AD_____

AWARD NUMBER: DAMD17-03-1-0314

TITLE: Computer-Aided Detection of Mammographic Masses in Dense Breast Images

PRINCIPAL INVESTIGATOR: Lisa M. Kinnard, Ph.D.

CONTRACTING ORGANIZATION: Howard University
Washington, DC 20059

REPORT DATE: June 2006

TYPE OF REPORT: Annual Summary

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-06-2006	Annual Summary	1 Jun 2003 – 31 May 2006

4. TITLE AND SUBTITLE

Computer-Aided Detection of Mammographic Masses in Dense Breast Images

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0314

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Lisa M. Kinnard, Ph.D.

E-Mail: l_kinnard@yahoo.com

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Howard University
Washington, DC 20059

8. PERFORMING ORGANIZATION REPORT NUMBER

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

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This document describes the research tasks and educational activities in which the PI has been engaged during the second phase of this post-doctoral work. The PI has selected 300 dense breast masses for testing using an automated segmentation algorithm which combines region growing with cost function analysis. This method has been validated on all cases using overlap, accuracy, sensitivity, specificity, Dice Similarity Index, and kappa statistics by three expert radiologists. The method has also been compared to a second algorithm developed by a group at The Johns Hopkins University. The PI has engaged in a number of technical development activities, including attending meetings in her research area, engaging in oral presentations describing her path through graduate school and through her post-doctoral award, reviewing grants and journal submissions, learning proper interviewing techniques, and teaching CAD methods to wide audiences.

15. SUBJECT TERMS
image segmentation, validation methods, CAD

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	56	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Cover……1

SF 298……..……2

Introduction……..……....4

Body……4

Key Research Accomplishments……4, 10, 15

Reportable Outcomes……4, 10, 15

Conclusions……..16

References……17

Appendices……17

I. INTRODUCTION

Breast mass segmentation is arguably one of the most difficult tasks in the development of Computer-Aided Diagnostic (CAD_x) systems. The main objective of this research is to develop an image segmentation method for mammograms that contain dense tissue as well as for mammograms that contain dense/fatty tissue, while its second objective is to incorporate the segmentation method into a CAD_x system. Specifically, we intended to do the following: (1) To develop an automatic image segmentation scheme to separate clinically occult breast masses from surrounding tissue (2) To evaluate the method by comparing the ROIs with mammographers' drawings and (3) To separate masses from glandular tissues using the Multiple Circular Path Convolution Neural Network (MCPCNN) classifier. The following is a summary of the PI's research and training activities during the grant period.

II. BODY

During the past 36 months the PI has tested and validated an automatic image segmentation algorithm on a set of dense breast mass cases. This section of the final summary provides a detailed description of the research and training tasks on a year-by-year basis. Part A summarizes the activities that occurred during months 1-12, Part B summarizes the activities that occurred during months 13-24, and Part C summarizes the activities that occurred during months 25-36.

A. Year 1, Months 1-12

During the first year, the PI performed the initial database collection, coordinated ground truth tracing sessions with two expert radiologists, attended medical image conferences, attended local medical image meetings, and team taught an imaging technologies course at the Catholic University of America.

A.1 Key Research Accomplishments – Year 1

1. Expanded database to 300 images collected from Digital Database for Screening Mammography (DDSM)
   - Cases have American College of Radiology (ACR) density ratings of 3 and 4
   - Collected Georgetown University Medical Center (GUMC) data for expansion of current database

2. Tested current segmentation method on 198 images

3. Conducted expert radiologist trace sessions with first radiologist
   - first radiologist traced 298 masses
   - second radiologist has agreed to trace masses

A.2 Reportable Outcomes – Year 1

Manuscripts

1. Published manuscript in proceedings of International Symposium on Biomedical Imaging (ISBI) 2004 meeting: "Likelihood Function Analysis for Segmentation of Mammographic Masses for Various Margin Groups"

2. Submitted manuscript to Journal of Medical Physics: "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study"manuscript is currently undergoing 2nd review by editors

Oral Presentation

"Likelihood Function Analysis for Segmentation of Mammographic Masses for Various Margin Groups", ISBI Meeting, Arlington, VA

Technical Development Activities:

1. Attended two cancer imaging workshops conducted by the Washington Academy of Biomedical Engineering:
   - 11/12/03: "Cancer Imaging for the Operating Room of 2020" (Georgetown University)
   - 9/29/03: "Individualized Treatment Using Pharmaco-Genomics & Functional Imaging" (George Washington University)
2. Attended weekly cancer workshops conducted by the Howard University Cancer Center (made oral presentation in December of 2003)
3. Attended International Symposium on Biomedical Imaging (ISBI) 2004 meeting
4. Attended SPIE Medical Imaging Meeting
5. Taught "Computer-Aided Diagnosis" portion of "Introduction to Imaging Technologies" course, The Catholic University of America, course number ENGR552

B. Year 2, Months 13-24

During the second year the PI has tested and validated an automatic image segmentation algorithm on a set of dense breast mass cases for both non-processed and background trend corrected images. The following is a detailed description of the experiments and is divided into the following sections (B.1) Experiments: (B.1.1) Segmentation Method – an overview of the automated image segmentation method (please see Appendix for detailed description of method) (B.1.2) Database and Experiments – description of masses used and experiments performed (B.1.3) Results – statistical and graphical results of the experiment and (B.1.4) Discussion of Results; (B.2) Key Research Accomplishments; and (B.3) Reportable Outcomes.

B.1 Experiments

B.1.1 Segmentation Method

The segmentation method used in this study evaluates the steepest changes within a probabilistic cost function in an effort to determine the computer segmented contour which is most closely correlated with expert radiologist manual traces. It segments breast masses by combining region growing with the analysis of a probability-based function [1]. Once a set of contours is grown using region growing the probability density functions inside and outside the contours are found. A function, which is the logarithm of these probability density functions, is then constructed. The function is then searched for possible steep change locations, i.e., sharp changes in the logarithm values, and the intensities corresponding to those locations are likely to produce contours which are highly correlated with expert traces. A detailed description of the method is provided in the manuscripts located in the appendix of this document [2, 3].

B.1.2 Database and Experiments

Three-hundred forty-two cases have been selected from the University of South Florida's Digital Database for Screening Mammography (DDSM) [2], where 175 of these cases are cancerous masses and 167 of the cases are benign masses. The densities of all cases from the DDSM have been rated according to the American College of Radiology's (ACR) density scale, which ranges from 1-4. A breast containing a great deal of fatty tissue would receive a rating of 1 and a breast containing a great deal of dense tissue would receive a rating of 4. The current database contains 242 cases with a density rating of 3 and 100 cases with a density rating of 4. In the current experiment the cost likelihood function threshold values (TV_1 and TV_2) were set to 1800 and 1300, respectively. Approximately 300 of the cases were manually traced by two expert radiologists. All cases have been validated by both radiologists, where the validation measures are overlap, accuracy, sensitivity, specificity, Dice Similarity Index (DSI), and kappa statistics as described in the literature [3,4] and manuscripts [5-7]. Initially, the images were not pre-processed in order to preserve the true mass borders. In hopes of attaining higher validation statistical values, the PI applied the background trend correction technique to the entire dataset and ran a second segmentation experiment on the pre-processed images.

B.1.3. Results

Statistical Results

Tables 1-4 contain p-values for Analysis of Variance (ANOVA) tests, in which a set of intra-observer experiments were performed to determine the value of pre-processing on segmentation results. Specifically, the PI tested non-processed versus pre-processed datasets for all statistical measures, and both expert radiologists. A table entry containing "NS" implies that there were no statistically significant differences for a particular test. The computer produces the three traces which it feels are the closest contours to those traced by the expert radiologists, so the results shown in the table contain results for tests for all three groups. Further, the maximum values of statistical measures for a subset of cancer cases were found to find the proximity between the optimal region-growing trace as determined by the computer and the region-growing trace with the highest possible value for a particular measure.

Table 1 – ANOVA test P-values for Intra-observer Experiment: Non-Processed vs. Pre-Processed Cancer Cases (Expert A)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	$2.2\text{x}10^{-6}$	NS	$1.4\text{x}10^{-6}$	$3.4\text{x}10^{-3}$	$4.5\text{x}10^{-7}$	$1.4\text{x}10^{-3}$
Group 2 Trace	$4.0\text{x}10^{-4}$	NS	$1.3\text{x}10^{-5}$	$3.8\text{x}10^{-6}$	$9.4\text{x}10^{-5}$	$3.5\text{x}10^{-2}$
Group 3 Trace	$4.3\text{x}10^{-6}$	NS	$1.5\text{x}10^{-5}$	$2.7\text{x}10^{-4}$	$1.1\text{x}10^{-5}$	$2.8\text{x}10^{-2}$

Table 2 – ANOVA test P-values for Intra-observer Experiment: Non-Processed vs. Pre-Processed Benign Cases (Expert A)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	$1.37\text{x}10^{-6}$	NS	$2.0\text{x}10^{-6}$	NS	$3.8\text{x}10^{-7}$	$2.9\text{x}10^{-5}$
Group 2 Trace	$2.2\text{x}10^{-3}$	NS	$1.6\text{x}10^{-5}$	$3.4\text{x}10^{-4}$	$4.9\text{x}10^{-4}$	$1.5\text{x}10^{-2}$
Group 3 Trace	NS	NS	$5.1\text{x}10^{-6}$	$4.6\text{x}10^{-5}$	NS	NS

Table 3 – ANOVA test P-values for Intra-observer Experiment: Non-Processed vs. Pre-Processed Cancer Cases (Expert B)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	$3.5\text{x}10^{-5}$	NS	$2.0\text{x}10^{-6}$	$1.2\text{x}10^{-3}$	$1.1\text{x}10^{-5}$	$2.8\text{x}10^{-3}$
Group 2 Trace	NS	NS	$1.3\text{x}10^{-4}$	$6.4\text{x}10^{-8}$	$3.2\text{x}10^{-2}$	NS
Group 3 Trace	NS	$2.2\text{x}10^{-2}$	$7.0\text{x}10^{-4}$	$3.7\text{x}10^{-6}$	NS	NS

Table 4 – ANOVA test P-values for Intra-observer Experiment: Non-Processed vs. Pre-Processed Benign Cases (Expert B)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	$9.8\text{x}10^{-7}$	NS	$1.7\text{x}10^{-6}$	NS	$2.3\text{x}10^{-7}$	$9.0\text{x}10^{-6}$
Group 2 Trace	$1.8\text{x}10^{-3}$	NS	$4.1\text{x}10^{-6}$	$1.3\text{x}10^{-4}$	$3.9\text{x}10^{-4}$	$6.8\text{x}10^{-3}$
Group 3 Trace	NS	NS	$3.7\text{x}10^{-7}$	$1.2\text{x}10^{-5}$	NS	NS

Table 5 – Mean Statistical Values Non-Processed Cases: Expert A, Cancer Cases

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	0.18	0.72	0.18	1.0	0.27	0.22
Group 2 Trace	0.34	0.76	0.37	0.997	0.47	0.39
Group 3 Trace	0.36	0.76	0.46	0.95	0.51	0.40

Table 6 – Mean Statistical Values Non-Processed Cases: Expert B, Cancer Cases

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	0.36	0.81	0.39	0.97	0.50	0.42
Group 2 Trace	0.50	0.84	0.63	0.92	0.64	0.54
Group 3 Trace	0.47	0.81	0.70	0.86	0.62	0.50

Table 7 – Mean Statistical Values Pre-Processed Cases: Expert A, Cancer Cases

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	0.17	0.72	0.18	1.0	0.27	0.22
Group 2 Trace	0.34	0.76	0.37	0.99	0.47	0.39
Group 3 Trace	0.36	0.75	0.46	0.95	0.51	0.40

Table 8 – Mean Statistical Values Pre-Processed Cases: Expert B, Cancer Cases

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**	**kappa**
Group 1 Trace	0.25	0.83	0.26	1.0	0.37	0.33
Group 2 Trace	0.45	0.86	0.49	0.99	0.57	0.53
Group 3 Trace	0.43	0.84	0.59	0.94	0.58	0.51

Table 9 – Mean Values for Contour Yielding Maximum Value vs. Computer Choice Contours

	Mean Maximum Overlap Value	Mean Group 1 Overlap Value	Mean Group 2 Overlap Value	Mean Group 3 Overlap Value
Expert A	0.62	0.28	0.45	0.48
Expert B	0.60	0.47	0.50	0.36

Visual Results

Figures 1-4 show segmentation results for both the pre-processed and non-processed mass cases

.

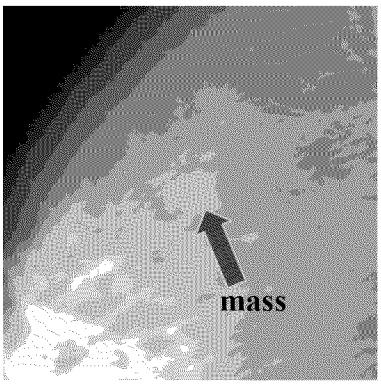


Figure 1a – Original Image
(Cancer Case, Density=3)

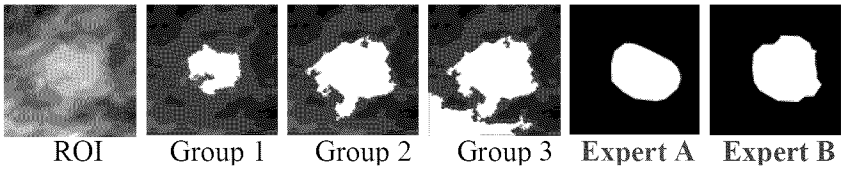


Figure 1b – Cropped original With Computer Results (Non-Processed Image)

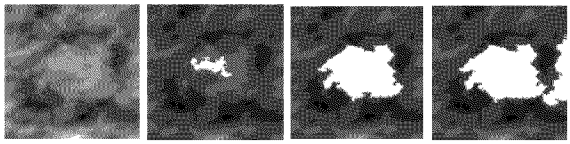


Figure 1c – Cropped original With Computer Results (Pre-Processed Image)

Figure 1: Computer Segmentation Results for a Cancerous Mass

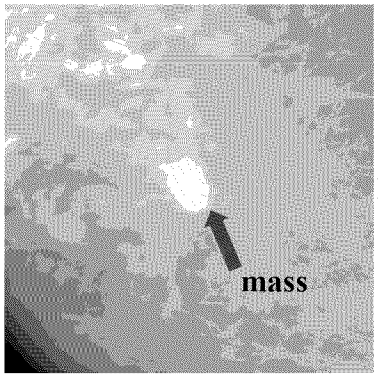


Figure 2a – Original Image
(Cancer Case, Density=3)

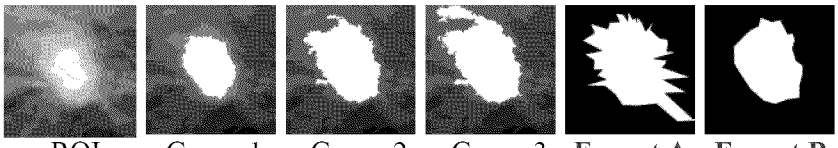


Figure 2b – Cropped original With Computer Results (Non-Processed Image)

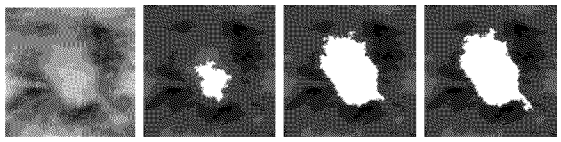


Figure 2c – Cropped original With Computer Results (Pre-Processed Image)

Figure 2: Computer Segmentation Results for a Cancerous Mass

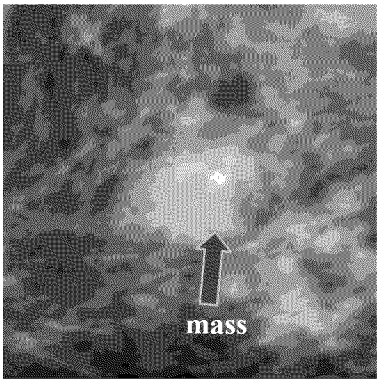
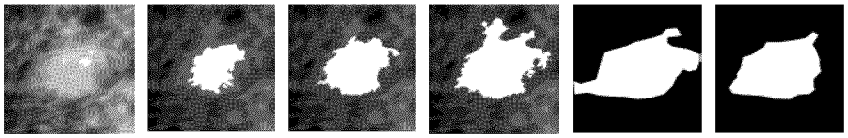
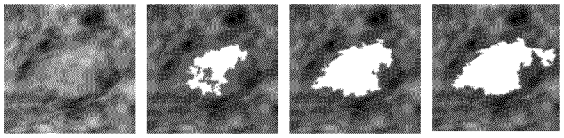


Figure 3a – Original Image
Benign Case, Density=3)



ROI Group 1 Group 2 Group 3 **Expert A** **Expert B**

Figure 3b – Cropped original With Computer Results (Non-Processed Image)



ROI Group 1 Group 2 Group 3

Figure 3c – Cropped original With Computer Results (Pre-Processed Image)

Figure 3: Computer Segmentation Results for a Benign Mass

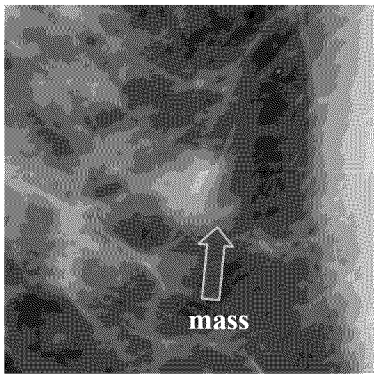
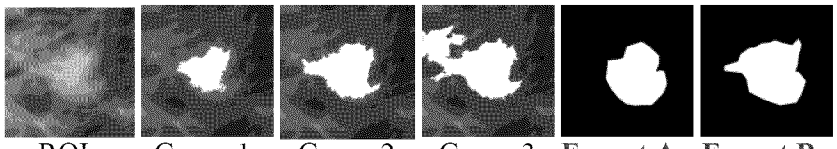
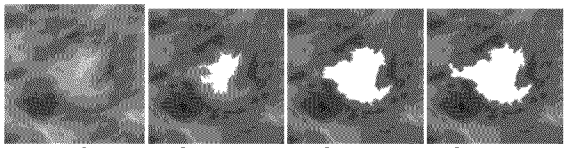


Figure 4a – Original Image
Benign Case, Density=3)



ROI Group 1 Group 2 Group 3 **Expert A** **Expert B**

Figure 4b – Cropped original With Computer Results (Non-Processed Image)



ROI Group 1 Group 2 Group 3

Figure 4c – Cropped original With Computer Results (Pre-Processed Image)

Figure 4: Computer Segmentation Results for a Benign Mass

B.1.4. Discussion of Results

It has been observed that the segmentation algorithm produces better results using the non-processed images as inputs rather than using the pre-processed images as inputs, under the given set of parameters. As stated previously, the intensity corresponding to the location where the steep likelihood changes occur is likely to produce the contour that matches closely with the expert radiologist traces. The steep change location is determined by a set of threshold values determined by the user. The background trend correction process generally causes dark areas in the image to become darker, therefore, the contrast between the mass and background is higher for some cases. This, in turn creates more steep changes in the likelihood functions, which may have formerly been smooth. Therefore, the computer is likely to choose higher intensity values, consequently the contours will be small.

