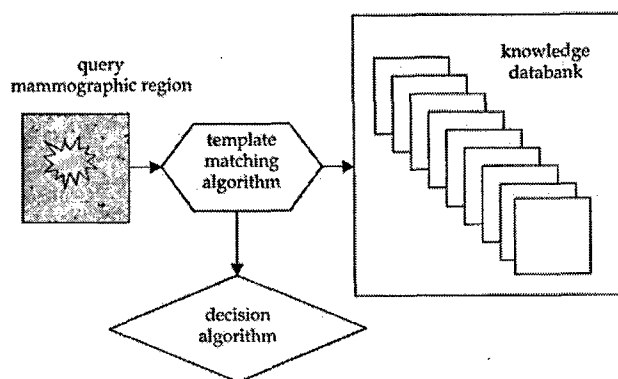
FIG. 1. Overview of the KB-CAD scheme.

sity rating of their corresponding mammogram. The ACR density rating is part of the associated patient information that is provided in the DDSM database.

B. Overview of the CAD scheme

Figure 1 highlights the three critical components of our KB-CAD scheme: (1) the knowledge databank, (2) the template matching algorithm, and (3) the knowledge-based decision algorithm. The knowledge databank contains mammographic ROIs that depict masses of known truth or normal tissue. Each ROI stored in the databank serves as a template. A query suspicious mammographic region is compared to the stored templates using the template-matching algorithm. Based on their information content, all similar templates in the databank are retrieved. A decision algorithm effectively combines the similarity indices and known truth of the retrieved templates into a prediction regarding the presence of a mass in the suspicious query mammographic region.

C. The template-matching algorithm

This section describes the algorithm employed in the study to measure the similarity between a query mammographic region and a template ROI stored in the knowledge databank. The algorithm utilizes mutual information (MI), a similarity index borrowed from information theory.[29]

Mutual information is a measure of general interdependence between two random variables x and y.[29] The MI concept can be easily extended to images. Given two images X and Y, their MI $I(X;Y)$ is expressed as

$$I(X;Y) = \sum_x \sum_y P_{XY}(x,y) \log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}, \quad (1)$$

where $P_{XY}(x,y)$ is the joint probability density function (pdf) of the two images based on their corresponding pixel values.[29] Equation (1) assumes that the image pixel values are samples of two random variables x and y, respectively. $P_X(x)$ and $P_Y(y)$ are the marginal pdfs. The basic idea is that when two images are similar, pixels with a certain intensity value in one image should correspond to a more clustered distribution of the intensity values in the other image.[30] The more the two images are alike, the more information X

provides for Y and vice versa. Therefore, the MI can be thought as an intensity-based measure of how much two images are alike. In the template-matching context, the MI increases when the query image X and the template image Y depict similar structures. Then, the pixel value in image X is a good predictor of the pixel value at the corresponding location in image Y.

Theoretically, MI is a more effective and robust similarity metric than traditional correlation.[32] Correlation techniques assume a linear relationship between the intensity values in the two images. MI measures general dependence without making any *a priori* assumptions.

The MI estimation of two mammographic ROIs requires computation of the joint and marginal pdfs as shown in Eq. (1). There are two published methods for the task: (1) Parzen windows,[33] and (2) the histogram approach.[34] We followed the histogram approach since it is quick and easy to implement. Time efficiency is very important for a knowledge-based CAD system.

According to the histogram method, a pdf is approximated using a histogram. For each histogram bin, the probability is estimated by counting the number of pixels that fall into a particular bin and dividing that number by the total number of pixels. Then, the MI of two images X and Y can be computed according to Eq. (1).

