Improvement of mammographic mass characterization using spiculation measures and morphological features

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

(Received 4 December 2000; accepted for publication 3 May 2001)

We are developing new computer vision techniques for characterization of breast masses on mammograms. We had previously developed a characterization method based on texture features. The goal of the present work was to improve our characterization method by making use of morphological features. Toward this goal, we have developed a fully automated, three-stage segmentation method that includes clustering, active contour, and spiculation detection stages. After segmentation, morphological features describing the shape of the mass were extracted. Texture features were also extracted from a band of pixels surrounding the mass. Stepwise feature selection and linear discriminant analysis were employed in the morphological, texture, and combined feature spaces for classifier design. The classification accuracy was evaluated using the area A_z under the receiver operating characteristic curve. A data set containing 249 films from 102 patients was used. When the leave-one-case-out method was applied to partition the data set into trainers and testers, the average test A_z for the task of classifying the mass on a single mammographic view was 0.83 ± 0.02, 0.84 ± 0.02, and 0.87 ± 0.02 in the morphological, texture, and combined feature spaces, respectively. The improvement obtained by supplementing texture features with morphological features in classification was statistically significant ($p = 0.04$). For classifying a mass as malignant or benign, we combined the leave-one-case-out discriminant scores from different views of a mass to obtain a summary score. In this task, the test A_z value using the combined feature space was 0.91 ± 0.02. Our results indicate that combining texture features with morphological features extracted from automatically segmented mass boundaries will be an effective approach for computer-aided characterization of mammographic masses. [DOI: 10.1118/1.1381548]

Key words: computer-aided diagnosis, mammography, breast mass characterization, segmentation, morphological features

I. INTRODUCTION

Mammography is currently the only proven and cost-effective method to detect early breast cancer. Masses are important indicators of malignancy on mammograms. However, only a small percentage of masses found on mammograms are malignant. Many benign conditions, such as cysts and fibroadenomas are detected as breast masses. Some benign masses may look suspicious enough for the radiologist to recommend biopsy. In three studies, it was found that only 20%–30% of mammographically suspicious nonpalpable breast masses that underwent biopsy were malignant.[1–3] In order to reduce costs and patient discomfort, it is important to reduce the number of benign biopsies without missing any malignant masses. Computer-aided diagnosis has the potential to assist the radiologists in the characterization of mammographic masses.[4]

In recent years, many researchers have investigated the use of computer-extracted image features for classification of breast masses as malignant or benign. The features were extracted from the gray-level and morphological characteristics of the lesion. Kilday *et al.*[5] extracted mass shapes using interactive gray-level thresholding, and classified them into cancer, cyst, and fibroadenoma categories using morphological features and patient age. Pohlman *et al.*[6] segmented masses using an adaptive region growing algorithm, whose parameters were interactively adjusted. After mass segmentation, features related to tumor shape and boundary roughness were automatically extracted and used for the classification of the lesions. They found that their tumor boundary roughness feature provided slightly inferior classification accuracy compared to two experienced radiologists who specialized in mammography. Rangayyan *et al.*[7] used a measure of the diffusion of a mass into the surrounding mammogram termed edge acutance, as well as a number of shape factors, including Fourier descriptors, moments, and compactness, to classify masses. They found the edge acutance measure to be superior to the other features extracted from the mass shape. Using the acutance measure alone, they were able to correctly classify 93% of masses in a database of 54 cases. Viton *et al.*[8] characterized the degree of spiculation and the presence of fuzzy areas in the region surrounding a mass by means of polar and pseudopolar representations of this region. Huo *et al.*[9] extracted features related to the margin and the density of the masses for classification. They designed and tested a two-stage hybrid classifier consisting of a rule-based stage and an artificial neural network stage on a data

set of 95 mammograms. The hybrid classifier achieved an area under the receiver operating characteristic (ROC) curve of 0.94 for their data set. Sahiner *et al.* and Chan *et al.* used texture features extracted from transformed images for characterization of breast masses,[10] and investigated the effect of their computer-aided diagnosis (CAD) method on radiologists' rating of breast masses.[4] They showed that their CAD method could significantly improve radiologists' accuracy in characterization of masses, and thereby might reduce unnecessary biopsies.

A second class of techniques for computer aided characterization of breast lesions use the computer to combine mammographic features extracted by a radiologist into a malignancy rating. Getty *et al.* designed a classifier based on 12 mammographic features extracted by radiologists, and showed that the classifier could substantially increase the radiologists' diagnostic accuracy.[11] Lo *et al.* and Baker *et al.* designed a neural network classifier based on BI-RADS features of the American College of Radiology, and the personal and family history of the patient.[12–14] The neural network classifier had significantly higher specificity at high sensitivity levels compared to radiologists.[14]

In the clinical evaluation of a mammographic mass, its shape and margin characteristics are very important.[15] We previously introduced a rubber-band straightening transform to analyze the margin characteristics of a mass.[10] In the present study, our aim is to include features related to the shape of the mass to improve the characterization accuracy. In order to obtain an accurate delineation of mass boundaries, we have developed a fully automated three-stage segmentation method. The first stage of our segmentation method is based on a clustering technique that we previously investigated. Clustering is used to find the general outline of the mass shape. This general outline is refined using an active contour method in the second stage. In the third stage, spiculations are detected and segmented based on image gradient directions. After segmentation, morphological features are extracted from the mass shape, and are combined with the texture features that we have previously utilized for characterization of breast masses.

II. METHODS

A. Data set

The mammograms used in this study were randomly selected from the files of patients in the Radiology Department at the University of Michigan who had undergone biopsy. All mammograms were acquired with dedicated mammographic systems. The criteria for inclusion of a mammogram in the data set were that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 249 mammograms from 102 patients. The mammograms contained a total of 122 benign and 127 malignant masses. The true pathology of the masses was determined by biopsy and histologic analysis. Six of the benign masses, and 63 of the malignant masses were characterized as spiculated by a radiologist experienced in mammographic interpretation. Out of the 249 mammograms, 223 were acquired six months or less before biopsy, and 26 were acquired more than six months before biopsy. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist on a scale of 1 (most benign mammographic appearance) to 10 (most malignant mammographic appearance). The distribution of the malignancy ranking of the masses on each view is shown in Fig. 1. Note that the malignant and benign masses overlap over the entire range of suspicion for malignancy, indicating that the malignant or benign features of these masses could not be easily distinguished by radiologists. This is consistent with the fact that all these masses had undergone biopsy. The size of the masses in our data set ranged from 5 to 29 mm (mean size $=12.5$ mm). The distribution of the size for malignant and benign masses is shown in Fig. 2. It is observed that the distribution of the size for malignant masses is similar to that for benign masses.