The ANOVA test results show that there were statistically significant differences between the non-processed and pre-processed images for both expert radiologists, for most statistics, where the mean values were higher for non-processed vs. pre-processed images for most statistics. These results imply

that it may not be necessary to pre-process the images, but rather to use different parameters for the automated selection process of finding optimal contours. Preliminary work has been done to determine how close the statistical values of the computer chosen contours are to those of the contours which obtain the greatest statistical values (see Table 9).

B.2. Key Research Accomplishments

1. Completed expert radiologist tracing of 300 masses
2. Tested the efficacy of background trend correction upon segmentation improvement
3. Added Dice Similarity Index (DSI) and kappa statistics as validation measures
4. Validated masses using all validation measures
5. Reviewed literature concerning inter-observer variability

B.3 Reportable Outcomes

Manuscripts:

1. L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study," *J. of Medical Physics*, vol. 31, no. 10, 2004, pp. 2796-2810.
2. L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study," *Virtual Journal of Biophysics*, Vol. 8, Issue 7, Oct. 1, 2004, http://www.vjbio.org/bio/ (selected across several medical and biophysics journals).
3. L. Kinnard, S.-C. B. Lo, E. Duckett, E. Makariou, M.T. Freedman, and M. Chouikha, "Mass Segmentation of Dense Breasts on Digitized Mammograms: Analysis of probability-based function," *Medical Imaging 2005: Image Processing, February, 2005, Proceedings of SPIE,* vol. 5747, pp. 1813-1823.

Poster Presentation:

1. L. Kinnard, S.-C. B. Lo, E. Duckett, E. Makariou, M.T. Freedman, and M. Chouikha, "Mass Segmentation of Dense Breasts on Digitized Mammograms: Analysis of probability-based function," *Medical Imaging 2005: Image Processing, February, 2005, Proceedings of SPIE,* vol. 5747, pp. 1813-1823.

Oral Presentations:

1. "The Post-Doctoral Experience: A Year in Review", *Preparing for the Postdoctoral Institute*, August, 2004, Howard University and The University of Texas at El Paso.
2. "Computer-Aided Diagnosis and Image Segmentation of Mammographic Masses", *Symposium on Translational Research for Cancer Detection, Diagnosis, Prevention, and Treatment*, The Howard University Cancer Center and the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, November, 2004.

Technical Development Activities:

1. Attended meetings and one workshop of the Washington Academy of Biomedical Engineering (WABME)
2. Attended cancer workshops conducted by the Howard University Cancer Center
3. Attended SPIE Medical Imaging Meeting (February, 2005, San Diego, CA)
4. Served as the Faculty Retreat Committee Chair, for which the theme was a grant proposal writing contest. The PI also served as the PI of her group, and the group placed 2nd out of eight groups.

C. Year 3 – Months 25-36

During the final year of the grant period, the PI performed several dense breast segmentation experiments, attended several conferences and meetings, gave oral presentations to graduate students, and interviewed for various research and teaching positions. Section (C.1.1) describes the experiments performed during the final year, section (C.1.2) gives results for these experiments, section (C.1.3) provides a discussion of results, section (C.2) lists key research accomplishments, and section (C.3) lists reportable outcomes. The segmentation algorithm and image database have been described in sections B.1.1-B.1.2.

C.1. Experiments

C.1.1. Experiment Descriptions

For all tables in this section, a table entry containing the abbreviation "NS" means "No Significant" difference, so there was no statistically significant difference for a particular test. All tables contain intra-observer experiments, or, comparisons between the computer traces and two expert radiologists, namely, Expert A and Expert B. The probabilistic-likelihood method narrows a set of 200-500 traces to a set of three possible choices that will best match the radiologist traces, namely, group 1, group 2, and group 3 traces. Typically, the group 1 trace encapsulates the mass body, the group 2 trace encapsulates the mass body + the mass borders that extend into surrounding fibroglandular tissue, and the group 3 trace encapsulates the mass body + the mass borders that extend into surrounding fibroglandular tissue + additional tissue that may not belong to the mass.

Experiment 1

During the second year of the grant period the PI began an experiment which compared the segmented results to the maximum achievable values for each validation statistic, namely, the overlap, accuracy, sensitivity, specificity, and Dice Similarity Index (DSI) statistics. Tables 10-17 contain results for these experiments.

Experiment 2

In previous studies the PI and colleagues determined that the computer algorithm was capable of narrowing a set of 200-500 possible contour traces to the trace which would closely match manual ground truth traces provided by expert radiologists. In the case of dense breast masses this optimal trace is more difficult to determine due to the masses' unclear borders, therefore the set of 200-500 possible contour traces were narrowed to two possible optimal traces. The PI added yet a third expert radiologist trace to see if this person could serve as a "tie-breaker", and would therefore strongly agree with Expert A or Expert B. The details of this experiment can be found in the PI's submission to the ISBI 2006 conference, located in the appendix of this document.

Experiment 3

In a third experiment the PI compared the probabilistic-likelihood method (the algorithm used throughout the research study) to a Gradient Vector Flow (GVF) algorithm developed by a research group at The Johns Hopkins University. The details of the GVF algorithm are described in a summary which is a portion of a manuscript comparing the two algorithms to be submitted to the Journal of Physics and Medicine in Biology. Tables 18-25 contain the results of this third experiment.

C.1.2. Results

Maximum Value Experiment Results

Cancerous Mass Case Results

Table 10 – ANOVA test P-values:
Max Values vs. Computer Choice Cancer Cases (Expert A)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	5.0×10^{-12}	5.3×10^{-4}	4.1×10^{-31}	NS	2.3×10^{-11}
Group 2 Trace	4.6×10^{-3}	4.4×10^{-2}	1.4×10^{-14}	NS	5.2×10^{-3}
Group 3 Trace	2.7×10^{-4}	2.1×10^{-3}	7.4×10^{-11}	6.3×10^{-5}	4.7×10^{-4}

Table 11 – Mean Values of Computer Choice and Max Value
Statistical Measurements (Expert A, Cancer Cases)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	0.31	0.75	0.34	0.97	0.45
Group 2 Trace	0.49	0.80	0.58	0.93	0.63
Group 3 Trace	0.48	0.79	0.65	0.88	0.63
Max Values	0.60	0.88	0.92	0.97	0.73

Table 12 – ANOVA test P-values:
Max Values vs. Computer Choice Cancer Cases (Expert B)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	3.85×10^{-8}	4.1×10^{-3}	2.9×10^{-30}	NS	9.3×10^{-8}
Group 2 Trace	NS	NS	3.1×10^{-14}	NS	NS
Group 3 Trace	4.4×10^{-3}	1.4×10^{-3}	4.4×10^{-12}	1.2×10^{-4}	6.9×10^{-3}

Table 13 – Mean Values of Computer Choice and Max Value
Statistical Measurements (Expert B, Cancer Cases)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	0.37	0.81	0.41	0.96	0.51
Group 2 Trace	0.52	0.85	0.65	0.92	0.67
Group 3 Trace	0.49	0.82	0.72	0.87	0.64
Max Values	0.60	0.88	0.95	0.96	0.73

Benign Mass Case Results

Table 14 – ANOVA test P-values:
Max Values vs. Computer Choice Benign Cases (Expert A)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	3.8×10^{-10}	3.5×10^{-3}	1.4×10^{-30}	NS	1.2×10^{-9}
Group 2 Trace	1.0×10^{-2}	3.3×10^{-2}	9.6×10^{-15}	1.6×10^{-3}	7.5×10^{-3}
Group 3 Trace	4.2×10^{-4}	1.8×10^{-4}	1.9×10^{-10}	2.8×10^{-9}	2.8×10^{-4}

Table 15 – Mean Values of Computer Choice and Max Value
Statistical Measurements (Expert A, Benign Cases)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	0.35	0.83	0.39	0.99	0.48
Group 2 Trace	0.50	0.86	0.62	0.95	0.64
Group 3 Trace	0.50	0.83	0.74	0.88	0.64
Max Values	0.60	0.90	0.97	0.99	0.74

Table 16 – ANOVA test P-values:
Max Values vs. Computer Choice Benign Cases (Expert B)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	$3.8x10^{-10}$	$3.5x10^{-3}$	$1.4x10^{-30}$	NS	$1.2x10^{-9}$
Group 2 Trace	$1.0x10^{-2}$	$3.8x10^{-3}$	$9.6x10^{-15}$	$1.6x10^{-3}$	$7.5x10^{-3}$
Group 3 Trace	$4.2x10^{-4}$	$1.8x10^{-4}$	$1.7x10^{-3}$	$2.8x10^{-9}$	$2.8x10^{-4}$

Table 17 – Mean Values of Computer Choice and Max Value
Statistical Measurements (Expert B, Benign Cases)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
Group 1 Trace	0.35	0.83	0.39	0.99	0.48
Group 2 Trace	0.50	0.86	0.62	0.95	0.64
Group 3 Trace	0.50	0.83	0.74	0.88	0.64
Max Values	0.60	0.90	0.97	0.99	0.74

Probabalistic-Likelihood Algorithm vs. GVF Algorithm Results
Cancerous Mass Case Results

Table 18: Probabalistic-Likelihood Algorithm vs. GVF Algorithm Results, Cancer Cases (Expert A)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
GVF vs. group 1	NS	NS	NS	NS	NS
GVF vs. group 2	$5.92x10^{-11}$	0.02	$1.72x10^{-12}$	$1.09x10^{-05}$	$4.1x10^{-10}$
GVF vs. group 3	$5.37x10^{-15}$	0.02	$9.72x10^{-22}$	$5.17x10^{-12}$	$8.59x10^{-15}$

Table 19: Mean Values of Statistical Measurements (Expert A, Cancer Cases)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
GVF	0.27	0.70	0.29	0.99	0.41
group 1	0.27	0.70	0.29	0.98	0.40
group 2	0.45	0.76	0.52	0.94	0.59
group 3	0.46	0.75	0.60	0.89	0.62

Table 20: Probabalistic-Likelihood Algorithm vs. GVF Algorithm Results, Cancer Cases (Expert B)

	Overlap	**Accuracy**	**Sensitivity**	**Specificity**	**DSI**
GVF vs. group 1	NS	NS	NS	NS	NS
GVF vs. group 2	$3.28x10^{-10}$	NS	$3.28x10^{-15}$	$8.94x10^{-08}$	$7.07x10^{-09}$
GVF vs. group 3	$3.04x10^{-08}$	NS	$1.43x10^{-23}$	$8.85x10^{-18}$	$1.1x10^{-07}$

Table 21: Mean Values of Statistical Measurements (Expert B, Cancer Cases)

	Overlap	Accuracy	Sensitivity	Specificity	DSI
GVF	0.35	0.82	0.38	0.98	0.50
group 1	0.36	0.81	0.39	0.97	0.50
group 2	0.51	0.84	0.64	0.91	0.65
group 3	0.48	0.81	0.71	0.86	0.63

Benign Mass Case Results

Table 22: Probabalistic-Likelihood Algorithm vs. GVF Algorithm Results, Benign Cases (Expert A)

	Overlap	Accuracy	Sensitivity	Specificity	DSI
GVF vs. group 1	NS	NS	NS	NS	NS
GVF vs. group 2	2.23×10^{-07}	NS	1.05×10^{-09}	1.7×10^{-05}	5.03×10^{-06}
GVF vs. group 3	6.6×10^{-07}	NS	1.48×10^{-22}	8.62×10^{-17}	3.73×10^{-06}

Table 23: Mean Values of Statistical Measurements (Expert A, Benign Cases)

	Overlap	Accuracy	Sensitivity	Specificity	DSI
GVF	0.34	0.82	0.37	0.99	0.49
group 1	0.32	0.81	0.34	0.99	0.45
group 2	0.48	0.84	0.57	0.96	0.61
group 3	0.47	0.80	0.71	0.86	0.61

Table 24: Probabalistic-Likelihood Algorithm vs. GVF Algorithm Results, Benign Cases (Expert B)

	Overlap	Accuracy	Sensitivity	Specificity	DSI
GVF vs. group 1	NS	NS	NS	NS	NS
GVF vs. group 2	4.29×10^{-08}	NS	4.75×10^{-11}	6.84×10^{-06}	1.84×10^{-06}
GVF vs. group 3	1.48×10^{-05}	0.02	5.93×10^{-25}	4.97×10^{-18}	6.93×10^{-05}

Table 25: Mean Values of Statistical Measurements (Expert B, Benign Cases)

	Overlap	Accuracy	Sensitivity	Specificity	DSI
GVF	0.37	0.85	0.41	0.99	0.52
group 1	0.35	0.84	0.37	0.99	0.49
group 2	0.52	0.87	0.63	0.99	0.65
group 3	0.48	0.82	0.77	0.85	0.62

C.1.3 Discussion of Results

For the maximum value experiment (Experiment 1) there were statistically significant differences for Expert A for nearly all statistical measurements, and for all three group traces. This means that according to Expert A, there is more work that needs to be done. This was the case for both cancerous and benign masses. However for Expert B there were statistically significant differences for the group 1 and group 3 traces, but only one statistically significant difference (occurred for sensitivity) for the group 2 trace. This result is encouraging because it reveals that for the group 2 trace, while the values of the statistical measurements are lower than the maximum achievable values, the values are not significantly lower than the maximum achievable values. This was the case for the cancerous masses but not for the benign masses.

For the Probablistic-Likelihood vs. GVF experiment there were no statistically significant differences between the GVF trace and the group 1 trace for all statistical measurements. This is an expected result because the GVF traces had a tendency to be small, and the group 1 traces were also small because they typically encapsulated the mass body, which is also a small area. This result was consistent between observers and for both cancerous and benign masses. There were statistically significant differences for the group 2 traces vs. GVF traces, and for the group 3 traces vs. GVF traces for all statistical measurements except for the accuracy measurement. The mean values for the probabilistic-likelihood method were consistently higher than those of the GVF method.

C.2 Key Research Accomplishments

1. Compared probabilistic-likelihood trace choices to traces for which the statistical measurements had maximum values
2. Added a third observer to attempt to find a consensus among observers
3. Compared probabilistic-likelihood algorithm to GVF algorithm
4. Performed study which analyzed inter-observer variability, using the STAPLE algorithm (results do not appear in this document, but will appear in the manuscript)

C.3 Reportable Outcomes

Conferences and Meetings:

1. Intercultural Cancer Council Annual Meeting, April 2006
2. Southern Regional Education Board (SREB) Compact for Faculty Diversity, October 2005
3. 134th Meeting of the Cancer Advisory Board, June 2005
4. Department Of Defense CDMRP-Howard University Reverse Site Visit Meeting, April 2006

Technical and Professional Development Activities:

1. Associate Editor (Referee) for Journal of Medical Physics submission
2. Served on National Science Foundation (NSF) grant panel
3. Attended Georgetown University Post-doctoral meeting: Finding, Writing, and Husbanding Research Grants, by Bill Sansalone
4. Taught Computer Aided Detection and Diagnosis portion of "Biomedical Device Discovery & Development" course taught at the Food and Drug Administration (FDA) Staff College, Gaithersburg, MD, Fall, 2005.
5. Served as a judge for the University of Maryland College Park (UMCP) - University of Maryland Baltimore County (UMBC) AGEP conference

Poster Presentation:

"Mass Segmentation on Dense Breasts on Digitized Mammograms", L. Kinnard, S.-C. B. Lo, E. Duckett, E. Makariou, M.T. Freedman, and M. Chouikha, *Department of Defense Era-Of-Hope Meeting*, June, 2005, Philadelphia, PA.