The number of bins selected for histogram approximation is a critical issue.[35] More bins allow for more detailed representation of the pdfs. However, these details may be nothing more than noise caused by the small sample size in each bin. The potential estimation error can substantially alter the results of a study.[36] Since the DDSM images considered in our study are 12 bit images, the 4096×4096 2D histogram required for the estimation of the joint pdf of two mammographic ROIs will be very sparse leading to serious MI estimation errors. Following typical practices of image registration applications, the pdfs were estimated using a reduced number of 256 equal-sized intensity bins for the histogram approximation technique. Furthermore, since the distribution for the pixel values can vary substantially among ROIs we applied the following rules. For each ROI, the mean μ and standard deviation σ of the ROI pixel values were calculated. Then, the interval $[\mu-2\sigma,\mu+2\sigma]$ was divided into the pre-selected number of equal segments (i.e., 256). Any rare pixel values falling outside the predetermined interval were assigned to the extreme left or right bins when calculating the histograms. The above rules were followed consistently for all ROIs.

D. The knowledge-based decision index

The knowledge-based decision index was computed using the level of similarity and the ground truth of the best-matched templates. Two experiments were performed to determine the most effective way to use the CAD system as a computer aid for the detection of mammographic masses. In the first experiment, the knowledge databank included only

TABLE I. ROC performance of the CAD scheme for two decision indices (D_1, D_2) and for varying number of the top matches considered ($k = 1$, 10, 50, 100, 200, 400, ALL). The MI calculations were based on 256 histogram bins and the full resolution 512×512 ROIs.

	1	10	50	100	200	400	ALL
D_1	0.71±0.01	0.71±0.01	0.71±0.01	0.72±0.01	0.73±0.01	0.74±0.01	0.75±0.01
D_2	0.71±0.01	0.79±0.01	0.84±0.01	0.85±0.01	0.85±0.01	0.86±0.01	0.87±0.01

mammographic ROIs that contained a mass. In the second experiment, the knowledge databank included both normal and mass ROIs.

Experiment 1: Given a query mammographic ROI Q_i, a decision index was calculated based on the MI between the query ROI and each known *mass* M_j in the knowledge databank. The decision index $D_1(Q_i)$ was the average MI of the k best mass matches:

$$D_1(Q_i) = \frac{1}{k} \sum_{j=1}^{k} \mathrm{MI}(Q_i, M_j). \tag{2}$$

Theoretically, a query ROI depicting a mass should match better with the databank of mass ROIs than a query ROI depicting normal breast tissue, thus resulting in a higher $D_1(Q_i)$.

Experiment 2: Given a query mammographic ROI Q_i, a decision index $D_2(Q_i)$ was calculated as the difference of two terms. The first term measures the average MI between the query ROI and its k best mass matches M_j. Similarly, the second term measures the average MI between the query ROI and its k best normal N_j matches,

$$D_2(Q_i) = \frac{1}{k} \sum_{j=1}^{k} \mathrm{MI}(Q_i, M_j) - \frac{1}{k} \sum_{j=1}^{k} \mathrm{MI}(Q_i, N_j). \tag{3}$$

Theoretically, a query ROI depicting a mass should have a higher $D_2(Q_i)$.

E. Performance evaluation

The diagnostic performance of the CAD system was evaluated using a leave-one-out sampling scheme.[37] Given the database of 1465 mammographic ROIs, each ROI was excluded once to serve as a query case. In experiment 1, the remaining mass cases were used to establish the knowledge databank. In experiment 2, the remaining 1464 cases were used to establish the databank. The experiments were repeated until every ROI served as a query ROI. The calculated decision indices D_1 and D_2 were analyzed based on Receiver Operating Characteristic (ROC) analysis methodology. The ROCKIT software package developed by Metz *et al.* (available at www-radiology.uchicago.edu/krl/toppage11.htm) was used to fit ROC curves to the two decision indices implemented in this study. For both indices, ROC performance was estimated for varying values of the top matches (parameter k) considered.