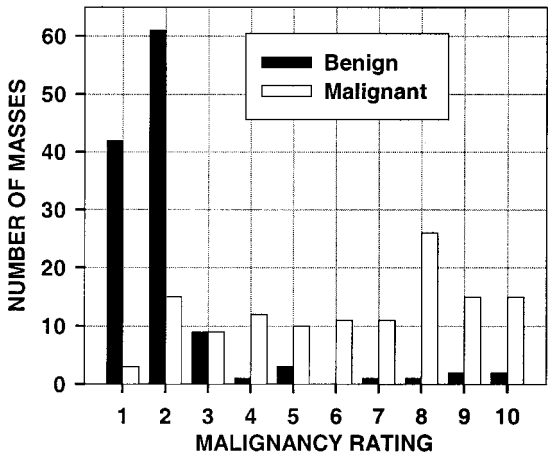

FIG. 1. The distribution of the malignancy rating of the masses in our data set, by an experienced radiologist: (1) very likely benign, (10) very likely malignant.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel size of 100 μm$\times$100 μm and

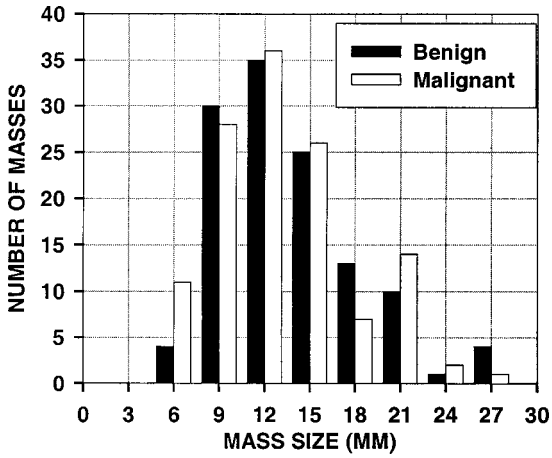

FIG. 2. The distribution of the mass size for the 249 masses in our data set. Mass sizes were measured as the longest dimension of the mass by an experienced radiologist.

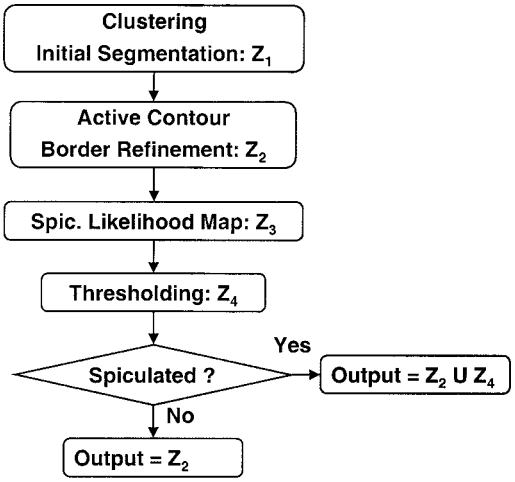

FIG. 3. The block diagram for the mass segmentation algorithm. All images Z_k, for $k \neq 3$, are binary images, with a nonzero value indicating an object pixel.

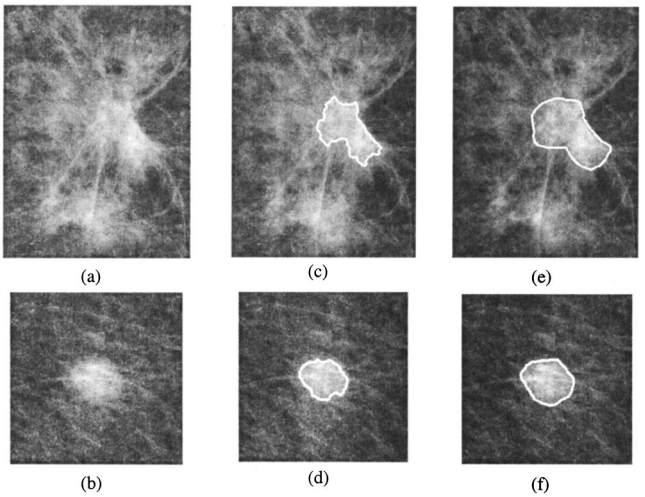

FIG. 4. The mass ROI, the initial contour, and the final contour of the active contour model for a spiculated mass [(a), (c), and (e)] and a nonspiculated mass [(b), (d), and (f)].

4096 gray levels. The digitizer was calibrated so that gray level values were linearly proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of 0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually, with the OD range extending to 3.5. The pixel values were linearly converted before they were stored on the computer so that a high pixel value represented a low optical density.

The location of the biopsied mass was identified by the radiologist, and a region of interest (ROI) containing the mass was extracted for computerized analysis. The size of the ROI was chosen such that the radiologist-marked lesion and a band of about 50-pixel-wide surrounding background were included in the ROI.

Before any processing, the ROIs were first processed with a background correction algorithm. The goal of background correction is to reduce the nonuniform background caused by the overlapping breast structures and the location of the lesion on the mammogram. The nonuniform background is not related to mass malignancy, but may affect the segmentation and feature extraction results used in our computerized analysis. Details and examples of our background correction technique can be found in the literature.[16,17]

B. Mass segmentation

We used a fully automated segmentation method to extract the mass shape. The block diagram for our mass segmentation algorithm is shown in Fig. 3, and the individual steps of the segmentation algorithm are explained in the following.

1. Initial mass segmentation

The mass segmentation method employed in this study started with the initial detection of a mass shape within a ROI using a pixel-by-pixel *K*-means clustering algorithm, which was discussed in detail in the literature.[18,19] The parameters of the segmentation algorithm were chosen so that the segmented region was slightly smaller than the apparent size of the mass. This choice prevented most of the masses from merging into neighboring objects. After clustering, one to several objects would be segmented in the ROI. If more than one object was segmented, the largest connected object was selected. The selected object was then filled, grown in a local neighborhood, and eroded and dilated with morphological operators. In the resulting binary image, a nonzero value indicated an object pixel, and zero value indicated a background pixel. The implementation details of these steps have been described in the literature.[10] Figures 4(a)–4(d) show examples of a spiculated mass and a nonspiculated mass and the results of the first stage segmentation.

2. Active contour segmentation

Although initial mass segmentation resulted in reasonable mass shapes for most of the masses, further refinement was necessary before detection and segmentation of the spiculations. We used an active contour model for mass shape refinement.

