SANDIA REPORT

SAND97-8233 • UC–408
Unlimited Release
Printed February 1997

Decision Trees and Integrated Features for Computer Aided Mammographic Screening

W. Philip Kegelmeyer, Jr., Bennett Groshong, Mark Allmen, and Kevin Woods

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94551
for the United States Department of Energy
under Contract DE-AC04-94AL85000

Approved for public release; distribution is unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.
NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of the contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors or subcontractors.

This report has been reproduced from the best available copy.

Available to DOE and DOE contractors from:

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge TN 37831

Prices available from (615) 576-8401, FTS 626-8401.

Available to the public from:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.
Springfield, VA 22161

SAND 97-8233

Distribution
Category UC-408

Unlimited Release
Printed February, 1997

Decision Trees and Integrated Features for Computer Aided Mammographic Screening

by

W. Philip Kegelmeyer, Jr, Bennett Groshong, Mark Allmen, and Kevin Woods
Center for Computational Engineering
Sandia National Laboratories
Livermore, California 94551

Abstract

Breast cancer is a serious problem, which in the United States causes 43,000 deaths a year, eventually striking 1 in 9 women. Early detection is the only effective countermeasure, and mass mammography screening is the only reliable means for early detection. Mass screening has many shortcomings which could be addressed by a computer-aided mammographic screening system.

Accordingly, we have applied the pattern recognition methods developed in earlier investigations of spiculated lesions in mammograms to the detection of microcalcifications and circumscribed masses, generating new, more rigorous and uniform methods for the detection of both those signs. We have also improved the pattern recognition methods themselves, through the development of a new approach to combinations of multiple classifiers.

Acknowledgments

This effort was supported by the United States Department of Energy through the Sandia National Laboratories LDRD program, contract # DE-AC04-76DO00789, and by the Department of the Army research grants #DAMD17-94-J-4328 and #DAMD17-94-J-4015. Additional general computing resources were donated by the Hancock Family Breast Cancer Foundation.

In the context of the microcalcification detection work, we gratefully acknowledge the aid of Nico Karssemeijer, of the University Hospital Nijmegen, who provided the input test data, its cluster truth data, and detection results for the pixel truth data, and that of Laura Mascio, of Lawrence Livermore National Laboratory, who also provided detection results.

In the context of the circumscribed masses work, we are grateful to the Division of Radiologic Sciences at the Bowman-Gray School of Medicine for its support in acquiring a mammogram database. In particular, we wish to thank Pete Santago in Medical Engineering for his assistance in arranging the project, and Jeff Carr in Radiology for his help in designing the database and acquiring films. In addition, Rita Fremanis, David Brown, and Mark Hiatt in Radiology provided instrumental assistance in selecting, acquiring, and truthing cases, and Bill Hinson in Medical Engineering provided support for the digitizing equipment.

Contents

1 Introduction **13**
1.1 Computer Aided Mammographic Screening . . . 13
1.2 Context and History . . . 14
1.3 An Overview of this Report . . . 15

2 Statement of Problem and Method **17**
2.1 The Basic Mammographic Cancer Signs . . . 17
2.2 Why Current and Previous Work is Insufficient . . . 17
2.3 How To Address Those Deficiencies . . . 18

3 Microcalcification Detection **21**
3.1 Problem Statement . . . 21
3.2 Hypothesis . . . 21
3.3 Methods . . . 22
3.3.1 Test Data and Performance Criteria . . . 22
3.3.2 Initial Segmentor Set . . . 23
3.3.3 The Initial Feature Set . . . 25
3.3.4 Feature Set Reduction . . . 26
3.4 Results . . . 28
3.5 Conclusions and Future Work . . . 31

4 Mass Detection **33**
4.1 Problem Statement . . . 33
4.2 Hypothesis . . . 33
4.3 Methods . . . 35
4.3.1 Subsampling and Segmentation . . . 35
4.3.2 Boundary brightness Gradient Flattening . . . 36
4.3.3 Contrast Variability Reduction . . . 38
4.3.4 Suppression of linear structures . . . 38
4.3.5 Edge detection . . . 39
4.3.6 Edge Sorting and Thresholding . . . 39
4.3.7 Circular Hough transform . . . 39
4.3.8 Hough peak normalization . . . 41