F. Influence of implementation parameters

In a knowledge-based system, comparing a query case with every archived case can be computationally expensive. This is certainly a concern with image databases and global similarity metrics such as the mutual information. One way to reduce the computation time is by reducing the number of histogram bins employed for the MI estimation. We repeated the previous experiments estimating the MI using 64 and 128 histogram bins to evaluate the impact of this implementation parameter on the overall performance of the CAD scheme. In addition, we studied the effect of image sub-sampling. Since a knowledge-based CAD system requires individual comparisons of the query ROI with all stored ROIs, it can be computationally more effective if the comparisons are performed on reduced-resolution ROIs. We repeated the above-mentioned experiments with sub-sampled ROIs (256×256, 128×128, and 64×64) to determine if the CAD detection rate degrades for sparsely sampled ROIs.

III. RESULTS

The experimental results are presented in two sections. Each section addresses an important issue: (1) overall ROC performance, (2) influence of the implementation parameters on performance and time efficiency of the proposed CAD scheme.

A. Overall ROC performance of the CAD scheme

No particular trend was observed in obtaining higher MI values with template ROIs extracted from the same mammogram as the query ROI. Therefore, the overall detection performance of the KB-CAD scheme was analyzed on a per ROI basis, not on a per-case basis. Table I shows the performance of the CAD system as measured by the ROC area index (A_z) for each one of the decision indices D_1, D_2 and for varying number of the top matches considered (parameter k).

Several observations can be made based on Table I. The performance of the KB-CAD scheme varied substantially depending on the decision algorithm. Overall, the CAD system had a significantly better ROC performance when the decision index was calculated using the knowledge databank that includes both mass and normal templates (D_2). Furthermore, CAD performance improved as more matched cases were considered in the calculation of the decision index D_2. The CAD system achieved its best ROC performance $(A_z = 0.87 \pm 0.01)$ when all archived cases were included in the calculation of D_2. However, when the detection decision was
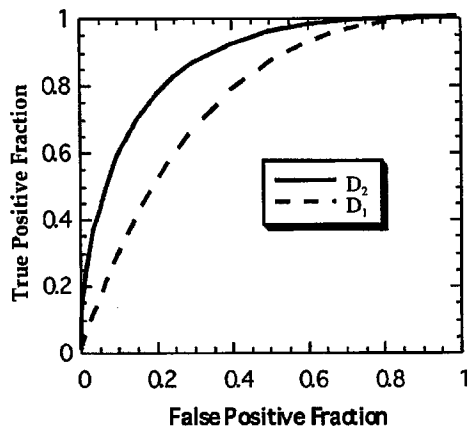
FIG. 2. ROC curves of the KB-CAD scheme based on the two decision algorithms (D_1 and D_2). The calculation of the two decision indices includes all archived templates.

based only on the best mass matches (D_1), the ROC area index was statistically significantly lower ($A_z = 0.75 \pm 0.01$) but substantially less dependent on the parameter k. Figure 2 shows the corresponding ROC curves of the CAD system based on the two decision algorithms (D_1 and D_2) for k = ALL. As the figure shows the best performing knowledge-based CAD scheme achieved 90% sensitivity while safely eliminating 65% of the normal regions.

The best performing CAD scheme was analyzed in more detail. First, detection accuracy was evaluated separately for malignant and benign masses. The CAD scheme showed robust performance among the two groups of masses: A_z(malignant masses versus normal)$=0.88 \pm 0.01$ and A_z(benign masses versus normal)$=0.86 \pm 0.01$. A small subset of mammographic ROIs (57 out of 809 mass regions) contained both a mass and microcalcifications. Significant degradation in ROC performance was observed for this subset ($A_z = 0.80 \pm 0.04$) compared to the remaining set of mass regions ($A_z = 0.89 \pm 0.01$).

To assess the effect of case difficulty, the best performing CAD scheme was further analyzed for each subgroup of masses according to their DDSM subtlety rating. The mass subtlety rating is not a BI-RADS standard. It is simply a subjective impression of the DDSM radiologist on the subtlety of the lesion. A higher subtlety rating indicates a more obvious lesion. Table II shows that the overall ROC area index of the CAD tool is fairly robust regardless of the reported subtlety of the mass ROIs. The only exception is the subgroup of masses with Subtlety rating 2. For this subgroup, the KB-CAD had a statistically significantly lower ROC performance than the other subgroups.