An active contour is a deformable continuous curve, whose shape is controlled by internal forces (the model, or *a priori* knowledge about the object to be segmented) and external forces (the image).[20] The internal forces impose a smoothness constraint on the contour, and the external forces push the contour toward salient image features, such as edges. To solve a segmentation problem, an initial boundary is iteratively deformed so that the energy due to internal and external forces is minimized along the contour. The energy terms used in our implementation are described in the literature.[21] We used the shape segmented by our first stage segmentation method as the initial boundary. To minimize the contour energy, we used an iterative algorithm proposed by Williams and Shah.[22] The details of our active contour model have been described elsewhere.[23] Figures 4(c)–4(f) show the initial and final contours of the model for a spiculated mass and a nonspiculated mass, respectively. A binary image, denoted by Z_2 in the schematic shown in Fig. 3, is

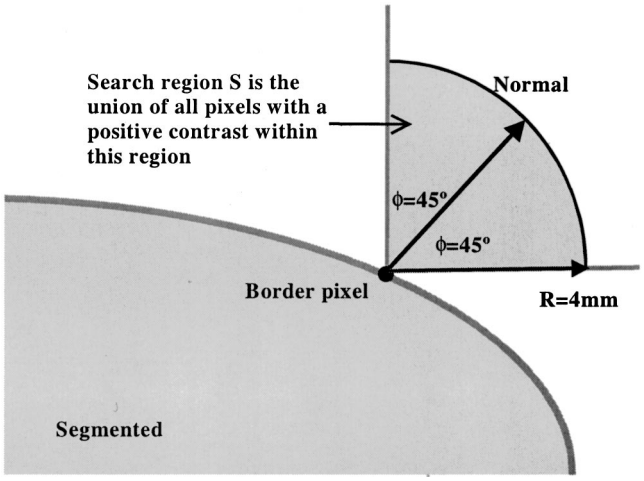

FIG. 5. The definition of the search region for a given border pixel.

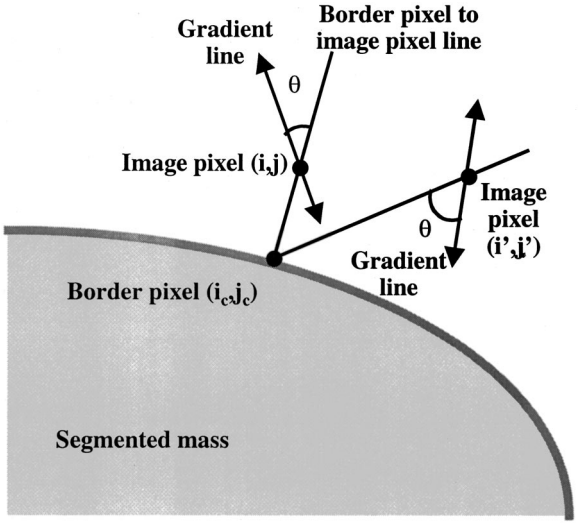

FIG. 6. The definition of the angular difference θ.

produced by filling the interior of the resulting contour, such that any pixel within the object has a pixel value of 1, and any background pixel has a pixel value of 0.

3. Segmentation of spiculations

Spiculations on mammograms appear as linear structures with a positive image contrast, and they usually lie in a radial direction to the mass. As a result of their linearity, the gradient directions at image pixels on or close to the spiculation are more or less in the same orientation relative to that of the spiculation. Karssemeijer *et al.* have used this property for detecting spiculated lesions on mammograms.[24] In this study, we developed a method for determining whether a pixel (i_c, j_c) on the mass contour lies on the path of a spiculation, and to segment the spiculation if it does.

For a pixel (i_c, j_c) on the mass boundary, a search region $S(i_c, j_c)$ is defined as the set of all image pixels that (i) lie outside the mass; (ii) have a positive contrast; (iii) are at a distance less than 4 mm from (i_c, j_c); and (iv) are within $\pm\pi/4$ of the normal to the mass contour at (i_c, j_c) (Fig. 5). At each image pixel (i,j) in $S(i_c, j_c)$, the obtuse angle θ between two lines is computed, where the first line is defined by the gradient direction at (i,j), and the second line joins the pixel (i,j) to the mass boundary pixel (i_c, j_c) (Fig. 6). We have used a method based on convolution with Gaussian derivatives[25] for computing the gradients. The spiculation measure $x(i_c, j_c)$ at a mass boundary pixel (i_c, j_c) is defined as the average value of θ in the search region $S(i_c, j_c)$. If the pixel (i_c, j_c) lies on the path of a spiculation, then θ will be close to $\pi/2$ whenever the image pixel (i,j) is on the spiculation, and hence the mean of the spiculation measure will be high.

For the segmentation task, we computed $x(i_c, j_c)$ for a sequence of 30 contours. The first contour in the sequence is that provided by the active contour model. The following contours in the sequence are obtained by expanding the previous contour by one pixel at a time, so that x is computed in a 30-pixel-wide band around the mass. The resulting image in the 30-pixel-wide band around is referred to as the spiculation likelihood map, and is denoted by Z_3 in Fig. 3. Figure 7 shows the spiculation likelihood map for the two masses used in Fig. 4. The spiculation likelihood map Z_3 is used for both detecting whether a mass is spiculated, and for segmenting the spiculations. To detect whether a mass is spiculated, a binary image Z_4 is produced by thresholding Z_3, at a threshold T. After initial experimentation, the value of T was chosen to be 0.85. This threshold was kept constant in the segmentation algorithm for all images used in the study.

After thresholding, all connected objects in Z_4 are detected. The number of the objects is used as an estimate of the number of possible spiculations. The ratio of the total area of the objects in Z_4 to the mass area is used as an indication of the relative size of the spiculations. The product of the two features above (number of objects and the size ratio) is used as a *spiculation detection variable* to classify the mass as spiculated or nonspiculated. The choice of the threshold for this classification is discussed in Sec. II D. If the mass is classified as spiculated, then the algorithm combines the binary image that represents the mass outline detected by the active contour model (Z_2) and the binary image

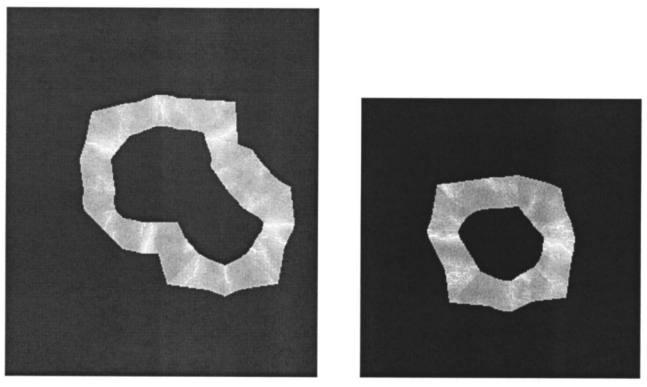

FIG. 7. The spiculation likelihood maps for the spiculated and the nonspiculated masses shown in Fig. 4: (a) spiculated, (b) nonspiculated.