4.3.9 Hough peak qualification . . . 41
4.3.10 Contrast rectification . . . 42
4.3.11 Feature mapping . . . 43
4.4 Experimental Results . . . 43
4.4.1 Results for the MIAS Database . . . 43
4.4.2 Results for the BGSM Database . . . 47
4.5 Conclusions . . . 56
4.6 Discussion . . . 56

5 Combination of Multiple Classifiers 57
5.1 Introduction . . . 57
5.2 Context and Related Work . . . 57
5.3 Our Method Compared to Related Approaches . . . 58
5.3.1 The Proposed Approach: DCS-LA . . . 59
5.3.2 The Behavior-Knowledge Space Approach . . . 59
5.3.3 The Classifier Rank Approach . . . 59
5.3.4 A Modified Classifier Rank Approach . . . 59
5.4 Methods . . . 60
5.4.1 Data Sets . . . 60
5.4.2 ROC Analysis . . . 60
5.4.3 Individual Classifiers . . . 61
5.4.4 DCS-LA Implementation and Application . . . 62
5.5 Results . . . 62
5.6 Conclusions . . . 66

6 Discussion 67
6.1 An Outline of Accomplishments . . . 67
6.2 Future Work . . . 68

7 Bibliography 71

A General Notes on the Appendices 77

B Microcalcification Issues 79

C Mass Feature Issues 81
C.1 Notes and Experiments . . . 81
C.2 Unexplored Avenues . . . 91
C.3 Feature Analysis Methods and Experiments . . . 91

D Miscellaneous Issues 93
D.1 Classification Coding . . . 93
D.2 Analysis Tools . . . 94

List of Figures

1.1 Densities to Features to Probabilities 15

3.1 The left column shows a portion of a mammogram with three calcifications arranged in a triangle. It is shown twice to facilitate easier viewing along rows. The top row shows the results of the local segmentors (from left to right) MES, Peak, and Adaptive Threshold. The second row shows the results of the global segmentors (from left to right) Boundary, Woods and Outlier. 27
3.2 Initial ROC Curve Comparing Random Field Model and Dense Feature Map Performance 29
3.3 Results on the Few, Some, and All Datasets 30
3.4 Results on the Few, Some, and All Datasets 30

4.1 Subimages of selected MIAS masses 34
4.2 Circumscribed mass detection algorithm flowchart 35
4.3 Image and boundary histogram examples 37
4.4 Linear feature suppression function shapes 38
4.5 Edge point qualification 39
4.6 Circular Hough Transform Accumulator 40
4.7 Mass contrast schematic and image mapping functions 42
4.8 Peak, original, and summed disk feature images for selected mammograms from the MIAS database. 45
4.9 Plots of features across the 44 mammogram MIAS data set. 46
4.10 Performance Results as a ROC Curve for the MIAS data set. 47
4.11 Peak, original, and summed disk feature images for selected mammograms from the BGSM mammogram data set. 49
4.12 Plots of features across part of the BGSM mammogram data set. 50
4.13 Performance Results as a ROC Curve for the BGSM data set. 51

5.1 A) Partial ROC curves for 6 individual classifiers, and their AUCs. B) Composite ROC curve for the individual classifiers, and the ROC curve for an Oracle classifier. 63
5.2 The composite and Oracle ROC curves for the 6 individual classifiers compared to the results for the DCS-LA, Behavior Knowledge Space, Classifier Rank, and Modified Classifier Rank methods. 65

