Computerized image analysis: Estimation of breast density on mammograms

Chuan Zhou, Heang-Ping Chan,[a] Nicholas Petrick, Mark A. Helvie, Mitchell M. Goodsitt, Berkman Sahiner, and Lubomir M. Hadjiiski
Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0030

An automated image analysis tool is being developed for the estimation of mammographic breast density. This tool may be useful for risk estimation or for monitoring breast density change in prevention or intervention programs. In this preliminary study, a data set of 4-view mammograms from 65 patients was used to evaluate our approach. Breast density analysis was performed on the digitized mammograms in three stages. First, the breast region was segmented from the surrounding background by an automated breast boundary-tracking algorithm. Second, an adaptive dynamic range compression technique was applied to the breast image to reduce the range of the gray level distribution in the low frequency background and to enhance the differences in the characteristic features of the gray level histogram for breasts of different densities. Third, rule-based classification was used to classify the breast images into four classes according to the characteristic features of their gray level histogram. For each image, a gray level threshold was automatically determined to segment the dense tissue from the breast region. The area of segmented dense tissue as a percentage of the breast area was then estimated. To evaluate the performance of the algorithm, the computer segmentation results were compared to manual segmentation with interactive thresholding by five radiologists. A ''true'' percent dense area for each mammogram was obtained by averaging the manually segmented areas of the radiologists. We found that the histograms of 6% (8 CC and 8 MLO views) of the breast regions were misclassified by the computer, resulting in poor segmentation of the dense region. For the images with correct classification, the correlation between the computer-estimated percent dense area and the ''truth'' was 0.94 and 0.91, respectively, for CC and MLO views, with a mean bias of less than 2%. The mean biases of the five radiologists' visual estimates for the same images ranged from 0.1% to 11%. The results demonstrate the feasibility of estimating mammographic breast density using computer vision techniques and its potential to improve the accuracy and reproducibility of breast density estimation in comparison with the subjective visual assessment by radiologists. © *2001 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1376640]

I. INTRODUCTION

Breast cancer is one of the leading causes for cancer mortality among women.[1] One in every eight women will develop breast cancer at some point in their lives. The most successful method for the early detection of breast cancer is screening mammography. Currently, mammograms are analyzed visually by radiologists. Because of the subjective nature of visual analysis, qualitative responses may vary from radiologist to radiologist. Therefore, a computerized method for analyzing mammographic features would be useful as a supplement to the radiologist's assessment. Previous research efforts in computer-aided diagnosis (CAD) for breast cancer detection mainly concentrated on detection and characterization of masses and microcalcifications on mammograms by using computer vision techniques. It has been demonstrated that an effective CAD algorithm can improve the diagnostic accuracy of breast cancer characterization on mammograms, which, in turn, may reduce unnecessary biopsies. In this work, we are studying the feasibility of developing a CAD system for an analysis of breast density on mammograms. Studies have shown that there is a strong positive correlation between breast parenchymal density on mammograms and breast cancer risk.[2–9] The relative risk is estimated to be about 4 to 6 times higher for women whose mammograms have parenchymal densities over 60% of the breast area, as compared to women with less than 5% of parenchymal densities.

An important difference between breast density as a risk factor and most other risk factors is the fact that breast tissue density can be changed by dietary or hormonal interventions.[6,10,11] Although there is no direct evidence that changes in mammographic breast densities will lead to changes in breast cancer risk, the strong correlation between breast density and breast cancer risk has prompted researchers to use mammographic density as an indicator for monitoring the effects of intervention as well as for studying breast cancer etiology.[6,11–13]

Different methods have been used for the evaluation of

mammographic breast density. Earlier studies used a subjective visual assessment of the breast parenchyma primarily based on the four patterns described by Wolfe[2] (N1 is comprised entirely of fat; P1 has up to 25% nodular densities; P2 has over 25% nodular mammographic densities; DY contains extensive regions of homogeneous mammographic densities). The subjectivity in classifying the mammographic patterns introduced large variability in the risk estimation. Later studies used more quantitative estimates, such as planimetry, to measure the dense area in the breast manually outlined by radiologists on mammograms.[3,7] These studies indicate that the percentage (%) of mammographic densities relative to the breast area can predict the breast cancer risk more accurately than a qualitative assessment of mammographic patterns. Warner *et al.*[15] conducted a meta-analysis of the studies published between 1976 and 1990 to investigate the effect of different methods of classification on estimates of cancer risk. They found that the mammographic parenchymal pattern does correlate with the breast cancer risk. The magnitude of the risk varies according to the method used to evaluate the mammograms. With the quantitative estimates of mammographic density, the difference in risk between the highest and the lowest risk category is substantial and is greater than the risks associated with most other risk factors for breast cancer. More recent studies used fractal texture and the shape of the gray level histogram[14] to quantify the parenchymal pattern or used interactive thresholding on digitized mammograms to segment the dense area.[11,15] It was reported that the thresholding method provided a higher risk value than the texture measure or the histogram shape.[16] Other researchers have attempted to calculate a breast density index to model the radiologists' perception.[17]

In clinical practice, radiologists routinely estimate the breast density on mammograms by using the BI-RADS lexicon as recommended by the American College of Radiology[18] in order to provide a reference for mammographic sensitivity. Because of the lack of a quantitative method for breast density estimation, researchers often use the BI-RADS rating for monitoring responses to preventive or interventional treatment and the associated changes in breast cancer risk.[19] We have found that there is a large interobserver variability in the BI-RADS ratings among experienced mammographers.[20,21] An automated and quantitative estimation, as investigated in this study, will provide not only an efficient means to measure mammographic density, but also a reproducible estimate that will reduce the inter- and intraobserver variability of mammographic density measurements. This image analysis tool will therefore allow researchers to study more definitively the relationship of mammographic density to breast cancer risk, detection, prognosis, and mammographic sensitivity, and to better monitor the response of a patient to preventive or interventional treatment of breast cancers.

In this paper, we will describe the image processing techniques used in our automated breast density segmentation algorithm. The performance of the computer segmentation was evaluated by a comparison with the average segmentation by 5 radiologists using interactive thresholding in the same data set.

II. MATERIALS AND METHODS

A. Database

A data set consisting of 260 mammograms of 65 patients was used for the development of the histogram analysis method in this study. Each case contains the craniocaudal (CC) view and the mediolateral oblique (MLO) view of both breasts of the patient. The first 50 mammograms were consecutive screening cases from the patient files in the Radiology Department at the University of Michigan. After data analysis, it was found that there were very few dense breasts in the initial data set. An additional 15 cases visually judged by radiologists to be dense breasts were then randomly selected and mixed with the initial set. The images were processed individually without knowing their BI-RADS categories. The mammograms were acquired with mammography systems approved by the Mammography Quality Standards Act (MQSA) and were digitized with a LUMISYS 85 laser film scanner with a pixel size of $50\ \mu\text{m} \times 50\ \mu\text{m}$ and 4096 gray levels. The gray levels are linearly proportional to optical densities (O.D.) from 0.1 to greater than 3 O.D. units. The nominal O.D. range of the scanner is 0–4 with large pixel values in the digitized mammograms corresponding to low O.D. The full resolution mammograms were first smoothed with a 16×16 box filter and subsampled by a factor of 16, resulting in $800\ \mu\text{m} \times 800\ \mu\text{m}$ images of approximately 225×300 pixels in size for small films and 300×375 pixels for large films.