TABLE III. Effect of mammographic density on the ROC performance of the KB-CAD scheme.

Mammographic density	No. of mass ROIs	No. of normal ROIs	A_z
1: fatty breast	193	96	0.98 ± 0.01
2: fibroglandular breast	362	272	0.91 ± 0.01
3: heterogeneous breast	195	208	0.87 ± 0.02
4: dense breast	59	80	0.64 ± 0.05

Since mass detection is more challenging in dense breasts, we also analyzed the CAD performance for each subgroup of ROIs based on the DDSM density rating of the mammogram from which they were extracted. Table III summarizes those results. Table III shows that the ROC area varied significantly, starting from almost perfect performance in fatty breasts ($A_z = 0.98 \pm 0.01$) and progressively degrading in fibroglandular ($A_z = 0.91 \pm 0.01$) and heterogeneous breasts ($A_z = 0.87 \pm 0.02$). The CAD performance was dramatically lower for dense mammograms ($A_z = 0.64 \pm 0.05$) than for all remaining categories. Since the ROIs extracted from dense mammograms comprised only 10% (139/1465) of the whole data set, it is unclear if the inferior performance can be partially contributed to the lower representation of dense ROIs in the knowledge databank.

B. Influence of implementation parameters on CAD performance

Tables IV and V demonstrate the impact of two implementation parameters on the overall ROC area index of the KB-CAD scheme. The first parameter is the number of histogram bins used in the calculation of the MI between two ROIs. The second parameter is the sub-sampling factor of the mammographic regions. Table IV shows the impact of both parameters on decision index D_1. Table V corresponds to decision index D_2. The calculation of both decision indices was based on all archived cases (k = ALL).

Two important conclusions can be drawn from the above-mentioned tables. First, when estimating the MI between two ROIs, the number of histogram bins should be selected carefully. CAD performance can be significantly degraded as the number of histogram bins increases. The degradation is particularly strong with the coarser ROIs; using a large number of bins introduces serious estimation errors due to the smaller number of pixels available in each bin. However, there is no such concern with the full-resolution ROIs. Sec-

TABLE IV. Effect of image sub-sampling ($256 \times 256, 128 \times 128, 64 \times 64$) and the number of histogram bins (64,128,256) on the overall ROC area index of the KB-CAD scheme for decision index D_1. The full resolution ROIs are 512×512. The reduced size ROIs were created by sub-sampling accordingly the full resolution ROIs.

	512×512	256×256	128×128	64×64
64 bins	0.75 ± 0.01	0.75 ± 0.01	0.75 ± 0.01	0.72 ± 0.01
128 bins	0.75 ± 0.01	0.75 ± 0.01	0.73 ± 0.01	0.59 ± 0.01
256 bins	0.75 ± 0.01	0.73 ± 0.01	0.71 ± 0.01	0.51 ± 0.01

TABLE II. Effect of mass subtlety rating on the overall ROC performance of the KB-CAD scheme.

	Subtlety=1	Subtlety=2	Subtlety=3	Subtlety=4	Subtlety=5
ROC A_z	0.87 ± 0.04	0.79 ± 0.03	0.86 ± 0.02	0.85 ± 0.01	0.89 ± 0.01

TABLE V. Effect of image sub-sampling (256×256, 128×128, 64×64) and the number of histogram bins (64,128,256) on the overall ROC area index of the KB-CAD scheme for decision index D_2. The full resolution ROIs are 512×512. The reduced size ROIs were created by sub-sampling accordingly the full resolution ROIs.

	512×512	256×256	128×128	64×64
64 bins	0.87±0.01	0.87±0.01	0.87±0.01	0.87±0.01
128 bins	0.87±0.01	0.87±0.01	0.86±0.01	0.84±0.01
256 bins	0.87±0.01	0.86±0.01	0.84±0.01	0.81±0.01

ond, decision index D_2 appears to be more robust to the above effects. D_2 is basically the difference of two terms. If both terms are over- or underestimated, their difference can still reasonably maintain its relative discriminant power. Our experimental results support this hypothesis.