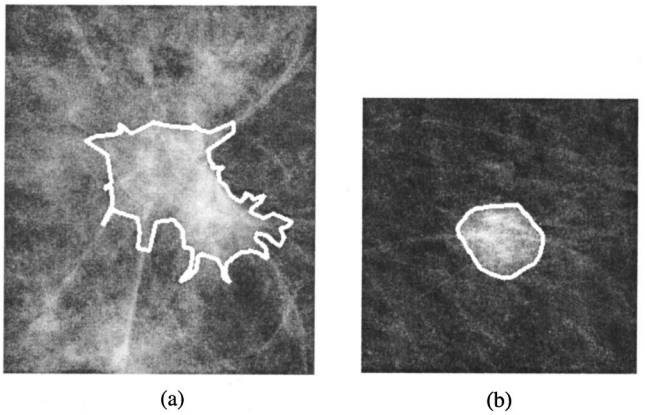

FIG. 8. The result of the final segmentation for the spiculated and the nonspiculated masses shown in Figs. 4 and 7: (a) spiculated, (b) nonspiculated.

that represents the result of thresholding (Z_4) to segment the spiculations (Fig. 3). If the mass is classified as nonspiculated, then the output of the segmentation is Z_2. Figure 8 shows the result of spiculation detection and segmentation for the masses used in Figs. 4 and 7.

C. Feature extraction

1. Extraction of morphological features

Malignant masses tend to have more irregular contours than benign masses. In addition, spiculation is a strong indication for malignancy. Therefore, features related to the segmented mass shape are expected to yield useful information for characterization of breast masses. In this study, thirteen morphological features were extracted from the final mass outline. A list of these thirteen features, as well as their accuracy in classifying each mass in our data set as malignant or benign, are shown in Table I. In this section, we describe these morphological features. The classification accuracy is discussed in Sec. IV.

The first five morphological features listed in Table I are based on the normalized radial length (NRL), defined as the Euclidean distance from the object's centroid to each of its edge pixels and normalized relative to the maximum radial length for the object.[5] In our previous studies, we found that NRL mean, standard deviation, entropy, area ratio, and zero crossing count were useful for discriminating between objects containing masses and normal tissue.[26]

TABLE I. The list of the morphological features used in this study, and the area A_z under the ROC curve when each feature is used alone for classification.

Morphological feature name	Classification accuracy A_z
Fourier descriptor	0.82
Convexity	0.79
Rectangularity	0.75
Perimeter	0.75
NRL mean	0.72
Contrast	0.71
NRL entropy	0.69
Circularity	0.67
NRL area ratio	0.66
NRL standard deviation	0.65
NRL zero crossing count	0.64
Perimeter-to-area ratio	0.63
Area	0.60

The sixth feature, convexity, is defined as the ratio of the area of the segmented object to the area of the smallest convex shape that contains the object. If the object is convex, as is the case with many benign masses, then this feature will attain its maximum value of unity. If the object shape is highly nonconvex, as is the case with many spiculated or malignant masses, then the value of this feature will be small.

The seventh feature, Fourier descriptor (FD), is based on the Fourier transform of the object boundary sequence. To compute the Fourier transform of the object boundary sequence, the x and y coordinates of each border pixel m is represented as a complex number, $z(m)=x(m)+jy(m)$, where $0m<N$, and $z(m)$ is a periodic sequence with period N. Let $c(k)$ denote the Fourier coefficients of the periodic sequence $z(m)$, and let $d(k)$ be a periodic sequence with period N, defined in the interval $0k<N$ as

$$d(k)=\begin{cases}0 & k=0\\ |c(k)/c(1)| & k\neq 0.\end{cases} \tag{1}$$

It can be shown that $d(k)$ is independent of rotation, translation, and scaling of the object, and the choice of the initial point $z(0)$ on the object contour sequence.[27] Objects with irregular contours have more high-frequency components than those with smooth contours. The following summary Fourier descriptor measure[28] which emphasizes low-frequency components of $d(k)$ is therefore useful in discriminating between shapes with smooth and irregular contours

$$\mathrm{FD}=\frac{\sum_{k=-N/2+1,k\neq 0}^{N/2} d(k)/|k|}{\sum_{k=-N/2+1,k\neq 0}^{N/2} d(k)}. \tag{2}$$

For computational efficiency, all contours were interpolated to a large integral power of 2, (2^{12}) before the computation of the Fourier series.

The remaining six features were also shown to be useful in discriminating between objects containing masses and normal tissue.[26] These features include the perimeter, area, perimeter-to-area ratio, circularity, rectangularity, and contrast of the object. The definition of these features can be found in the literature.[26]

2. Extraction of texture features

The texture of the region surrounding the mass can yield important features for its classification. Since possible spiculations and the gradient of the opacity caused by the mass are approximately radially oriented, the texture of the region surrounding a mass is expected to have a radial dependence. However, most texture extraction methods are designed for texture orientations in a uniform direction (horizontal, verti-

TABLE II. The list of the texture features used in this study, and the area A_z under the ROC curve when each feature is used alone for classification. For each measure, the range of A_z values for different pixel-pair distances and directions is shown.

Spatial gray-level dependence (SGLD) feature measure	Classification accuracy A_z	Run-length statistics (RLS) feature measure	Classification accuracy A_z
Difference average	0.52–0.66	Long runs emphasis	0.63–0.66
Difference entropy	0.53–0.66	Run percentage	0.59–0.65
Inverse difference moment	0.50–0.66	Gray level nonuniformity	0.59–0.62
Difference variance	0.52–0.65	Run length nonuniformity	0.55–0.57
Inertia	0.53–0.65	Short runs emphasis	0.50–0.56
Correlation	0.50–0.61		
Inf. measure of correlation 1	0.50–0.61		
Inf. measure of correlation 2	0.50–0.59		
Energy	0.54–0.59		
Entropy	0.54–0.58		
Sum variance	0.52–0.58		
Sum entropy	0.51–0.57		
Sum average	0.55–0.56		

cal, or at a certain angle between these two directions). To be able to extract meaningful texture features from the region surrounding a mass, we have designed a rubber band straightening transform (RBST) that maps a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region).[10,29,30] In the transformed image, the border of the mass is expected to appear approximately as a horizontal edge, and spiculations are expected to appear approximately as vertical lines.

The mass outline produced by the first stage segmentation discussed previously is used for defining the RBST image. The mass object produced by this stage is usually slightly smaller than what can be visually discerned on the mammogram. Thus, a thin border region along the mass margin is included in the RBST image. Important texture and gradient information at the mass margin is therefore included in the analysis of the region surrounding the mass. A 40-pixel-wide region, corresponding to a 4 mm band is used to determine the RBST image.

The texture features extracted from the RBST images include 13 texture measures, each calculated at 4 directions and 10 distances, from the spatial gray-level dependence (SGLD) matrices, and 5 run-length statistics (RLS) measures, each calculated at four directions, as described in our previous work.[10] A list of the SGLD and RLS texture measures is shown in Table II. Also shown in Table II are the classification accuracies when each measure is used alone to distinguish between malignant and benign ROIs. For conciseness, the range of classification accuracy (over four directions and ten distances for SGLD measures, and over four directions for RLS measures) of each texture measure is shown. The definition of these features[31,32] and the parameters used in this study can be found in the literature.[10]

D. Classification

The classifier in this study was designed to classify the masses on each available view. The same mass imaged on the CC and MLO views, and any additional views received different classification scores for each view. To assess the classifier accuracy, we considered both film-based and case-based methods. In the film-based method, the purpose was to classify the mass on each view as malignant or benign. In the case-based method, the purpose was to classify each mass as malignant or benign, using the information from all available views. To merge the information from different views of a lesion, we considered two methods. In the first method, the scores from different views were averaged. In the second method, the maximum malignancy score among all views was used as the score of the mass. The second method corresponds to calling a mass malignant if it appears to be malignant on any view, whereas the first method gives equal weight to each view to predict malignancy.

