

Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study

Lisa Kinnard

ISIS Center, Georgetown University Medical Center, Washington, DC 20057-1479,
Department of Electrical and Computer Engineering, Howard University, Washington, DC,
and Biomedical NMR Laboratory, Department of Radiology, Howard University, Washington, DC

Shih-Chung B. Lo^{a)}

ISIS Center, Georgetown University Medical Center, Washington, DC 20057-1479

Erini Makariou

ISIS Center, Georgetown University Medical Center, Washington, DC 20057-1479

Teresa Osicka

ISIS Center, Georgetown University Medical Center, Washington, DC 20057-1479
and Department of Electrical Engineering and Computer Science, The Catholic University of America,
Washington, DC

Paul Wang

Biomedical NMR Laboratory, Department of Radiology, Howard University, Washington, DC

Mohamed F. Chouikha

Department of Electrical and Computer Engineering, Howard University, Washington, DC

Matthew T. Freedman

ISIS Center, Georgetown University Medical Center, Washington, DC 20057-1479

(Received 5 January 2004; revised 16 April 2004; accepted for publication 22 June 2004;
published 17 September 2004)

Our purpose in this work was to develop an automatic boundary detection method for mammographic masses and to rigorously test this method via statistical analysis. The segmentation method utilized a steepest change analysis technique for determining the mass boundaries based on a composed probability density cost function. Previous investigators have shown that this function can be utilized to determine the border of the mass body. We have further analyzed this method and have discovered that the steepest changes in this function can produce mass delineations that include extended projections. The method was tested on 124 digitized mammograms selected from the University of South Florida's Digital Database for Screening Mammography (DDSM). The segmentation results were validated using overlap, accuracy, sensitivity, and specificity statistics, where the gold standards were manual traces provided by two expert radiologists. We have concluded that the best intensity threshold corresponds to a particular steepest change location within the composed probability density function. We also found that our results are more closely correlated with one expert than with the second expert. These findings were verified *via* Analysis of Variance (ANOVA) testing. The ANOVA tests obtained *p*-values ranging from 1.03×10^{-2} – 7.51×10^{-17} for the single observer studies and 2.03×10^{-2} – 9.43×10^{-4} for the two observer studies. Results were categorized using three significance levels, i.e., $p < 0.001$ (extremely significant), $p < 0.01$ (very significant), and $p < 0.05$ (significant), respectively. © 2004 American Association of Physicists in Medicine. [DOI: 10.1118/1.1781551]

Key words: mass boundary detection, mammography, probability-based cost function

I. INTRODUCTION

In the United States, breast cancer accounts for one-third of all cancer diagnoses among women and it has the second highest mortality rate of all cancer deaths in women.¹ Breast cancer studies are therefore essential for its ultimate eradication. Several studies show that only 13%–29% of suspicious masses are determined to be malignant,^{2–4} indicating that there are high false positive rates for biopsied breast masses. A higher predictive rate is anticipated by combining the mammographer's interpretation and the computer analysis.

Other studies show that 7.6%–14% of the patients have mammograms that produce false negative diagnoses.^{5,6} Alternatively, a Computer Assisted Diagnosis (CAD_x) system can serve as a clinical tool for the radiologist and consequently lower the rate of missed breast cancer.

Generally, CAD_x systems consist of three major stages, namely, segmentation, feature calculation, and classification. Segmentation is arguably one of the most important aspects of CAD_x—particularly for masses—because a strong diagnostic predictor for masses is shape. Specifically, many ma-

lignant masses have ill-defined, and/or spiculated borders and many benign masses have well-defined, rounded borders. Furthermore, breast masses can have unclear borders and are sometimes obscured by glandular tissue in mammograms. During the search for suspicious areas masses of this type may be overlooked by radiologists. When a specific area is deemed to be suspicious, the radiologist analyzes the overall mass, including its shape and margin characteristics. The margin of a mass is defined as the interface between the mass and surrounding tissue, and is regarded by some as one of the most important factors in determining its significance.⁷ Specifically, a spiculated mass consists of a central mass body surrounded by fibrous extensions, hence the resulting stellate shape. In this context, “extension” refers to those portions of the mass containing ill-defined borders, spiculations, fibrous borders, and projections. Although the diameters of these cancers are measured across the central portion of the mass, microscopic analysis of the extensions also reveals associated cancer cells, in other words, the extended projections may contain active mass growth.^{7,8} In addition, the features of the extended projections and ill-defined borders are highly useful for identifying masses. Hence, proper segmentation—including the body and periphery—is essential for the computer to analyze, and in turn, determine the malignancy of the mass in mammographic CAD_x systems.

Te Brake and Karssemeijer⁹ implemented a discrete dynamic contour model, a method similar to snakes, which begins as a set of vertices connected by edges (initial contour) and grows subject to internal and external forces. Li¹⁰ developed a method that employs *k*-means classification to categorize pixels as belonging to the region of interest (ROI) or background. Petrick *et al.*¹¹ developed the Density Weighted Contrast Enhancement (DWCE) method, in which series of filters are applied to the image in an attempt to extract masses. Pohlman *et al.*¹² developed an adaptive region growing method whose similarity criterion is determined from calculations made in 5 × 5 windows surrounding the pixel of interest. Mendez *et al.*¹³ developed a method, which combined bilateral image subtraction and region growing.

Several studies have also used probability-based analysis to segment digitized mammograms. Li *et al.*¹⁴ developed a segmentation method that first models the histogram of mammograms using a finite generalized Gaussian mixture (FGGM) and then uses a contextual Bayesian relaxation labeling (CBRL) technique to find suspected masses. Furthermore, this method uses the Expectation-Maximization (EM) technique in developing the FGGM model. Comer *et al.*¹⁵ utilized an EM technique to segment digitized mammograms into homogeneous texture regions by assigning each pixel to one of a set of classes such that the number of incorrectly classified pixels was minimized. Kupinski and Giger¹⁶ developed a method, which combines region growing with probability analysis to determine final segmentation. In their method, the probability-based function is formed from a specific composed probability density function, determined by a set of image contours produced by the region growing method. This method is a highly effective one and it was

implemented by Te Brake and Karssemeijer in their work⁹ that compared the results of a model of the discrete dynamic contour model with those of the probability-based method. For this reason, we chose to investigate its use as a possible starting point from which a second method could be developed. Consequently for our implementation of this work we discovered an important result, i.e., the steepest changes of a cost function composed from two probability density functions of the regions. It appears that in many cases this result produces contour choices that encapsulate important borders such as mass spiculations and ill-defined borders.

Several CAD_x classification techniques have been developed. They are described here to underscore the importance of accurate segmentation in CAD_x studies. Lo *et al.*¹⁷ developed an effective analysis method using the circular path neural network technique that was specifically designed to classify the segmented objects, and it can certainly be extended for the applications related to mass classification. Polakowski *et al.*¹⁸ used a multilayer perceptron (MLP) neural network to distinguish malignant and benign masses. Both Sahiner *et al.*¹⁹ and Rangayyan *et al.*²⁰ used linear discriminant analysis to distinguish benign masses from malignant masses. While many CAD_x systems have been developed, the development of fully-automated image segmentation algorithms for breast masses has proven to be a daunting task.

II. METHODS

A. Segmentation method—Maximum change of cost function as a continuation of probability-based function analysis

As a point of clarification, the function used to find optimal region growing contours in the Kupinski and Giger study¹⁶ is referred to as the probability-based function and our function is referred to as the cost function. The two functions are similar, however they differ in terms of the images used in their formation. As an initial segmentation step, the region growing is used to aggregate the area of interest,^{12,13,21} where grayscale intensity is the similarity criterion. This phase of the algorithm starts with a seed point whose intensity is high, and nearby pixels with values greater than or equal to this value are included in the region of interest. As the intensity threshold decreases, the region increases in size, therefore there is an inverse relationship between intensity value and contour size. In many cases the region growing method is extremely effective in producing contours that are excellent delineations of mammographic masses. However, the computer is not able to choose the contour that is most highly correlated with the experts' delineations, specifically, those masses that contain ill-defined margins or margins that extend into surrounding fibroglandular tissue. Furthermore, the task of asking a radiologist to visually choose the best contour would be both time intensive and extremely subjective from one radiologist to another.