C.1 Plot of radial edge location error vs angle around circle for a radius of 30 pixels. . 84
C.2 Plot of radial edge angle error vs angle around circle for a radius of 30 pixels. . . 85
C.3 Plot comparing radial edge angle error for the “Canny” and “glmax” edge detectors. 85
C.4 Plot comparing radial edge angle error for the “glmax” edge detector operating from a Gaussian blur of 1 and 2 for a circle radius of 30 86
C.5 Plot comparing radial edge angle error for the “glmax” edge detector operating from a Gaussian blur of 1 for circle radii of 10, 30, and 60. 87
C.6 Maximum pixel error wrt circle center for radii 2-60 and Gauss filter sigma = 1 and 2 88
C.7 angle error wrt circle center for radii 10,30,60 and Gauss filter sigma = 1 89
C.8 angle error wrt circle center for radii 10,30,60 and Gauss filter sigma = 2 89

List of Tables

3.1 The features computed for each object (except for Laws) returned by a segmentor, with a partial list of researchers who have used the feature or one similar. 26
3.2 The Reduced Set of Selected Features 28

4.1 Key to column captions and numerical table entries. 51
4.2 BGSM dataset performance analysis - part 1 52
4.3 BGSM dataset performance analysis - part 2 53
4.4 BGSM dataset performance analysis - part 3 54
4.5 BGSM dataset performance analysis - part 4 55

5.1 CMC results with individual classifiers set to TP rates as close to 80% as possible. All classifiers agree for 22,552 of the samples, or about 89.5% of the time. 66

C.1 Circle and Hough accumulator peak radius correlation 82
C.2 Comparison of peaks for clean and noisy circle images. 83
C.3 Edge accuracy statistics for images filtered with a Gaussian $\sigma = 1, 2$ 87

Acronyms and Abbreviations

ANN	Artificial Neural Network
AUC	Area Under the ROC curve
BDT	Binary Decision Tree
BGSM	Bowman-Gray School of Medicine
BKS	Behavior-Knowledge Space
CAMS	Computer Aided Mammographic Screening
CART	Classification And Regression Trees
CCD	Charge Coupled Device
CDF	Cumulative Distribution Function
cdf-m	cumulative distribution function minus mean
CMC	Combination of Multiple Classifiers
DCS	Dynamic Classifier Selection
DCS-LA	Dynamic Classifier Selection by Local Accuracy
DFM	Dense Feature Map(ping)
FAR	False Alarm Rate
FN	False Negative
FP	False Positive
KNN	K Nearest Neighbor
MC	MicroCalcification
MES	Minimal Error Segmentor
MIAS	Mammographic Image Analysis Society
MSE	Mean Squared Error
P_d	Probability of Detection
PDF	Probability Distribution Function
QSFS	Quasi-Sequential Feature Selection
RER	Redistribution Error Rate
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SD	Standard Deviation
SE	Standard Error
TN	True Negative
TP	True Positive

Chapter 1

Introduction

1.1 Computer Aided Mammographic Screening

Breast cancer is a serious problem, which in the United States causes 43,000 deaths a year, striking 1 in 9 women[1]. Even by age 45, one woman in 93 has contracted breast cancer. Early detection has a *dramatic* impact, however, raising the survival rate from 50% to 95%. Early intervention also permits the use of much less extreme and expensive medical remedies, and so improves quality of life while greatly reducing health care costs.

Mass radiological screening is the only reliable means for early detection. Unfortunately, since only one out of each hundred asymptomatic women has a lesion, examining the resulting mammograms is tedious and inefficient, which leads to missed cancers. Further, there is an intense need for many more exams than are currently conducted. The American population is currently significantly undertested, with only 10% of the screenings recommended by the American Cancer Society actually being performed. There are a handful of reasons for this, but an important one is that there are simply not enough radiologists. If *every* radiologist of every discipline were to convert over to solely reading mammograms, we'd still have only a third of the radiologists required[1].

These problems can be addressed by making mammographers more efficient, and mammography more attractive. As it stands, the probability of detection in mammographic mass screening is relatively low, around 85%. This is largely due to fatigue and the volume of cases that have to be considered, as the sensitivity goes to almost 100% if the mammographer is given enough time to consider the case. A reliable and tireless computer system that can assist the radiologist, bringing to their attention lesions they may have overlooked, thus helps in a number of ways. It provides for early detection and will reduce the amount of radiologist time spent per case. This last fact is particularly important in the economic sense, as radiologist's time accounts for approximately half the cost of mammograms.