B. Breast segmentation and image enhancement

The breast image is first segmented from the surrounding image background by boundary detection. The detected boundary separated the breast from other background features such as the directly exposed area, patient identification information, and lead markers. The density analysis was performed only within the breast region. An automated breast boundary tracking technique developed previously[22,23] was modified to improve its performance. Briefly, the technique used a gradient-based method to search for the breast boundary. The background of the image was estimated initially by searching for the largest background peak from the gray level histogram of the image. After subtracting this background level from the breast region, a simple edge was found by a line-by-line gradient analysis from the top to the bottom of the image. The criterion used in detecting the edge points was the steepness of the gradient of four adjacent pixels along the horizontal direction. The steeper the gradient, the greater the likelihood that an edge existed at that corresponding image point. The simple edge served as a starting point for a more accurate tracking algorithm that followed. The tracking of the breast boundary started from approximately the middle of the breast image and moved upward and downward along the boundary. The direction to search for a new edge point was guided by the previous edge points. The edge
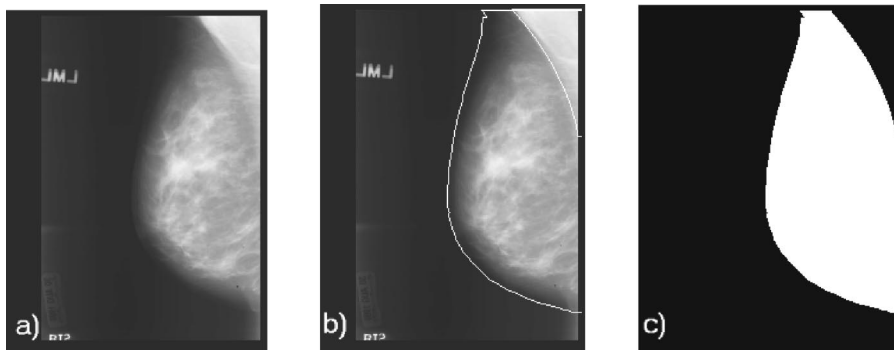
FIG. 1. (a) A mammogram from our image database; (b) the image superimposed with the detected breast boundary and pectoral muscle boundary; (c) the binary map of the segmented breast region.

location was again determined by searching for the maximum gradient along the gray level profile normal to the tracking direction. Since the boundary tracking was guided by the simple edge and the previously detected edge points, it could steer around the breast boundary and was less prone to diversion by noise and artifacts. The accuracy of the boundary tracking technique was evaluated in our previous study[23] by quantifying the root-mean-square differences between the detected and manually identified breast boundaries. In the current study, the performance of the boundary tracking technique for this data set was determined by superimposing the detected boundary on the breast image and visually judged if the detected boundary coincided with the perceived breast boundary. The breast image and its boundary were displayed by appropriately adjusting the contrast and brightness. Incomplete, jagged and mistracked boundaries were considered incorrect tracking.

The unexposed film area around the film edges was detected automatically. After the breast boundary was found, a region growing algorithm was used to fill the enclosed breast region. The result was a binary map that distinguished the breast region from the background areas. An example of the tracked breast boundary and the breast binary map is shown in Figs. 1(a)–1(c).

For the MLO view mammograms, an additional step has to be performed for segmentation of the pectoral muscle. The initial edge in the pectoral region was found as the maximum gradient point by a line-by-line gradient analysis from the chest wall to the breast boundary. The false pectoral muscle edge points were discarded by an edge validation process. First, a straight line was fitted to the initial edge points, and the points that did not lie close to the fitted line were removed. Second, the remaining edge points that were connected were identified by an 8-connectivity criterion. An edge segment was removed if its direction was inconsistent with the pectoral edge direction relative to the breast image. Finally, a second order curve was fitted to the remaining edge points to separate the pectoral muscle from the breast region. The pixels in the pectoral muscle region were excluded from the histogram analysis and breast area calculation. The accuracy of the pectoral muscle detection was also judged visually in this study, similar to the method used for the breast boundary described above. Figure 1 shows the pectoral muscle trimming result for an MLO view mammogram.

To facilitate histogram analysis, a dynamic range compression method was developed to reduce the gray level range of the histograms. With our digitization, the gray levels of the dense tissue are higher than those of the adipose tissue. Because of variations in exposure condition and breast thickness near the periphery, the gray level distribution corresponding to the breast parenchymal pattern is superimposed on a low frequency background that mainly represents the global variations in exposure. This low frequency background distorts the characteristic features of the histogram due to the density pattern. To reduce the distortion, an adaptive dynamic range compression technique was applied to the breast image. For a given breast image, $F(x,y)$, which contains low frequency background and higher frequency breast tissue structures, a smoothed image, $F_B(x,y)$, was obtained by applying a large-scale box filter to $F(x,y)$ to remove the high frequency components while retaining the low frequency components. The image $F_B(x,y)$ was then compressed by a scale factor k:

$$F_C(x,y) = k F_B(x,y). \tag{1}$$

To reconstruct the high frequency components, $F_C(x,y)$, was subtracted from a constant gray level G, and added to the original image, $F(x,y)$:

$$F_D(x,y) = G - F_C(x,y), \tag{2}$$

$$F_E(x,y) = F_D(x,y) + F(x,y). \tag{3}$$

Histogram analysis was applied to the dynamic-range-compressed image $F_E(x,y)$. Figure 2 shows an example of the resulting images and gray level histograms obtained from this procedure, where the size of box filter is 35×35, the scale factor k is 0.5, and the constant gray level G is the maximum gray level of the compressed image $F_C(x,y)$. The values of these parameters were chosen experimentally as a balance between reducing the dynamic range and preserving the image features in the compressed image.

C. Breast density segmentation and estimation

A rule-based threshold technique was developed to segment the dense areas from the breast background. The histogram of the breast region on the dynamic-range-compressed mammogram was generated and smoothed. The histograms of these images in the database were analyzed to formulate an automatic thresholding routine. The histograms were
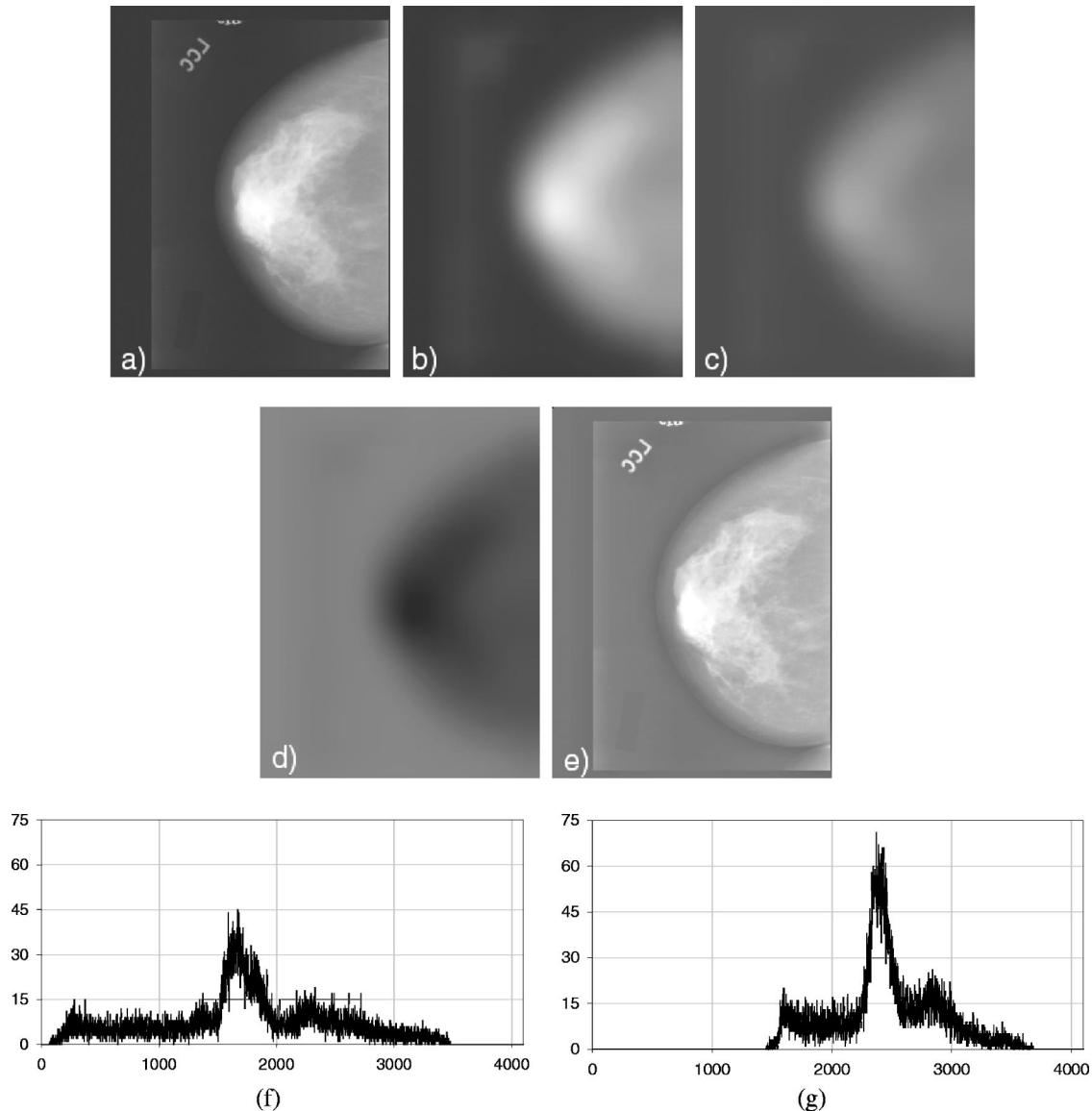
FIG. 2. (a) A typical mammogram from our image database; (b) the low frequency image $F_B(x,y)$ obtained by an 35×35 box filter; (c) the compressed image $F_C(x,y)$; (d) the inverted image $F_D(x,y)$; (e) the enhanced image $F_E(x,y)$; (f) the gray level histogram within the breast region of the original image $F(x,y)$; and (g) the gray level histogram of the breast region of the enhanced image $F_E(x,y)$.