The segmentation technique described in this work attempts to solve and automate this process by adding a two-dimensional (2-D) shadow and probability-based compo-

nents to the segmentation algorithm. Furthermore, we have devised a steepest descent change analysis method that chooses the best contour which delineates the mass body contour as well as its extended borders, i.e., extensions into spiculations and areas in which the borders are ill-defined or obscured. It has been discovered that the probability-based function is capable of extracting the central portion of the mass density as demonstrated by the previous investigators,¹⁶ and in this work the method has been advanced further such that it can include the extensions of the masses. The enhanced method can produce contours, which closely match expert radiologist traces. Specifically, it has been observed that this technique can select the contour that accurately represents the mass body contour for a given set of parameters. However, a further analysis of the cost function composed from the probability density functions inside and outside of a given contour revealed that the computer could choose a set of three segmentation contour choices from the entire set of contour choices, and latter make a final decision from these three choices.

1. Region growing and preprocessing

Initially, a 512×512 pixel area surrounding the mass was cropped. The region growing technique^{12,13,21} to aggregate the region of interest was employed, where the similarity criterion for our region growing algorithm is grayscale intensity. To start the growth of the first region, a seed point was placed at the center of the 512×512 ROI. The region growing process continues by decreasing the intensity value until we have grown a sufficiently large set of contours.

Next, the image is multiplied by a 2-D trapezoidal membership function with rounded corners whose upper base measures 40 pixels and lower base measures 250 pixels (1 pixel = 50 microns). This function was chosen because it is a good model of the mammographic mass' intensity distribution. Since the ROI's have been cropped such that the mass' center was located at the center of the $512 \text{ pixel} \times 512 \text{ pixel}$ area, shadow multiplication emphasizes pixel values at the center of the ROI and suppresses background pixels. The image to which the shadow has been applied is henceforth referred to as the "processed" image. The original image and its processed version were used to compute the highest possibility of its boundaries. The computation method is comprised of two components for a given boundary: (1) formulation of the composed probability as a cost function and (2) evaluation of the cost function.

The contours were grown using the original image as opposed to the processed image, and this choice accounts for a major difference between the current implementation and that of the previous investigators.¹⁶ By using contours generated from the original image, a cost function composed from the probability density functions inside and outside of the contours was produced. In many situations, the greatest changes in contour shape and size occur at sudden decreases within the function. In analyzing these steep changes it was observed that the intensity values corresponding to the steep changes typically produced contours that encapsulated both

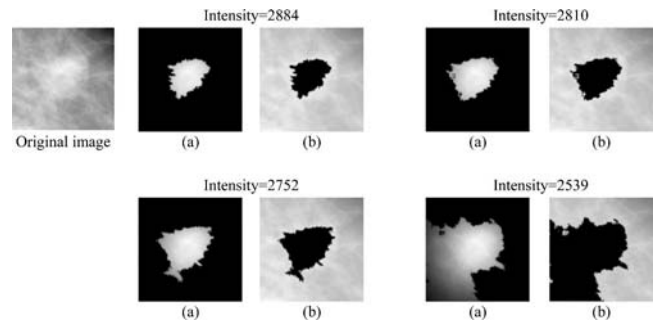


FIG. 1. Four grown contours used to construct the cost function: starts from high intensity thresholds and moves towards low intensity thresholds. Each contour separates the ROI into two parts: (a) Segmented image (based on processed image) used to compute density function $p(f_i(x,y)|S_i)$ and (b) masked image (based on the nonprocessed original image) used to compute density function $p(m_i(x,y)|S_i)$ for four intensity threshold values.

the mass body as well as its spiculated projections or ill-defined margins. This phenomenon would be suppressed if the processed image was used to generate the contour. A more detailed discussion of steep changes within the cost function is forthcoming in Sec. II A 2 C.

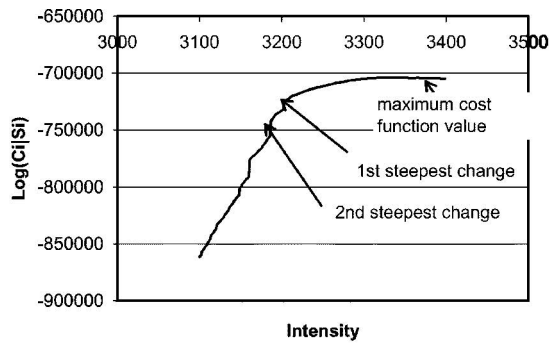
The processed image was mainly used to construct the cost function. A common technique used in mass segmentation studies is to pre-process the images using some type of filtering mechanism^{11,16} in an effort to separate the mass from surrounding fibroglandular tissue. This method could be particularly beneficial to the region growing process because it would aid in preventing the regions from growing into surrounding tissue. Alternatively, the filtering process could impede our goal of attempting to encapsulate a mass's extended borders as well as borders that are ill-defined due to the filtering process's a tendency to create rounded edges on margins that are actually jagged or spiculated. This phenomenon could potentially defeat the goal of extracting mass borders. For these reasons, we have chosen to aggregate the contours using the original ROI rather its processed version.

2. Formulation of the composed probability as a cost function

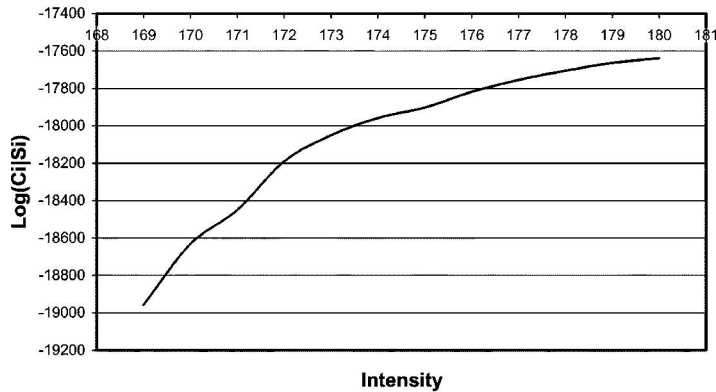
In the context of this work, the composed probability is defined as the probability density functions of the pixels inside and outside a contour using a processed and nonprocessed version of an image. Specifically, for a contour (S_i), the composed probability (C_i) is calculated:

$$C_i|S_i = \prod_{j=0}^h p(f_i(x,y)|S_i) \times \prod_{j=0}^h p(m_i(x,y)|S_i). \quad (1)$$

The quantity $f_i(x,y)$ is the set of pixels, which lie inside the contour S_i [see Fig. 1(a)], and this area contained processed pixel values. The quantity $p(f_i(x,y)|S_i)$ is the probability density function of the pixels inside S_i ($f_i(x,y)$), where "i" is the intensity threshold used to produce the contours given by the region growing step, and "h" is the maximum intensity value. The quantity $m_i(x,y)$ is the set of pixels, which lie outside the contour S_i [see Fig. 1(b)], and this area contained nonprocessed pixels. The quantity $p(m_i(x,y)|S_i)$ is



(a)



(b)

FIG. 2. (a) Example of cost function with steepest change location indicators. (b) Example of a probability-based function without an obvious steepest change location.

the probability density function of the pixels outside S_i , where “ i ” is the intensity threshold used to produce the contours given by the region growing step, and “ h ” is the maximum intensity value. For implementation purposes, the logarithm of the composed probability of the two regions, C_i was used:

$$\begin{aligned} \text{Log}(C_i|S_i) = & \log\left(\prod_{j=0}^h p(f_i(x,y)|S_i)\right) \\ & + \log\left(\prod_{j=0}^h p(m_i(x,y)|S_i)\right). \end{aligned} \quad (2)$$