Stepwise feature selection and linear discriminant analysis were used for classifier design, and an *N*-fold cross-validation resampling scheme was used for partitioning the data into design and test sets. In a first set of experiments, we used tenfold cross validation. The data set was partitioned into ten random partitions such that all mammograms from one patient were grouped into the same partition. Nine of the partitions were used for feature selection and classifier training, and the remaining partition was used for testing. The purpose of grouping all mammograms of one patient into the same partition was to ensure that the test data were independent from training. Without this type of partitioning, one mammogram from a patient may be used for training a classifier that will be tested on another mammogram of the same patient, which may bias the test results because the training and test sets may not be completely independent. The test partition was rotated in a round-robin manner so that all partitions served as a test partition once and only once. The discriminant scores were analyzed using ROC methodology, using the LABROC program of Metz *et al.*[33] For each test partition, the classification accuracy was evaluated as the area A_z under the ROC curve. A mean A_z value for the data set was obtained by averaging these ten A_z values. In a second set of experiments, we used a leave-one-case-out method for data partitioning. This method is similar to ten-fold cross validation discussed previously, with the differences that, in

the leave-one-case-out method, each partition consisted of films from one and only one patient, and that the scores from all ROIs were accumulated for the ROC analysis. Since there were 102 patients, this corresponded to 102-fold cross validation. The statistical significance of the difference between ROC curves obtained with classifiers using different feature spaces (texture, morphological, or combined) was tested using the CLABROC program of Metz *et al.*[34]

Classifier training consisted of three stages, and was based on the training set alone for all of these three stages. The first stage was related to mass segmentation. As discussed in Sec. II B, the decision to classify a mass as spiculated or nonspiculated was based on thresholding a spiculation detection variable obtained from the spiculation likelihood map. The value of this threshold was determined from the training set such that the sum of correct decision percentages for the spiculated and nonspiculated masses was maximized for the training set. Classification of a mass as spiculated or nonspiculated determined if the spiculation segmentation step would be applied to the mass (see Fig. 3). This affected the morphological features extracted and selected in the second stage of classifier training. The second stage of the training involved stepwise feature selection,[35,36] which has been used for classifier design in many of our CAD applications.[10,17,37,38] Stepwise feature selection iteratively enters features into or removes features from the group of selected features based on a feature selection criterion. In this study, the feature selection criterion was based on the Wilks' lambda,[39] obtained using the trainers alone. The number of features in stepwise feature selection was controlled by the *F*-to-enter and *F*-to-remove thresholds, which were evaluated over a range from 5.0 to 2.0. In the third stage, the coefficients of the linear classifier were determined based on the training set. By making these three decisions independent of the test set, we aimed at improving the generalizability of our classification results to unknown cases in the patient population.

III. RESULTS

Figure 9 shows the distribution of the detection variable used for the classification of a mass as spiculated or nonspiculated. It is observed that by properly choosing the threshold, more than 30% (60/180) of the nonspiculated masses can be correctly identified without misclassifying any spiculated masses. At the selected threshold for the spiculation detection variable (see the earlier paragraph) 77% (53/69) of the spiculated masses and 78% (140/180) of the nonspiculated masses were correctly identified. Since there are six spiculated but benign masses in our data set, we did not use this variable for the classification of the masses as malignant or benign.

For both the tenfold cross validation and leave-one-case-out data partitioning methods, we investigated the classification of the masses as malignant or benign in the morphological feature space alone, texture feature space alone, and the combined morphological and texture feature space.

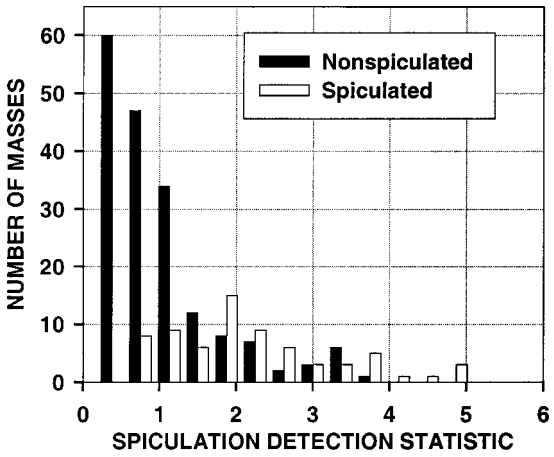

FIG. 9. The distribution of the spiculation detection variable for the spiculated and the nonspiculated masses.

A. Tenfold cross validation

The average number of selected features was 2, 10, and 14 in the morphological, texture, and combined feature spaces. The resulting A_z values for each of the ten partitions are shown in Table III. It is observed that combining the morphological and texture feature spaces improves the classification accuracy. The average A_z value for the ten partitions in this study was 0.85 for either the texture or the morphological features used alone. Using the combined feature space, the average test A_z value for the ten partitions reached 0.89.

B. Leave-one-case-out

The average number of selected features was 4, 8, and 10 in the morphological, texture, and combined feature spaces. The resulting A_z values were 0.84 ± 0.02, 0.83 ± 0.02, and 0.87 ± 0.02 in the morphological, texture, and combined feature spaces, respectively. The ROC curves for classification in these three feature spaces is shown in Fig. 10. For classification in the combined feature space ($A_z=0.87\pm0.02$), the distribution of the classifier scores for the 249 masses is shown in Fig. 11. This distribution represents film-based classification results, in the sense that the mass on each film

TABLE III. The test A_z values for each partition using linear discriminant analysis with morphological, texture, and combined feature spaces.