Oral Presentations:

1. "Key Components for a Successful Post-Doc", *Preparing for the Postdoctoral Institute*, August, 2005, Howard University and The University of Texas at El Paso.
2. "Educational Paths and Decisions: The Road Less Traveled", The University Of Iowa College of Engineering's Ethnic Inclusion Seminar Series, November, 2005
3. "Educational Paths and Decisions: The Road Less Traveled", North Carolina State University, Department of Statistics, February, 2006

Grant Proposals Submitted:

1. American Cancer Society Mentored Research Scholar Grant in Applied and Clinical Research: Research Proposals Directed at Poor and Underserved Populations
   - Title: "Breast Cancer Diagnostic Image Querying System for Minority Women"
   - Initial submission date: 4/1/05; re-submission date 10/15/05
2. National Institutes of Health (NIH) National Cancer Institute (NCI) Mentored Career Development Award for Underrepresented Minorities (K01)
   - Title: "A Content-Based Image Retrieval System for Breast Masses: General and Minority Populations"
   - Initial submission date: 6/1/05
3. NIH Cancer Bioinformatics Grid (CaBIG) Imaging Group: co-wrote this proposal with colleagues; proposal was accepted

Interviews:

1. U.S. Patent and Trademark Office (CAD group): Patent Investigator – received an offer
2. Philips (CAD group): Research and Development Engineer
3. Food and Drug Administration (FDA)/NIH: Research Fellow - received and accepted an offer
   - This position is a joint relationship between The FDA's Center for Devices and Radiological Health (CDRH) Division of Imaging and Applied Mathematics (DIAM) and the NIH's National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the NCI. The PI will study the effect of drug treatment upon lung cancer tumors using statistical area measurements. Furthermore the PI hopes to continue work in Breast CAD because there are other researchers within the DIAM group who have ongoing projects in this area.
4. Temple University: Assistant Professor
5. Morgan State University: Assistant Professor

Manuscripts:

The Probabilistic Likelihood and Gradient Vector Flow Algorithms: A Comparison Study for Dense Breast Mass Segmentation (In preparation for submission to Physics and Medicine in Biology)

III. CONCLUSIONS

The initial research question for the maximum value experiment was: Are the computer choice statistical values significantly lower than the maximum achievable values given by region growing? According to Expert B, the answer is yes for group 1 and 3 traces but no for group 2 trace, for cancer cases. This result is encouraging because it means that it may possible to conclude that the group 2 trace is the optimal choice of the possible 200-500 contour choices per mass. The initial research question for the probabilistic-likelihood vs. GVF experiment was: Are there statistically significant differences between the two methods for a set of statistical measurements, and if so, which method achieves better results? We proved with statistical significance that for the current data set the probabilistic-likelihood method performed better. The GVF method worked very well for contours that were well-defined, however in our experiment it encountered difficulties for masses with ill-defined borders.

During this research phase of the award the PI gained a great appreciation for the difficulty of segmenting objects with ill-defined borders, and the importance of proper segmentation in the development of Computer-Aided Diagnostic systems. Since shape is such an important factor in diagnostic radiology proper segmentation is of paramount importance. During the technical and professional development phase of the award the PI gained immeasurable experience by attending meetings in her research area, taking on leadership roles in two activities, engaging in oral presentations describing her path through graduate school and through her post-doctoral award, reviewing grants and journal submissions, learning proper interviewing techniques, and teaching Computer-Aided Diagnostic techniques to audiences with a wide range of educational backgrounds. During the interview process the

post-doctoral experience was well-received by companies and universities alike, and the PI is greatly appreciative to have been given this opportunity. Fortunately, this award enabled her to continue work in the medical imaging field and to therefore continue the fight to reduce the cancer mortality rates all over the world.

IV. REFERENCES

1. M.A. Kupinski, M.L. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms", IEEE Trans. on Med. Imag., vol. 17, no. 4, 510-517 (1998).
2. M. Heath, K.W. Bowyer, D. Kopans et al., "Current status of the digital database for screening mammography", Digital Mammography, Kluwer Academic Publishers, 457-460 (1998).
3. J.R. Landis, G.G. Koch, "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, vol. 33, no. 1, pp. 159-174, 1977.
4. J.L. Fleiss, B. Levin, M.C. Paik, *Statistical Methods for Rates and Proportions*, Wiley, pp. 598-604, 2003.
5. L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Likelihood Function Analysis for Segmentation of Mammographic Masses for Various Margin Groups", Proceedings of the IEEE Symposium on Biomedical Imaging, 113-116 (2004).
6. L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study," *J. of Medical Physics*, vol. 31, no. 10, 2004, pp. 2796-2810.
7. L. Kinnard, S.-C. B. Lo, E. Duckett, E. Makariou, M.T. Freedman, and M. Chouikha, "Mass Segmentation of Dense Breasts on Digitized Mammograms: Analysis of probability-based function," *Medical Imaging 2005: Image Processing, February, 2005, Proceedings of SPIE,* vol. 5747, pp. 1813-1823.
8. C. Xu and J.L. Prince, "Snakes, Shapes, and Gradient Vector Flow", *IEEE Transactions on Image Processing, 359-369, March, 1998.*

V. APPENDIX

The appendix of this document contains mansuscripts written during the award period, the manuscript abstract for the DOD Era-Of-Hope meeting 2005, and a summary of the GVF algorithm.

MASS SEGMENTATION OF DENSE BREASTS ON DIGITIZED MAMMOGRAMS

L. Kinnard[1,2], S.-C. B. Lo[2], E. Duckett[3], E. Makariou[2], M.T. Freedman[2], and M. Chouikha[1]
[1]Department of Electrical and Computer Engineering, Howard University, Washington, D.C.
[2]ISIS Center, Georgetown University Medical Center, Washington, D.C.
[3]Advanced Radiology, Glen Burnie, MD
e-mail: kinnard@isis.imac.georgetown.edu

In this study a segmentation algorithm based on steepest changes of a probabilistic cost function is tested on non-processed and pre-processed dense breast images in an attempt to determine the efficacy of pre-processing for dense breast masses. The pre-processing method is a background trend correction (BTC) technique.

The segmentation method used in this study evaluates the steepest changes within a probabilistic cost function in an effort to determine the computer segmented contour which is most closely correlated with expert radiologist manual traces. This method segments breast masses by combining region growing with probability-based function analysis. Based on this analysis the three best contours are chosen and a final selection is made from these three choices. Typically, the Group 1 trace encapsulates the central portion of the mass, the Group 2 trace encapsulates the central mass and borders extending into surrounding tissue (e.g. – spiculations), and the Group 3 trace encapsulates the area covered by the Group 2 trace and surrounding fibroglandular tissue. The BTC method alters intensity values of the Region of Interest (ROI) using a polynomial fitting function. This method was tested on 71 dense cancerous masses. The computer-segmented results were manually traced by two expert radiologists for validation purposes. The overlap (O), accuracy (A), sensitivity (SE), specificity (SP), and Dice Similarity Coefficient (DSC) statistics were calculated, where a DSC value greater than 0.7 implies strong agreement between the computer segmented result and the expert radiologist trace. Tables 1-2 contain mean values for all statistics, and Figures 1-2 show computer segmented results.

Generally, the BTC method worsened the computer segmented results for Experts A and B regarding overlap, DSC, and sensitivity statistics. These results conflict with visual inspection of the BTC processed ROI's because this method sometimes creates a crater-like effect around the mass borders in areas where it was formerly difficult to separate mass borders from surrounding tissue. Further, some light areas are lightened by background trend correction which causes areas outside the mass to be joined with areas inside the mass. This phenomenon subsequently causes the region to grow too much. We feel that the computer-segmentation results can be improved by changing the parameters used to determine the intensities that will produce the contours that best match expert radiologist traces. The purpose of this work is to facilitate breast cancer screening using digitally automated segmentation method capable of locating mass borders embedded in dense breasts.

Table 1 – Statistical Results for Non-Processed and Processed ROI's (Expert A)

	Expert A (non-processed ROI)					**Expert A (BTC processed ROI)**				
	O	**A**	**SE**	**SP**	**DSC**	**O**	**A**	**SE**	**SP**	**DSC**
Group 1	0.3	0.73	0.32	0.98	0.44	0.18	0.71	0.19	1	0.28
Group 2	0.46	0.78	0.56	0.93	0.6	0.34	0.76	0.36	0.99	0.46
Group 3	0.47	0.77	0.63	0.88	0.64	0.34	0.75	0.44	0.95	0.49

Table 2 – Statistical Results for Non-Processed and Processed ROI's (Expert B)

	Expert A (non-processed ROI)					Expert A (BTC processed ROI)				
	O	**A**	**SE**	**SP**	**DSC**	**O**	**A**	**SE**	**SP**	**DSC**
Group 1	0.38	0.82	0.4	0.97	0.52	0.26	0.83	0.27	1	0.38
Group 2	0.52	0.84	0.65	0.91	0.66	0.44	0.86	0.49	0.99	0.57
Group 3	0.48	0.81	0.72	0.86	0.63	0.41	0.84	0.57	0.94	0.57

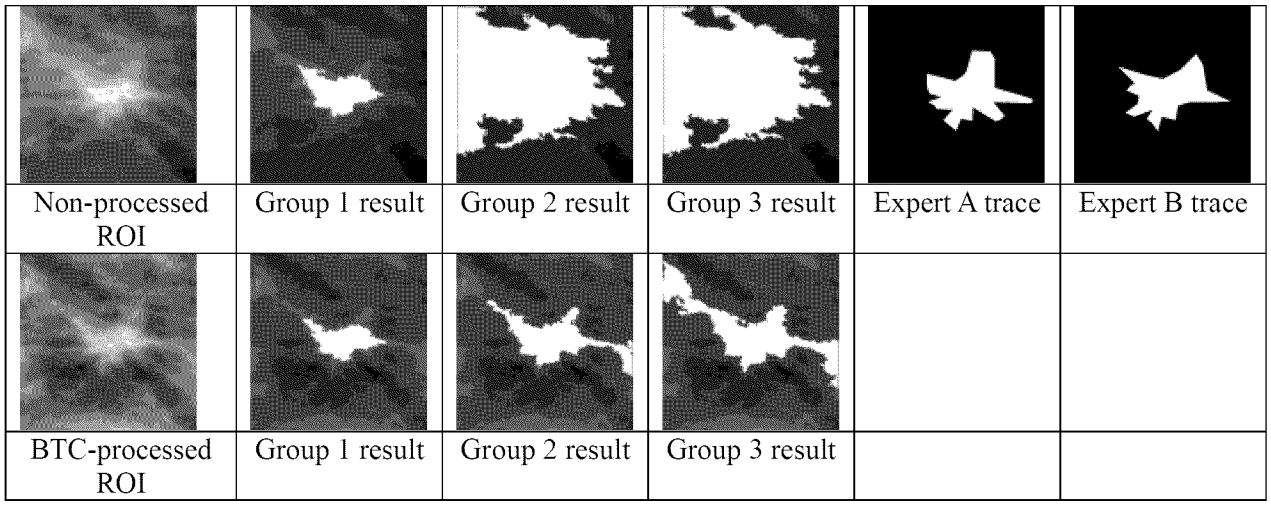


Figure 1 – A Cancerous Mass Showing Improved Results Due to BTC Processing

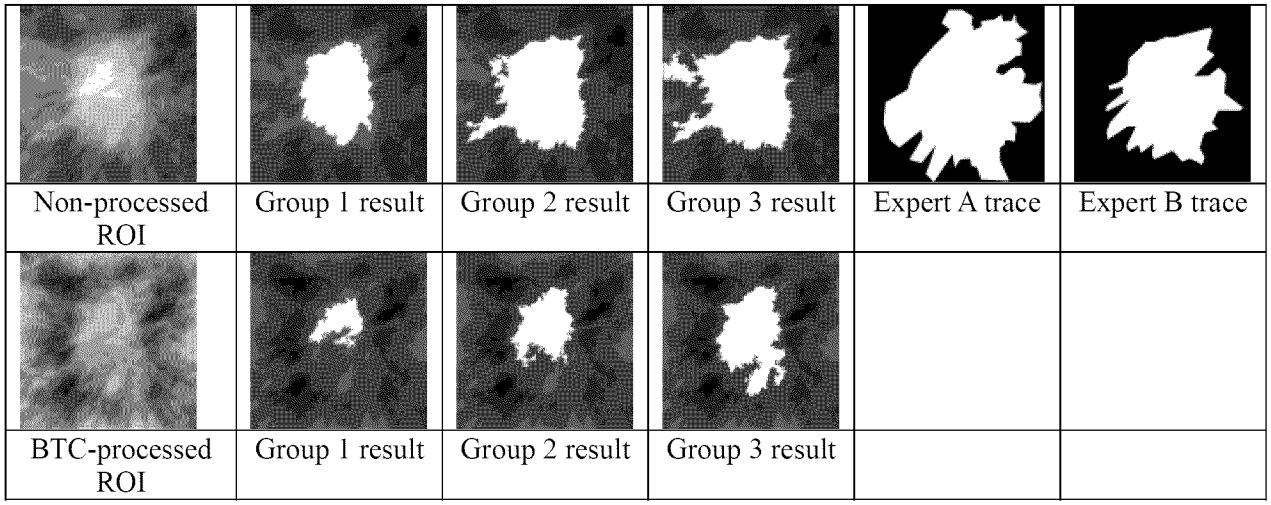


Figure 2 – A Cancerous Mass Showing Worsened Results Due to BTC Processing

The U.S. Army Medical Research and Materiel Command under DAMD17-0301-0314 supported this work.

CHAPTER 1. COMPARISON OF SEGMENTATION METHODS: REGION-GROWING AND GRADIENT VECTOR FLOW

1.1. Introduction and Snake Background

Although our region-growing method achieved better results on the mixed density breast images, it appears to have worked reasonably well on the chosen set of dense breast images. During several of my talks and interviews over the past few months I have often been asked if the method had been compared to another method. In response to these requests I thought that it would be worth our time to compare the computer results of our method to the results of Gradient Vector Flow (GVF), a method implemented by Xu and Prince of Johns Hopkins University. The GVF method is an extension of the snake method, developed by Kass and Witkin. It differs from the snake because can grow into concave areas (see figure 1.1).

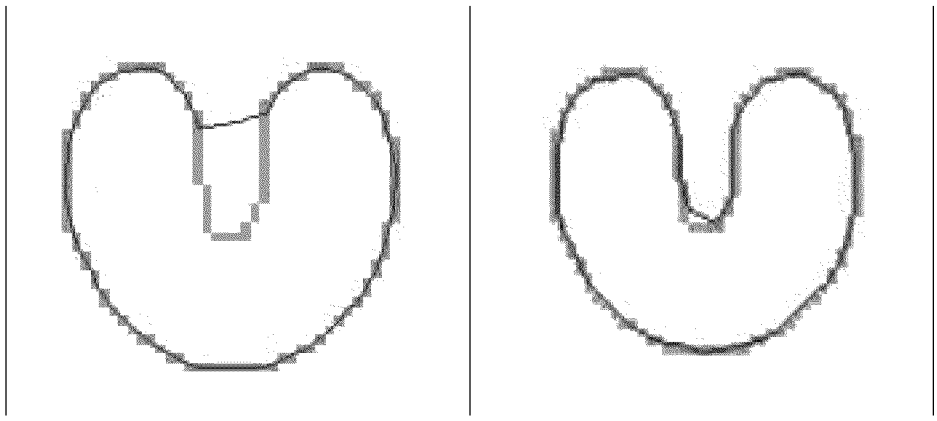


Figure 1.1: The Letter 'U' on a Homogeneous Background: (a)Traditional Snake (b)GVF Snake

If we define the snake as v(s) = (x(s),y(s)) where x(s) and y(s) are coordinates along the contour $s \in$=[0,1] (see figure 1.2).

The snake is defined as an energy minimizing spline, where the goal is to move it towards the borders of a Region Of Interest (ROI) by minimizing the energy. Initially the snake is shaped like a circle, is placed near the borders of the ROI and it shrinks (or expands) until it reaches

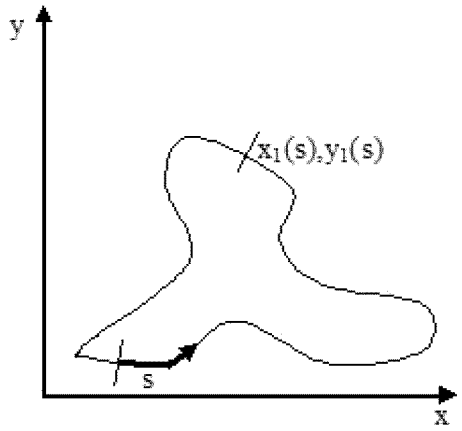


Figure 1.2: Visual of Parametric Representation of Snake

the borders. The energy function to be minimized is defined as:

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{snake}(v(s))ds \\ &= \int_0^1 E_{int}(v(s)) + E_{image}(v(s)) + E_{con}(v(s))ds \end{aligned} \tag{1.1}$$

where E_{int} is the internal energy of the snake due to bending, E_{image} refers to image forces, and E_{con} refers to external constraint forces.

Xu and Prince define typical external energies as:

$$E^1_{ext}(x,y) = -|\nabla I(x,y)|^2 \tag{1.2}$$

$$E^2_{ext}(x,y) = -|\nabla(G_\sigma(x,y) * I(x,y))|^2 \tag{1.3}$$

where $I(x,y)$ is the image, $G_\sigma(x,y)$ is a 2D Gaussian function, σ is standard deviation, and, ∇ is the gradient operator.

1.2. Gradient Vector Flow Field

The authors defined an irrotational external force field called the gradient vector flow (GVF) field. The GVF field points toward the object boundary when it is near to the boundary, but varies slowly over homogeneous image regions. The process begins by defining an edge map,

$f(x,y)$, which comes from the image $I(x,y)$. It is stronger near edge boundaries and is defined as:

$$f(x,y) = -E^i_{ext}(x,y) \tag{1.4}$$

for i = 1 or 2. The field ∇f has vectors pointing toward the edges, but $\nabla f = 0$ in homogeneous regions. The GVF field is defined as the vector field $v(x,y) = (u(x,y), v(x,y))$ that minimizes the energy function:

$$E = \iint \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |v - \nabla f|^2 dxdy \tag{1.5}$$

where μ is a regularization parameter that governs a tradeoff between the first and second terms.