3. The cost function based on the composed probability density functions

To select the contour that represents the fibrous portion of the mass, it is appropriate to examine the maximum value of the cost function:

$$\arg \max(\text{Log}(C_i|S_i); S_i, i = 1, \dots, n). \quad (3)$$

It has been assessed (also by other investigators^{9,16}) that the intensity value corresponding to this maximum value is the optimal intensity needed to delineate the mass body contour. However, in the current implementation it was discovered that the intensity threshold corresponding to the maximum value confines the contour to the fibrous portion of the mass, or, the mass body. In this study many of these contours did not include the extended borders. It is therefore hypothesized that the contour representing the mass extended borders may

well be determined by assessing the greatest changes of the cost function, or locating the steepest value changes within the function

$$\frac{d}{di} (\text{Log}(C_i|S_i); S_i, i = 1, \dots, n). \quad (4)$$

Based on this assumption, cost functions associated with masses were analyzed. The analysis reveals that the most likely boundaries of masses associated with expert radiologist traces are usually produced by the intensity value corresponding to the first or second steepest change of value immediately following the maximum value on the cost function [see Fig. 2(a)]. The description of this discovery is given below. It is followed by a validation study described in Sec. II B and by results shown in Sec. III. The overarching goal of the steep descent method is to determine whether a certain contour is the best contour, and whether it represents the mass and its extended borders.

4. The definition of steepest change

The term “steepest change” is rather subjective. In this work we define it as a location between two or more points in the cost function where the values experience a significant change. When the values are plotted as a function of intensity, these significant changes are often visible in the function. In some cases the cost function increases at a slow rate, therefore a potential steepest change location could be missed. The algorithm design compensates for this issue by

calculating the difference between values in steps over several values and comparing the results to two threshold values. The difference equation is given by

$$d(t) = f(z - wt) - f(z - w(t + 1)), \quad t = 0, m, \quad (5)$$

where f is the cost function, z is the maximum intensity, w is the width of the interval over which the cost function differences are calculated (e.g.—for $w=5$ differences are calculated every 5 points), and m is the total number of points in the searchable area divided by w . Note that “ wt ” is associated with a specific contour “ i ” described earlier. If the value of $d(t)$ yields a value greater than or equal to a given threshold, then the intensity corresponding to this location is determined to be a steepest change location. The threshold algorithm occurs as follows:

If $(d(t) \geq TV_1); \quad t = 0, \dots, m$

Then choice 1 = intensity where that condition is satisfied.

If $(d(t) \geq TV_2); \quad t = m, \dots, z$

Then choice 2 = intensity where that condition is satisfied.

where TV_1 and TV_2 are pre-defined threshold values, m is the location in the function where the choice 1 condition is satisfied, and z is the location in the function where the choice 2 condition is satisfied. During the examination of the contour growth with respect to the cost function, the first steepest change [$d(t)_{MC1}$ as choice 1] is determined by TV_1 immediately after the location of the maximum cost function value (corresponding to the mass body discussed earlier). The second the steepest change [$d(t)_{MC2}$ as choice 2] is determined by TV_2 after the first steepest change has been established.

Figure 1(a) illustrates how the algorithm is carried out. In this figure, the maximum value on the cost function occurs for a grayscale intensity value of approximately 3330. The searching process begins from this maximum point and it is discovered that the first steepest change [$d(t)_{MC1}$ as choice 1] occurs for a grayscale intensity value approximately equal to 3200. From this point the searching process continues and it is discovered that the second steepest change [$d(t)_{MC2}$ as choice 2] occurs for a grayscale intensity value approximately equal to 3175. In summary, intensity values of 3330, 3200, and 3175 can be used to grow 3 potential mass delineation candidates, and the large set of intensity choices has been narrowed to 3 choices. The following scenarios occurred when the three contour choices produced by the (1) maximum intensity value on the cost function (2) the intensity corresponding to the first steepest change on the cost function, and (3) the intensity corresponding to the second steepest change on the cost function.

(1) Intensity corresponding to the maximum value on the cost function: The central body of the mass was encapsulated.

- (2) Intensity corresponding to the first steepest change on the cost function: The central body of the mass + some of its extended borders (i.e., projections and spiculations) was encapsulated.
- (3) Intensity corresponding to the second steepest change on the cost function: The central body of the mass + more extended borders + surrounding fibroglandular tissue was encapsulated.

The intensity corresponding to the first steepest change provides the best choice, and an examination of this observation is shown and discussed in Secs. III and IV of this work.

As stated previously, the steep changes within the cost function would be suppressed if the processed image was used to generate the contour; therefore, the function would be relatively smooth. Figure 2(b), which shows a probability-based function produced by contours that were grown using a processed ROI, demonstrates this issue.

B. Validation method

In several segmentation studies the results were validated using the overlap statistic alone, however, it was necessary to analyze the performance of the steepest change algorithm on the basis of four statistics to verify that the algorithm is indeed capable of categorizing mass and background pixels correctly. This type of analysis provides helpful information regarding necessary changes for the algorithm’s design and can possibly aid in its optimization.

The segmentation method was validated on the basis of overlap, accuracy, sensitivity, and specificity.^{22,23} These statistics are calculated as follows:

$$\text{Overlap} = \frac{N_{TP}}{N_{FN} + N_{TP} + N_{FP}}, \quad (6)$$

$$\text{Accuracy} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \quad (7)$$

$$\text{Sensitivity} = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad (8)$$

$$\text{Specificity} = \frac{N_{TN}}{N_{TN} + N_{FP}}, \quad (9)$$

where N_{TP} is the true positive fraction (part of the image correctly classified as mass), N_{TN} true negative fraction (part of the image correctly classified as surrounding tissue), N_{FP}

TABLE I. Distribution of DDSM masses studied according to their subtlety ratings.

Subtlety category	Cancer	Benign
Number of masses with a rating=1	5	3
Number of masses with a rating=2	12	12
Number of masses with a rating=3	18	17
Number of masses with a rating=4	9	23
Number of masses with a rating=5	15	10

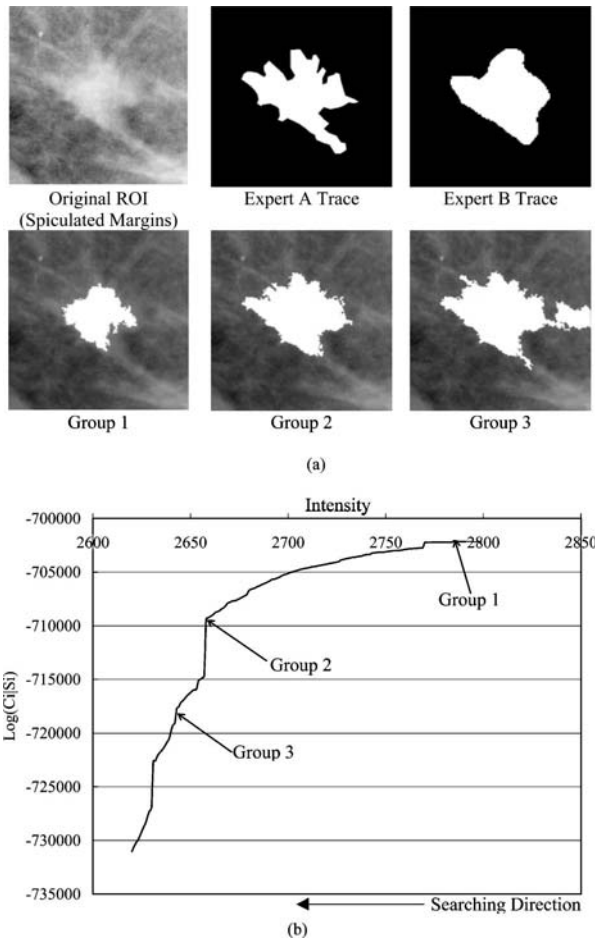


FIG. 3. (a) Segmentation results for a malignant mass with spiculated margins (subtlety=2) (b) the corresponding cost function.

is the false positive fraction (part of the image incorrectly classified as mass), and N_{FN} is the false negative fraction (part of the image incorrectly classified as surrounding tissue). This method requires a gold standard, or a contour to which the segmentation results can be compared. The gold standards for the experiments performed in this work were mass contours, which have been traced by expert radiologists.