Accordingly, Sandia's LDRD Program, under its Biomedical Engineering Strategic Initiative area, funded a related three-year research project. The goal of the project was to apply prior work in the general area of computer-aided mammographic screening to additional cancer signs, generating a full single-sign solution integrated with a uniform pattern recognition method.

This SAND report is the documentation for that project.

1.2 Context and History

The work reported here is not Sandia's first effort in this field. From 1989 until 1993 there was an intermittently funded effort, under Sandia's Technology Maturation Process, to investigate spiculated lesions, the most virulent of the single cancer signs. When spotted, they are associated with a malignant cancer 95% of the time, and so are always immediately biopsied. They are also, unfortunately, among the trickiest to detect, as they are a lesion without an obvious center or mass, and manifest generally only as a disturbance in the usual interior structure of the breast, a phenomenon sometimes referred to as "architectural distortion".

Nonetheless, the spiculated lesion detection methods were sufficiently successful that they were tested in a retrospective clinical study at Scott and White Research Clinic in Temple, TX[2]. 85 mammogram cases were analyzed by computer, and 4 mammographers then read those cases themselves, with and without the aid of the computer. The result was that the radiologist probability of detection increased 9.7%, from 80.6% to 90.3%. This result has a statistical p-value of 0.005, which means that there is only one chance in 200 that it was a statistical fluke. Further, and almost as importantly, there was *no* increase in false alarm rate: the increase in detection sensitivity came with no penalty.

Though there were some innovative features developed in the spiculated lesion work, most of the dramatic results were due to the new ideas in pattern recognition that were developed concurrently. Detection of any property in images, lesions included, is basically pattern recognition. Whenever one designs a pattern recognition algorithm the solution generally breaks into two pieces: the "inference engine" which makes the classifications, and the image measurements, called "features", on which the classifications are based. The classification method tends to be problem independent, while the features are tailored to the phenomena to be detected and the sort of images they arise in.

A "feature" can any measurement on an image. One of the innovations in the spiculated lesion work was the development of the "dense feature map" (DFM) approach to feature extraction. In the DFM method, all features are "pixel-based", which means that they are each associated with a particular pixel, and are extracted from some local window centered on that pixel. The basic idea is to convert each pixel to a handful of measures which are indicative of cancer or normal tissue. These measures are then analyzed by a classification method to assign each pixel a probability of being cancerous; a visual representation is in Figure 1.1. The output of this process, the "probability image", was the second innovation of the prior work. The third main feature was the use of binary decision trees (BDTs) as the classification method. BDTs, with their ability to ignore useless features and their capacity for rapidly processing large amounts of data, were the only pattern recognition method that would have made possible application of the DFM method to the huge images common to digital mammography.

According, the aim of the current research was to apply these same pattern recognition methods to the detection of the other two important cancer signs: microcalcifications and circumscribed masses.