LIKELIHOOD FUNCTION ANALYSIS FOR SEGMENTATION OF MAMMOGRAPHIC MASSES FOR VARIOUS MARGIN GROUPS

Lisa Kinnard[a,b,c], *Shih-Chung B. Lo*[a], *Erini Makariou*[a], *Teresa Osicka*[a,d], *Paul Wang*[c], *Matthew T. Freedman*[a], *Mohamed Chouikha*[b]

[a]ISIS Center, Dept. of Radiology, Georgetown University Medical Center, Washington, D.C., USA
[b]Department of Electrical and Computer Engineering, Howard University, Washington, D.C., USA
[c]Biomedical NMR Laboratory, Department of Radiology, Howard University, Washington, D.C., USA
[d]Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington DC, USA

ABSTRACT

The purpose of this work was to develop an automatic boundary detection method for mammographic masses and to observe the method's performance on different four of the five margin groups as defined by the ACR, namely, spiculated, ill-defined, circumscribed, and obscured. The segmentation method utilized a maximum likelihood steep change analysis technique that is capable of delineating ill-defined borders of the masses. Previous investigators have shown that the maximum likelihood function can be utilized to determine the border of the mass body. The method was tested on 122 digitized mammograms selected from the University of South Florida's Digital Database for Screening Mammography (DDSM). The segmentation results were validated using overlap and accuracy statistics, where the gold standards were manual traces provided by two expert radiologists. We have concluded that the intensity threshold that produces the best contour corresponds to a particular steep change location within the likelihood function.

1. INTRODUCTION

In a CAD_x system, segmentation is arguably one of the most important aspects – particularly for masses – because strong diagnostic predictors for masses are shape and margin type [2,9]. The margin of a mass is defined as the interface between the mass and surrounding tissue [2]. Furthermore, breast masses can have unclear borders and are sometimes obscured by glandular tissue in mammograms. A spiculated mass consists of a central mass body surrounded by fibrous projections, hence the resulting stellate shape. For the aforementioned reasons, proper segmentation - to include the body and periphery - is extremely important and is essential for the computer to analyze, and in turn, determine the malignancy of the mass in mammographic CAD_x systems.

Over the years researchers have used many methods to segment masses in mammograms. Petrick [7] et al. developed the Density Weighted Contrast Enhancement (DWCE) method, in which series of filters are applied to the image in an attempt to extract masses. Comer et al. [1] segmented digitized mammograms into homogeneous texture regions by assigning each pixel to one of a set of classes such that the number incorrectly classified pixels was minimized via Maximum Likelihood (ML) analysis. Li [5] developed a method that employs k-means classification to classify pixels as belonging to the region of interest (ROI) or background.

Kupinski and Giger developed a method [4], which uses ML analysis to determine final segmentation. In their method, the likelihood function is formed from likelihood values determined by a set of image contours produced by the region growing method. This method is a highly effective one that was also implemented by Te Brake and Karssemeijer in their comparison between the discrete dynamic contour model and the likelihood method [9]. For this reason we chose to investigate its use as a possible starting point from which a second method could be developed. Consequently in our implementation of this work we discovered an important result, i.e., the maximum likelihood steep change. It appears that in many cases this method produces contour choices that encapsulate important borders such as mass spiculations and ill-defined borders.

2. METHODS

2.1 Initial Contours

As an initial segmentation step, we followed the overall region similarity concept to aggregate the area of interest [1, 4]. Used alone, a sequence of contours representing the mass is generated; however, the computer is not able to choose the contour that is most closely correlated with the experts' delineations. Furthermore, we have devised an ML function steep change analysis method that chooses the best contour that delineates the mass body as well as its extended borders, i.e., extensions into spiculations and areas in which the borders are ill-defined or obscured. This method is an extension of the method developed by Kupinski and Giger [4] that uses ML function analysis to select the contour which best represents the mass, as compared to expert radiologist traces. We have determined that this technique can select the contour that accurately represents the mass body contour for a given set of parameters; however, further analysis of the likelihood function revealed that the computer could

0-7803-8388-5/04/$20.00 ©2004 IEEE

choose a set of three segmentation contour choices from the entire set of contour choices, and then make a final decision from these three choices.

The algorithm can be summarized in several steps. Initially, we use an intensity based thresholding scheme to generate a sequence of grown contours (S_i), where gray value is the similarity criterion. The image is also multiplied by a 2D trapezoidal membership function (2D shadow), whose upper base measures 40 pixels and lower base measures 250 pixels (1 pixel = 50 microns). The image to which the shadow has been applied is henceforth referred to as the "fuzzy" image. The original image and its fuzzy version were used to compute the likelihood of the mass's boundaries. The computation method is comprised of two components for a given boundary: (1) formulation of the composite probability and (2) evaluation of likelihood.

In addition, we chose to aggregate contours using the original image. This accounts for the major difference from that implemented by the previous investigators. Since smoother contours were not used, the likelihood function showed greater variations. In many situations, the greatest variations occurred when there was a sudden increase of the likelihood, and this was strongly correlated with the end of the mass border growth. This phenomenon would be suppressed if the fuzzy image was used to generate the contours. The fuzzy image was used mainly to construct the likelihood function.

2.2 Composite Probability Formation

For a contour (S_i), the composite probability (C_i) is calculated:

$$C_i|S_i = p(f_i(x,y)|S_i) \times p(m_i(x,y)|S_i) \qquad (1)$$

The quantity $f_i(x,y)$ is the area to which the 2D shadow has been multiplied, $p(f_i(x,y)|S_i)$ is the probability density function of the pixels inside S_i where 'i' is the region growing step associated with a given intensity threshold. The quantity $m_i(x,y)$ is the area outside S_i (non-fuzzy), and $p(m_i(x,y)|S_i)$ is the probability density function of the pixels outside S_i. Next we find the logarithm of the composite probability of the two regions, C_i:

$$Log(C_i|S_i) = \log(p(f_i(x,y)|S_i)) + \log(p(m_i(x,y)|S_i)) \qquad (2)$$

2.3 Evaluation of Likelihood Function

The likelihood that the contour represents the fibrous portion of the mass, i.e., mass body is determined by assessing the maximum likelihood function:

$$\arg\max(Log(C_i|S_i)), S_i, i = 1,..n \qquad (3)$$

Equation (3) intends to find the maximum value of the aforementioned likelihood values as a function of intensity threshold. It has been assessed (also by other investigators [4]) that the intensity value corresponding to this maximum likelihood value is the optimal intensity needed to delineate the mass body contour. However, in our implementation it was discovered that the intensity threshold corresponding to the maximum likelihood value confines the contour to the mass body. In our study many of these contours did not include the extended borders. We, therefore, hypothesize that the contour represents the mass's extended borders may well be determined by assessing the maximum changes of the likelihood function, i.e., locate the steepest likelihood value changes within the function:

$$\frac{d}{di}(Log(C_i|S_i)), S_i, i = 1,\ldots,n \qquad (4)$$

Based on this assumption, we have carefully analyzed the behavior of maximum likelihood function. The analysis reveals that we have successfully discovered that the most accurate mass delineation is usually obtained by using the intensity value corresponding to the first or second steep change locations within the likelihood function immediately following the maximum likelihood value on the likelihood function.

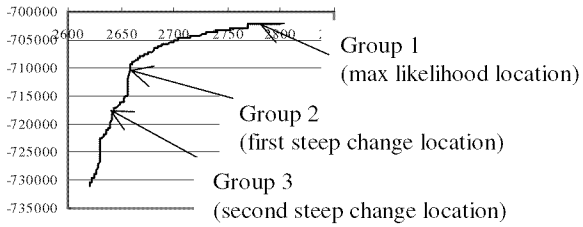


Figure 1: A likelihood function with steep change indicators

2.4 Steep change definition

The term "steep change" is rather subjective and can defined as a location between two or more points in the function where the likelihood values experience a significant change. In some cases the likelihood function increases at a slow rate. The algorithm design accounts for this issue by calculating the difference between likelihood values in steps over several values and comparing the results to two thresholds. The difference equation is given by:

$$h(t) = f(z - wt) - f(z - w(t+1)), \quad t = 0,\ldots,N \qquad (5)$$

where f is the likelihood function, z is the maximum intensity, w is the width of the interval over which the likelihood differences are calculated (e.g. – for w=7 differences are calculated every 7 points), and N is the total number of points in the searchable area divided by w. If the calculation in question yields a value greater than or equal to a given threshold, then the intensity corresponding to this location is considered to be a steep change location. The threshold algorithm occurs as follows:

If $(h(t)_{ML} \geq ML_{T1})$; t=0,…,m
Then choice 1 = intensity where that condition is satisfied
If $(h(t)_{ML} \geq ML_{T2})$; t=m,…,z
Then choice 2 = intensity where that condition is satisfied

where $h(t)_{ML}$ is the steep change value given by equation (5), ML_{T1} and ML_{T2} 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. Once the condition is satisfied for the first threshold value (ML_{T1}) then its corresponding intensity value is used to produce the segmentation contour for the first steep change location. Once the condition is satisfied for ML_{T2} then its corresponding intensity value is used to produce the segmentation contour for the second steep change location.

2.5 Validation

The segmentation method was validated on the basis of overlap and accuracy [8,10]:

$$Overlap = \frac{N_{TP}}{N_{FN} + N_{TP} + N_{FP}} \qquad (6)$$

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

where N_{TP} is the true positive fraction, N_{TN} true negative fraction, N_{FP} is the false positive fraction, and N_{FN} is the false negative fraction. The gold standards used for the validation study were mass contours, which have been traced by expert radiologists.

Our experiments produced contours for the intensity values resulting from three locations within the likelihood functions: (1) The intensity for which a value within the likelihood function is maximum (group 1 contour) (2) The intensity for which the likelihood function experiences its first steep change (group 2 contour) and (3) The intensity for which the likelihood function experiences its second steep change (group 3 contour). We have observed that the intensity for which the likelihood function experiences its first steep change produces the contour trace that is most highly correlated with the gold standard traces, regarding overlap and accuracy.

3. EXPERIMENTS AND RESULTS

Here we describe the database used, describe the experiments, provide visual results obtained by the algorithm, as well as report the results obtained by the ANOVA test.

3.1 Database

For this study, a total of 122 masses were chosen from the University of South Florida's Digital Database for Screening Mammography (DDSM) [3]. The films were digitized at resolutions of 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. The images were of varying subtlety ratings. The first set of expert traces was provided by an attending physician of the 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.

3.2 Experiments and Results

As mentioned previously, the term "steep change" is very subjective and therefore a set of thresholds needed to be set in an effort to define a particular location within the likelihood function as a "steep change location". For this study the following thresholds were experimentally chosen: ML_{T1}=1800, ML_{T2}=1300, where ML_{T1}= threshold for steep change location 1 for the likelihood function, and ML_{T2} = threshold for steep change location 2 for the likelihood function. We performed a number of experiments in an effort to prove that the intensity for which the likelihood function experiences the first steep change location produces the contour trace, which is most highly correlated with the gold standard traces regarding overlap and accuracy.

First we present segmentation results for two malignant cases followed segmentation results for two benign cases. Each figure contains an original image, traces for Experts A and B, and computer segmentation results for groups 1, 2, and 3. Second, we present data that plots the mean values for various margin groups for both overlap and accuracy measurements. The plots present data for the spiculated and ill-defined groups of malignant masses, and ill-defined and circumscribed groups of benign masses. Data was not presented for the other categories because there was not a sufficient amount of cases.

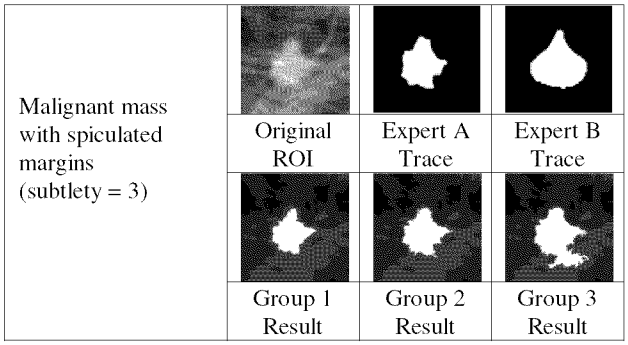


Figure 2: Segmentation Results: Spiculated Malignant Mass

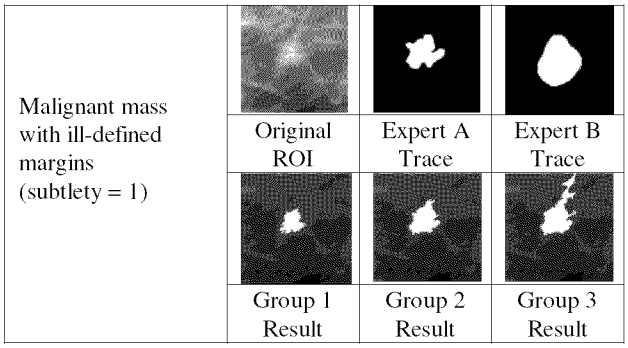


Figure 3: Segmentation Results: Ill-defined Malignant Mass

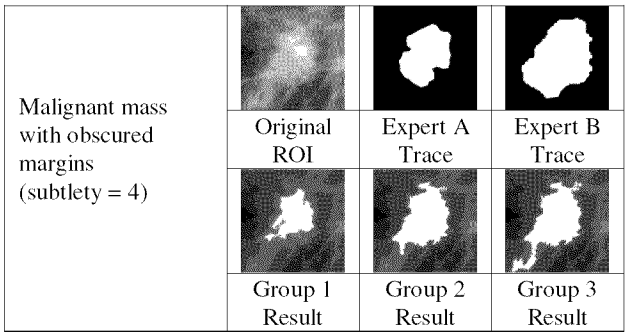


Figure 4: Segmentation Results: Obscured Malignant Mass

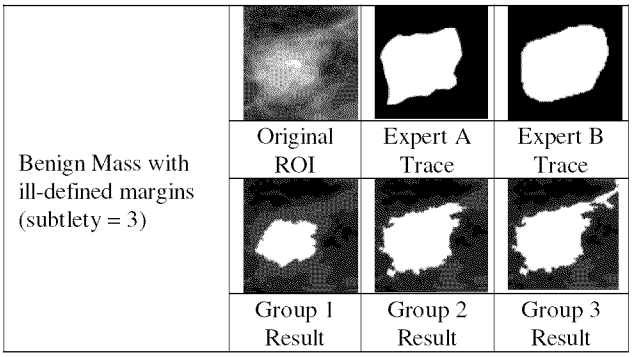


Figure 5: Segmentation Results: Ill-defined Benign Mass

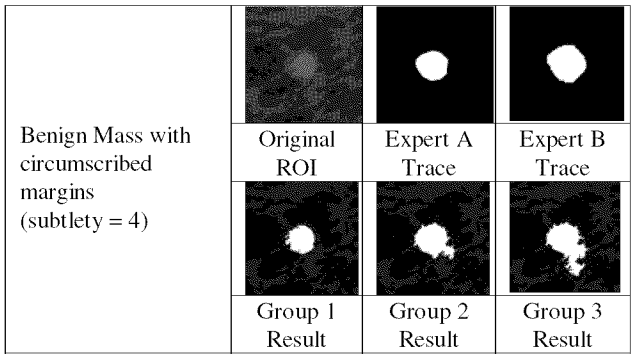


Figure 6: Segmentation Results: Circumscribed Benign Mass

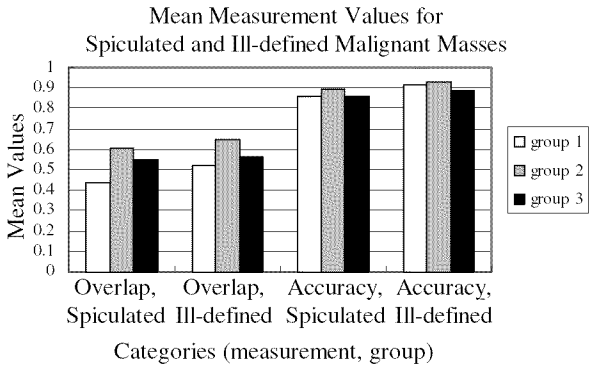


Figure 7: Mean Measurement Values (Malignant Masses)

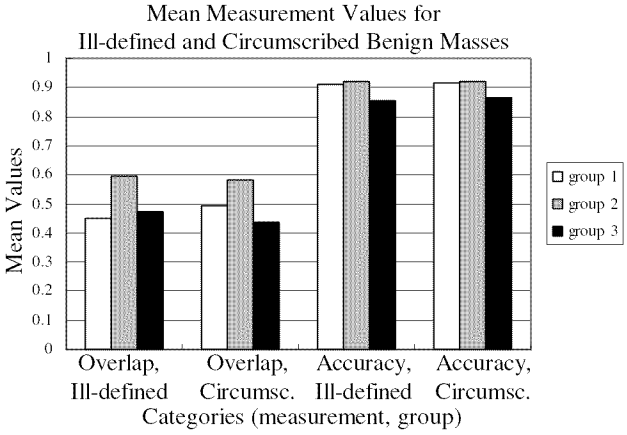


Figure 8: Mean Measurement Values (Benign Masses)

4. DISCUSSION AND CONCLUSION

The visual results (see Figures 2-6) reveal that the group 2 trace appears to delineate the masses better than the group 1 and group 3 contours in most cases. Visually, it appears that the method has performed equally well on all margin groups. This is an encouraging result because some of the more difficult masses to segment are typically those that are spiculated, obscured, and those that have ill-defined borders. The plots shown in Figures 7-8 confirm that the group 2 trace performs better than the other groups on the basis of overlap and accuracy for all margin groups, therefore supporting our visual observations.

In future work, a worthwhile study would be to test gather more data for all margin groups in an effort to see if the various groups require different parameter values to maximize the algorithm's robustness. Our ultimate goal is to optimize its performance for those masses falling in the ill-defined and obscured margin groups because segmentation of masses falling into those categories is exceedingly difficult.

5. REFERENCES

[1] M.L. Comer, E.J. Delp, "The EM/MPM algorithm for segmentation of textured images: Analysis and further experimental results", *Proceedings of the 1995 IEEE ICIP*, Lausanne, Switzerland, September 16-19, 1996.

[2] J.R. Harris, M.E. Lippman, M. Morrow, S. Hellman, "Diseases of the breast", Lippincott-Raven Publishers, Philadelphia, PA, pp. 80-81, 1996.

[3] M. Heath, K.W. Bowyer, D. Kopans et al., "Current status of the digital database for screening mammography", Digital Mammography, Kluwer Academic Publishers, pp. 457-460 , 1998.