The experiments produced contours for the intensity values resulting from three locations within the cost functions: (1) The intensity of the maximum value within the cost function; (2) the intensity for which the cost function experiences its first steepest change; and (3) the intensity for which the cost function experiences its second steepest change. It has been observed that the intensity for which the cost function experiences its first steepest change produces the contour trace that is most highly correlated with the gold standard traces, regarding overlap and accuracy. In cases for which better results occur at the second steepest change location, there is no significant difference between these results and the results calculated for the first steepest change location. Second, it has been observed that the results are more closely correlated with one expert than with the second expert. These hypotheses were tested using the one-way Analysis of Vari-

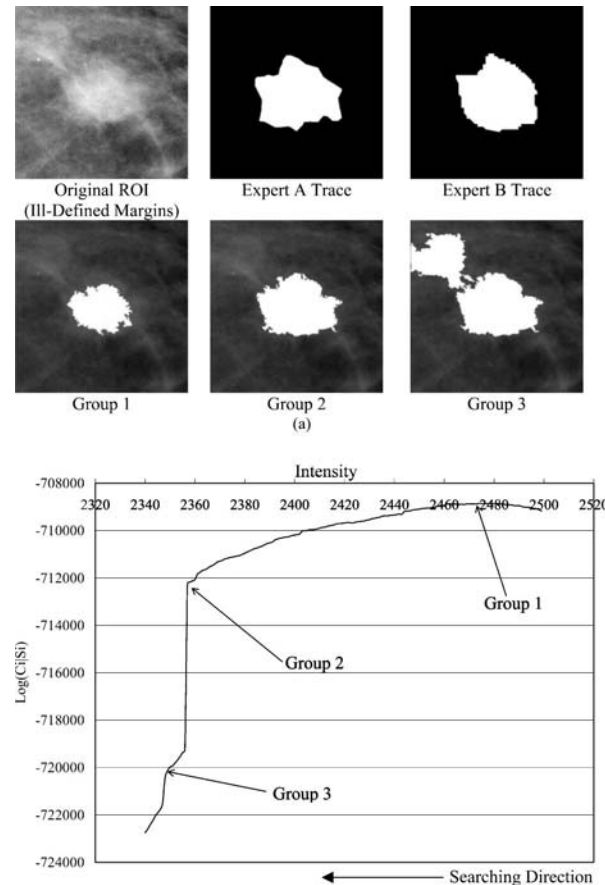


FIG. 4. (a) Segmentation results for a malignant mass with ill-defined margins (subtlety=3); (b) the corresponding cost function.

ance (ANOVA) test.^{24,25} In this study, three significance levels (i.e., $p < 0.001$, $p < 0.01$, and $p < 0.05$) were used to categorize the ANOVA results as described in the next section.

III. EXPERIMENTS AND RESULTS

The following sections describe the database and experiments, and provide segmentation results and ANOVA test results.

A. Database

For this study, a total of 124 masses were chosen from the University of South Florida's Digital Database for Screening Mammography (DDSM).²⁶ The DDSM films were digitized at 43.5 or 50 μm 's using either the Howtek or Lumisys digitizers, respectively. The DDSM cases have been ranked by expert radiologists on a scale from 1 to 5, where 1 represents the most subtle masses and 5 represents the most obvious masses. Table I lists the distribution of the masses studied according to their subtlety ratings. The images were of varying contrasts and the masses were of varying sizes.

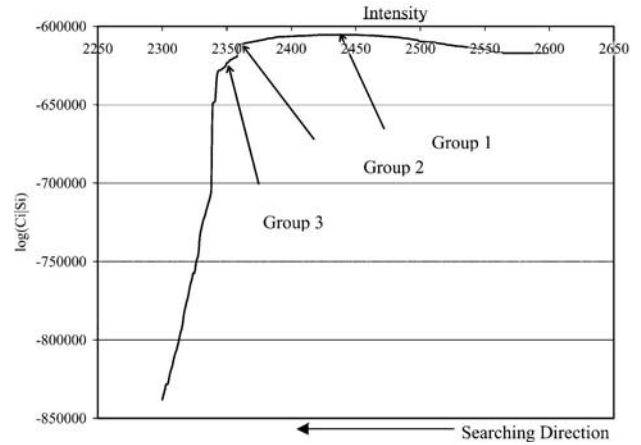
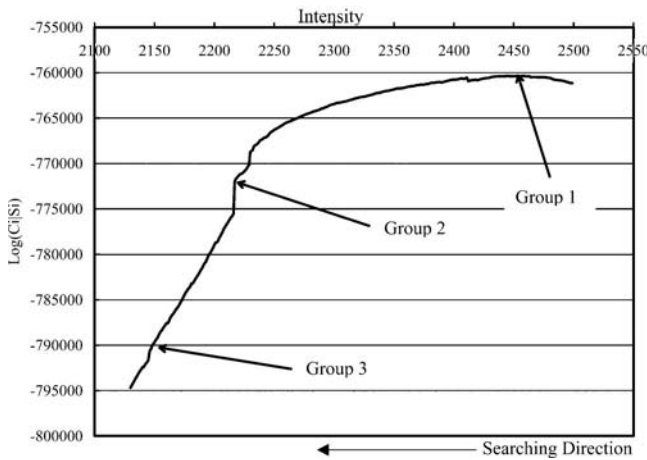
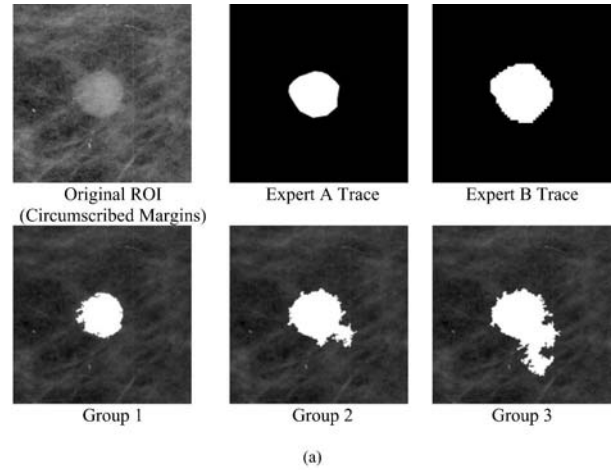
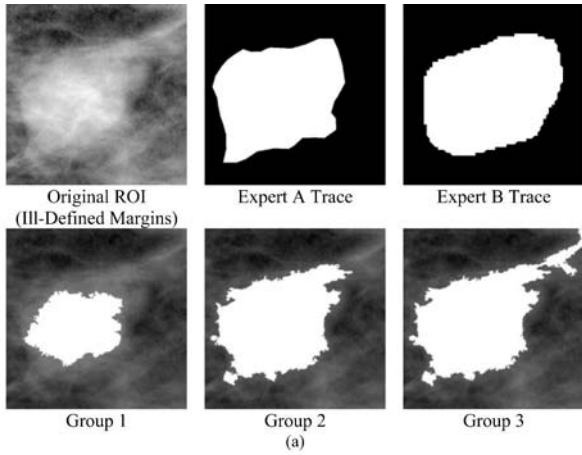


Fig. 5. (a) Segmentation results for a benign mass with ill-defined margins (subtlety=3); (b) the corresponding cost function.

Fig. 6. (a) Segmentation results for a benign mass with circumscribed margins (subtlety=4); (b) the corresponding cost function.