grouped into four classes based on the characteristic shapes of their histograms. It was observed that the grouping corresponded approximately to the four BI-RADS breast density ratings: Class I corresponded to breasts of almost entirely fat, Class II corresponded to scattered fibroglandular densities, Class III corresponded to heterogeneously dense and Class IV corresponded to extremely dense breasts. Examples of typical histograms for these four classes are shown in Fig. 3. The histograms seemed to follow two basic patterns. In one pattern, there was only one dominant peak, which represented most of the breast structures in the breast region. In the other pattern, in addition to a large peak in the histogram, there was one or two smaller peaks on the right or left side of the large peak. In a majority of the cases, the smaller peak was distinguishable from the large one when the random fluctuation on the histogram was smoothed.

1. Peak detection and feature description

The gray level histogram within the breast area was generated and normalized, and passed through an averaging window to smooth out the random fluctuations. We estimated the window size to be in the range of 30 to 50 gray levels by experimentally evaluating the histogram shapes and density segmentation at different window sizes. Too small a window size cannot smooth out the fluctuation and too large a window size will blur the useful features. A window size of 30 was used in this study. The second derivative of every point on the histogram curve was computed. An example of the histogram and its second derivative curve are shown in Fig. 4. The zero crossing locations were detected by scanning for the positive-to-negative and negative-to-positive changes on the latter curve. If the second derivative was negative be-
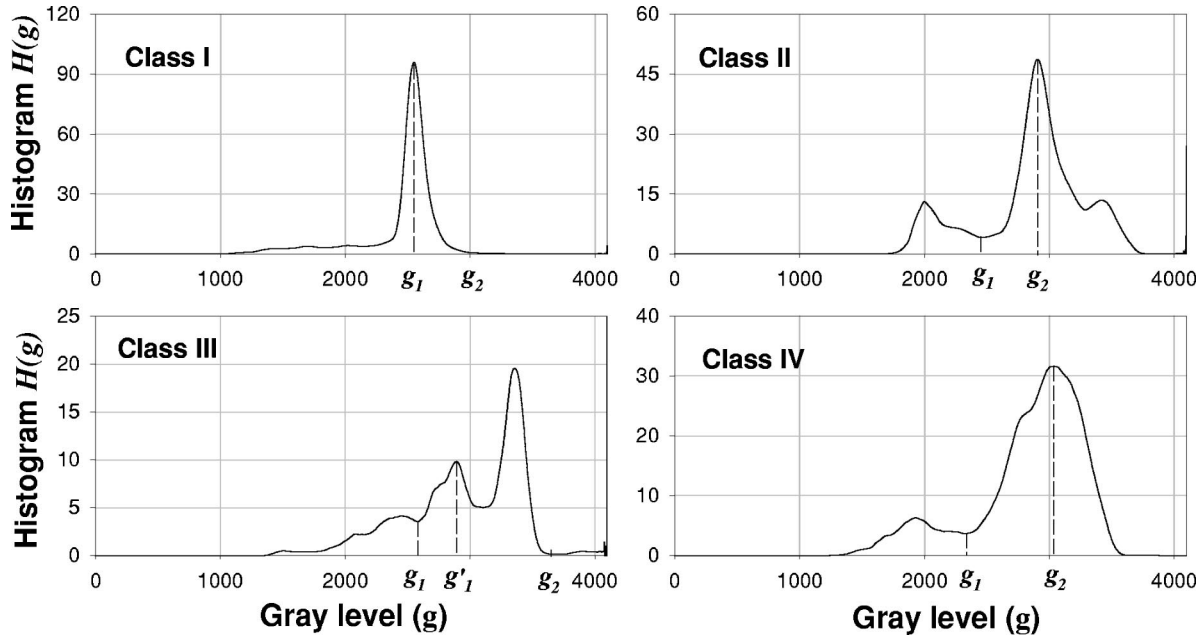
FIG. 3. Four typical classes of histograms and the setting of gray level interval $[g_1, g_2]$ for the threshold calculation.

tween two zero crossing points, it indicated that a peak existed between these two points on the histogram. Normally, as shown in Fig. 4, a peak included the peak point P_0 and two valley points P_1 and P_2 located on the two sides of the peak point. The peak point P_0 was determined by searching for the maximum histogram value between the zero crossing points Z_2 and Z_3, and the P_1 and P_2 points were obtained by searching for the point with minimum histogram value between zero crossing points Z_1, Z_2 and Z_3, Z_4, respectively.

The following peak features can be defined by peak point P_0 and valley points P_1 and P_2:

Energy: $\quad E = \dfrac{1}{A} \sum_{i=P_1}^{P_2} f(i)*f(i),$ (4)

left-side energy: $\quad E_L = \dfrac{1}{A} \sum_{i=P_1}^{P_0} f(i)*f(i),$ (5)

right-side energy: $\quad E_R = \dfrac{1}{A} \sum_{i=P_0}^{P_2} f(i)*f(i),$ (6)

likelihood: $\quad L = E/E',$ (7)

where $f(\cdot)$ is the histogram, A is the total energy of the entire histogram and $A = \sum_{i=0}^{N} f(i)*f(i), N$ is the maximum gray level of the histogram. E' is the energy calculated by approximating the histogram in the interval $[P_1, P_2]$ using two straight lines, $P_1 P_0$ and $P_0 P_2$. The energy E of the peak is used to compare the sizes of the peaks on the histogram, higher energy means bigger size of the peak. E_L and E_R split the energy E into two parts from the peak point for calculating the ratio of the energy in these two parts. The likelihood L describes how close the real peak is to the triangle represented by the three points P_0, P_1 and P_2.

2. Rule-based histogram classification

A rule-based histogram classifier was developed to classify the gray level histogram of the breast area into four classes. As shown in Fig. 3, a typical Class I breast is almost entirely fat, it has a single narrow peak on the histogram. Class II has scattered fibroglandular densities, it has two peaks, other than the tail part on the left, on the histogram, with the smaller peak on the right of the bigger one. Class III is heterogeneously dense, it also has two peaks, but the smaller peak is on the left of the bigger one. Class IV is extremely dense, which has a single dominant peak on the histogram, but it is wider compared with the peak in the Class I histogram, and a second small peak sometimes occurs to the left of the main peak.

The classification is performed in two steps. In the first step, the computer determines whether there is only one single peak in the histogram. The biggest peak (main peak)
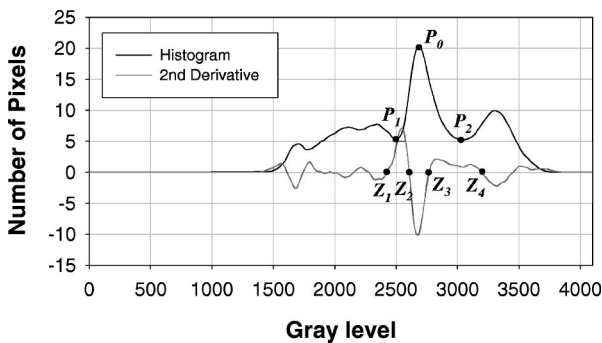
FIG. 4. The gray level histogram (solid curve) and the second derivative (dot) curve. P_0 is the peak point, P_1 and P_2 are the valley points of the peak on the two sides of the peak point P_0. Points Z_1, Z_2, Z_3 and Z_4 are zero crossing points on the second derivative curve, which are used for searching the points P_0, P_1 and P_2.