[4] M.A. Kupinski, M.L. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms", *IEEE Trans. on Med. Imag.*, vol. 17, no. 4, pp. 510-517, 1998.

[5] L. Li, Y. Zheng, L. Zhang, R. Clark, "False-positive reduction in CAD mass detection using a competitive classification strategy", *Med. Phys.*, vol. 28, pp. 250-258, 2001.

[6] J.E. Martin, "Atlas of mammography: histologic and mammographic correlations (second edition)", Williams and Wilkins, Baltimore, MD, p. 87, 1988.

[7] N. Petrick, H-P Chan, B. Sahiner, D. Wei, "An Adaptive Density-Weighted Contrast Enhancement Filter for Mammographic Breast Mass Detection", *IEEE Trans. on Med. Imag.*, vol. 15, no. 1, pp. 59-67, 1996.

[8] J. Suckling, D.R. Dance, E. Moskovic, D.J. Lewis, S.G. Blacker, "Segmentation of mammograms using multiple linked self-organizing neural networks", *Med. Phys.*, vol. 22, pp. 145-152, 1995.

[9] G.M. te Brake, N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms", *Med. Phys.*, vol. 28, no. 2, pp. 259-266, 2001.

[10] B. Van Ginneken, "Automatic segmentation of lung fields in chest radiographs", *Med. Phys.*, 27, pp. 2445-2455, 2000.

6. ACKNOWLEDGMENTS

This work was supported by US Army Grant numbers DAMD17-01-1-0267, DAMD 17-00-1-0291, DAAG55-98-1-0187, and DAMD 17-03-1-0314.

Mass Segmentation of Dense Breasts on Digitized Mammograms: Analysis of a Probability-Based Function

L. M. Kinnard[a,b], S.-C. B. Lo[a], E. Duckett[c], E. Makariou[a], T. Osicka[d], Matthew T. Freedman[a], and M. F. Chouikha[b]

[a]ISIS Center, Department of Radiology, Georgetown University, Washington DC, USA;
[b]Howard University Department of Electrical and Computer Engineering, Washington DC, USA;
[c]Advanced Radiology, Glen Burnie, MD, USA;
[d]The Catholic University of America Department of Electrical Engineering and Computer Science, Washington DC, USA

ABSTRACT

In this study, a segmentation algorithm based on the steepest changes of a probabilistic cost function was tested on non-processed and pre-processed dense breast images in an attempt to determine the efficacy of pre-processing for dense breast masses. Also, the inter-observer variability between expert radiologists was studied. Background trend correction was used as the pre-processing method. The algorithm, based on searching the steepest changes on a probabilistic cost function, was tested on 107 cancerous masses and 98 benign masses with density ratings of 3 or 4 according to the American College of Radiology's density rating scale. The computer-segmented results were validated using the following statistics: overlap, accuracy, sensitivity, specificity, Dice similarity index, and kappa. The mean accuracy statistic value ranged from 0.71 to 0.84 for cancer cases and 0.81 to 0.86 for benign cases. For nearly all statistics there were statistically significant differences between the expert radiologists.

Keywords: mass segmentation, inter-observer variability, digitized mammograms

1. 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] Several studies have shown 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 have shown that 7.6% - 14% of the patients have mammograms that produce false negative diagnoses.[5,6] More accurate prediction can be achieved by combining a mammographer's interpretation with that of a Computer Assisted Diagnosis (CAD_x) system, which can analyze masses for key diagnostic indicators such as shape. For example, many malignant masses have ill-defined, and/or spiculated borders and many benign masses have well-defined, rounded borders. Furthermore, the borders of breast masses are sometimes obscured in mammograms by glandular tissue. A CAD_x system can help physicians identify these areas more accurately through a process called segmentation in which the computer automatically separates a region of interest from surrounding tissue.

Mass segmentation has prompted the development of many techniques and it continues to be one of the most closely studied areas in CAD_x today. Te Brake and Karssemeijer[7] have implemented a discrete dynamic contour model, a method similar to snakes, that begins as a set of vertices connected by edges (initial contour) and grows subject to internal and external forces. Li[8] has developed a method that employs k-means classification to assign pixels to the region of interest (ROI) or to the background. Petrick et al.[9] have developed the Density Weighted Contrast Enhancement (DWCE) method, in which a series of filters are applied to the image in an attempt to extract masses. Comer et al.[10] have utilized an EM technique to segment digitized mammograms into

Further author information: (Send correspondence to Lisa M. Kinnard)
Lisa M. Kinnard: E-mail: kinnard@isis.imac.georgetown.edu, Telephone: 1 202 687 1572

homogeneous texture regions by assigning each pixel was to one of a set of classes so that the number incorrectly classified pixels is minimized. Kupinski and Giger[11] have developed a method, that combines region growing with probability analysis to determine final segmentation. In this method, the probability-based function is formed from a specific composed probability density function that is determined by a set of image contours produced by the region growing method.

2. METHODS

2.1. Segmentation and Pre-processing

Our method evaluates the steepest changes within a probabilistic cost function in an effort to determine the computer segmented contour that is most closely correlated with expert radiologist manual traces.[11,12] This method segments breast masses by combining region growing with probability-based cost function analysis. For each cost function there are a number of steepest changes in likelihood (see Figures 2e and 3e), where a steepest change location is defined by a set of thresholds. In most cases the trace which is most likely to enclose the mass in its entirety is produced by the intensity corresponding to that steepest change location. For example, a steepest change location in Figure 2e is located at the intensity = 3100. The intensity corresponding to the maximum value on the cost likelihood function is most likely to enclose the mass's central body. Based on this analysis the three best contours are chosen and the computer makes a final selection from these three choices. Typically, the Group 1 trace encapsulates the central portion of the mass (intensity corresponds to maximum value on likelihood function), the Group 2 trace encapsulates the central mass and borders extending into surrounding tissue (intensity corresponds to first steepest change location), and the Group 3 trace encapsulates the area covered by the Group 2 trace and surrounding fibroglandular tissue (intensity corresponds to second steepest change location).

The masses used in this study were exceedingly difficult to segment due to the surrounding dense tissue. We therefore thought that a contrast enhancement method - background trend correction in this experiment - would help the segmentation process. The background correction technique is based on a two-dimensional third order polynomial fit given by:

$$BC(x,y) = \sum_{j=0}^{n} a_j x^{s_j} y^{t_j}, \tag{1}$$

where n=3. Hence, the corrected image ($f_c(x,y)$) is obtained by subtracting the background trend ($BC(x,y)$) from the original image $f(x,y)$:

$$f_c(x,y) = f(x,y) - BC(x,y). \tag{2}$$

2.2. Statistical Methods

All masses were manually traced by two expert radiologists and the overlap, accuracy, sensitivity, specificity, Dice Similarity Index (DSI), and kappa (κ) statistic were calculated.[13,14] All statistics are formulated using the following terms: N_{TP} = the number of true positive pixels (pixels that are actually mass, N_{TN} = the number of true negative pixels (pixels that are actually background), N_{FP}= the number of pixels the computer interprets as mass which are actually background, and N_{FN} = the number of pixels the computer interprets as background which are actually mass (see Figure 1).

$$Overlap = \frac{N_{TP}}{N_{TP} + N_{FP} + N_{FN}}, \tag{3}$$

$$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \tag{4}$$

$$Sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{5}$$

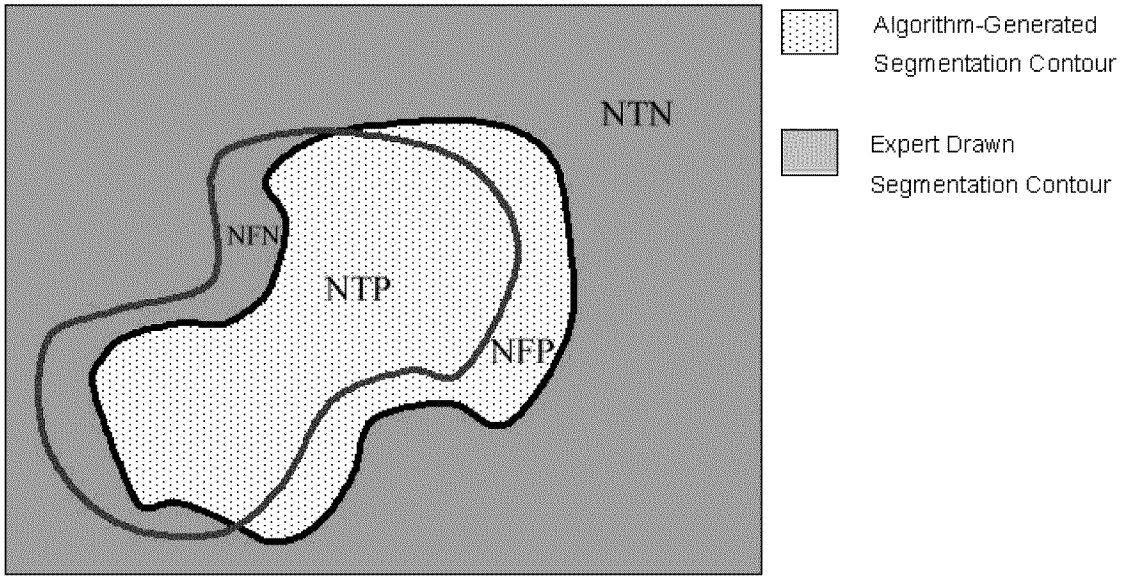


Figure 1. This figure is an example of a mass traced by an expert radiologist superimposed with the computer interpretation

$$Specificity = \frac{N_{TN}}{N_{TN} + N_{FP}}, \tag{6}$$

$$DSI = \frac{2N_{TP}}{N_{FN} + 2N_{TP} + N_{FP}}, \tag{7}$$

$$\kappa = \frac{2(N_{TP}N_{TN} + N_{FP}N_{FN})}{(N_{TP} + N_{FN})(N_{FN} + N_{TN}) + (N_{TP} + N_{FP})(N_{FP} + N_{TN})}. \tag{8}$$

Specifically, Landis and Koch[15] have developed a six-point scale with which the kappa statistic can be analyzed (see table 1).

Table 1. Six-point Scale Indicating the Performance of the Kappa Statistic.

κ	**Strength of Agreement**
< 0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0,80	Substantial
0.81 - 1.00	Almost Perfect

The statistics have values ranging from 0 to 1, where a value of 0 indicates no agreement and a value of 1 indicates perfect agreement. While these statistics measure the performance of segmentation algorithms reasonably well, it is possible that the algorithm in question can be biased toward one expert radiologist. To examine this issue we used a two-tailed T-test which was performed using the SPSS^{TM} statistical package.

Table 2. T-test Results for all Statistics: Expert A vs. Expert B (Non-Processed Cancerous Masses)

Hypothesis	**P-value (Group 1)**	**P-value (Group 2)**	**P-value (Group 3)**
Difference between Experts A and B (overlap)	0.000	0.000	NS
Difference between Experts A and B (accuracy)	0.000	0.000	0.000
Difference between Experts A and B (sensitivity)	0.000	0.000	0.000
Difference between Experts A and B (specificity)	NS	0.000	0.000
Difference between Experts A and B (DSI)	0.000	0.000	NS
Difference between Experts A and B (κ)	0.000	0.000	0.000

Table 3. Mean (μ) Values for all Statistics: Experts A and B (Non-Processed Cancerous Masses)

Statistic	**μ-value (Exp. A, Group 1)**	**μ-value (Exp. B, Group 1)**	**μ-value (Exp. A, Group 2)**	**μ-value (Exp. B, Group 2)**	**μ-value (Exp. A, Group 3)**	**μ-value (Exp. B, Group 3)**
Overlap	0.28	0.36	0.46	0.52	0.47	0.49
Accuracy	0.72	0.82	0.78	0.85	0.77	0.82
Sensitivity	0.30	0.39	0.54	0.65	0.61	0.72
Specificity	0.98	0.98	0.94	0.92	0.89	0.86
DSI	0.41	0.51	0.60	0.66	0.62	0.64
κ	0.32	0.43	0.48	0.57	0.47	0.53

2.3. Database and Experiments

The cases for this work were obtained from the University of South Florida's Digital Database for Screening Mammography (DDSM).[16] The densities of all cases were rated using the American College of Radiology (ACR) scale which ranges from 1 to 4. A breast containing a great deal of dense tissue would receive a rating of 4. Approximately two-thirds of the cases used in this work received a density rating of 3 while the remaining cases received a density rating of 4.

We performed two experiments in which we calculated the statistics between the computer results and manual traces from both expert radiologists. In the first experiment the masses were unprocessed and in the second experiment they were processed using background trend correction.

3. RESULTS

3.1. Statistical Results

Tables 2- 9 show p-values for the t-tests which analyzed inter-observer variability as well as the mean values of all statistics for both expert radiologists. The significance level is $p < 0.05$. A table entry whose value is 0.000 implies that there the p-value for a particular test was less than 0.000 and a table entry of "NS" implies that there was no significant difference for a particular test.

Table 4. T-test Results for all Statistics: Expert A vs. Expert B (Non-Processed Benign Masses)

Hypothesis	**P-value (Group 1)**	**P-value (Group 2)**	**P-value (Group 3)**
Difference between Experts A and B (overlap)	0.000	0.002	NS
Difference between Experts A and B (accuracy)	0.000	0.007	0.040
Difference between Experts A and B (sensitivity)	0.000	0.000	0.000
Difference between Experts A and B (specificity)	NS	0.025	NS
Difference between Experts A and B (DSI)	0.000	0.003	NS
Difference between Experts A and B (κ)	0.000	0.001	NS

Table 5. Mean (μ) Values for all Statistics: Experts A and B (Non-Processed Benign Masses)

Statistic	**μ-value (Exp. A, Group 1)**	**μ-value (Exp. B, Group 1)**	**μ-value (Exp. A, Group 2)**	**μ-value (Exp. B, Group 2)**	**μ-value (Exp. A, Group 3)**	**μ-value (Exp. B, Group 3)**
Overlap	0.32	0.36	0.49	0.52	0.48	0.49
Accuracy	0.81	0.84	0.84	0.86	0.81	0.83
Sensitivity	0.36	0.40	0.60	0.66	0.72	0.77
Specificity	0.98	0.98	0.94	0.93	0.86	0.86
DSI	0.46	0.51	0.63	0.66	0.63	0.63
κ	0.40	0.45	0.55	0.59	0.52	0.53

Table 6. T-test Results for all Statistics: Expert A vs. Expert B (Background Trend Corrected Cancerous Masses)

Hypothesis	**P-value (Group 1)**	**P-value (Group 2)**	**P-value (Group 3)**
Difference between Experts A and B (overlap)	0.000	0.000	0.000
Difference between Experts A and B (accuracy)	0.000	0.000	0.000
Difference between Experts A and B (sensitivity)	0.000	0.000	0.000
Difference between Experts A and B (specificity)	NS	0.024	0.003
Difference between Experts A and B (DSI)	0.000	0.000	0.000

Table 7. Mean (μ) Values for all Statistics: Experts A and B (Background Trend Corrected Cancerous Masses)

Statistic	**μ-value (Exp. A, Group 1)**	**μ-value (Exp. B, Group 1)**	**μ-value (Exp. A, Group 2)**	**μ-value (Exp. B, Group 2)**	**μ-value (Exp. A, Group 3)**	**μ-value (Exp. B, Group 3)**
Overlap	0.19	0.26	0.38	0.47	0.38	0.44
Accuracy	0.73	0.83	0.78	0.87	0.77	0.85
Sensitivity	0.20	0.27	0.41	0.52	0.49	0.61
Specificity	1.00	1.00	0.99	0.99	0.94	0.93
DSI	0.29	0.38	0.51	0.60	0.53	0.59

Table 8. T-test Results for all Statistics: Expert A vs. Expert B (Background Trend Corrected Benign Masses)

Hypothesis	**P-value (Group 1)**	**P-value (Group 2)**	**P-value (Group 3)**
Difference between Experts A and B (overlap)	0.000	0.000	0.002
Difference between Experts A and B (accuracy)	0.001	0.002	0.006
Difference between Experts A and B (sensitivity)	0.000	0.000	0.000
Difference between Experts A and B (specificity)	NS	0.049	0.010
Difference between Experts A and B (DSI)	0.000	0.000	0.003

Table 9. Mean (μ) Values for all Statistics: Experts A and B (Background Trend Corrected Benign Masses)

Statistic	**μ-value (Exp. A, Group 1)**	**μ-value (Exp. B, Group 1)**	**μ-value (Exp. A, Group 2)**	**μ-value (Exp. B, Group 2)**	**μ-value (Exp. A, Group 3)**	**μ-value (Exp. B, Group 3)**
Overlap	0.21	0.24	0.41	0.45	0.44	0.47
Accuracy	0.80	0.83	0.84	0.87	0.83	0.85
Sensitivity	0.21	0.24	0.44	0.48	0.56	0.62
Specificity	1.00	1.00	0.99	0.99	0.94	0.94
DSI	0.31	0.34	0.53	0.57	0.59	0.62

3.2. Visual Results

Figures 2 - 3 contain the following parts: (a) original image (b) cropped ROI and its computer segmented results (non-processed image) (c) cropped ROI and its computer segmented results (background trend corrected image) (d) manually traced expert delineations and (e) cost likelihood functions. Again, the Group 1 trace encapsulates the central portion of the mass (intensity corresponds to maximum value on likelihood function), the Group 2 trace encapsulates the central mass and borders extending into surrounding tissue (intensity corresponds to first steepest change location), and the Group 3 trace encapsulates the area covered by the Group 2 trace and surrounding fibroglandular tissue (intensity corresponds to second steepest change location).