The first set of expert traces was provided by an attending physician at Georgetown University Medical Center (GUMC), and is hereafter referred to as the Expert A traces. The second set of expert traces was provided by the DDSM, and is hereafter referred to as the Expert B traces.

B. Experiments

As mentioned previously, the term “steepest change” is very subjective. Therefore, a set of thresholds needed to be set in an effort to define a particular location within the cost function as a “steepest change location.” For this study the following thresholds were experimentally chosen: $TV_1 = 1800$, $TV_2 = 1300$, where TV_1 equals the threshold for steepest change location 1 for the cost function, and TV_2 equals the threshold for steepest change location 2 for the cost function. A number of experiments were performed in an effort to prove that (1) the intensity for which the cost function experiences the first steepest change location produces the contour trace, which is most highly correlated with the gold standard traces with regard to overlap and accuracy. In cases for which the second steepest change location achieves better results, there are no significant differences between the values obtained from the first steepest change

location and the second steepest change location. The experiments linked with these hypotheses comprise the studies for a single observer. We have also set out to prove that (2) our results are more closely correlated with one expert than with the second expert. The experiments linked with this hypothesis comprise the studies between two observers. First segmentation results for two malignant cases are presented, followed by segmentation results for two benign cases. Second, the ANOVA results for a set of hypotheses are presented. The contours produced by the maximum value as well as by the steepest change locations within the cost functions are labeled as follows: (1) group 1: The intensity for which a value within the cost function is maximum; (2) group 2: The intensity for which the cost function experiences its first steepest change; (3) group 3: The intensity for which the cost function experiences its second steepest change.

C. Results

Figures 3–6 display the results for two malignant cases accompanied by their cost functions as well as results for two benign cases accompanied by their cost functions. The ANOVA results appear in a set of tables (Secs. II–IV), where each table lists the hypothesis tested along with p -values and their corresponding categorizations. The p -values are catego-

rized in the following way: not significant (NS for $p > 0.05$), significant (S for $p < 0.05$), very significant (VS for $p < 0.01$), and extremely significant (ES for $p < 0.001$). Each p -value table is followed by a second table, which contains the mean values of overlap, accuracy, sensitivity, and specificity for each group. Sections II and III are identical regarding the experiments, however, the pathologies of the masses

are different (Sec. II—malignant masses, Sec. III—benign masses). Although the experiments are identical they have been separated for clarity purposes.

A larger set of segmentation results has been placed in an image gallery containing 7 malignant mass results (Fig. 7) and 7 benign mass results (Fig. 8). These figures are located in the Appendix.

1. Segmentation results

2. ANOVA test results for comparison of contour groups with single observer: Malignant cases

TABLE II. Single observer results (expert A gold standard, malignant masses).

ANOVA test	P -value (group 1 vs group 2)	P -value (group 2 vs group 3)	P -value (group 1 vs group 3)
Difference between groups (overlap)	1.78×10^{-4} (ES)	2.91×10^{-2} (S)	NS
Difference between groups (accuracy)	NS	3.14×10^{-2} (S)	NS
Difference between groups (sensitivity)	1.88×10^{-9} (ES)	NS	1.85×10^{-13} (ES)
Difference between groups (specificity)	5.12×10^{-4} (ES)	2.40×10^{-3} (VS)	2.71×10^{-9} (ES)

TABLE III. Mean values for overlap, accuracy, sensitivity, and specificity (expert A gold standard, malignant masses).

Measurement	Mean value (group 1)	Mean value (group 2)	Mean value (group 3)
Overlap	0.47	0.60	0.53
Accuracy	0.88	0.90	0.87
Sensitivity	0.49	0.75	0.81
Specificity	0.99	0.94	0.88

TABLE IV. Single observer results (expert B gold standard, malignant masses).

ANOVA test	P -value (group 1 vs group 2)	P -value (group 2 vs group 3)	P -value (group 1 vs group 3)
Difference between groups (overlap)	3.96×10^{-6} (ES)	NS	1.58×10^{-4}
Difference between groups (accuracy)	NS	NS	NS
Difference between groups (sensitivity)	4.88×10^{-8} (ES)	4.31×10^{-2} (S)	4.25×10^{-12} (ES)
Difference between groups (specificity)	2.70×10^{-4} (ES)	4.36×10^{-4} (ES)	1.44×10^{-7} (ES)

TABLE V. Mean values for overlap, accuracy, sensitivity, and specificity (expert B gold standard, malignant masses).

Measurement	Mean value (group 1)	Mean value (group 2)	Mean value (group 3)
Overlap	0.38	0.54	0.51
Accuracy	0.83	0.86	0.84
Sensitivity	0.38	0.56	0.60
Specificity	1.00	0.98	0.94

3. ANOVA test results for comparison of contour groups with single observer: Benign cases

TABLE VI. Single observer results (expert A gold standard, benign masses).

ANOVA test	<i>P</i> -value (group 1 vs group 2)	<i>P</i> -value (group 2 vs group 3)	<i>P</i> -value (group 1 vs group 3)
Difference between groups (overlap)	3.19×10^{-4} (ES)	8.38×10^{-4} (ES)	NS
Difference between groups (accuracy)	NS	4.73×10^{-3} (VS)	2.51×10^{-3} (VS)
Difference between groups (sensitivity)	1.14×10^{-9} (ES)	1.89×10^{-2} (S)	7.51×10^{-17} (ES)
Difference between groups (specificity)	8.93×10^{-3} (VS)	1.24×10^{-3} (VS)	3.32×10^{-10} (ES)

TABLE VII. Mean values for overlap, accuracy, sensitivity, and specificity (expert A gold standard, benign masses).

Measurement	Mean value (group 1)	Mean value (group 2)	Mean value (group 3)
Overlap	0.46	0.58	0.45
Accuracy	0.90	0.91	0.85
Sensitivity	0.49	0.73	0.82
Specificity	0.99	0.94	0.86

TABLE VIII. Single observer results (expert B gold standard, benign masses).

ANOVA test	<i>P</i> -value (group 1 vs group 2)	<i>P</i> -value (group 2 vs group 3)	<i>P</i> -value (group 1 vs group 3)
Difference between groups (overlap)	8.82×10^{-5} (ES)	NS	1.62×10^{-2} (S)
Difference between groups (accuracy)	NS	2.62×10^{-2} (S)	2.48×10^{-2} (S)
Difference between groups (sensitivity)	1.61×10^{-7} (ES)	NS	3.14×10^{-12} (ES)
Difference between groups (specificity)	1.18×10^{-2} (S)	1.27×10^{-2} (S)	1.25×10^{-7} (ES)

TABLE IX. Mean values for overlap, accuracy, sensitivity, and specificity (expert B gold standard, benign masses).

Measurement	Mean value (group 1)	Mean value (group 2)	Mean value (group 3)
Overlap	0.36	0.51	0.44
Accuracy	0.88	0.89	0.83
Sensitivity	0.36	0.61	0.69
Specificity	0.99	0.94	0.86

4. ANOVA test results for comparison of contour groups between two observers

TABLE X. Two observer results: expert A vs expert B, malignant masses.

ANOVA test	<i>P</i> -value (group 1 vs group 2)	<i>P</i> -value (group 2 vs group 3)	<i>P</i> -value (group 1 vs group 3)
Expert A vs expert B (overlap)	3.12×10^{-3} (VS)	3.32×10^{-2} (S)	NS
Expert A vs expert B (accuracy)	1.20×10^{-2} (S)	4.46×10^{-2} (S)	NS
Expert A vs expert B (sensitivity)	9.43×10^{-4} (ES)	3.38×10^{-4} (ES)	3.67×10^{-4} (ES)
Expert A vs expert B (specificity)	NS	NS	NS

TABLE XI. Mean values for overlap, accuracy, sensitivity, and specificity (expert A vs expert B, malignant masses).