P_M and its location are detected by comparing the energy of the peaks on the histogram. The single peak feature is mainly determined by the energy E under the main peak and the features E_L and E_R. If the histogram is found to have a single-peak pattern, in general, a narrow peak corresponds to very fatty breast (Class I), and a wider peak corresponds to very dense breast (Class IV). However, in some cases, the histogram of these two classes is very similar, as discussed below (Fig. 9), and it is difficult to distinguish them by their gray level histogram distributions. Two additional image features were analyzed to classify very fatty and very dense breasts. One feature is the gray level standard deviation (Std) in the entire breast area, defined as

$$\text{Std} = \left(\frac{1}{N} \sum_{x \in \text{MAP}} \sum_{y \in \text{MAP}} (f(x,y) - \bar{f}(x,y))^2 \right)^{1/2}, \quad (8)$$

where MAP is the breast binary map region, N is the pixel numbers within MAP. Another feature is the number of single pixels and single pixel-size holes (NSH) counted in the breast area of a segmented binary image using the biggest histogram peak point P_M as a threshold. For a very fatty mammogram, the breast mainly consists of a fatty background with some fibrous structures and fibroglandular tissue scattered in the breast area. The NSH value was found to be larger (greater than 50 pixels on average), and Std smaller (less than 500 on average), compared with a mammogram of a very dense breast.

In the second step, if the histogram is found to have more than one peak, decision rules are used to decide if the second major peak is on the left side or on the right side of P_M by the features E, E_L, E_R and L, and the relative position of the two peaks. If the second major peak is on the right, then the histogram is classified to be Class II; otherwise, it is classified to be Class III.

3. Gray level thresholding

Gray level thresholding is essentially a pixel classification problem. Its objective is to classify the pixels of a given image into two classes: one includes pixels with gray values that are below or equal to a certain threshold; the other includes those with gray values above the threshold. Thresholding is a popular tool for image segmentation, a variety of techniques have been proposed over the years. In our study, two threshold selection methods are used: one is the Discriminant Analysis (DA) method[24] and the other is the Maximum Entropy Principle (MEP) based method.[25] The DA method assumes that the image gray levels can be classified into two classes by a threshold. To estimate the threshold, a discriminant criterion based on the within-class variance and between-class variance is introduced. An optimal threshold is selected by the discriminant criterion to maximize the separability of the resultant classes in terms of gray levels. This method is well-suited for the cases where the gray level histogram is bimodal. In an ideal situation, the histogram has a deep and sharp valley between the two peaks representing objects and background, respectively, and the optimum corresponds to the gray level at the bottom of this valley. A

more detailed description of the DA method can be found in Appendix A.

For the MEP method, the optimal threshold value is determined by maximizing the *a posteriori* entropy subject to certain inequality constraints that are derived by means of special measures characterizing the uniformity and the shape of the regions in the image. As is well-known,[26] the maximum *a posteriori* probability can serve as a criterion to select *a priori* probability distributions when very little is known about the probability distribution. Compared with the DA method, MEP can provide a better thresholding result if the gray level histogram does not have a bimodal distribution. A more detailed description of the MEP method can be found in Appendix B.

The gray level histograms of the mammograms in our study are very complex, the histogram may be unimodal, bimodal or multi-modal. It is difficult to select an appropriate threshold by one general threshold selection method. Therefore, we combined both the DA and the MEP methods, to select a threshold according to the characteristic features of the histogram that has been classified into one of the four classes. Suppose $f(g)$ is the gray level histogram of the breast area. Let $T = \text{Method}(f(g)|g_1 < g < g_2)$ represent the threshold, T, that is selected by use of Method in the interval $[g_1, g_2]$ of the histogram $f(g)$, where Method can be either the DA or MEP method. The settings of the interval $[g_1, g_2]$ for the four classes are discussed below and shown in Fig. 3.

Class I: The histogram is unimodal so that the threshold is selected as

$$T = \text{MEP}(f(g)|g_1 < g < g_2),$$

where, g_1 is the main peak point; g_2 is the valley point on the right side of main peak.

Class II: The histogram is not unimodal and the histogram is classified as Class II; the threshold is selected by averaging two thresholds that are computed in two different intervals of the histogram by the DA method:

$$T_1 = \text{DA}(f(g)|g > g_1),$$

$$T_2 = \text{DA}(f(g)|g > g_2),$$

$$T = (T_1 + T_2)/2,$$

where g_1 is the valley on the left of the main peak; g_2 is the main peak point.

Class III: The histogram is not unimodal; there are two possibilities in the histogram distribution: there is a valley between the main peak and its left side peak, as shown in Fig. 3, or no obvious valley exists between the main peak and its left side peak. In two different intervals of the histogram, two thresholds are computed as

$$T_1 = \text{DA}(f(g)|g_1 < g < g_2),$$

$$T_2 = \text{DA}(f(g)|g_1' < g < g_2),$$

where g_1 is the left valley point of the left-side peak (P_{LM}) of the main peak, g_1' is the peak point of P_{LM} and g_2 is right valley point of the main peak. If there is an obvious valley, $T = (T_1 + T_2)/2$, otherwise $T = T_1$.

Class IV: Since the histogram is considered unimodal, the threshold is computed by the MEP method, $T = \text{MEP}(f(g)|g_1 < g < g_2)$, where, g_1 is the left valley point of the main peak; g_2 is the main peak point.

D. Radiologists' segmentation of dense breast tissue

In order to evaluate the accuracy of the computer segmentation method, the computer segmentation results were compared to radiologists' manual segmentation in the data set of 65 patient cases. Details of the observer study for estimation of the breast density and statistical analysis of the results were discussed elsewhere.[27] Briefly, a graphical interface was developed for displaying the mammograms and recording the observer's evaluation. The CC-view and MLO-view mammograms for a given breast were displayed side-by-side; a radiologist observer examined the mammograms and gave a BI-RADS rating and a visual estimation of the percent breast density with 10% increments. After the subjective evaluation, each view was displayed sequentially, together with the histogram of the dynamic-range-compressed image. The radiologist would interactively choose a threshold by moving a slider along the abscissas of the histogram plot. The segmented binary image, displayed side-by-side with the mammogram, would change instantaneously when the threshold was changed. The radiologist could inspect if the segmented area corresponded to the dense area on the mammogram. Once the radiologist was satisfied with the segmentation of the dense area, the gray level threshold and the percent dense area derived from this threshold were recorded. The display then moved to the next view of the same breast for evaluation. The mammograms of the other breast for the same patient would then be displayed and evaluated in the same way. The entire process was repeated for each patient until all patients in the data set were evaluated.

Five MQSA-approved radiologists participated in the experiment. To familiarize the radiologists with the procedures and to assist them in their visual estimation of the percent breast density, we had them trained on a separate set of 25 patient cases prior to the evaluation of the actual data set. During the training session, the computer displayed the percent breast dense area to the radiologist, which was obtained by the radiologist's interactive thresholding of the image. The radiologist could then compare the manually segmented percentage with their visually assessed percent density for the image. This feedback helped ''calibrate'' the radiologists' visual estimates of the percent dense breast area. The percent dense area obtained by interactive thresholding was not displayed during the actual study.

III. RESULTS

An example of a typical mammogram from each of the four classes and its corresponding enhanced image, its histogram, the selected threshold and the segmented image are shown in Figs. 5(a)–5(d), respectively.

The average percent breast density obtained from manual segmentation by the five trained radiologists for each mammogram was used as the ''true standard'' of the percent breast density for that mammogram. The breast region was segmented by the breast boundary tracking technique, and the pectoral muscle was trimmed for the MLO-view mammograms. The breast boundary was accurately tracked on 92.3% (240/260) of the mammograms, and the pectoral muscle was correctly trimmed on 74.6% (97/130) of the MLO views. The histograms of 6% (8 CC views and 8 MLO views) of the breast regions did not exhibit the typical characteristic features of the four classes and were misclassified by the computer, resulting in poor segmentation of the dense region.

Figure 6 shows a comparison of the percent breast density visually estimated by radiologists against the true standard for the 94% of the 260 mammograms that were classified correctly by the computer. Table I summarizes the comparison of the radiologists' visual estimates with the true standard. The ''difference'' between the estimated % breast density and the true standard was calculated for each case, and the mean and the standard deviation of this difference over all cases were estimated for each radiologist and shown in the table. Therefore, the mean difference was the average bias of the estimated % breast density from the true standard over all images in the data set. It can be seen that almost all radiologists had a positive bias, on average, when they visually estimated mammographic density, except for Radiologist 5 who had a small negative average bias on the CC-view reading. For a given radiologist, the over-estimation increased as the breast density increased. Although the correlation coefficients were high, ranging from 0.90 to 0.95, the deviations from the diagonal line were systematic. The average bias from the true standard varied from less than 1% to 11%, depending on the radiologist. The root-mean-square (RMS) errors of the five radiologists relative to the true standard ranged from 7.5% to 16.3%.