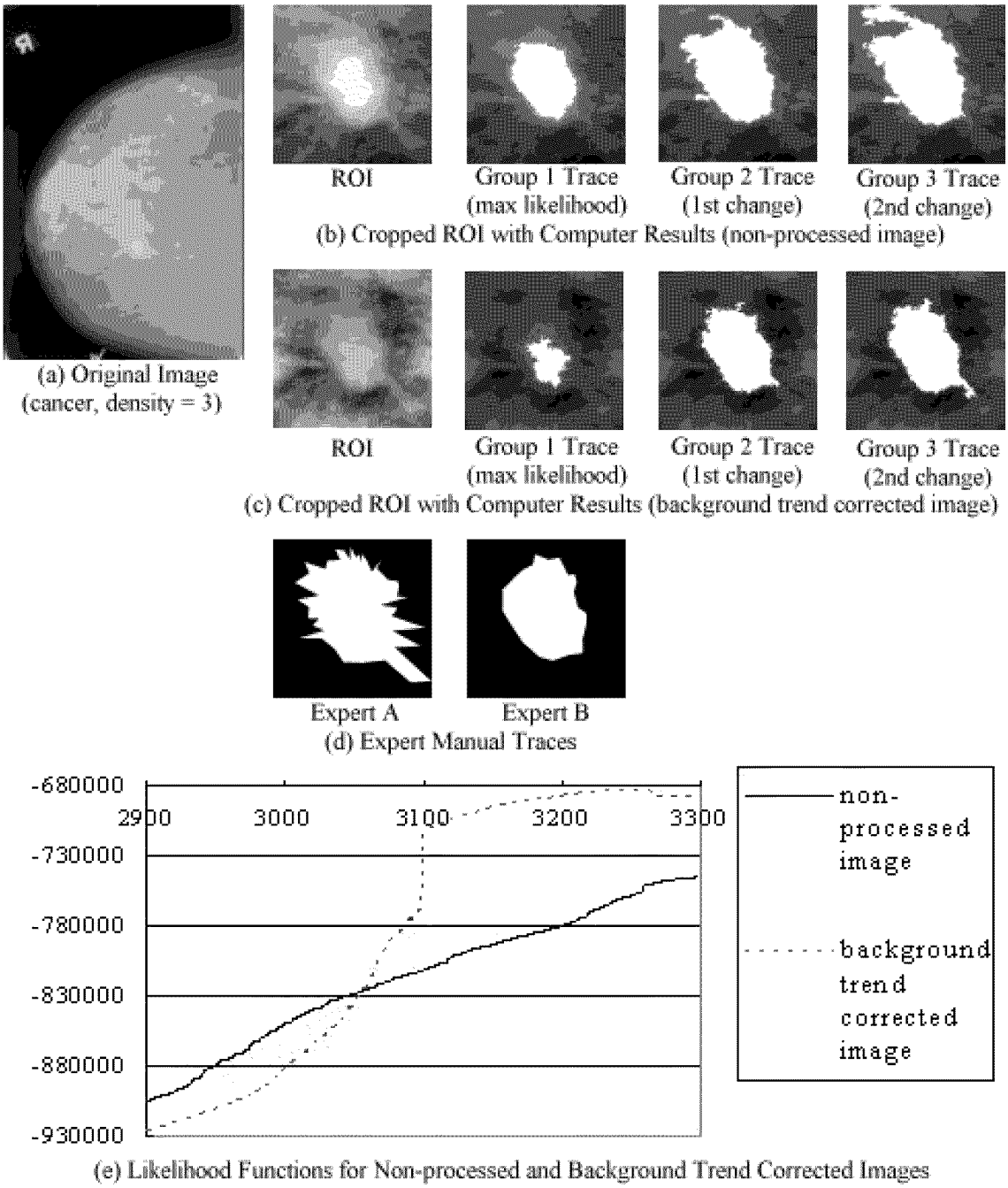


Figure 2. Cancerous mass: (a) original image (b) cropped ROI and its computer segmented results (non-processed image) (c) cropped ROI and its computer segmented results (background trend corrected image) (d) manually traced expert delineations and (e) cost likelihood functions

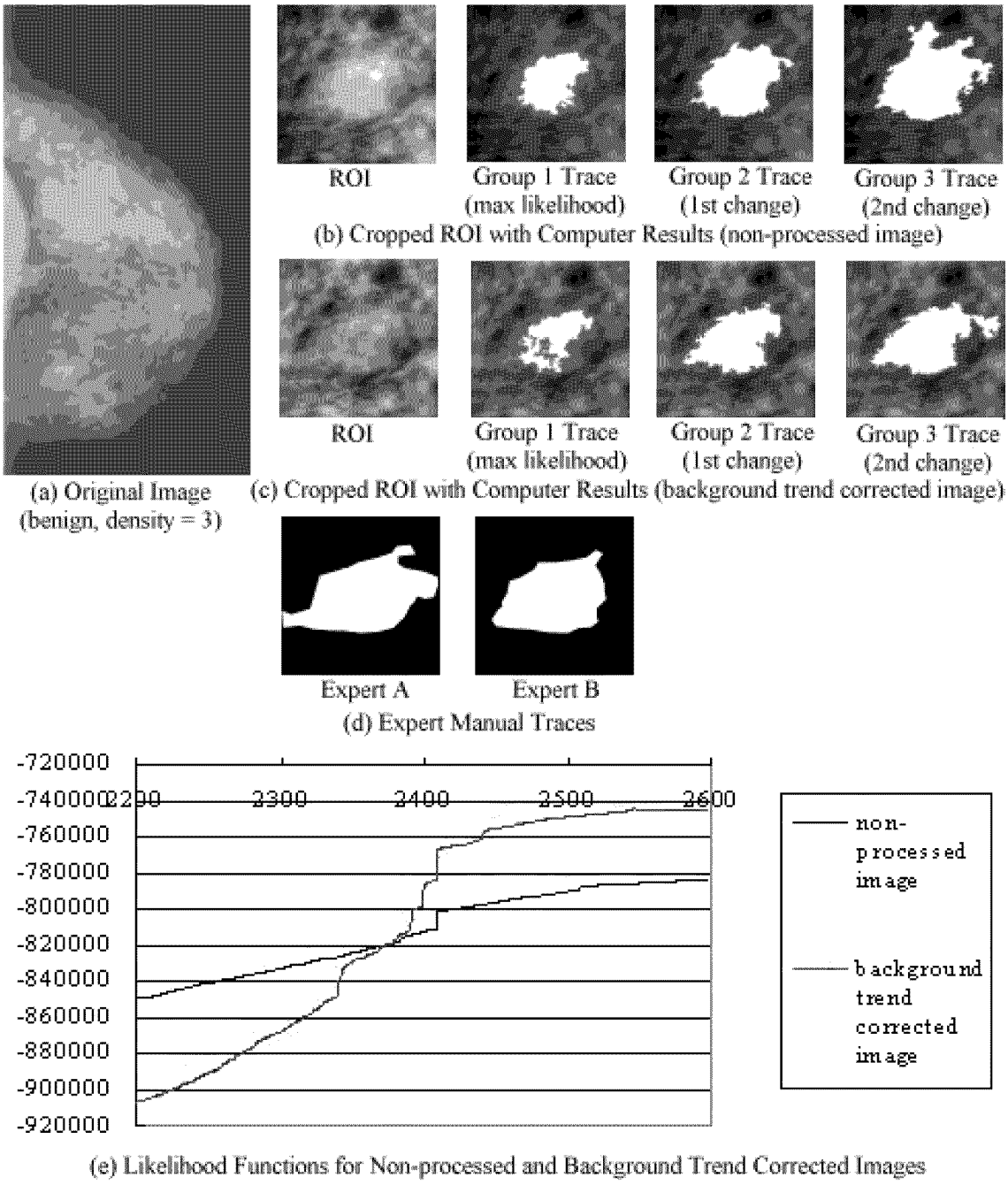


Figure 3. Benign mass: (a) original image (b) cropped ROI and its computer segmented results (non-processed image) (c) cropped ROI and its computer segmented results (background trend corrected image) (d) manually traced expert delineations and (e) cost likelihood functions

4. DISCUSSION AND CONCLUSION

As the visual and statistical results demonstrate, the background trend correction pre-processing method does not seem to have improved the performance of the automated segmentation algorithm. From a visual standpoint, background trend correction seems to have caused some areas inside the masses to become darker, and thus, the region growing portion of the algorithm would not grow into these areas. Simultaneously, for some cases this darkening effect caused a sharper contrast between the mass and surrounding tissue, making the mass boundaries easier to see.

For most statistics there were statistically significant differences between both radiologists. In general, the group 2 and group 3 traces achieved better performance values than the group 1 traces for both radiologists. The mean values for Expert B were greater than those for Expert A, which reveals that there was stronger agreement between the computer and Expert B than between the computer and Expert A.

Background trend correction causes the likelihood cost functions to incur more steepest changes as opposed to the cost likelihood functions for the non-processed images, which are typically smooth. In turn, the computer makes its decisions earlier in the steep change searching process, consequently, the mass contours encapsulate smaller areas. In future work it will be necessary to change the steepest change parameters to account for the change. The inter-observer variability implies that in future work we should also investigate the possibility of obtaining a consensus opinion between the two existing radiologists. An alternative method would obtain more radiologist traces.

ACKNOWLEDGMENTS

This work was supported by U.S. Army grants DAMD17-03-1-0314, DAMD-17-01-1-0267, and DAMD17-00-1-0267. The content of this paper does not necessarily reflect the position or policy of the government.

REFERENCES

1. J.V. Lacey Jr, S.S. Devesa, L.A. Brinton, "Recent trends in breast cancer incidence and mortality, *Environmental and Molecular Mutagenesis*, **39(2-3)**, pp. 82-88, 2002.
2. J.E. Meyer, D.B. Kopans, P.C. Stomper, K.K. Lindfors, "Occult Breast Abnormalities: Percutaneous Preoperative Needle Localization, *Radiology*, **150(2)**, pp. 335-337, 1984.
3. A.L. Rosenberg, G.F. Schwartz, S.A. Feig, A.S. Patchefsky, "Clinically occult breast lesions: localization and significance, *Radiology*, **162**, pp. 167-170, 1987.
4. B.C.Yankaskas, M.H. Knelson, M.L. Abernethy, J.T. Cuttino, R.L. Clark, "Needle localization biopsy of occult lesions of the breast, *Investigative Radiology*, **23**, pp. 729-733, 1988.
5. J.A. Harvey, L.L. Fajardo, C.A. Innis, "Previous mammograms in patients with impalpable breast carcinoma: retrospective vs. blinded interpretation, *American Journal of Roentgenology*, **161**, pp. 1167-1172, 1993.
6. J.E. Martin, M. Moskowitz, J.R. Milbrath, "Breast cancer missed by mammography, *American Journal of Roentgenology*, **132**, pp. 737-739, 1979.
7. G.M. te Brake, N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms, *Medical Physics*, **28**, pp. 259-266, 2001.
8. L. Li, Y. Zheng, L. Zhang, R. Clark, "False-positive reduction in CAD mass detection using a competitive classification strategy, *Medical Physics*, **28**, pp. 250-258, 2001.
9. N. Petrick, H.-P. Chan, B. Sahiner, D. Wei, "An Adaptive Density-Weighted Contrast Enhancement Filter for Mammographic Breast Mass Detection, *IEEE Transactions on Medical Imaging*, **15**, pp. 59-67, 1996.
10. M.L. Comer, S. Liu, 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).
11. M.A. Kupinski, M.L. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms", *IEEE Transactions on Medical Imaging*, **17(4)**, pp. 510-517, 1998.
12. L.M. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Steepest changes of a probability-based cost function for delineation of mammographic masses: a validation study", *Medical Physics*, **31(10)**, pp. 2741, 2004.

13. B. Van Ginneken, B.M. ter Häar Romeny, "Automatic Segmentation of Lung Fields in Chest Radiographs", *Medical Physics*, **27(10)**, pp. 2445-2455, 2000.
14. J.L. Fleiss, B. Levin, M.C. Paik, *Statistical Methods for Rates and Proportions*, Wiley, pp. 598-604, 2003.
15. J. R. Landis, G. G. Koch, "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, **33(1)**, pp. 159-174, 1977.
16. M. Heath, K.W. Bowyer, D. Kopans et al., "Current Status of the Digital Database for Screening Mammography", *Digital Mammography*, Kluwer Academic Publishers, 1998.

INTER-OBSERVER DIFFERENCES IN VALIDATION FOR DENSE BREAST MASS SEGMENTATION

Lisa Kinnard[a,b], Shih-Chung B. Lo[a], Erini Makariou[a], Eva Duckett[c], Mohamed Chouikha[b]

[a]ISIS Center, Dept. of Radiology, Georgetown University Medical Center, Washington, D.C., USA
[b]Department of Electrical and Computer Engineering, Howard University, Washington, D.C., USA
[c]Advanced Radiology, Glen Burnie, MD, USA

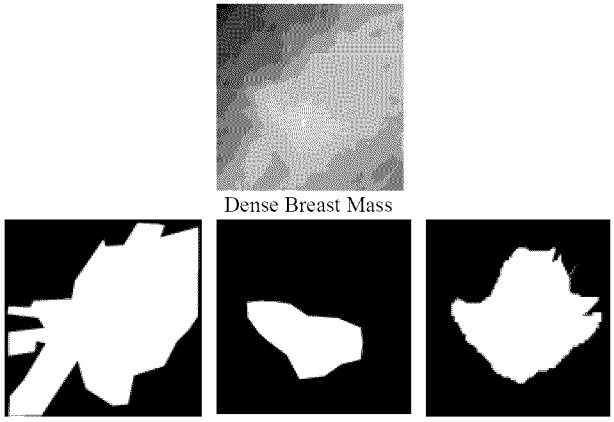
ABSTRACT

Validation of breast mass image segmentation algorithms is a key component of their success. However, in cases where the masses are embedded in dense tissue it is difficult to obtain consistent gold standard traces among expert radiologists. In this study we examined inter-observer variability by performing ANOVA tests ($p<0.05$) on a set of three segmentation traces, in efforts to decide upon the best trace. We used the overlap, accuracy, sensitivity, specificity, and Dice Similarity Index to validate the traces and discovered statistically significant results between one trace and the second and third traces. The p-values ranged between $1.4x10^{-2}$ and $4.86x10^{-25}$.

1. INTRODUCTION

One of the greatest challenges in validation of segmentation algorithms is inter-observer variability among gold standard traces. Typical studies use one to three observers when validating their algorithms, and strong agreement between these observers is desirable. Sahiner et. al. compared an automated breast mass segmentation method to manual traces of two expert radiologists and analyzed the degree of agreement between the two observers [1]. They calculated the minimum Euclidean distance, the Haussdorf distance, and the overlap measure and determined if the difference between the computer-segmented trace and the expert trace fell within the range of variation between observers. Pasquerault et. al. compared three segmentation algorithms for mammographic microcalcifications with an expert radiologist and three experienced scientists by independently rating the accuracy of each algorithms and then determining which method was preferred by each expert [2]. In both evaluation studies, intra-observer variability was addressed by allowing the observers to randomly view cases more than once. Zheng et. al. compared the performance of three digitized mammography CAD schemes after the images in question were rotated and resampled [3]. Specifically, their multiple image-based scheme matched regions by comparing the distance between centers of gravity of two Regions Of Interest (ROI) and the maximum radial length of either ROI. Zhou et. al. developed an automated nipple identification system, where two expert radiologists identified nipple locations on a set of digitized mammograms and the images either contained clearly identifiable nipples or invisible nipples [4]. For the invisible nipple locations one radiologist estimated their locations once, a second radiologist estimated their locations twice, and the three estimates were averaged.

Strong agreement between observers can be difficult to achieve due to an ROI's unclear borders (see Figure 1). Specifically, dense breast masses on digitized mammograms are difficult to observe and are therefore difficult to trace. It is also important that the segmentation algorithm is not biased toward a particular observer so we must incorporate as many observers as possible into a validation study. In this work we attempted to determine optimal computer segmentation masses for dense breast masses by studying inter-observer variability between a set of three expert radiologists.



Dense Breast Mass

Expert A Trace | Expert B Trace | Expert C Trace

Figure 1: Malignant Dense Breast Image With Three Expert Traces

2. METHOD

In previous work -- and in the current study -- we utilized a segmentation algorithm which combines region growing with likelihood function analysis [5,6]. This method narrows a large set of computer-segmented contours to three possible choices, and the ultimate goal is to choose the best contour from these three choices. In this study we are observing inter-observer variability between experts. We visually observed moderate to strong agreement between a pair of observers on breast masses with easily identifiable borders, however, for dense breast cases we observed that the agreement was not as strong. Furthermore, a colleague

pointed out large differences between observers and cited these differences as a critical area to be addressed in subsequent studies. We have performed a set of intra-observer studies that used the Analysis of Variance (ANOVA) test to compare the three final computer-segmented results to manual traces provided by three expert radiologists. The database, validation methods, and experiments are described in the next several sections.

2.1. Database

The database is a set of 124 malignant cases and 135 benign cases provided by the University of South Florida's Digital Database for Screening Mammography [7]. A set of expert radiologists manually traced the ROIs, where the first two observers were expert radiologists from Advanced Radiologists corporation (Expert A) and the Georgetown University Medical Center (Expert B), respectively. The third radiologist trace data (Expert C) was provided by the DDSM project, a collaborative effort between several hospitals. It appears that the DDSM expert data was provided by several expert radiologists, as some traces are tightly drawn around the ROI and other traces are not tightly drawn around the ROI. Since Experts A and B were instructed to trace the ROI borders as closely as possible, it was necessary to use the tightly drawn DDSM contours for the current study. There were approximately 40 DDSM tightly drawn traces for malignant masses and 26 tightly drawn traces for benign masses.

The three computer-segmented traces are henceforth referred to as: a) Group 1 trace: the trace encapsulating the central mass body, b) Group 2 trace: the trace encapsulating the central mass body and its extended borders (spiculations and projections, for example), and c) Group 3 trace: the trace encapsulating the mass body, its extended borders, and surrounding fibroglandular tissue which may or may not belong to the mass.