Measurement	Mean value, expert A (group 1)	Mean value, expert B (group 1)	Mean value, expert A (group 2)	Mean value, expert B (group 2)	Mean value, expert A (group 3)	Mean value, expert B (group 3)
Overlap	0.49	0.38	0.62	0.55	0.55	0.51
Accuracy	0.89	0.83	0.91	0.87	0.87	0.84
Sensitivity	0.52	0.38	0.75	0.60	0.82	0.68
Specificity	0.99	1.00	0.95	0.97	0.89	0.91

TABLE XII. Two observer results: expert A vs expert B, benign masses.

ANOVA test	<i>P</i> -value (group 1 vs group 2)	<i>P</i> -value (group 2 vs group 3)	<i>P</i> -value (group 1 vs group 3)
Expert A vs expert B (overlap)	NS	NS	NS
Expert A vs expert B (accuracy)	NS	NS	NS
Expert A vs expert B (sensitivity)	3.56×10^{-2} (S)	4.90×10^{-2} (S)	2.03×10^{-2} (S)
Expert A vs expert B (specificity)	NS	NS	NS

TABLE XIII. Mean values for overlap, accuracy, sensitivity, and specificity: expert A vs expert B, benign masses.

Measurement	Mean value, expert A (group 1)	Mean value, expert B (group 1)	Mean value, expert A (group 2)	Mean value, expert B (group 2)	Mean value, expert A (group 3)	Mean value, expert B (group 3)
Overlap	0.42	0.35	0.57	0.50	0.48	0.44
Accuracy	0.90	0.88	0.91	0.89	0.85	0.83
Sensitivity	0.44	0.36	0.71	0.61	0.79	0.69
Specificity	0.99	0.99	0.94	0.94	0.86	0.86

IV. DISCUSSION

A. Segmentation results

The ROI's shown in Figs. 3 and 4 demonstrate that the intensity produced by the maximum value is capable of accurately delineating the mass body contour, and in some cases this intensity corresponding to the maximum value produces a contour, which falls inside the mass body contour. This situation can be problematic because low segmentation sensitivities can produce large errors during the feature calculation and classification phases of CAD_x. Of the three available segmentation choices for each mass, it appears that the first steepest change location produces the contours with the strongest correlation in comparison to both gold standards. These contours appear to cover both the mass body contour as well as the extended borders. In some instances the region grows into some areas that are not declared as mass areas by the gold standards—we call this flooding—and fails to grow into other areas that have been declared as mass areas. Finally, the second steepest change location produces contours that also cover both the mass body contour as well as the extended borders, and, these contours tend to also include surrounding fibroglandular tissue; hence, the flooding phenomenon is a common occurrence. In the cases shown, it is clear that steepest change location 1 produces the best contours, in comparison to the gold standards, however, the ANOVA test results allow us to make such a claim. The following discussion is divided into five sections: single observer malignant results, single observer benign results, and two observer results (malignant and benign), algorithm performance, and an additional discussion on methods.

B. Malignant cases with single observer

For both the expert A and expert B gold standards, Tables II–V show a statistically significant difference between groups 1 and 2 on the basis of overlap and sensitivity, where the mean values of group 2 were higher than the mean values of group 1 for these statistics. These results are expected because as shown in the figures, the group 2 contours consistently covered more of the mass area (and correctly covered this mass area) as compared to the group 1 contours, according to both experts. There was a statistically significant difference in sensitivity between group 1 and group 3, where the mean of group 3 was higher than the mean of group 1. This difference is an expected result because out of all the groups, group 3 contours consistently covered the most mass area. For the expert B gold standard there was a statistically significant difference in overlap between group 1 and group 3, where the mean of group 3 was higher than the mean of group 1. This difference is also an expected result because, out of all the groups, the group 3 contours covered the most mass area correctly.

C. Benign cases with single observer

For the expert A traces there were statistically significant differences between the group 2 and group 3 traces on the

basis of overlap, accuracy, and sensitivity, where the group 2 mean values for overlap and accuracy were higher than those of group 3 (see Tables VI–IX). This difference is an expected result because it is likely that many of the group 3 contours contained flooded areas, which cause both of these values to be lower than those values of contours without flooded areas. The overlap and sensitivity values for group 2 were significantly higher than those of group 1. This difference is also an expected result because the group 2 contours not only covered more mass area but also covered this area correctly. Finally, the group 3 accuracy and sensitivity values were significantly higher than those for group 1. Again this difference is an expected result because the group 3 contours not only covered more mass area but they also covered this area correctly.

For the expert B gold standard there were statistically significant differences between the group 2 and group 3 traces on the basis of accuracy and sensitivity, where the group 2 mean values for overlap and accuracy were higher than those of group 3. This difference is an expected result because it is likely that many of the group 3 contours contained flooded areas, which cause both of these values to be lower than contours without flooded areas. There were statistically significant differences between group 1 and group 2 on the basis of overlap and sensitivity, where the mean values for group 2 were higher than the mean values for group 1. This is an expected result because the group 2 contours not only covered more mass area but they also covered this area correctly. There were statistically significant differences between group 3 and group 1 on the basis of overlap and sensitivity, where the mean values for group 3 were higher than those of group 1. Again this difference is an expected result because the group 3 contours not only covered more mass area but they covered this area correctly.

In nearly all cases for the single observer studies, it was expected that the specificity values for group 1 would always be higher than those for groups 2 and 3 because this contour always covered the smallest mass area; consequently its background was always highly correlated with the background areas dictated by the gold standards. Moreover, in some cases the group 2 and group 3 contours grew into areas that were not regarded as mass, but rather were regarded as background; therefore, their specificity values had a lower correlation with the gold standard as compared to the group 1 contours.

D. Malignant and benign cases with two observers

For the two observer studies, comparisons were made between experts A and B on a group-by-group basis in an effort to prove that there were significant differences between the two radiologists on the basis of overlap, accuracy, sensitivity, and specificity (see Tables X–XIII). For the malignant masses, there were statistically significant differences between the two experts on the basis of overlap, accuracy, and sensitivity. There was a statistically significant difference between the two experts for group 3 on the basis of sensitivity. For the benign masses, there were statistically significant differences between the two experts for all three groups on the

basis of sensitivity. For all cases, expert A's values were consistently higher than those of expert B. These statistically significant differences between the experts were expected due to their differences in opinion. The fact that expert A's mean values were higher than those for expert B, however, does not warrant the conclusion that expert A is a more reliable expert; however, it does warrant the conclusion that there is stronger agreement between the computer's results and expert A's traces. Furthermore, there were less statistically significant differences for the benign cases than for the malignant cases. This result is expected because, in general, benign masses have better defined borders, and thus the two experts were more likely to agree.

E. Algorithm performance

Apparently the chosen thresholds produce first steepest change location intensities that generate contours closely correlated with the expert traces. In some instances the second steepest change location is extremely far from the first steepest change location, which implies that the function in question increases very slowly; moreover, many of the second steepest change location intensities produce contours with flooded areas. For the majority of the cases in which the second steepest change location contour achieves a higher sensitivity value, but not a significantly higher sensitivity value, we can still choose the first steepest change location contour because the difference between the two contours is likely to be negligible.

In analyzing the probability-based cost functions, we found that those functions with very steep changes are typically associated with masses that have well-defined borders while those functions that increase slowly are associated with masses that have ill-defined borders. This phenomenon may make it necessary to develop an adaptive threshold process for the steepest change evaluation such that the functions are grouped into various categories (e.g., smooth versus steep), because a threshold value that is optimal for a steep function may not be optimal for a smooth function.

F. Additional discussion on methods used

In this study the steepest descent method appears to have the advantage of locating ill-defined margins as well as extensions such as malignant spiculations and projections for mammographic masses. If solely the human eye is used, it can be difficult to separate the mass from the surrounding fibroglandular tissue. Therefore, this method has the potential to complement the process of reading mammographic films. One of the downfalls of the method is its dependence upon the assumption that masses are generally light in color. This assumption impedes the region growing process because masses that contain darker areas and are surrounded on one or more sides by bright tissue can cause contours to flood into areas that are not actual mass tissue. Typically, this situation occurs for the mass located on the border of the breast region on a mammogram.