The above-mentioned experiments were performed on a Sun Sparc Ultra-80 workstation with 4 450 MHz processors (Sun Microsystems, Mountain View, CA). Using a single processor, the time requirements to calculate the mutual information between two mammographic ROIs ranged from 0.01 to 0.21 s depending on the ROI size and number of histogram bins selected for the MI estimation. Therefore, the proposed knowledge-based CAD scheme can be easily translated into a real-time CAD system. It takes 2.5 min to compare a mammographic region with 1000 archived cases. The above-mentioned calculation assumes 512×512 ROIs and 64 histogram bins. If the comparison is made using sub-sampled ROIs (64×64), the CAD response time can be reduced to 10 s per query mammographic ROI.

IV. DISCUSSION

In this study, we presented a knowledge-based mass detection scheme for screening mammograms. The proposed CAD scheme is designed to provide a prediction regarding the presence or absence of a mass in a query mammographic region based on similar cases stored in the system's knowledge databank. In its present state, the CAD scheme can function as an interactive tool to help radiologists analyze mammographic regions that attract visual attention. However, the proposed algorithm could be combined with other mass detection schemes for evidence-based reduction of false positive CAD cues. Based on our study, the system was able to maintain 90% sensitivity while effectively eliminating 65% of the normal regions. The performance was consistent for both malignant and benign masses. Since breast masses span a wide range of shapes, sizes, and contrast, the performance of a knowledge-based CAD scheme can be easily compromised if its knowledge databank is limited. Our CAD scheme was developed and evaluated based on a large number of examples from a publicly available database. It has been reported that the database contains really challenging cases.[38] Overall, the estimated performance of our CAD scheme compares favorably with published results from other CAD systems.[14,28] However, direct comparison is not feasible since the results were obtained from different databases. Further studies are needed to evaluate our approach in

other data sets and larger populations of screened women. Furthermore, since the proposed CAD capitalizes on continuously depositing cases in the databank, it is important to assess the impact of the digitization process. The present study was based on DDSM cases digitized with the same digitizer. Studies are under way evaluating how well the CAD system can generalize to other DDSM cases digitized using a different digitizer.

The reported CAD performance was fairly robust regardless of the mass subtlety rating. However, analysis according to breast density showed that CAD performance degrades substantially in dense breasts as it is clinically known. This issue needs investigation due to the lower representation of dense mammograms in the dataset. It is possible that augmenting the knowledge databank with more examples from dense mammograms will improve the CAD performance. Another potentially promising strategy is to design the KB-CAD scheme so that each query ROI is only compared to archived ROIs that were extracted from mammograms with similar density rating as the query ROI. We acknowledge that although indexing the ROIs according to their mammographic density may improve the overall performance of the knowledge-based scheme during the development stage, it may also introduce a serious bias. Observer variability in the reporting of BI-RADS findings is a well-documented issue. Specifically, a study indicated that the overall agreement across observers for the BI-RADS reporting of the mammographic density is only moderate.[39] The same study also showed very poor agreement among observers in use of the category "heterogeneous" breast. Since the DDSM density rating was reported by several different radiologists at various clinical sites, it is expected that any CAD tool developed on the data set will be more fault-tolerant than a CAD tool developed based on cases collected from a single site and read by a single radiologist. However, this issue needs careful investigation.

The main innovation of this study is the application of the mutual information as the similarity metric in a knowledge-based system. MI is a statistical tool that measures to what degree one image can be predicted from another. In image databases, similarity is typically feature-based and often demands substantial image preprocessing. In contrast, the MI between two images is calculated directly without the burden and potential variability of segmentation, object recognition, and feature selection. Therefore, critical CAD issues such as optimized feature selection and merging are bypassed in the proposed KB-CAD system. Considering the difficulty of the mass detection task, the presented concept could be generalizable to other imaging modalities and diagnostic tasks. However, special attention is required when selecting certain implementation parameters. Our study showed that parameters such as the image sub-sampling factor and the number of histogram bins used to estimate the MI affect the overall performance of the detection scheme. For the detection of mammographic masses, if the number of histogram bins is kept reasonably low, then the overall ROC performance of the system remains very robust to image sub-sampling. Continuing research on the formulation of information-theoretic

similarity metrics can be a promising alternative to feature-based CAD techniques.