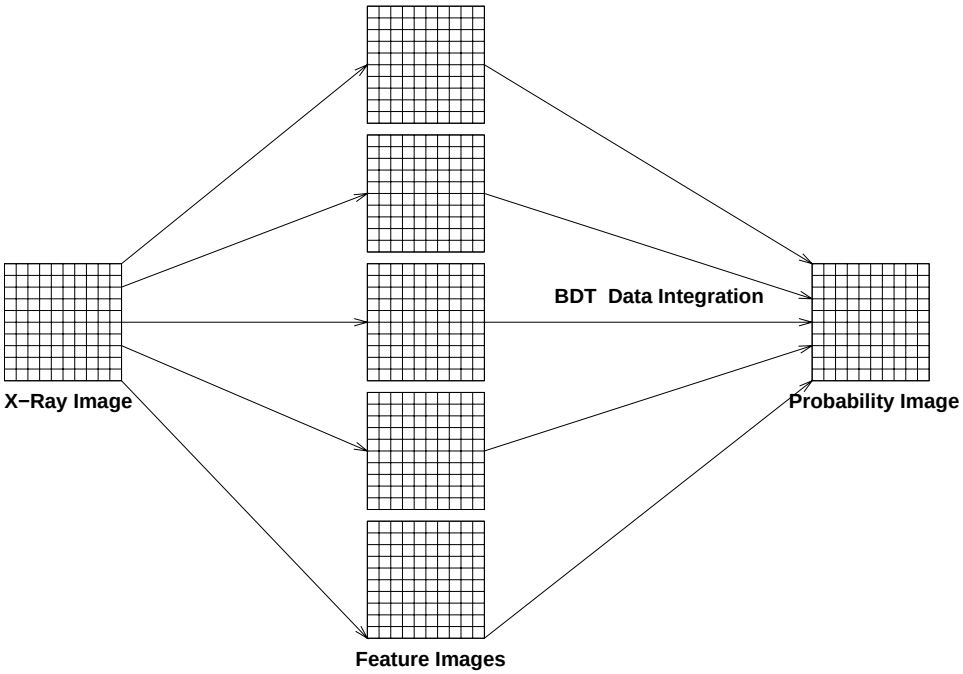


Figure 1.1: Densities to Features to Probabilities

1.3 An Overview of this Report

The next section, Chapter 2, will discuss in more detail the nature of mammographic data and the motivation behind the approach we took to its analysis. Chapter 3 presents methods, analysis, and results for our investigation into the detection of microcalcifications, and Chapter 4 does the same for the detection of circumscribed masses. Chapter 5 discusses our work on the second half of the detection process: improving the classification methods themselves. Chapter 6 ends the main portion of the report with a review of the accomplishments and some thoughts on how to follow up on our results. Finally, we close with a string of appendices of interest mainly to someone who might wish to continue this work, as they list the various loose ends and interesting but incompletely explored avenues that we accumulated in the course of these investigations.

Chapter 2

Statement of Problem and Method

2.1 The Basic Mammographic Cancer Signs

In America, the basic mammographic exam involves two x-ray views of each breast, resulting in a total of four films. There is a view taken from above (the "cranio-caudal" or "CC" view) and one from an angled side view (the "medio-lateral oblique" or "MLO" view). In both cases the breast is compressed between paddles during the procedure; this allows for a reduction in radiation dose, improved image quality, and some separation of overlapping tissues. The radiologist typically looks at the current four films in conjunction with whatever films may be available from earlier exams.

There are six basic signs that radiologists use to detect cancers in mammograms. Three, spiculated lesions, circumscribed masses, and calcifications, apply to single mammogram views and have all received some attention in the literature. Three apply to pairs of images; these are the detection of asymmetry between the right and left breast, the presence or absence of a finding in both the top and side view of the same breast, and changes in mammographic breast appearance over time. Methods for these paired image signs are far less standard and straightforward, for both humans and computers.

A useful atlas of characteristic visual appearances for all of these cancer signs may be found in [3].

2.2 Why Current and Previous Work is Insufficient

Though challenging, mammogram interpretation is distinguished by the fact that it is a comparatively simple medical imaging problem, in that there is only one disease to look for. Further, though there are certainly great variations in mammogram appearance and structure, the breast can at least be imaged so that *only* the breast is involved. There are no bone, lung, or other obvious but irrelevant structures to confuse the interpretation.

As a result of this simplicity, investigation into computer aided mammographic screening dates back to 1972[4], though it was only in the late 1980's that computer and imaging technology had

advanced to the point where such research could blossom. Though work in this field to date has been successful enough in some specific domains, it has been hindered by the following deficiencies:

- *Point solutions.* That is, techniques that address one and only one of the single image cancer signs; there were no methods for considering even all of the single-view signs, much less all six. There are many approaches to microcalcification, circumscribed mass, or spiculated lesion detection individually (Sandia's earlier work falls in this category), but none which attempt to take a unified look at all three.
- *Poor information integration.* The problem of pattern recognition in images can be conceptually divided into image features and analysis of these features by an inference engine. The first half is the question of the individual features, the image characteristics one uses as low level clues as to the presence or absence of the pattern in question. The second is the information calculus, the means by which these clues are assembled into a decision concerning the existence of the pattern. Other researchers have unfortunately generally combined clever and insightful features with pattern recognition based on ad hoc heuristics, subjective decision boundaries, or parametric approaches whose assumptions are knowingly violated by the problem at hand.