Figure 7 shows the comparison of the percent breast density between the computer segmentation and the true standard for the 94% of mammograms whose histograms were considered to be correctly classified. There was a trend of over-estimation in the very fatty breasts. In the medium dense range, the variances from the true standard were high. Some images had a large deviation from the diagonal line, indicating that the threshold was incorrectly determined. Table II summarizes the comparison between the computer performance and the true standard. For the CC views with correct histogram classification, the correlation between the computer-estimated percent dense area and the true percent breast density was 0.94, and between the computer and the radiologists' average visual estimate was 0.87 (not plotted). These correlation coefficients were 0.91 and 0.82, respectively, for the MLO views with correct classification. Although the correlation coefficients of the computer segmentation with the true standard were not better than those of the visual estimates, the average biases of the computer segmentation from the true standard were less than 2%, which were substantially less than those of visual estimates (Table I). This indicates that computerized segmentation is a good alternative to manual segmentation although variances of the automated method will need to be further reduced. The RMS
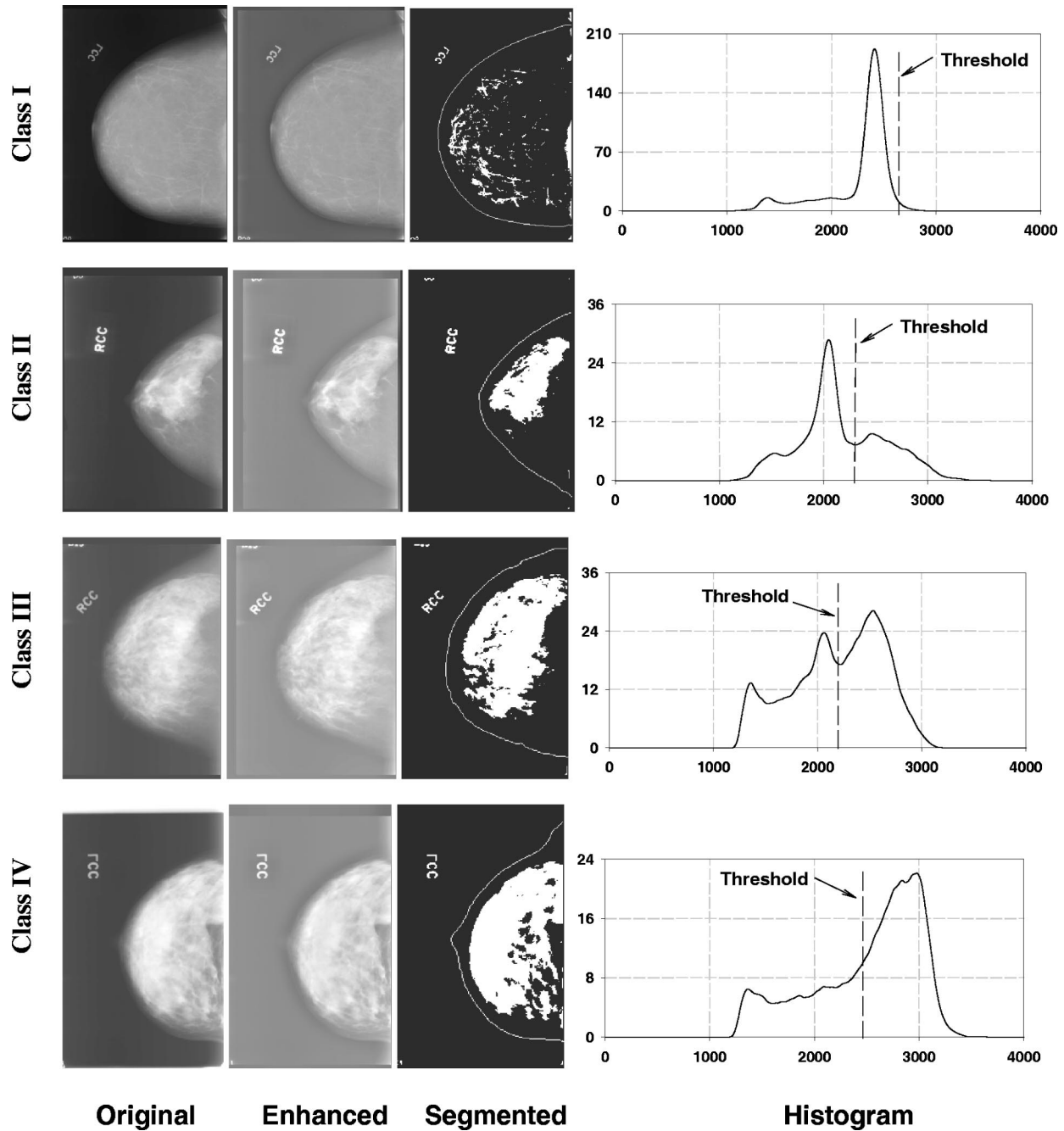
F IG. 5. Four classes of typical mammograms and corresponding enhanced and segmented image, histogram and threshold.

errors of the computer segmentation were also less than those of the radiologists' visual estimates, at 6.1% and 7.2%, respectively, for the CC view and MLO view, when the histograms were correctly classified. The biases and RMS errors for the different subsets of images are also shown in Table II. It can be seen that correct histogram classification was the most important factor in reducing the biases and the RMS errors. The contributions by breast boundary detection and pectoral muscle segmentation were minor, on average, for improving the estimation of the percent dense breast area.

Figure 8 shows the comparison of the individual radiologists' manual segmentation against the true standard. For CC views, the RMS difference in the percent breast density be-

tween an individual radiologist's manual segmentation and the true standard varied from 2.9% to 5.9% among the five radiologists. For MLO views, the RMS difference varied from 2.8% to 6.2%. The average biases of the five radiologists ranged from −2.8% to 2.2% for the CC views and from −3.1% to 3.0% for the MLO views. The maximum biases of the five radiologists varied from 4.4% to 22.6% for the CC views and from 5.2% to 23% for the MLO views.

The five radiologists provided BI-RADS density ratings for each breast. Although the BI-RADS ratings exhibited large inter-observer variations,[20] it is interesting to compare the computer's histogram classification with the BI-RADS ratings. Since there were 260 images, each with 5 radiolo-
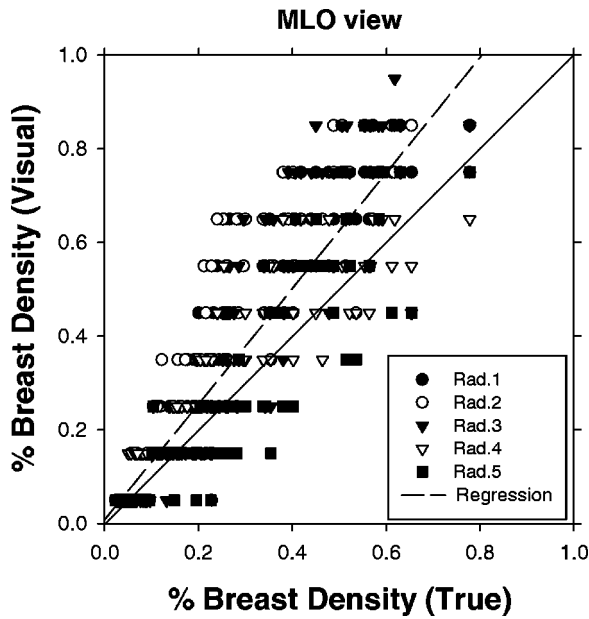
MLO view

FIG. 6. A comparison of the percent breast density between five radiologists' visual estimates and the true standard. The dashed line represents the linear regression of all data points on the plot. The MLO view is shown. The trend for the CC view is similar.

gists' ratings, there were a total of 1300 rating comparisons. The comparison of the computer and the radiologists' BI-RADS ratings is shown in Table III. It was found that 87.4% of Class I classification have BI-RADS ratings 1 or 2, 92.0% of Class II classifications have density ratings 2 or 3, 83.4% of Class III classifications have density ratings 3 or 4

and 57.1% of Class IV classifications have density rating 4. More detailed analysis of the variability of radiologists' BI-RADS ratings was discussed by Martin *et al.*[21]

IV. DISCUSSION

Radiologists routinely estimate mammographic breast density using the four BI-RADS categories. In studies that require breast density estimation, radiologists' visual estimates of mammographic density were often used as the density measure. Our observer study indicates that interobserver variation between the BI-RADS ratings of five experienced radiologists ranged from -1 to $+1$. The subjectively estimated percent dense area can deviate from the true standard by as much as 40%, as shown in Fig. 6. These results indicate the need to develop an objective method for the estimation of mammographic breast density in order to improve the accuracy and reproducibility of the estimation. A computerized image analysis method for mammographic breast density estimation will be a useful tool for study of breast cancer risk factors and for monitoring the change of breast cancer risk with preventive or interventional treatments.