All of the segmentation methods surveyed in the introduction of this paper are excellent solutions for the problems

their authors set out to solve, however, in some cases it is difficult to make comparisons between different methods without the availability of a set of several visual results. In some studies, the focus was either to detect masses or to distinguish malignant from benign masses. Thus, the validation process did not take the form of a comparison with expert radiologist manual traces; but rather, features were calculated on the potential mass candidates and they were later classified as being mass tissue or normal tissue.¹⁰⁻¹³ The purpose of Li's study¹⁴ was to distinguish between normal and abnormal tissue; thus the authors did not provide any statistics such as overlap or accuracy. Nevertheless, the study contains a figure of 60 masses that contain both computer and radiologist annotations to give the reader an idea of the computer algorithm's performance. Te Brake and Karssemeijer's study⁹ used the overlap statistic to test the efficacy of their method. They indicated that the central mass area was delineated by the radiologist and their computer results were compared to these annotations. The Kupinski and Giger study¹⁶ also used the overlap statistic to test the efficacy of their method and set a threshold for which the mass was considered to be successfully segmented. For example, masses whose overlap values are greater than 0.7 imply that there was successful segmentation.

The technical method presented herein shows that the results obtained from the maximization of the composed probability density function (i.e., the cost function) are equivalent to those obtained from previous methods presented by previous investigators. However, the steepest change of the composed probability density function is the closest to radiologists' determinations.

V. CONCLUSION

We have shown that our fully automatic boundary detection method for malignant and benign masses can effectively delineate these masses using intensities, that correspond to the first steepest change location within their cost functions. Additionally, the method appears to be more highly correlated with one set of expert traces than with a second set of expert traces, regarding the accuracy and overlap statistics. This result shows that inter-observer variability can be an important factor in segmentation algorithm design, and it has motivated us to seek the opinions of more expert radiologists to test the robustness of our algorithm. The second steepest change location intensity will always yield contours with higher sensitivity values, however, it behooves us to choose the first steepest change location intensity because it avoids the risk of choosing contours that contain substantial flooding. In future work, a worthwhile study would run the experiments for different threshold values in an effort to discover the possibility of deriving an optimal threshold procedure. We believe that such a procedure would improve the method of choosing optimal contours.

ACKNOWLEDGMENTS

This work was supported by U.S. Army Grant Nos. DAMD17-03-1-0314, DAMD17-01-1-0267, and DAMD

17-00-1-0291, and NIH Grant No. RCM/NCRR/NIH 2G12RR00348. The authors would also like to thank the referees for their constructive comments and recommendations.

APPENDIX A—GALLERY OF SEGMENTATION RESULTS

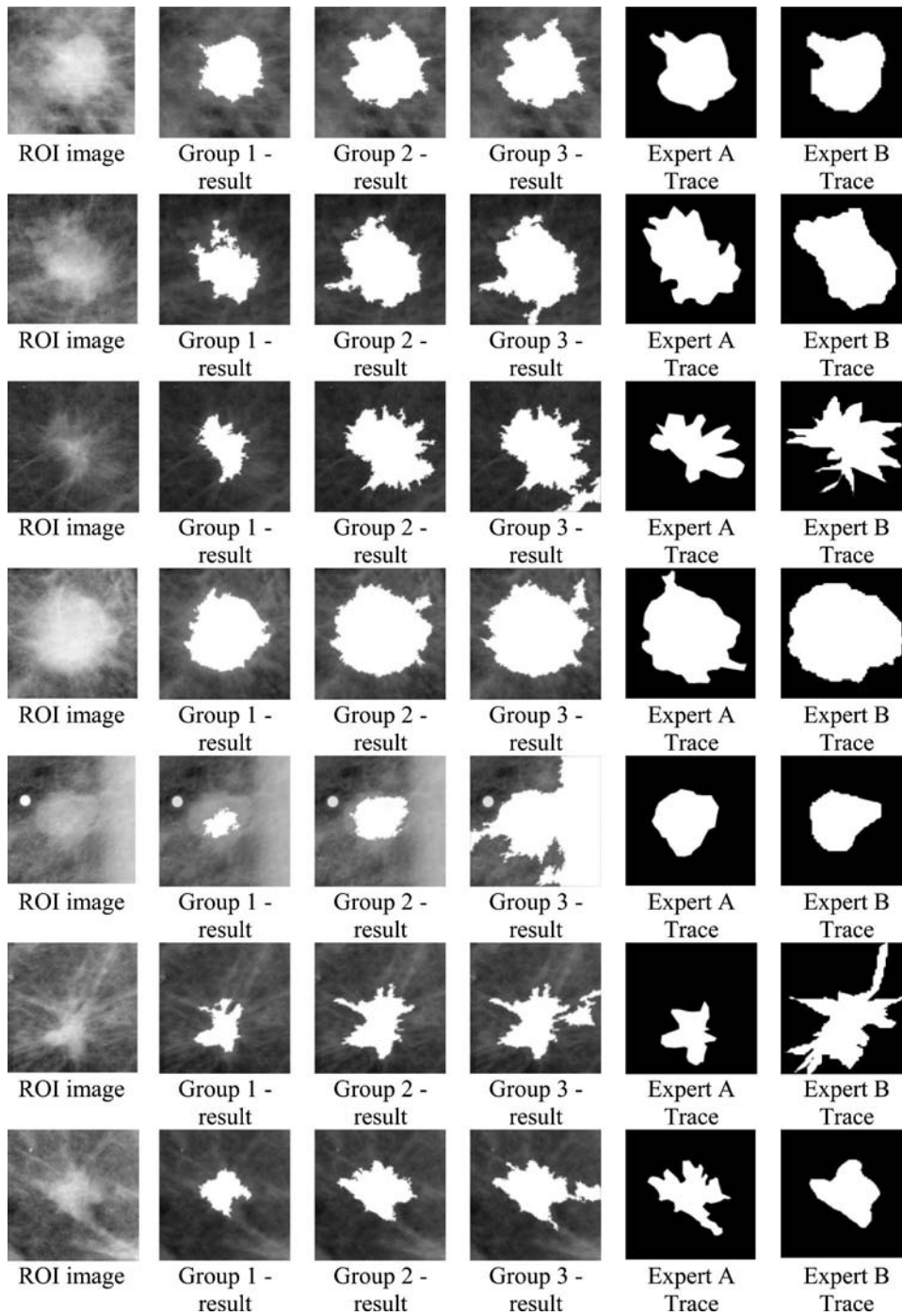


FIG. 7. Segmentation results for a set of malignant masses.