2.3 How To Address Those Deficiencies

Sandia's previous work in the field of mammographic image analysis has resulted in the invention of a powerful image feature for texture characterization, the discovery of a separate set of features that can effectively detect *normal* mammographic tissue, and the demonstration of the utility of dense feature representations. One of the results of the latter concept was a unique extension of binary decision tree pattern recognition theory in the form of "probability images", which are spatially dense characterizations of the likelihood of cancer in mammogram images.

Our approach in the current work is to simultaneously address the difficulties with point solutions and poor information integration by incorporating new or existing features for microcalcifications and circumscribed masses into the current pattern recognition system for spiculated lesions. We did this by exploiting the BDT's capacity for feature exploration as well as integration. That is, the BDT can be used as a tool to determine the relative value of a feature for separating classes, and can further suggest refinements to improve a good but not perfect feature. So the early methodology was the one of iterative refinement which worked so well with spiculated lesions; develop a variety of features, either invented from scratch or from the literature, use the BDT to determine which are the most promising, refine or expand on those, again use the BDT for evaluation, and continue until an effective subset of features has been determined.

The advantages of this approach are best appreciated by way of contrast with the previously standard method. The usual approach, in mammography and elsewhere, is to identify regions of interest (ROIs) and make measurements on these. If a lesion fails to be assigned an ROI, however, then some important information will be lost, no matter how clever the following processing. Dense, pixel-level feature processing requires considerably more computer processing, but provides a level of redundancy and flexibility that ROI processing cannot match. Further, it allows all the cancer

information for both single-image and paired-image detections to be consolidated into one powerful representation, the probability image.

A particularly satisfying aspect of this solution, and one that typifies the rigor possible with the use of the BDT, is that there is no human fudging or guessing. Humans are good at classification, but much less skilled at explaining what they are doing, at codifying *how* they do their classifications. As a result, human heuristics and decision procedures, once implemented for use by a computer, are notoriously brittle and easily broken by new, unfamiliar data.

The BDT is a *trained* classifier, however, and is trained in an automatic, hands-off fashion. As a result, it can be both statistically accurate and capable of robustly handling new data, as one simply adds samples from the new data to the training set. These properties, coupled with the thorough, detailed, and rich information representation afforded by the dense feature map paradigm, are the cornerstone of our approach to the development of a computer aided mammographic system.

Chapter 3

Microcalcification Detection

3.1 Problem Statement

Microcalcifications are the single most common of the breast cancer signs, occurring in around 40% of all cancers found mammographically. Unfortunately, they are also one of the least sensitive signs; their presence indicates a malignant cancer only about half of the time. The shape and distribution of the microcalcifications is important for accurate diagnosis; in particular, only clusters of microcalcifications are considered to be clinically significant.

The goal for a computer aided mammographic screening system then is to find clustered microcalcifications, which means finding the microcalcifications in the first place. Microcalcifications are small (generally less than 300 microns in diameter) dense objects of arbitrary shape; they can resemble scattered grains of rice or sand. Since they are dense, they can be high-contrast as well, but this is not guaranteed. They can appear anywhere in the breast, and as normal breast tissue is highly inhomogeneous and variable, it is quite easy for their appearance to be masked.

Still, since they have an easily described visual signature, and since "peak detection" has been a staple of signal processing since its inception as a discipline, microcalcifications have been one of the most thoroughly studied cancer signs, with many researchers and groups proposing varied image measurements designed to detect them.

3.2 Hypothesis

Accordingly, the intent of the current work was *not* to develop new and improved image features. It seems reasonable to assume that most of the best features have already been proposed, as is discussed in Section 3.3.3.

Therefore, the hypothesis investigated here is that it is improved *classification methods*, rather than feature construction, that will best advance the state of the art in microcalcification cluster detection. In particular, we hypothesize that we can improve detection performance by processing the most popular of the existent features with binary decision tree methods in the dense feature map context.