In this study, we used the average of the percent breast area obtained with interactive thresholding by five experienced radiologists as the true standard. The gray level thresholding method used in this study could achieve a reasonable segmentation of the dense areas on the mammogram because the image was preprocessed with dynamic range compression. The image-based analysis of breast density will not provide the actual percentage of fibroglandular tissue in the breast volume. However, the previous studies that estab-

TABLE I. A comparison of the radiologists' visual estimate of mammographic breast density with the true standard. The ''difference'' was defined as the difference between the estimated % breast density and the true standard for each case, and the mean and the standard deviation of this difference are tabulated.

| Image subsets	No. of images	Radiologist	Correlation	RMS error	Mean difference	Std. dev. of difference
CC view:						
All	130	Rad. 1	0.942	13.3%	6.9%	11.5%
		Rad. 2	0.931	14.5%	9.8%	10.7%
		Rad. 3	0.923	13.3%	6.3%	11.8%
		Rad. 4	0.934	7.5%	2.9%	7.0%
		Rad. 5	0.901	9.6%	−1.4%	9.6%
Histogram	122	Rad. 1	0.946	13.7%	7.2%	11.3%
correctly		Rad. 2	0.936	14.7%	10.3%	10.8%
classified		Rad. 3	0.929	14.2%	6.7%	11.6%
		Rad. 4	0.929	7.7%	3.1%	7.1%
		Rad. 5	0.900	9.7%	−1.3%	9.4%
MLO view:						
All	130	Rad. 1	0.933	14.5%	8.3%	12.0%
		Rad. 2	0.914	16.1%	11.2%	11.5%
		Rad. 3	0.915	14.4%	7.7%	12.2%
		Rad. 4	0.919	8.8%	4.3%	7.7%
		Rad. 5	0.910	9.2%	0.1%	9.2%
Histogram	122	Rad. 1	0.932	15.0%	8.3%	12.0%
correctly		Rad. 2	0.914	16.3%	10.9%	11.4%
classified		Rad. 3	0.919	14.7%	7.8%	12.2%
		Rad. 4	0.916	9.0%	4.3%	7.7%
		Rad. 5	0.909	9.4%	0.3%	9.2%
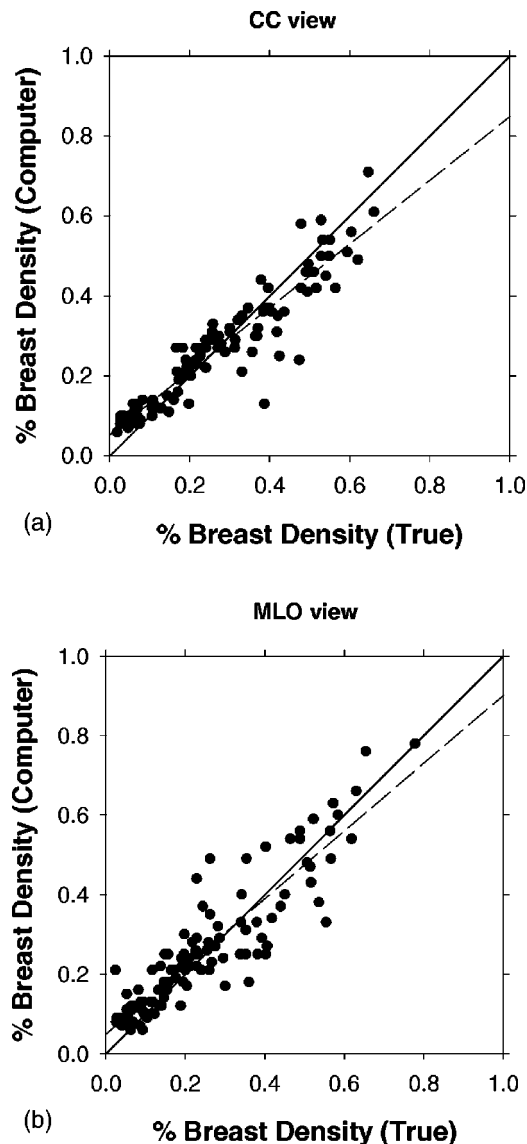
FIG. 7. A comparison of the percent breast density between the computer segmentation and the true standard. The dashed line represents the linear regression of the data on the plot. (a) CC view, (b) MLO view.

lished the correlation between breast density and breast cancer risk were all based on mammographic density. This indicated that mammographic density is a sufficiently sensitive marker for breast cancer risk, although it may be less accurate than volumetric density. An actual measurement of the percentage of fibroglandular tissue volume in the breast, for example, by x-ray penetration with correction for scatter and beam hardening, is difficult because it requires accurate x-ray sensitometry or phantom calibration for each image. These requirements will limit its use to a few laboratories that have specialized equipment and expert physicists. Magnetic resonance breast imaging can also provide volume measurement of dense tissue but it is expensive and not easily accessible. It can be expected that the estimation of mammographic breast density by a computerized image analysis method will be a more practical and viable approach, espe-

cially when direct digital mammography becomes more widely used in the future.

Our preliminary study indicates that breast density estimation can be performed automatically and accurately (Fig. 7). Although the accuracy of our current algorithm still needs to be improved, it can be seen that the computer segmentation can provide an estimate of the percent breast density with a very small bias (Table II). More importantly, computer segmentation will be more reproducible and consistent than visual estimates. This will improve the sensitivity of studies that depend on evaluation of the change in mammographic density over time or before and after a certain treatment.

In this study, we reduced the spatial resolution to a pixel size of $800 \, \mu\text{m} \times 800 \, \mu\text{m}$ for image processing. The small matrix size of the reduced images improves the computational efficiency. The reduction in resolution has two major effects: reducing the image noise and blurring the details. Since the significant dense tissue in the breast that contributes to the parenchyma is relatively large compared to 800 μm, it is not expected that processing at this pixel size will have a strong effect on the accuracy of the estimated percent breast density. Differences in the segmented area may occur mainly along the boundary of the dense tissue region, but the effect may be averaged out statistically along boundaries of reasonable lengths. The residual errors in the estimation of the dense area should not be substantial in comparison with the inter- and intra-radiologists' variations in their manual segmentation.

Successful segmentation of dense tissue depends strongly on whether a mammogram can be classified correctly into a proper class. A successful classification will likely result in the selection of a near optimal threshold. Conversely, if a mammogram is classified into a wrong class, the threshold will be selected incorrectly. For the mammograms of very fatty breasts, the gray level histogram has the characteristics of Class I, which contains one large single peak. These histograms can be distinguished relatively easily from most of the other classes of histograms if those histograms exhibit the typical features. For mammograms of BI-RADS category 2 or 3, there are scattered fibroglandular or heterogeneous densities in the breast. A small peak may be located on the left or on the right, or on both sides of the main peak on the histogram. The histogram could be classified into Class I if the small peak is not large enough and is not detected as a second peak. Otherwise, it would be classified into Class II or Class III, depending on the location of that small peak relative to the main peak of the histogram. For the two-peak pattern histogram, the DA threshold selection method is robust if there is an obvious valley between the two peaks. If the valley is flat or not obvious, averaging the two thresholds obtained by the DA method in two different intervals, as designed for this study, can reduce the chance of calculating an incorrect threshold that differs greatly from the optimum, but it also reduces the chance of finding the optimal threshold. Overall, the rules designed for classification of the two-peak patterns seem to perform consistently well for this data set. One of the difficult situations is to distinguish between

TABLE II. A comparison of computer segmentation with the true standard. The ''difference'' was defined as the difference between the estimated % breast density and the true standard for each case, and the mean and the standard deviation of this difference are tabulated.