Finally, an important component of a KB-CAD system is the decision algorithm that combines the similarity level and the truth files of the retrieved cases into a prediction about the query case. Our study showed that using a databank of both mass and normal cases results in a CAD system with a statistically significantly better performance. However, the study also showed that the overall CAD performance varies depending on the number of top matches considered in the calculation of the decision index. Based on the results, the CAD performance was optimal when all stored cases contributed equally in the derivation of the decision index. These findings are attributed to the fact that mutual information is primarily a shape and size-driven similarity measure. Although a mass ROI is expected to match better with other mass ROIs than normal ROIs, the opposite is not true. We will elaborate on that. Theoretically, MI should be able to capture the dissimilarity of two ROIs if one depicts a mass and one depicts just normal tissue since MI is affected by morphology. If, however, both ROIs depict normal parenchyma, then their probability of matching is smaller since the structures and patterns present in normal parenchyma are much more variable. Thus, the MI of two normal ROIs is generally expected to be low. Decision index D_2 capitalizes on this by taking the difference of two terms. If the query ROI contains a mass, then the difference reflects the substantial separation between the morphological properties of masses and normal regions. If the query ROI is normal, then the difference is small since the normal ROI should have low MI with either mass or normal cases. It is documented however, that in image registration MI can produce misleading matches in the presence of noise.[30] Nonetheless, the impact of the noise effect on D_2 should be minimized as more archived cases are considered in the calculation of the decision index.

To summarize, the recent emergence of multimedia digital libraries has increased the interest on comprehensive similarity metrics that can capture effectively the content of images without requiring elaborate image preprocessing. Such metrics can play an important role in knowledge-based CAD systems in an effort to facilitate evidence-based diagnostic interpretation of medical images. Our study showed that mutual information is a promising similarity metric in a knowledge-based CAD scheme for the detection of masses in screening mammograms.