Partition number	Morphological feature space	Texture feature space	Combined feature space
1	0.90 ± 0.06	0.92 ± 0.06	0.92 ± 0.07
2	0.92 ± 0.06	0.98 ± 0.03	1.000 000
3	0.83 ± 0.10	0.93 ± 0.06	0.94 ± 0.05
4	0.80 ± 0.08	0.83 ± 0.08	0.86 ± 0.08
5	0.94 ± 0.05	0.80 ± 0.16	0.92 ± 0.07
6	0.82 ± 0.08	0.66 ± 0.12	0.85 ± 0.08
7	1.000 000	1.000 000	0.96 ± 0.04
8	0.77 ± 0.10	0.71 ± 0.10	0.71 ± 0.11
9	0.64 ± 0.11	0.73 ± 0.10	0.74 ± 0.10
10	0.93 ± 0.05	0.91 ± 0.06	0.98 ± 0.03
Average	0.85	0.85	0.89

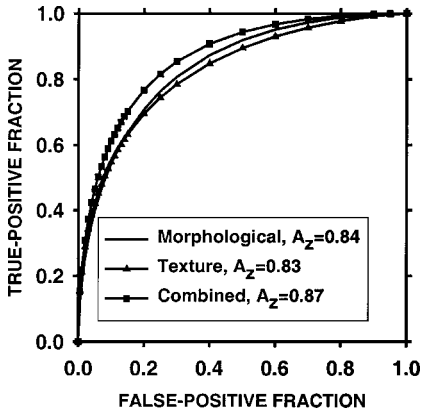

FIG. 10. ROC curves for classification of masses in the morphological, texture, and combined feature spaces.

is given a separate score, as discussed in Sec. II D. In practice, radiologists read different views of the same patient together. To simulate this condition, we combined the discriminant scores of different views of the same mass from the same year to obtain a single case-based score for each mass. This analysis resulted in 127 average scores for 102 patients, because some patients had mammograms spanning multiple years or from both breasts, and masses in different breasts or from different years were averaged separately. As described in Sec. II D, we compared using either the maximum malignancy score or the average malignancy score as the combination method. These two methods both resulted in ROC curves with $A_z = 0.91$. The distribution of the case-based scores using the averaging method is shown in Fig. 12. The ROC curves for film-based classification ($A_z = 0.87 \pm 0.02$) and case-based classification ($A_z = 0.91 \pm 0.02$) are shown in Fig. 13.

IV. DISCUSSION

Our results indicate that accurate segmentation of mammographic masses and the use of morphological features can

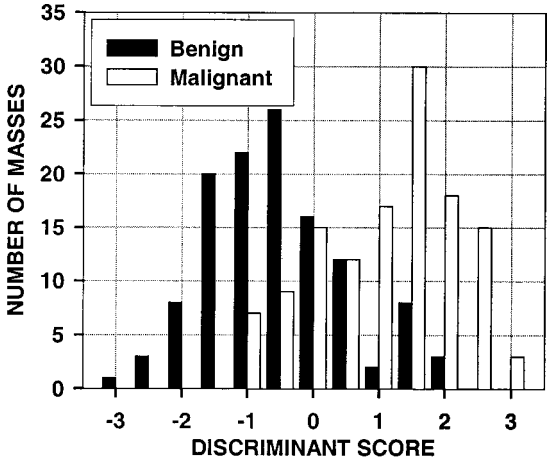

FIG. 11. The distribution of the film-based discriminant scores for leave-one-case-out classification of malignant and benign masses, using the combined feature space. The score of a mass on each film is considered independently.

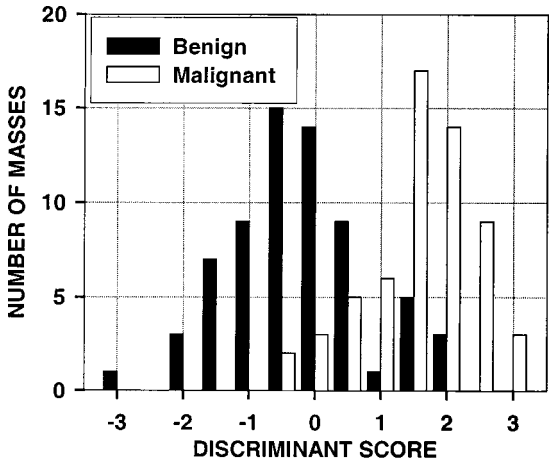

FIG. 12. The distribution of the case-based discriminant scores for leave-one-case-out classification of malignant and benign masses, using the combined feature space. The scores from the same mass of the same year have been averaged into a single score for the mass.

be effective in classifying breast masses as malignant or benign. When tenfold cross validation was used for data partitioning, the average classification accuracy with morphological features alone was equal to that with texture features alone ($A_z = 0.85$). The average classification accuracy improved to $A_z = 0.89$ when texture and morphological features were combined. In the tenfold cross-validation method, the test A_z values for each partition were computed separately. This meant that there were, on average, 24.9 films in each test partitioning. Due to the small number of cases used for computing the test ROC curves, the standard deviations of the A_z values were large, relative to those obtained using the leave-one-out method, as observed from Table III. As a result, the difference between the classifiers trained with the three different feature spaces did not reach statistical significance for any of the ten partitions shown in Table III. For the leave-one-case-out method, the scores from all ROIs were accumulated for the ROC analysis, as explained previously. This meant that the classification scores for all films were analyzed to obtain the test ROC curve. In this case, the classifier based on the combined feature space was significantly

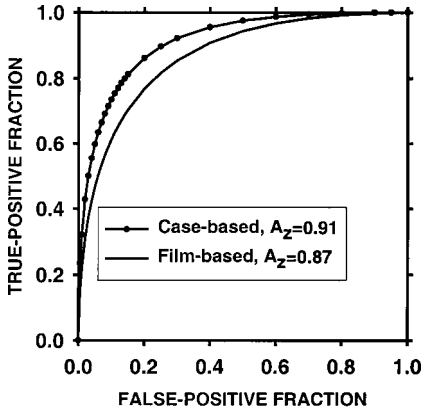

FIG. 13. Case-based and film-based ROC curves for classification of malignant and benign masses.

more accurate than that based on the texture feature space alone ($p=0.04$). The difference between the classifiers based on the combined and morphological feature spaces did not reach statistical significance.

We previously introduced a rubber-band straightening transform to analyze the margin characteristics of a mass in a texture feature space.[10] In this work, we developed a new three-stage segmentation method that consists of clustering, active contours, and spiculation detection; and evaluated the effectiveness of combining the morphological features extracted from the segmented mass and texture features for improving computerized breast mass classification. The morphological features used in this study were not novel;[5,26,28] and we had previously attempted to combine these features with texture features. However, with our previous mass segmentation method, we were unable to improve our texture-based classification results by including morphological features. This is a strong indication that the quality of segmentation is very important for morphological feature extraction.

The three-stage segmentation method used in this study adds two new stages to our previous segmentation method.[10] Previously, the clustering method was successful in segmenting the main portion of the mass from the background. However, one major limitation of clustering-based segmentation is that, even for well-circumscribed masses, the segmented shape contains many irregularities due to structured or random noises [see Fig. 4(d)]. Another limitation is that, to prevent merging with neighboring structures, the clustering parameters have to be chosen so that the segmented object is slightly smaller than the object that would visually be determined for a majority of the masses. Morphological features extracted from such a segmented mass may not adequately characterize the true morphology of the mass. The first new segmentation component of this study is the use of an active contour model for refining the clustering-based segmentation results. The second new component is the use of image gradient directions for detecting and segmenting spiculations. As shown in Fig. 9, the spiculation detection variable designed in this study was able to provide some separation between the spiculated and the nonspiculated masses. When the spiculation detection variable was used as the decision variable to classify the masses as spiculated or nonspiculated, the area A_z under the ROC curve was 0.85. However, this variable could not be directly used for the classification of the masses as malignant or benign, because almost half (64/127) of the malignant masses were visually characterized as nonspiculated by a radiologist experienced in mammographic interpretation.