2.2. Validation

The segmentation method was validated on the basis of overlap, accuracy, sensitivity, specificity, and Dice Similarity Index (DSI) [8, 9]:

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

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

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

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

$$DSI = \frac{2N_{TP}}{N_{FN} + 2N_{TP} + N_{FP}} \quad (10)$$

where N_{TP} is the true positive fraction, N_{TN} true negative fraction, N_{FP} is the false positive fraction, and N_{FN} is the false negative fraction. The gold standards used for the validation study were mass contours, which have been traced by expert radiologists.

2.3. Experiments

The current study attempts to determine the optimal contour from a set of three contour choices determined by an automated segmentation method. We performed a set of intra-observer studies, which compared the computer-segmentation trace to each individual expert. Specifically, we made the following comparisons for Experts A, B and C: (a) group 1 vs. group 2 (b) group 2 vs. group 3 and (c) group 1 vs. group 3. Next we performed a set of inter-observer experiments, which compared the preferences of each observer. Specifically, for groups 1, 2, and 3: (a) Expert A vs. Expert B (b) Expert B vs. Expert C, and (c) Expert A vs. Expert C.

3. RESULTS

The experiments have been performed for both malignant and benign masses, however, in the interest of brevity results are shown for the malignant masses. Tables 1-6 contain p-values ($p<0.05$) for the ANOVA tests of the intra-observer experiments described in section 2.3, and mean values for all statistical measurements. In cases where the result was not statistically significant, the table entry reads "NS". Tables 7-12 contain p-values ($p<0.05$) for the ANOVA tests of the intra-observer experiments described in section 2.3, and mean values for all statistical measurements. In cases where the result was not statistically significant, the table entry reads "NS". Figures 2-5 show a computer segmented results and expert traces for four malignant masses embedded in dense tissue.

3.1 Statistical Results

Table 1: Expert A Intra-observer Experiment, Malignant Cases

	Gr. 1 vs. Gr. 2	Gr. 2 vs. Gr. 3	Gr. 1 vs. Gr. 3
Overlap	$5.32\text{x}10^{-11}$	NS	$4.09\text{x}10^{-15}$
Accuracy	$1.4\text{x}10^{-2}$	NS	$2.2\ \text{x}10^{-2}$
Sensitivity	$4.48\text{x}10^{-14}$	$8.4\ \text{x}10^{-3}$	$4.86\text{x}10^{-25}$
Specificity	$8.1\ \text{x}10^{-3}$	$2.5\ \text{x}10^{-3}$	$1.19\text{x}10^{-10}$
DSI	$1.03\text{x}10^{-10}$	NS	$1.82\text{x}10^{-15}$

Table 2: Mean Measurement Values for all Statistical Measurements (Expert A)

	Group 1	Group 2	Group 3
Overlap	0.28	0.44	0.46
Accuracy	0.71	0.76	0.76
Sensitivity	0.30	0.52	0.60
Specificity	0.98	0.94	0.89
DSI	0.41	0.59	0.62

Table 3: Expert B Intra-observer Experiment, Malignant Cases

	Gr. 1 vs. Gr. 2	Gr. 2 vs. Gr. 3	Gr. 1 vs. Gr. 3
Overlap	$6.55\text{x}10^{-8}$	NS	$4.99\text{x}10^{-6}$
Accuracy	NS	NS	NS
Sensitivity	$1.63\text{x}10^{-15}$	$2.12\ \text{x}10^{-2}$	$6.63\text{x}10^{-24}$
Specificity	$8.43\text{x}10^{-5}$	$1.12\ \text{x}10^{-3}$	$3.77\text{x}10^{-13}$
DSI	$1.03\text{x}10^{-7}$	NS	$2.77\text{x}10^{-6}$

Table 4: Mean Measurement Values for all Statistical Measurements (Expert B)

	Group 1	Group 2	Group 3
Overlap	0.36	0.50	0.47
Accuracy	0.81	0.83	0.81
Sensitivity	0.39	0.63	0.70
Specificity	0.97	0.92	0.86
DSI	0.51	0.64	0.62

Table 5: Expert C Intra-observer Experiment, Malignant Cases

	Gr. 1 vs. Gr. 2	Gr. 2 vs. Gr. 3	Gr. 1 vs. Gr. 3
Overlap	1.37×10^{-5}	NS	1.68×10^{-5}
Accuracy	NS	NS	NS
Sensitivity	1.24×10^{-6}	2.62×10^{-2}	3.74×10^{-12}
Specificity	3.67×10^{-2}	1.54×10^{-3}	2.57×10^{-6}
DSI	2.25×10^{-5}	NS	2.14×10^{-5}

Table 6: Mean Measurement Values for all Statistical Measurements (Expert C)

	Group 1	Group 2	Group 3
Overlap	0.32	0.48	0.47
Accuracy	0.79	0.83	0.81
Sensitivity	0.33	0.53	0.61
Specificity	0.98	0.96	0.89
DSI	0.47	0.63	0.63

Table 7: Inter-observer Experiment Results: Group 1 Trace Malignant Masses

	Exp. A vs. Exp. B	Exp. B vs. Exp. C	Exp. A vs. Exp. C
Overlap	8.10×10^{-3}	3.76×10^{-3}	NS
Accuracy	3.43×10^{-3}	1.87×10^{-2}	NS
Sensitivity	8.43×10^{-3}	2.66×10^{-3}	NS
Specificity	NS	NS	NS
DSI	8.53×10^{-3}	8.59×10^{-3}	NS

Table 8: Inter-observer Mean Measurement Values for Group 1 Traces (Malignant Masses)

	Expert A	Expert B	Expert C
Overlap	0.33	0.44	0.33
Accuracy	0.77	0.86	0.80
Sensitivity	0.35	0.46	0.34
Specificity	0.98	0.98	0.99
DSI	0.47	0.59	0.48

Table 9: Inter-observer Experiment Results: Group 2 Trace Malignant Masses

	Exp. A vs. Exp. B	Exp. B vs. Exp. C	Exp. A vs. Exp. C
Overlap	9.90×10^{-3}	3.68×10^{-3}	NS
Accuracy	2.62×10^{-3}	1.41×10^{-2}	NS
Sensitivity	7.24×10^{-3}	1.28×10^{-3}	NS
Specificity	NS	NS	NS
DSI	8.11×10^{-3}	6.91×10^{-3}	NS

Table 10: Inter-observer Mean Measurement Values for Group 2 Traces (Malignant Masses)

	Expert A	Expert B	Expert C
Overlap	0.48	0.59	0.48
Accuracy	0.81	0.89	0.84
Sensitivity	0.54	0.68	0.53
Specificity	0.96	0.95	0.96
DSI	0.62	0.73	0.63

Table 11: Inter-observer Experiment Results: Group 3 Trace Malignant Masses

	Exp. A vs. Exp. B	Exp. B vs. Exp. C	Exp. A vs. Exp. C
Overlap	NS	NS	NS
Accuracy	3.61×10^{-2}	NS	NS
Sensitivity	4.68×10^{-3}	2.37×10^{-4}	NS
Specificity	NS	NS	NS
DSI	NS	NS	NS

Table 12: Inter-observer Mean Measurement Values for Group 3 Traces (Malignant Masses)

	Expert A	Expert B	Expert C
Overlap	0.48	0.53	0.47
Accuracy	0.80	0.85	0.81
Sensitivity	0.648	0.78	0.62
Specificity	0.90	0.88	0.89
DSI	0.63	0.68	0.63

3.2 Visual Results

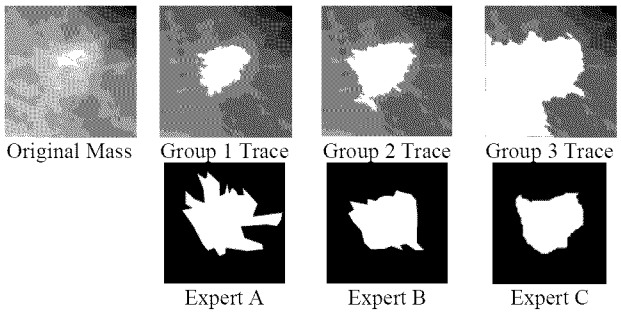


Figure 2: Malignant Mass Image with Computer Segmented Results and Expert Traces

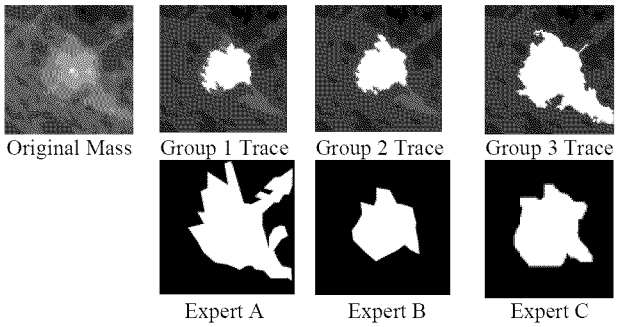


Figure 3: Malignant Mass Image with Computer Segmented Results and Expert Traces

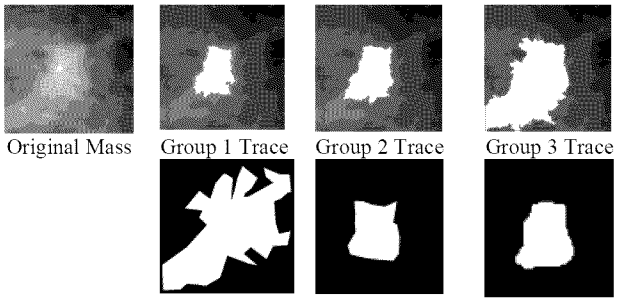


Figure 4: Malignant Mass Image with Computer Segmented Results and Expert Traces

4. DISCUSSION OF RESULTS AND CONCLUSION

4.1 Intra-observer Result Discussion

The statistical analysis shows that there were statistically significant differences for Expert A regarding the experiment that tested the group 1 traces versus the group 2 traces and for the experiment that tested the group 1 trace versus the group 3 traces for all statistical measurements. There were no statistically significant differences for the overlap, accuracy, and DSI measurements between the group 2 and group 3 traces, but the mean values for the group 3 traces were slightly higher than those of group 2. There were statistically significant differences for Expert B regarding the experiment that tested the group 1 traces versus the group 2 traces and for the experiment that tested the group 1 trace versus the group 3 traces for nearly all statistical measurements. There were no statistically significant differences for the overlap, accuracy, and DSI measurements between the group 2 and group 3 traces, but the mean values for the group 2 traces were slightly higher than those of group 3. There were statistically significant differences for Expert C regarding the experiment that tested the group 1 traces versus the group 2 traces and for the experiment that tested the group 1 trace versus the group 3 traces for nearly all statistical measurements. There were no statistically significant differences for the overlap, accuracy, and DSI measurements between the group 2 and group 3 traces, but again the mean values for the group 2 traces were slightly higher than or equal to those of group 3.

4.2 Inter-observer Result Discussion

The statistical analysis shows statistically significant differences in the experiments for Expert A versus Expert B and Expert A versus Expert C for nearly all statistical measurements for the group 1 and group 2 traces; however, for the group 3 trace there were few statistically significant differences between Experts.

4.3 Conclusion

The intra-observer results show that Experts B and C tend to favor the group 2 traces in comparison to the groups 1 and 3 traces. However, Expert A tends to favor the group 3 trace, in comparison to the groups 1 and 2 traces. These results are consistent with the fact that Expert A tends to draw larger traces, and the group 3 trace is always the largest of the three computer segmentation results. The inter-observer results show that the group 1 and group 2 traces are more closely correlated with Expert B, than with Experts A and C for nearly all statistical measurements. This is probably the case because Expert B appeared to have to traced the largest mass area which encapsulates the mass without including surrounding fibroglandular tissue.

Overall it appears that the group 2 trace may be the optimal contour trace for the aforementioned segmentation algorithm and in future work, we will test the effect of using the various segmentation results upon the results of a CAD_X system. If possible, we will also incorporate more expert radiologist traces.

5. ACKNOWLEDGMENTS

This work was supported by US Army Grant numbers DAMD17-01-1-0267, DAMD 17-03-1-0314, and NIH/NCI grant R21 CA102960-01.

6. REFERENCES

[1] B. Sahiner, N. Petrick, H.-P. Chan, L.M. Hadjiski, C. Paramagul, M.A. Helvie, M.N. Gurcan, "Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization", *IEEE Trans. on Med. Imag.*, vol. 20, no. 12, pp. 1275-1284, 2001.

[2] S. Pasquerault, L.M. Yarusso, J. Papaioannou, Y. Jiang, R.M. Nishikawa, "Radial gradient-based segmentation of mammographic microcalcifications: Observer evaluation and effect on CAD performance", Med. Phys., vol. 31, no. 9, pp. 2648-2657, 2004.

[3] B. Zheng, G.S. Maitz, M.A. Ganott, G. Abrams, J.K. Leader, D. Gur, "Performance and Reproducibility of a Computerized Mass Detection Scheme for Digitized Mammography Using Rotated and Resampled Images: An Assessment", *AJR*, vol. 185, pp. 194-198, 2005.

[4] C. Zhou, H.-P. Chan, C. Paramagul, M.A. Roubidoux, B. Sahiner, L.M. Hadjiski, N. Petrick, "Computerized nipple identification for multiple image analysis in computer-aided diagnosis", Med. Phys., vol. 31, no. 10, pp. 2871-2882, 2004.

[5] L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M.T. Freedman, M. Chouikha, "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study," *Med. Phys.*, vol. 31, no. 10, p. 2796-2810, 2004

[6] M.A. Kupinski, M.L. Giger, "Automated Seeded Lesion Segmentation on Digital Mammograms", *IEEE Trans. on Med. Imag.*, vol. 17, no. 4, pp. 510-517, 1998.

[7] M. Heath, K.W. Bowyer, D. Kopans et al., "Current status of the digital database for screening mammography", Digital Mammography, Kluwer Academic Publishers, pp. 457-460 , 1998.

[8] B. Van Ginneken, "Automatic segmentation of lung fields in chest radiographs", *Med. Phys.*, 27, pp. 2445-2455, 2000.

[9] M. Maddah, K.H. Zou, W.M. Wells, R. Kikinis, S.K. Warfield, "Automatic Optimization of Segmentation Algorithms Through Simultaneous Truth and Performance Level Estimation (STAPLE)", MICCAI 2004, pp. 274-282, 2004.

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 $1.03\times10^{-2}-7.51\times10^{-17}$ for the single observer studies and $2.03\times10^{-2}-9.43\times10^{-4}$ 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 Karssemeijer[9] 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. Li[10] 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 Giger[16] 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 work[9] 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 study[16] 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 pixel ×512 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 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.

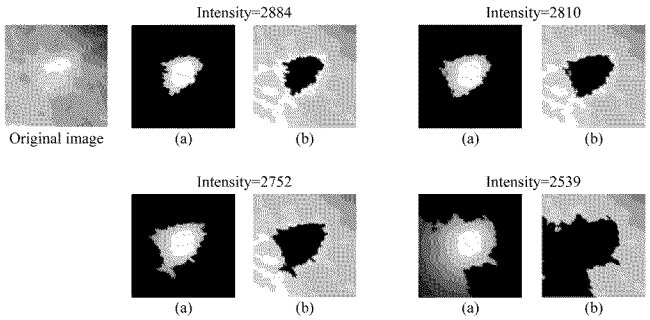


FIG. 1. Four grown contours used to construct the cost function: starts from high intensity thresholds and moves towards low intensity thresholds. Each contour separates the ROI into two parts: (a) Segmented image (based on processed image) used to compute density function $p(f_i(x,y)|S_i)$ and (b) masked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

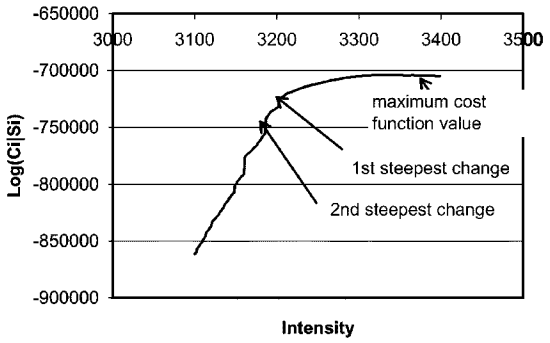
The processed image was mainly used to construct the cost function. A common technique used in mass segmentation studies is to pre-process the images using some type of filtering mechanism[11,16] in an effort to separate the mass from surrounding fibroglandular tissue. This method could be particularly 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

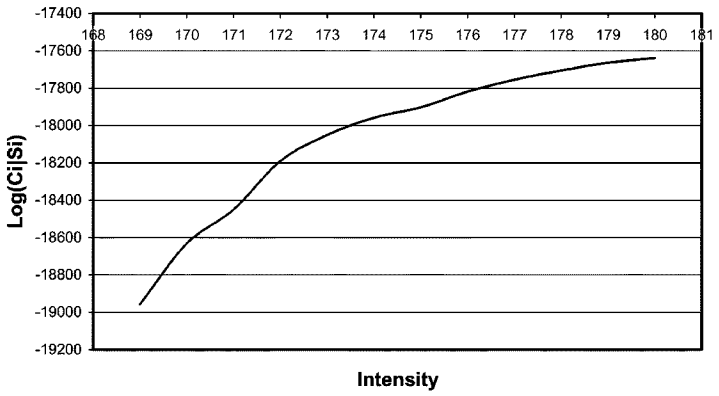
In the context of this work, the composed probability is defined as the probability density functions of the pixels inside and outside a contour using a processed and nonprocessed version of an image. Specifically, for a contour (S_i), the composed probability (C_i) is calculated:

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

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:

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

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:

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

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

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

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

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

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 $(d(t)\geqslant \mathrm{TV}_1); \quad t=0,\ldots,m$

Then choice 1 = intensity where that condition is satisfied.