ACKNOWLEDGMENTS

[a]Electronic mail: gt@deckard.mc.duke.edu

[1] R. T. Greenlee, M. B. Hill-Harmon, T. Murray, and M. Thun, "Cancer statistics, 2001," Ca-Cancer J. Clin. **51**, 15–36 (2001).
[2] U.S. D.H.H.S., "Healthy People 2010 (Conference Edition, in Two Volumes)," Washington, DC, 2000.
[3] L. Tabar *et al.*, "Reduction in mortality from breast cancer after mass screening with mammography," Lancet **1**, 829–832 (1985).
[4] R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).
[5] R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology **219**, 192–202 (2001).
[6] B. C. Yankaskas, M. J. Schell, R. E. Bird, and D. A. Desrochers, "Reassessment of breast cancers missed during routine screening mammography: A community-based study," AJR, Am. J. Roentgenol. **177**, 535–541 (2001).
[7] L. J. W. Burhenne *et al.*, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554–562 (2000).
[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] N. Petrick, B. Sahiner, H.-P. Chan, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis—Experience in 263 patients," Radiology **224**, 217–224 (2002).
[10] C. J. D'Orsi, "Computer-aided detection: There is no free lunch," Radiology **221**, 585–586 (2001).
[11] B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-copy mammographic readings with different computer-assisted detection cuing environments: Preliminary findings," Radiology **221**, 633–640 (2001).
[12] W. Qian, L. H. Li, and L. P. Clarke, "Image feature extraction for mass detection in digital mammography: Influence of wavelet analysis," Med. Phys. **26**, 402–408 (1999).
[13] B. Zheng, Y. H. Chang, X. H. Wang, W. F. Good, and D. Gur, "Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm," Acad. Radiol. **6**, 327–332 (1999).
[14] W. Qian, L. H. Li, L. P. Clarke, R. A. Clark, and J. Thomas, "Digital mammography: Comparison of adaptive and nonadaptive CAD methods for mass detection," Acad. Radiol. **6**, 471–480 (1999).
[15] H. P. Chan *et al.*, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," Med. Phys. **25**, 2007–2019 (1998).
[16] H. P. Chan *et al.*, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," Radiology **212**, 817–827 (1999).
[17] D. L. Thiele, C. Kimme-Smith, T. D. Johnson, M. McCombs, and L. W. Bassett, "Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes," Med. Phys. **23**, 549–545 (1996).
[18] Z. 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," Acad. Radiol. **5**, 155–168 (1998).
[19] F. Schmidt *et al.*, "An automatic method for the identification and interpretation of clustered microcalcifications in mammograms," Phys. Med. Biol. **44**, 1231–1243 (1999).
[20] M. A. Gavrielides, J. Y. Lo, R. Vargas-Voracek, and C. E. Floyd, "Segmentation of suspicious clustered microcalcifications in mammograms," Med. Phys. **27**, 13–22 (2000).
[21] S. Y. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," IEEE Trans. Med. Imaging **19**, 115–126 (2000).
[22] W. Qian, X. J. Sun, D. S. Song, and R. A. Clark, "Digital mammography: Wavelet transform and Kalman-filtering neural network in mass segmentation and detection," Acad. Radiol. **8**, 1074–1082 (2001).
[23] 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).
[24] J. J. Heine, S. R. Deans, R. P. Velthuizen, and L. P. Clarke, "On the statistical nature of mammograms," Med. Phys. **26**, 2254–2265 (1999).
[25] J. J. Heine and R. P. Velthuizen, "A statistical methodology for mammographic density detection," Med. Phys. **27**, 2644–2651 (2000).
[26] C. E. Floyd, Jr., J. Y. Lo, and G. D. Tourassi, "Breast biopsy: Case-based reasoning computer-aid using mammography findings for the breast biopsy decisions," AJR, Am. J. Roentgenol. **175**, 1347–1352 (2000).
[27] A. O. Bilska-Wolak and C. E. Floyd, Jr., "Development and evaluation of

a case-based reasoning classifier for prediction of breast biopsy outcome with BI-RADSTM lexicon," Med. Phys. **29**, 2090–2100 (2002).

[28] Y.-H. Chang, L. A. Hardesty, C. M. Hakim, T. S. Chang, B. Zheng, W. F. Good, and D. Gur, "Knowledge-based computer-aided mass detection on digitized mammograms: A preliminary assessment," Med. Phys. **28**, 455–461 (2001).

[29] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[30] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, *Medical Image Registration* (CRC, Boca Raton, FL, 2000).

[31] M. Heath *et al.*, "Current status of the digital database for screening mammography," in *Digital Mammography* (Kluwer Academic, Dordrecht, 1998).

[32] W. Li, "Mutual information functions versus correlation functions," J. Stat. Phys. **60**, 823–837 (1990).

[33] W. M. Wells, P. V. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," Med. Image Anal **1**, 35–51 (1996).

[34] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodal image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[35] A. Treves and A. Panzeri, "The upward bias in measures of information derived from limited data samples," Neural Comput. **7**, 399–407 (1995).

[36] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd, Jr., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," Med. Phys. **28**, 2394–2402 (2001).

[37] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability*, edited by D. R. Cox *et al.* (Chapman & Hall, New York, 1993).

[38] G. M. teBrake, N. Karssemeijer, and J. H. Hendricks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," Phys. Med. Biol. **45**, 2843–2857 (2000).

[39] W. A. Berg, C. Campassi, P. Langenberg, and M. J. Sexton, "Breast imaging reporting and data system: Inter- and intraobserver variability in feature analysis and final assessment," AJR, Am. J. Roentgenol. **174**, 1769–1777 (2000).