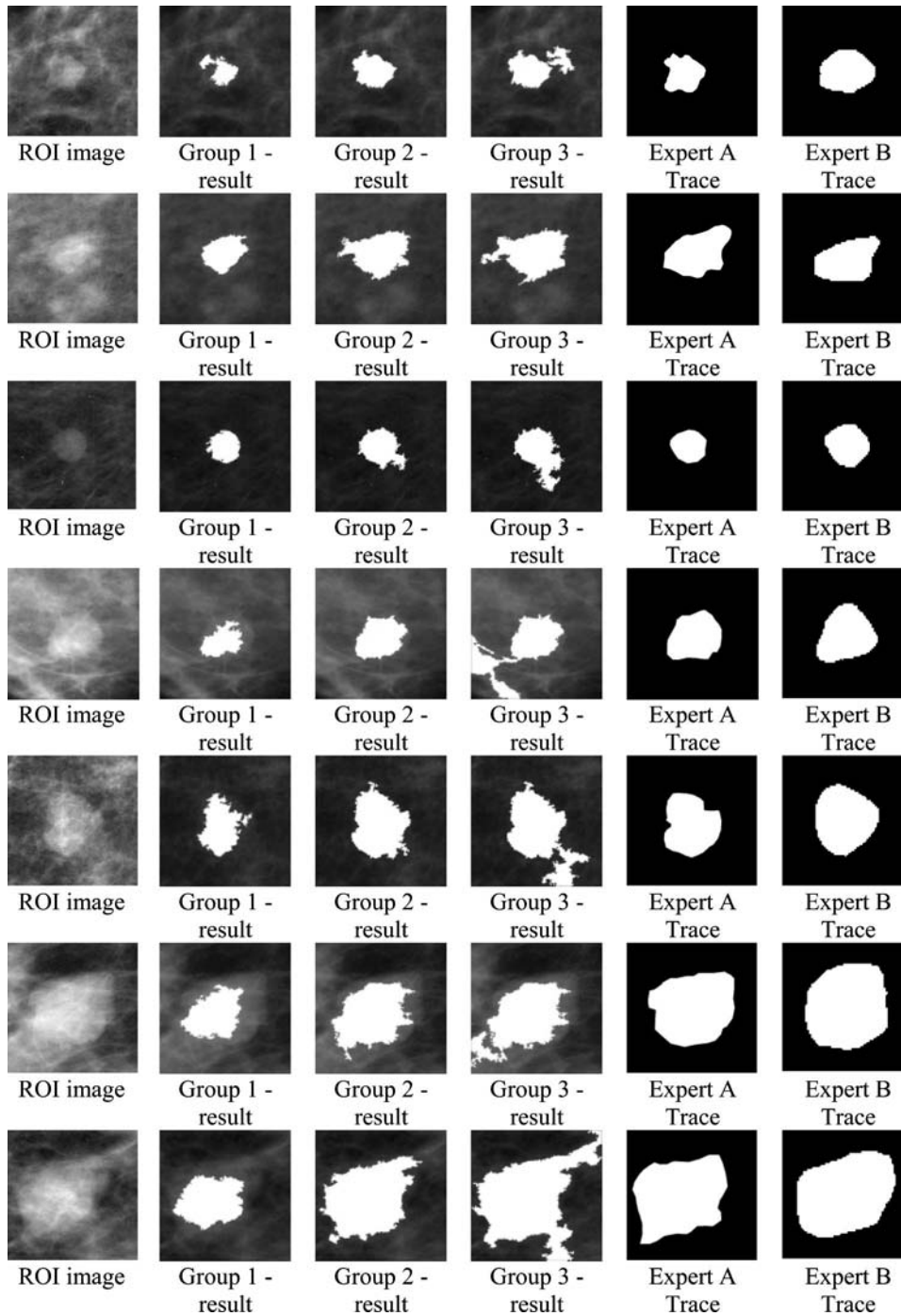


FIG. 8. Segmentation results for a set of benign masses.

^{a)} Author to whom correspondence should be addressed. Dr. Shih-Chung B. Lo, ISIS Center, Department of Radiology, Georgetown University, Box 571479, Washington, DC 20057-1479. Electronic mail: lo@isis.imac.georgetown.edu

¹ J. V. Lacey, Jr., S. S. Devesa, and L. A. Brinton, "Recent trends in breast cancer incidence and mortality," *Environ. Mol. Mutagen.* **39**, 82–88 (2002).

² J. E. Meyer, D. B. Kopans, P. C. Stomper, and K. K. Lindfors, "Occult breast abnormalities: percutaneous preoperative needle localization," *Radiology* **150**, 335–337 (1984).

³ A. L. Rosenberg, G. F. Schwartz, S. A. Feig, and A. S. Patchefsky, "Clinically occult breast lesions: localization and significance," *Radiology* **162**, 167–170 (1987).

⁴ B. C. Yankaskas, M. H. Knelson, M. L. Abernethy, J. T. Cuttino, and R.

L. Clark, "Needle localization biopsy of occult lesions of the breast," *Radiology* **23**, 729–733 (1988).

⁵ J. A. Harvey, L. L. Fajardo, and C. A. Innis, "Previous mammograms in patients with impalpable breast carcinoma: retrospective vs. blinded interpretation," *Am. J. Roentgenol., Radium Ther. Nucl. Med.* **161**, 1167–1172 (1993).

⁶ J. E. Martin, M. Moskowitz, and J. R. Milbrath, "Breast cancer missed by mammography," *Am. J. Roentgenol., Radium Ther. Nucl. Med.* **132**, 737–739 (1979).

⁷ J. R. Harris, M. E. Lippman, M. Morrow, and S. Hellman, *Diseases of the Breast* (Lippincott-Raven Publishers, Philadelphia, 1996), pp. 80–81.

⁸ J. E. Martin, *Atlas of Mammography: Histologic and Mammographic Correlations*, 2nd ed. (Williams and Wilkins, Baltimore, 1988), p. 87.

⁹ G. M. te Brake and N. Karssemeijer, "Segmentation of suspicious den-

- sities in digital mammograms," *Med. Phys.* **28**, 259–266 (2001).
- ¹⁰L. Li, Y. Zheng, L. Zhang, and R. Clark, "False-positive reduction in CAD mass detection using a competitive classification strategy," *Med. Phys.* **28**, 250–258 (2001).
- ¹¹N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Trans. Med. Imaging* **15**, 59–67 (1996).
- ¹²S. Pohlman, K. A. Powell, N. A. Obuchowski, W. A. Chilcote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," *Med. Phys.* **23**, 1337–1345 (1996).
- ¹³A. J. Méndez, P. G. Tahoces, M. J. Lado, M. Souto, and J. J. Vidal, "Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms," *Med. Phys.* **25**, 957–964 (1998).
- ¹⁴H. Li, Y. Wang, K. J. R. Liu, S.-C. B. Lo, and M. T. Freedman, "Computerized radiographic mass detection—part I: lesion site selection by morphological enhancement and contextual segmentation," *IEEE Trans. Med. Imaging* **20**, 289–301 (2001).
- ¹⁵M. L. Comer, S. Liu, and E. J. Delp, "Statistical segmentation of mammograms," *Digital Mammography '96: proceedings of the 3rd international workshop on digital mammography*, Chicago, IL, pp. 475–478, 9–12 June 1996.
- ¹⁶M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans. Med. Imaging* **17**, 510–517 (1998).
- ¹⁷S.-C. B. Lo, H. Li, Y. Wang, L. Kinnard, and M. T. Freedman, "A multiple circular path convolution neural network system for detection of mammographic masses," *IEEE Trans. Med. Imaging* **21**, 150–158 (2002).
- ¹⁸W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," *IEEE Trans. Med. Imaging* **16**, 811–819 (1997).
- ¹⁹B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Med. Phys.* **28**, 1455–1465 (2001).
- ²⁰R. M. Rangayyan, N. M. El-Faramawy, J. E. Leo Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imaging* **16**, 799–810 (1997).
- ²¹B. Sahiner, H.-P. Chan, D. Wei, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.* **23**, 1671–1684 (1996).
- ²²J. Suckling, D. R. Dance, E. Moskovic, D. J. Lewis, and S. G. Blacker, "Segmentation of mammograms using multiple linked self-organizing neural networks," *Med. Phys.* **22**, 145–152 (1995).
- ²³B. Van Ginneken, "Automatic segmentation of lung fields in chest radiographs," *Med. Phys.* **27**, 2445–2455 (2000).
- ²⁴D. Downing and J. Clark, *Statistics the Easy Way*, 2nd ed. (Barron's Educational Series, Hauppauge, 1989), pp. 184–206.
- ²⁵W. Hopkins, *A New View of Statistics: P Values and Statistical Significance*; available online at www.sportsci.org/resource/stats/pvalues.html.
- ²⁶M. Heath *et al.*, "Current status of the digital database for screening mammography," *Digital Mammography* (Kluwer Academic, New York, 1998), pp. 457–460.