If $(d(t)\geqslant \mathrm{TV}_2); \quad t=m,\ldots,z$

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 [$d(t)_{\mathrm{MC1}}$ 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 [$d(t)_{\mathrm{MC2}}$ 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 [$d(t)_{\mathrm{MC1}}$ 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 [$d(t)_{\mathrm{MC2}}$ 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:

$$\mathrm{Overlap}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{FN}}+N_{\mathrm{TP}}+N_{\mathrm{FP}}}, \tag{6}$$

$$\mathrm{Accuracy}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}, \tag{7}$$

$$\mathrm{Sensitivity}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}, \tag{8}$$

$$\mathrm{Specificity}=\frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}}+N_{\mathrm{FP}}}, \tag{9}$$

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

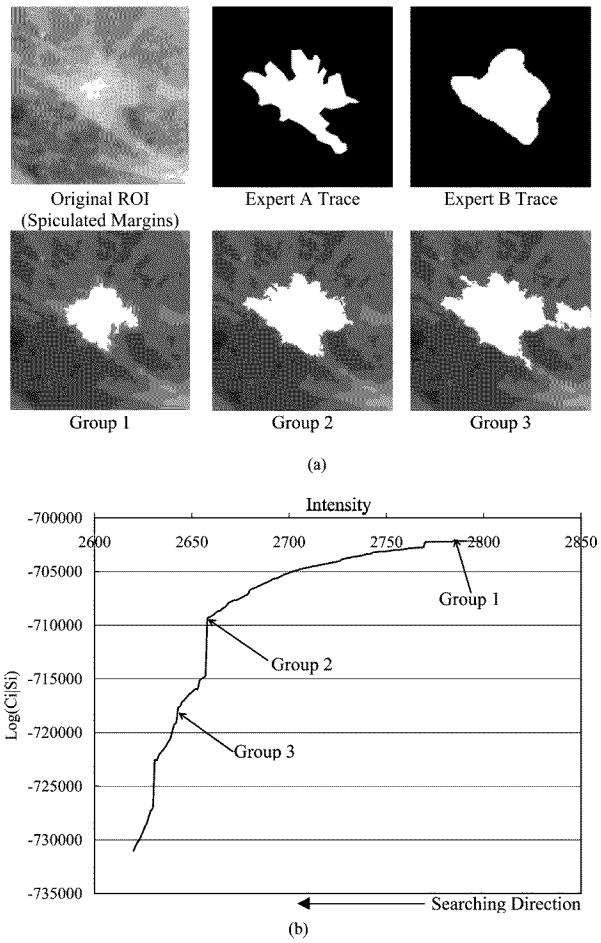


FIG. 3. (a) Segmentation results for a malignant mass with spiculated margins (subtlety=2) (b) the corresponding cost function.

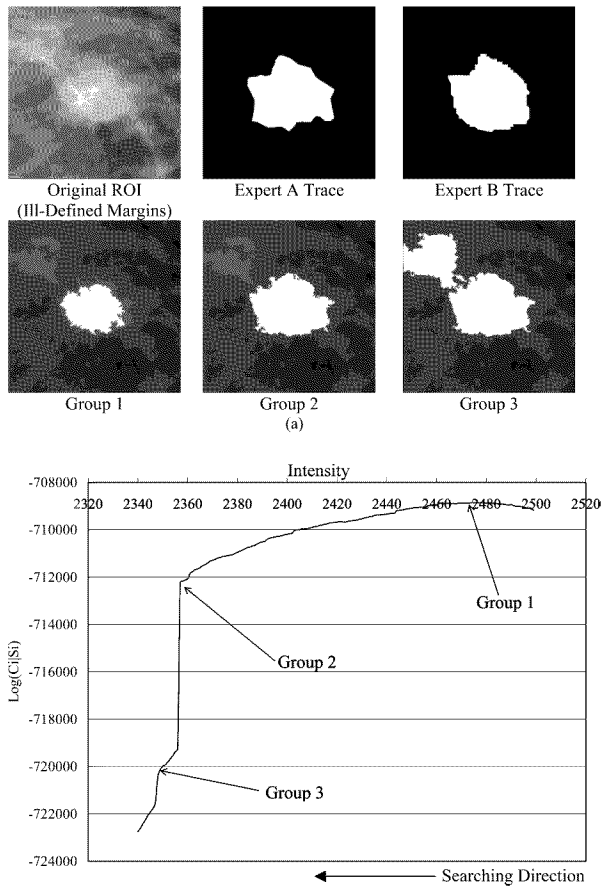


FIG. 4. (a) Segmentation results for a malignant mass with ill-defined margins (subtlety=3); (b) the corresponding cost function.

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.

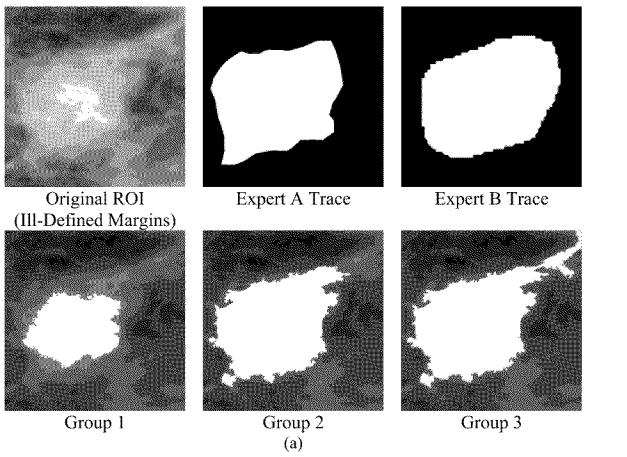
The experiments produced contours for the intensity values resulting from three locations within the cost functions: (1) The intensity of the maximum value within the cost function; (2) the intensity for which the cost function experiences its first steepest change; and (3) the intensity for which the 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 Variance (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.



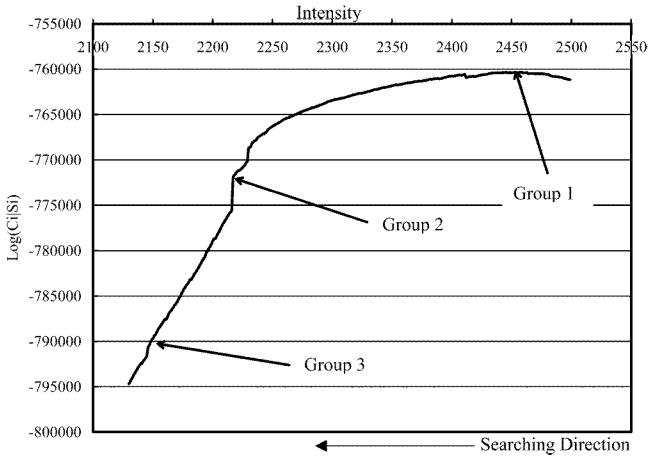
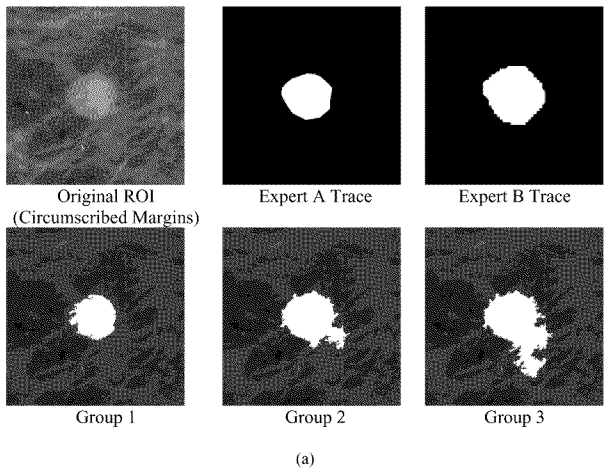


FIG. 5. (a) Segmentation results for a benign mass with ill-defined margins (subtlety=3); (b) the corresponding cost function.



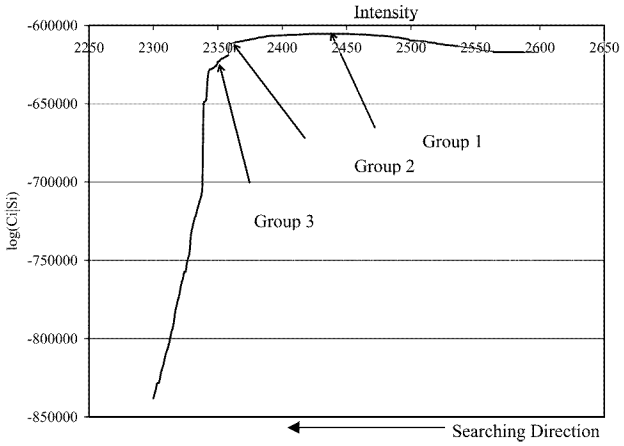


FIG. 6. (a) Segmentation results for a benign mass with circumscribed margins (subtlety=4); (b) the corresponding cost function.

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 be--nign 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	*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)	3.19×10^{-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	*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)	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	*P*-value (group 1 vs group 2)	*P*-value (group 2 vs group 3)	*P*-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	*P*-value (group 1 vs group 2)	*P*-value (group 2 vs group 3)	*P*-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 study[14] 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 study[9] 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 study[16] 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. RCMI/NCRR/NIH 2G12RR00348. The authors would also like to thank the referees for their constructive comments and recommendations.

APPENDIX A—GALLERY OF SEGMENTATION RESULTS

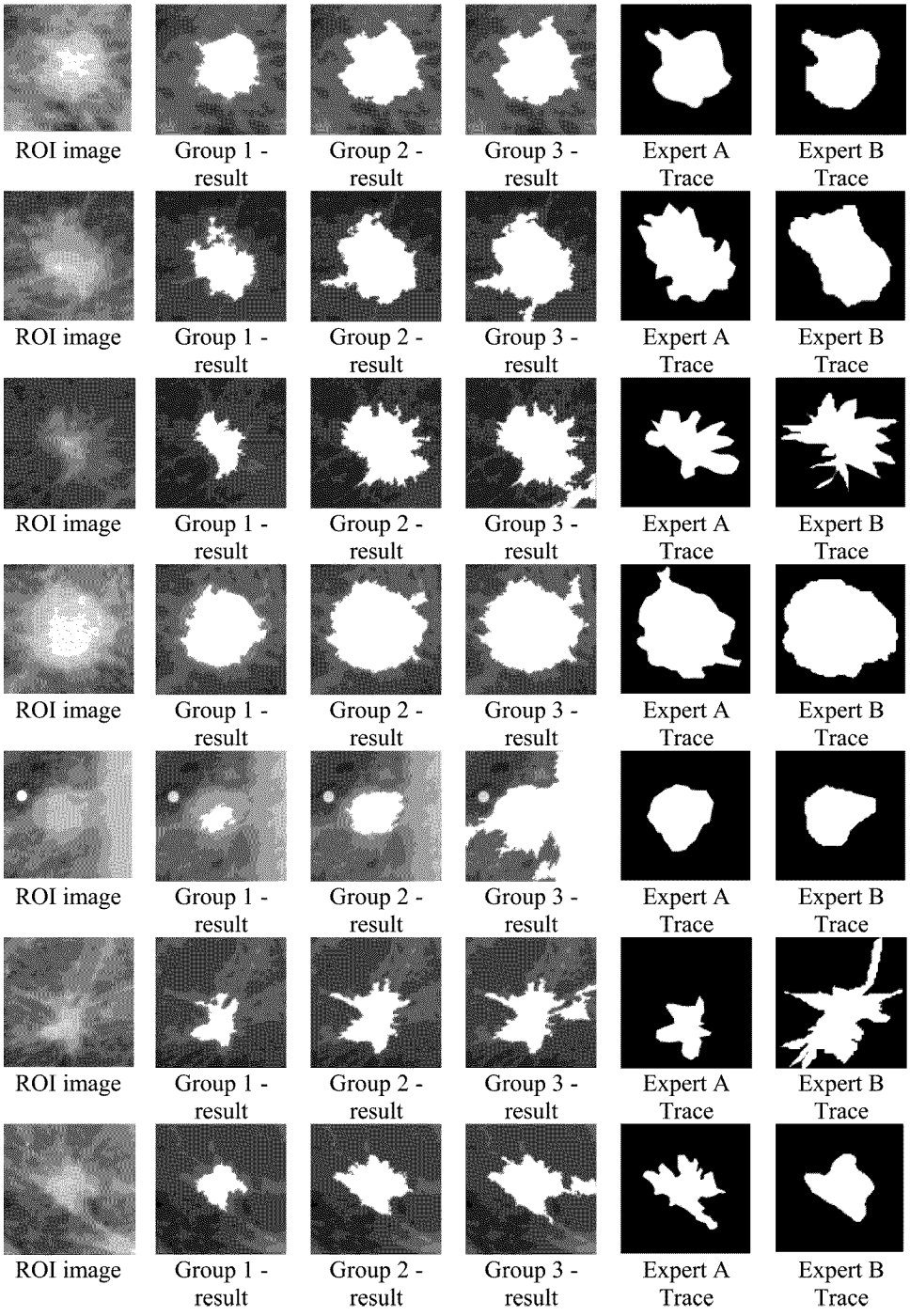


FIG. 7. Segmentation results for a set of malignant masses.

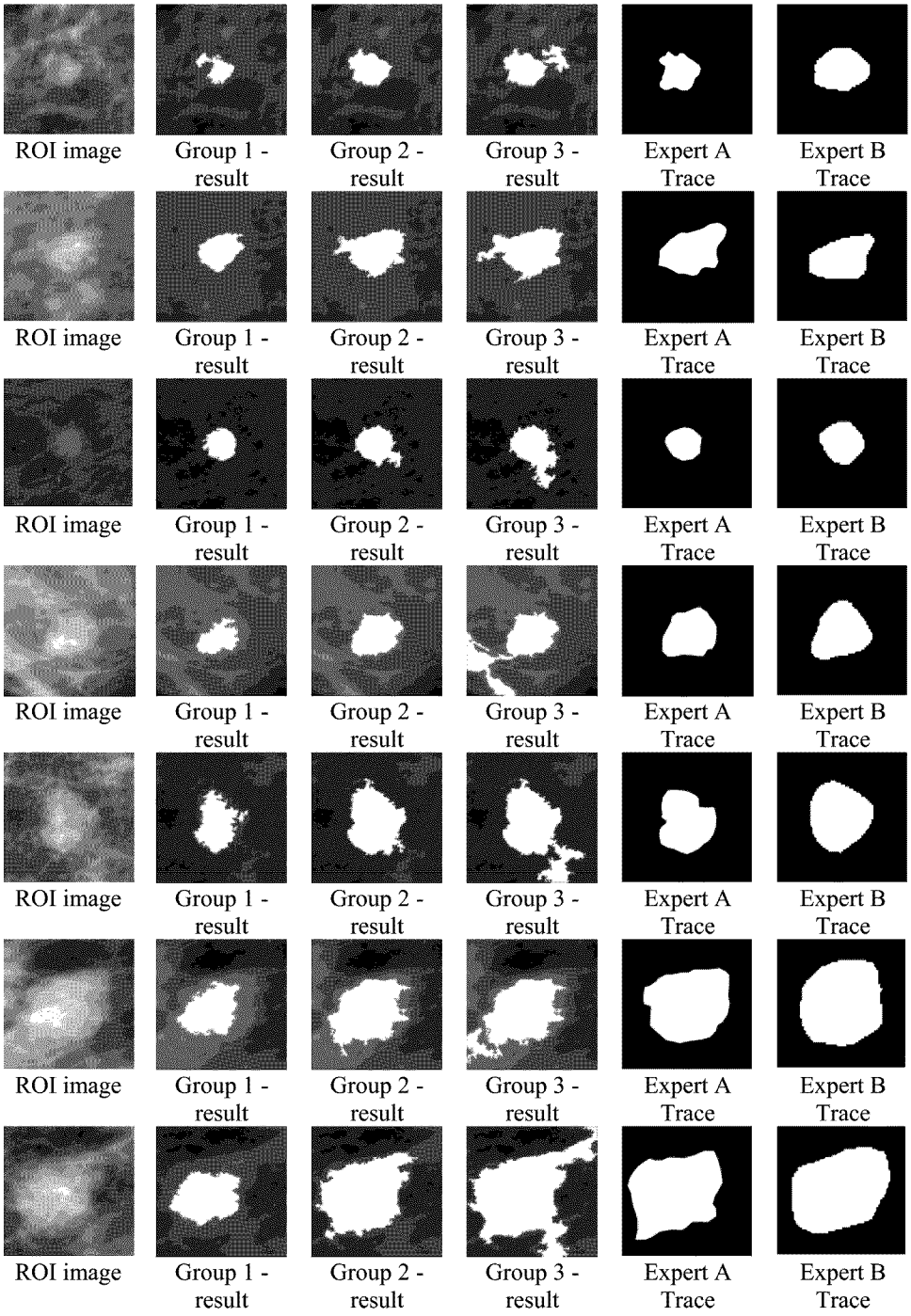


FIG. 8. Segmentation results for a set of benign masses.

[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).
[10] L. 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).
[11] N. 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).
[12] S. 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).
[13] A. 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).
[14] H. 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).
[15] M. 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.
[16] M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," IEEE Trans. Med. Imaging **17**, 510–517 (1998).
[17] S.-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).
[18] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers, M. P. DeSimio, D. W. Ruck, J. W. Hoffmeister, and R. A. Raines, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Trans. Med. Imaging **16**, 811–819 (1997).
[19] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," Med. Phys. **28**, 1455–1465 (2001).
[20] R. 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).
[21] B. 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).
[22] J. 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).
[23] B. Van Ginneken, "Automatic segmentation of lung fields in chest radiographs," Med. Phys. **27**, 2445–2455 (2000).
[24] D. Downing and J. Clark, *Statistics the Easy Way*, 2nd ed. (Barron's Educational Series, Hauppauge, 1989), pp. 184–206.
[25] W. Hopkins, *A New View of Statistics: P Values and Statistical Significance*; available online at www.sportsci.org/resource/stats/pvalues.html.
[26] M. Heath *et al.*, "Current status of the digital database for screening mammography," *Digital Mammography* (Kluwer Academic, New York, 1998), pp. 457–460.