Image subsets	No. of images	Correlation	RMS error	Mean difference	Std. dev. of difference
CC view:					
All	130	0.746	12.3%	1.3%	12.3%
Boundary correctly tracked	120	0.780	11.4%	1.4%	11.4%
Histogram correctly classified	122	0.943	6.1%	0.2%	6.2%
Boundary and histogram correctly done	113	0.953	5.6%	0.8%	5.6%
MLO view:					
All	130	0.780	11.6%	1.9%	11.5%
Boundary correctly tracked	120	0.766	11.9%	2.1%	11.7%
Histogram correctly classified	122	0.914	7.2%	1.5%	7.1%
Pectoral muscle correctly trimmed	97	0.733	11.6%	1.6%	11.6%
Boundary and histogram correctly done	112	0.912	7.2%	1.7%	7.1%
Boundary, histogram and pectoral muscle correctly done	83	0.891	7.1%	1.9%	6.8%

Class I and Class IV, when the histogram of a very dense breast mimics that of a very fatty breast, as shown in Fig. 9. This image was correctly classified with the additional features, Std and NSH. However, there were other cases that failed in spite of the additional criteria. The large difference in the optimal threshold locations between these two classes will lead to a large error in the estimated percent breast density if the histogram is misclassified. Further study is needed to more accurately distinguish these two classes.

The dynamic range reduction technique reduces the variability of the gray level histograms and enhances their characteristics. This pre-processing facilitates the classification of the image into the correct class. There are many image smoothing techniques published in the literature. Low-pass filtering with a box filter is the simplest choice. The effec-tiveness of background correction with a box filtered image depends on the box size. We found that a 35×35-pixel filter is a good balance between computation time and the capability to remove the high frequency components. The subtraction of the low-pass filtered image from the original image is a form of unsharp masking. The breast boundary is generally enhanced as shown in Fig. 2(e). The pixels at the enhanced breast boundary contribute a small peak to the left tail of the gray level histogram of the breast area. Moreover, if dense tissue is present close to the breast boundary, it may not be segmented correctly due to intensity reduction. Other low frequency estimation techniques such as wavelet decomposition will be investigated in future studies.

In this feasibility study, we used a small data set of mammograms to develop a rule-based classifier for the histogram analysis. Although a large fraction of the histograms manifest characteristic features that can be grouped into four classes, corresponding approximately to the four BI-RADS breast density ratings, there are many exceptions. One such example is shown in Fig. 9. This causes misclassification and incorrect thresholding by the histogram classifier. It will be
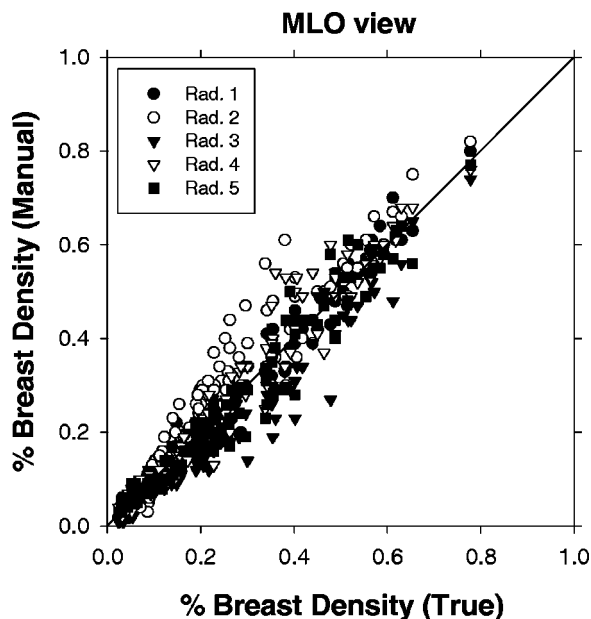

FIG. 8. A comparison of the percent breast density obtained from the five radiologists' manual segmentation with their average for the same mammograms. The MLO view is shown. The trend for the CC view is similar.

TABLE III. A comparison of computer classification and radiologists' BI-RADS breast density ratings.

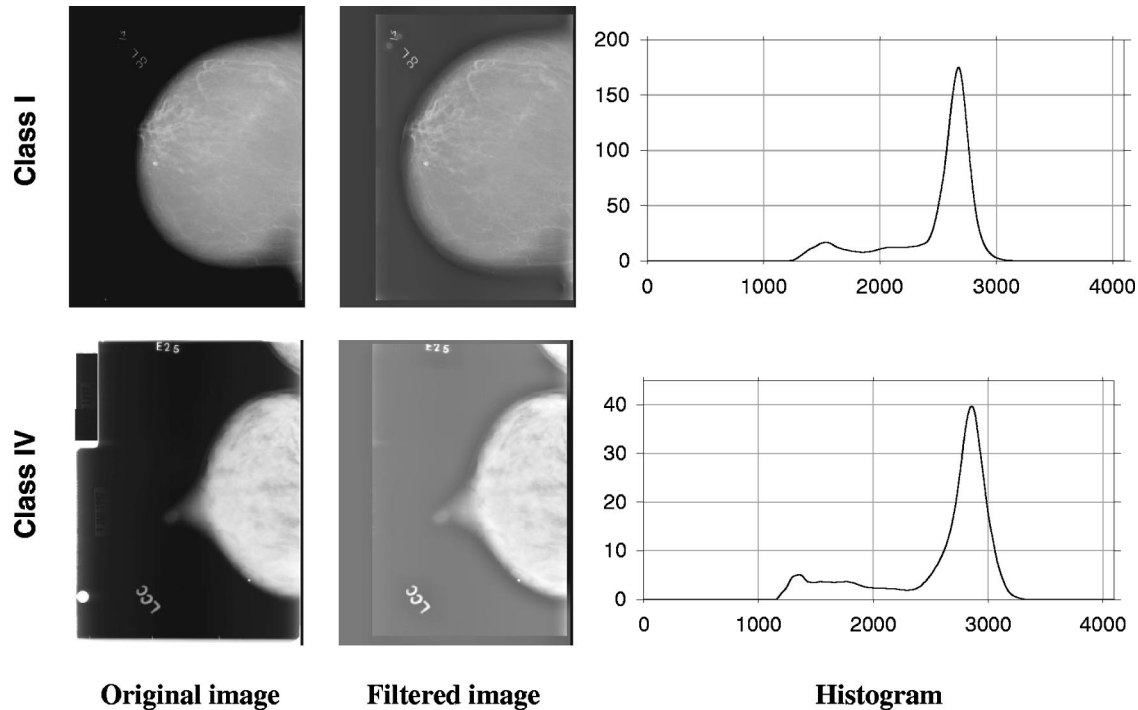| Computer classification	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	Total
Class I	210 (16.2%)	262 (20.2%)	52 (4%)	16 (1.2%)	540 (41.5%)
Class II	0 (0%)	92 (7.1%)	184 (14.2%)	24 (1.8%)	300 (23.1%)
Class III	1 (0.1%)	52 (4%)	167 (12.8%)	100 (7.7%)	320 (24.6%)
Class IV	5 (0.4%)	12 (0.9%)	43 (3.3%)	80 (6.2%)	140 (10.8%)
Total	216 (16.6%)	418 (32.2%)	446 (34.3%)	220 (16.9%)	1300 (100%)

FIG. 9. The gray level histograms of two mammograms classified by radiologists as BI-RADS rating 1 (upper mammogram) and BI-RADS rating 4 (lower mammogram). The shapes of the histograms are very similar and cannot be distinguished by our current histogram analysis method. These two examples were correctly classified with the additional Std and NSH criteria.

necessary to investigate if other classification strategies can be more effective than a rule-based method. Furthermore, we have not performed a systematic study to optimize the many parameters used in the segmentation algorithm. Further work will be required to investigate the dependence of the segmentation accuracy on the various parameters. The parameter selection and the performance of the computer classifier will have to be improved by training with a larger data set and its generalizability evaluated with unknown cases. The generalization of the algorithm to images acquired with other digitizers or direct digital mammography systems will also need to be investigated.

V. CONCLUSION

We are developing an image analysis method for automated segmentation of the dense area from mammograms and estimation of the percent mammographic density. Our preliminary study indicates the feasibility of our approach. The computer-estimated mammographic breast density correlate closely with the average manual segmentation by five experienced radiologists and the average bias is much less than that of the radiologists' visual estimation. We have found that correct classification of the histogram shapes is the most crucial step in our approach. The histograms of many mammograms have distinctive characteristics that can be recognized by a rule-based classifier. However, some histograms deviate from these rules and this can lead to misclassification. A further investigation will be needed to design more robust rules or classifiers to improve the classification accuracy. Despite these limitations, we have

demonstrated in this preliminary study that the estimation of mammographic density can be performed efficiently and accurately by the automated image analysis tool. The fully automated algorithm can provide an objective and reproducible quantitative estimation of mammographic breast density that is expected to be superior to subjective visual assessment and comparable to manual segmentation by radiologists.