The ability of each morphological feature to discriminate between the ROIs containing malignant and benign masses is shown in Table I in terms of the area A_z under the ROC curve. The A_z values indicate the accuracy of classifying the individual 249 ROIs as malignant or benign. The feature with the highest classification accuracy was the Fourier descriptor (FD). The stepwise method selected FD for all of the ten partitions shown in both the first and the last columns of Table III. When feature selection was performed using the morphological features alone, the contrast feature was selected, in addition to FD, for all of the ten partitions shown in the first column of Table III. The classification accuracy of the contrast feature is lower than those of several other features in Table I. However, contrast is the only feature in Table I that makes use of the gray scale information in the image. Therefore, compared to other morphological features, it seems to be able to introduce more complementary, and useful, information into the classifier when combined with FD.

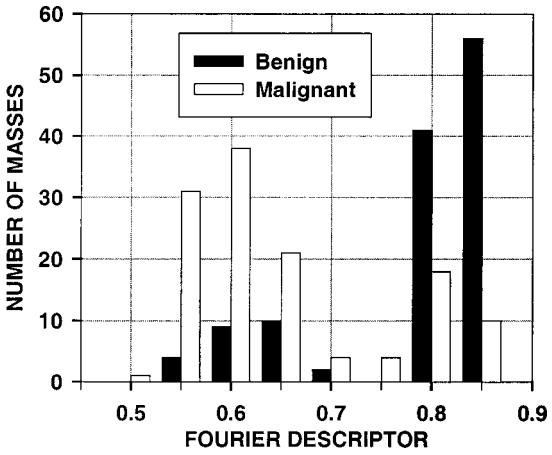

FIG. 14. The distribution of the Fourier descriptor feature for malignant and benign masses.

The ability of each texture measure to discriminate between malignant and benign masses is shown in Table II. It is observed that when used alone, the texture features are less effective than morphological features in classifying the masses in our data set. However, when texture features are combined using a linear classifier, the classification accuracy is comparable to classification using a linear classifier with morphological features alone. This may be an indication that the linear classifier is not as effective for combining these morphological features as for combining the texture features. We believe that a major reason for this is the distributions of the morphological features. It is known that the linear classifier is optimal for features with multivariate Gaussian class distributions with equal covariance matrices.[40] Due to the thresholding operation in segmentation (see the last paragraph of Sec. II B, and Fig. 3), the distributions of the morphological features in this study are very different from being Gaussian. As an example, the distribution of the Fourier descriptor feature is shown in Fig. 14. It can be observed that the distributions of both the benign and the malignant masses follow a bimodal distribution, very likely with the smaller peak corresponding to masses classified as spiculated, and the larger peak corresponding to those classified as nonspiculated. It is known that other types of classifiers, such as artificial neural networks or hybrid classifiers, perform better with non-Gaussian distributions. We will investigate the performance of other types of classifiers in these feature spaces in the future.

In a previous study, we had used the same texture features as those in this study, and had obtained an ROC area of 0.92

on a data set containing 238 masses.[4] The main reason for the lower accuracy with texture features in this study is the difference of the feature selection methods used in the two studies. In our previous study, the features were selected using the entire data set, as have been done in most studies in the CAD literature.[41–46] After feature selection, the data set was partitioned into training and test sets for formulation of the linear discriminant function. In the current study, both feature selection and classifier coefficient determination were performed on the training set. We have recently compared the effect of these two different approaches to feature selection on classifier performance prediction using a Monte Carlo simulation study.[39] We have found that, when feature selection is performed using the training set alone, the predicted test performance of the classifier is lower, in general, than that of a classifier trained with an infinite number of samples, as can be expected when a classifier is designed with a finite design sample set. However, when feature selection is performed using the entire set of available samples (training and test sets together), the predicted test performance can be higher or lower than that of a classifier trained with an infinite number of samples, depending on the number of available samples, the number of features, and the correlation between the features. The fact that the predicted performance of the classifier designed with a finite sample set can exceed that with an infinite sample set in the latter case indicates that feature selection using the entire available sample set can result in an overly optimistic prediction of the classifier performance. In studies with a clinical data set, there is no knowledge of the true class distributions, so it is difficult to predict which approach will be less biased. In order to provide a conservative prediction of the classifier performance for the general population, we chose to perform the feature selection on trainers alone in our current study.

Our data set contained 223 mammograms obtained less than six months before biopsy (preoperative mammograms) and 26 mammograms obtained more than six months prior to biopsy (prior mammograms). In order to obtain case-based average scores, we combined the scores from different years separately. Since the characteristics of the mass may change with time, combining scores across multiple years will not be reasonable. Similar to radiologists' interpretation,[4] case-based classification accuracy was higher than film-based accuracy, with $A_z = 0.91$ and $A_z = 0.87$ for the two methods, respectively.

An important feature of a CAD lesion classifier is its ability to alert radiologists to a suspicious lesion on a mammogram obtained at a time when the radiologist's suspicion level is not high enough to recommend biopsy. These prior mammograms, which are by definition more difficult to characterize, were included in our database because one would encounter such cases in clinical use or evaluation of a CAD system. If these 26 prior mammograms in our data set were excluded from the analysis, then case-based and film-based A_z values would be 0.94 and 0.88, respectively. Since the number of prior mammograms was small, we did not compare the classification accuracy of prior mammograms to that of preoperative mammograms in this study. When a larger set of prior mammograms is collected, it will be interesting and important to evaluate whether the computer classifier can predict the malignancy of the ''unsuspected'' masses in earlier years.

V. CONCLUSION

We have developed a fully automated three-stage segmentation method for delineation of mass boundary and detection and segmentation of spiculations. Morphological features describing the shape of the mass and texture features describing the margin characteristics of the mass were extracted from the segmented mass and a band of pixels surrounding the segmented mass, respectively. The data set was partitioned using a tenfold cross validation and a leave-one-case-out method for training and testing a classifier with stepwise feature selection followed by linear discriminant analysis. Using the combined feature space, the test classification accuracy was $A_z = 0.89$ and $A_z = 0.87$ for the tenfold cross validation and the leave-one-case-out methods, respectively. Case-based classification scores were obtained by averaging the test scores of the same mass from the same year. The area under the ROC curve for case-based classification was $A_z = 0.91$. Our results indicate that combining morphological features extracted from the automatically segmented mass boundary with texture features can significantly improve the accuracy for computer-aided characterization of mammographic masses.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award (B.S.) from the USAMRMC No. (DAMD 17-96-1-6012), USPHS Grant No. CA 48129, and a Whitaker Foundation Grant (N.P.). The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D., for providing the LABROC program.