APPENDIX A: GRAY-LEVEL THRESHOLDING—DISCRIMINANT ANALYSIS (DA) METHOD

Suppose the probability of the gray level n_i in an image with L gray levels can be estimated as

$$p_i = n_i/N, \quad N = \sum_{i=1}^{L} n_i. \tag{A1}$$

If the pixels in the image are classified into two classes C_0 and C_1 by the threshold k, then the probabilities of class occurrence and the class mean levels are given by

$$\omega_0 = \sum_{i=1}^{k} p_i = P(k), \quad \omega_1 = \sum_{i=k+1}^{L} p_i = 1 - P(k), \qquad \text{(A2)}$$

$$\mu_0 = \sum_{i=1}^{k} i p_i / \omega_0 = \mu(k) / \omega_0,$$

$$\mu_1 = \sum_{i=k+1}^{L} i p_i / \omega_1 = \frac{\mu_T - \mu(k)}{1 - P(k)}, \qquad \text{(A3)}$$

where

$$P(k) = \sum_{i=1}^{k} p_i, \quad \mu(k) = \sum_{i=1}^{k} i p_i \quad \text{and} \quad \mu_T = \sum_{i=1}^{L} i p_i, \quad \text{(A4)}$$

are the zeroth- and the first-order cumulative moments of the histogram up to the kth level, and the total mean level of original image, respectively.

The between-class variance is defined as

$$\sigma_B^2(k) = \omega_0(\mu_0 - \mu_T)^2 + \omega_1(\mu_1 - \mu_T)^2$$

$$= \omega_0 \omega_1(\mu_1 - \mu_0)^2 = \frac{[\mu_T P(k) - \mu(k)]^2}{P(k)[1 - P(k)]}, \qquad \text{(A5)}$$

and the optimal threshold k^* is given by

$$\sigma_B^2(k^*) = \max_{1 \leq k \leq L} \sigma_B^2(k). \qquad \text{(A6)}$$

APPENDIX B: GRAY-LEVEL THRESHOLDING—MAXIMUM ENTROPY PRINCIPLE (MEP) METHOD

Suppose the probability of the gray level n_i in an image with L gray levels can be estimated as

$$p_i = n_i / N, \quad N = \sum_{i=1}^{L} n_i. \qquad \text{(B1)}$$

After thresholding the image by threshold k, the *a posteriori* probability of the pixels with gray level value less than k, is given by

$$F(k) = \sum_{i=0}^{k} p_i. \qquad \text{(B2)}$$

And the *a posteriori* probability of all those pixels with values greater than or equal to k is $1 - F(k)$. Thus the Shannon entropy of the thresholded image is

$$H(F(k)) = -F(k)\log F(k) - (1 - F(k))\log(1 - F(k)). \qquad \text{(B3)}$$

The optimal threshold k maximizes $H(F(k))$.

a)Correspondence: Heang-Ping Chan, Ph.D., University of Michigan, Department of Radiology, 1500. Medical Center Drive, UHB1F510B, Ann Arbor, Michigan 48109-0030. Electronic mail: chanhp@umich.edu

[1] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, ''Cancer statistics, 1998,'' Ca-Cancer J. Clin. **48**, 6–29 (1998).

[2] J. N. Wolfe, ''Breast patterns as an index of risk for developing breast cancer,'' Am. J. Roentgenol. **126**, 1130–1139 (1976).

[3] J. N. Wolfe, A. F. Saftlas, and M. Salene, ''Evaluation of mammographic densities: A case-control study,'' Am. J. Roentgenol. **148**, 1087–1092 (1987).

[4] N. F. Boyd et al., ''Mammographic patterns and breast cancer risk: methodologic standards and contradictory results,'' J. Natl. Cancer Inst. **72**, 1253–1259 (1984).

[5] A. F. Saftlas and M. Szklo, ''Mammographic parenchymal patterns and breast cancer risk,'' Epidemiol. Rev. **9**, 146–174 (1987).

[6] J. Brisson, R. Verreault, A. S. Morrison, D. Tennina, and F. Meyer, ''Diet, mammographic features of breast tissue, and breast cancer risk,'' Am. J. Epidemiol. **130**, 14–24 (1989).

[7] A. F. Saftlas, R. N. Hoover, L. A. Brinton, M. Szklo, D. R. Olson, M. Salane, and J. N. Wolfe, ''Mammographic densities and risk of breast cancer,'' Cancer (N.Y.) **67**, 2833–2838 (1991).

[8] A. M. Oza and N. F. Boyd, ''Mammographic parenchymal patterns: A marker of breast cancer risk,'' Epidemiol. Rev. **15**, 196–208 (1993).

[9] C. Byne, ''Studying mammographic density: Implications for understanding breast cancer,'' J. Natl. Cancer Inst. **89**, 531–532 (1997).

[10] W. Leung, F. Goldberg, B. Zee, and E. Sterns, ''Mammographic density in women on postmenopausal hormone replacement therapy,'' Surgery (St. Louis) **122**, 669–674 (1997).

[11] N. F. Boyd, C. Greenberg, G. Lockwood, L. Little, L. Martin, J. Byng, Y. Martin, and D. Tritchler, ''Effects at two years of a low-fat, highcarbo-hydrate diet on radiologic features of the breast: Results from a randomized trial,'' J. Natl. Cancer Inst. **89**, 488–467 (1997).

[12] D. V. Spicer, G. Ursin, Y. R. Parisky, J. G. Pearce, D. Shoupe, A. Pike, and M. C. Pike, ''Changes in mammographic densities induced by a hormonal contraceptive designed to reduce breast cancer risk,'' J. Natl. Cancer Inst. **86**, 431–436 (1994).

[13] C. Byne, C. Schairer, J. N. Wolfe, N. Parekh, M. Salane, L. A. Brinton, R. Hoover, and R. Haile, ''Mammographic features and breast cancer risk: Effects with time, age, and menopause status,'' J. Natl. Cancer Inst. **87**, 1622–1629 (1995).

[14] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, ''Automated analysis of mammographic densities,'' Phys. Med. Biol. **41**, 909–923 (1996).

[15] S. J. Graham, M. J. Bronskill, J. W. Byng, M. J. Yaffe, and N. F. Boyd, ''Quantitative correlation of breast tissue parameters using magnetic resonance and x-ray mammography,'' Br. J. Cancer **73**, 162–168 (1996).

[16] M. J. Yaffe, N. F. Boyd, J. W. Byng, R. A. Jong, R. Fishell, G. A. Lockwood, L. E. Little, and D. L. Tritchler, ''Breast cancer risk and measured mammographic density,'' Eur. J. Cancer Prev. **7**, S47–S55 (1998).

[17] J. M. Boone, K. K. Lindfors, C. S. Veatty, and J. A. Seibert, ''A breast density index for digital mammograms based on radiologists' ranking,'' J. Digital Imaging **11**, 101–115 (1998).

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

[19] E. White, P. Velentgas, M. T. Mandelson, C. D. Lehman, J. G. Elmore, P. Porter, Y. Yasui, and S. H. Taplin, ''Variation in mammographic breast density by time in menstrual cycle among women aged 40–49 years,'' J. Natl. Cancer Inst. **90**, 906–910 (1998).

[20] C. Zhou, H. P. Chan, N. Petrick, B. Sahiner, H. M. A, M. A. Roubidoux, L. M. Hadjiiski, and M. M. Goodsitt, ''Computerized image analysis: Estimation of breast density on mammograms,'' Proc. SPIE **3979**, 1615–1624 (2000).

[21] K. E. Martin, M. A. Helvie, C. Zhou, M. A. Roubidoux, J. E. Baily, C. Paramagul, and H. P. Chan, ''Automatic computer-aided quantitative assessment of mammographic density: A validation study,'' 100th Annual Meeting of American Roentgen Ray Society, Washington, DC, 7–12 May 2000.

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

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

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

[25] A. K. C. Wong, ''A gray-level threshold selection method based on

maximum entropy principle,'' IEEE Trans. Syst. Man Cybern. **19**, 866–871 (1989).

[26] J. N. Kapur, ''Twenty-five years of maximum-entropy principle,'' J. Math. Phys. **17**, 103–156 (1983).

[27] K. E. Martin, M. A. Helvie, C. Zhou, M. A. Roubidoux, J. Bailey, P. C., C. Blane, K. Klein, S. Sonnad, and H. P. Chan, ''Mammographic density measured by an automatic computer-aided quantitative method: a feasibility study,'' Am. J. Roemtgend. (submitted).