[a]Electronic mail: berki@umich.edu

[1] G. Hermann, C. Janus, I. S. Schwartz, B. Krivisky, S. Bier, and J. G. Rabinowitz, ''Nonpalpable breast lesions: Accuracy of prebiopsy mammographic diagnosis,'' Radiology **165**, 323–326 (1987).
[2] F. M. Hall, J. M. Storella, D. Z. Silverstond, and G. Wyshak, ''Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography,'' Radiology **167**, 353 (1988).
[3] H. G. Jacobson and J. Edeiken, ''Biopsy of occult breast lesions: Analysis of 1261 abnormalities,'' J. Am. Math. Assoc. **263**, 2341–2343 (1990).
[4] H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, ''Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: An ROC study,'' Radiology **212**, 817–827 (1999).
[5] J. Kilday, F. Palmieri, and M. D. Fox, ''Classifying mammographic lesions using computer-aided image analysis,'' IEEE Trans. Med. Imaging **12**, 664–669 (1993).
[6] S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, ''Quantitative classification of breast tumors in digitized mammograms,'' Med. Phys. **23**, 1337–1345 (1996).
[7] R. M. Rangayyan, N. El-Faramawy, J. E. L. Desautels, and O. A. Alim, ''Discrimination between benign and malignant breast tumors using a

region-based measure of edge profile acutance,'' in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).
[8] J. L. Viton, M. Rasigni, G. Rasigni, and A. L. Llebaria, ''Method for characterizing masses in digital mammograms,'' Opt. Eng. (Bellingham) **35**, 3453–3459 (1996).
[9] Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, ''Automated computerized classification of malignant and benign masses on digitized mammograms,'' Acad. Radiol. **5**, 155–168 (1998).
[10] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, ''Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis,'' Med. Phys. **25**, 516–526 (1998).
[11] D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, ''Enhanced interpretation of diagnostic images,'' Invest. Radiol. **23**, 240–252 (1988).
[12] J. Y. Lo, J. A. Baker, P. J. Kornguth, and C. E. Floyd, ''Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features,'' Acad. Radiol. **2**, 841–850 (1995).
[13] J. A. Baker, P. J. Kornguth, J. Y. Lo, and C. E. Floyd, ''Artificial neural network: Improving the quality of breast biopsy recommendations,'' Radiology **198**, 131–135 (1996).
[14] J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, ''Breast cancer: Prediction with artificial neural network based on BI-RADS standardized lexicon,'' Radiology **196**, 817–822 (1995).
[15] C. J. D'Orsi and D. B. Kopans, ''Mammographic feature analysis,'' Semin. Roentgenol. **28**, 204–230 (1993).
[16] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images,'' IEEE Trans. Med. Imaging **15**, 598–610 (1996).
[17] 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).
[18] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue: An artificial neural network with morphological features,'' Proc. World Cong. Neural Net. **II**, 876–879 (1995).
[19] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, 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 on mammograms,'' Med. Phys. **23**, 1671–1684 (1996).
[20] M. Kass, A. Witkin, and D. Terzopoulos, ''Snakes: Active contour models,'' Int. J. Comput. Vis. **1**, 321–331 (1987).
[21] C. S. Poon and M. Braun, ''Image segmentation by a deformable contour model incorporating region analysis,'' Phys. Med. Biol. **42**, 1833–1841 (1997).
[22] D. J. Williams and M. Shah, ''A fast algorithm for active contours and curvature estimation,'' CVGIP: Image Understand. **55**, 14–26 (1992).
[23] H.-P. Chan, N. Petrick, and B. Sahiner, ''Computer-aided breast cancer diagnosis'' in *Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis*, edited by A. Jain, A. Jain, S. Jain, and L. Jain (World Scientific, River Edge, 2000), Chap. 6.
[24] N. Karssemeijer and G. te Brake, ''Detection of stellate distortions in mammograms,'' IEEE Trans. Med. Imaging **15**, 611–619 (1996).
[25] J. J. Koenderink and A. J. van Doorn, ''Generic neighborhood operators,'' IEEE Trans. Pattern Anal. Mach. Intell. **14**, 597–605 (1992).
[26] 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**, 1642–1654 (1999).
[27] S. Mori, H. Nishida, and H. Yamada, *Optical Character Recognition* (Wiley, New York, 1999).
[28] L. Shen, R. M. Rangayyan, and J. E. L. Desautels, ''Application of shape analysis to mammographic calcifications,'' IEEE Trans. Med. Imaging **13**, 263–274 (1994).
[29] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, G. M. M. , and D. D. Adler, ''Classification of masses on mammograms using a rubber-band straightening transform and feature analysis,'' Proc. SPIE **2710**, 44–50 (1996).
[30] B. Sahiner, H. P. Chan, N. Petrick, G. M. M., and M. A. Helvie, ''Characterization of masses on mammograms: Significance of the use of the rubber-band straightening transform,'' Proc. SPIE **3034**, 491–500 (1997).
[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Texture features for image classification,'' IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).
[32] M. M. Galloway, ''Texture classification using gray level run lengths,'' Comput. Graphics **4**, 172–179 (1975).
[33] C. E. Metz, B. A. Herman, and J. H. Shen, ''Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data,'' Stat. Med. **17**, 1033–1053 (1998).
[34] C. E. Metz, P. L. Wang, and H. B. Kronman, ''A new approach for testing the significance for differences between ROC curves measured from correlated data,'' in *Information Processing in Medical Imaging*, edited by F. Deconinck (Martinus Nijhoff, the Hague, 1984).
[35] N. R. Draper, *Applied Regression Analysis* (Wiley, New York, 1998).
[36] M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS, Chicago, IL, 1993).
[37] H. P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, ''Computerized analysis of mammographic microcalcifications in morphological and texture feature space,'' Med. Phys. **25**, 2007–2019 (1998).
[38] 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).
[39] B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, ''Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size,'' Med. Phys. **27**, 1509–1522 (2000).
[40] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).
[41] 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).
[42] K. G. A. Gilhuijs and M. L. Giger, ''Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging,'' Med. Phys. **25**, 1647–1654 (1998).
[43] B. S. Garra, B. H. Krasner, S. C. Horri, S. Ascher, S. K. Mun, and R. K. Zeman, ''Improving the distinction between benign and malignant breast lesions: The value of sonographic texture analysis,'' Ultrason. Imaging **15**, 267–285 (1993).
[44] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, ''Feature selection in the pattern classification problem of digital chest radiograph segmentation,'' IEEE Trans. Med. Imaging **14**, 537–547 (1995).
[45] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, ''Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence,'' Med. Phys. **19**, 1475–1481 (1992).
[46] Z. Huo, M. L. Giger, D. E. Wolverton, and W. Zhong, ''Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection,'' Med. Phys. **27**, 4–12 (2000).