3.3 Methods

3.3.1 Test Data and Performance Criteria

To evaluate any algorithm for calcification detection, we must describe the test data, its counterpart "truth" information (the location and extent of calcifications and clusters in the data) and an objective criteria for the identification of a proposed detection as a true detection or a false alarm. These issues are examined in the following sections.

Test Data

The mammograms used[1] in this phase of the research were recorded with a Kodak MIN-R/SO177 screen/film combination using various types of equipment. The films were digitized at a size of 2048 x 2048 with an Eikonix 1412 12 bits CCD camera. A sampling aperture of .05 mm in diameter and a 0.1 mm sampling distance was used, giving a resolution of 100 microns. The images were corrected for inhomogeneity of the light source (Gordon plannar 1417). A fixed calibration of the CCD camera was used. The optical density of 0.18 corresponds to a maximum gray-level output of 4095.

The dataset consists of 40 images from 21 different patients. All images show one or more clusters of microcalcifications. The position and size of the microcalcification clusters (annotated using circles) were marked by two expert radiologists, based on all patient data available (different views, magnifications).

Groundtruth Data

Building a BDT requires training data in which the true class of the feature vectors is known. For the current application, this requires images in which each pixel has been identified as being or not being a calcification. Such "truth" data was not available for the data set at hand; though the extent of the clusters had been identified by a radiologist, the individual calcifications and their extent had not been marked.

Therefore, we created pixel-level truth data using the radiologist's cluster circles and existent computer aided calcification detection methods. Nico Karssemeijer, of the University Hospital Nijmegen, and Laura Mascio, of Lawrence Livermore National Laboratory, supplied the image output of their respective calcification detection algorithms, both operated at their most sensitive parameter settings. All objects whose center was outside of all truth circles in a given image were discarded as false alarms. A union of all remaining objects formed a single result image. These result images were used as the pixel level truth data for subsequent training.

Visual inspection of the resulting images indicates that this procedure did indeed manage to capture most all of the true calcifications, at the expense of marking as calcifications some non-calcifications within the truth circles. Since some non-calcification behaviors would thus be flagged as calcification-like, the expected result is to increase the false alarm rate of the final classifier.

[1] Images were provided by courtesy of the National Expert and Training Centre for Breast Cancer Screening and the Department of Radiology at the University of Nijmegen, the Netherlands.

Note that these pixel-level truth images are used only for training the statistical classifier, *not* for evaluating the final performance of the detection scheme. The performance criteria, discussed in the next section, depends only on the radiologist-selected truth circles, not the individual calcifications.

Performance Criteria

In order for the current performance results to be entirely commensurable with some competing prior results on this database, we implemented the same performance criteria used to generate those results, [5]. A given cluster is considered to have been detected if two or more detected objects are found in its associated truth circle. In this fashion a probability of cluster detection (P_D) was determined for each image, and the overall P_D reported is the average of the individual P_Ds across the 40 test images.

All detections not within any truth circles are considered to be false alarms. To determine the number of false alarm clusters, a 0.5 cm radius circle is inscribed around the center of each false alarm, merging nearby objects. A false cluster is counted for each resulting contiguous region. The overall false alarm rate (FAR) is the total number of false alarms detected, averaged across the 40 test images.

3.3.2 Initial Segmentor Set

To detect a microcalcification cluster, one must first detect, and segment, its microcalcifications, where "segment" here just means to determine the shape of the microcalcification. Since the features are highly sensitive to the shape, the choice of segmentor strongly affects the quality of the features that are extracted.

Rather than commit to one single segmentor, we use a host of segmentors, computing the same features on the output of each. One segmentation algorithm may produce objects which, using the shape feature for example, can accurately distinguish microcalcifications (MCs), while another segmentor may produce objects where the shape feature provides no predictive power. With the DFM approach, it is expected that the BDT will discover which features of which segmentors are useful for distinguishing MC objects from non-MC objects. So that true MCs would not be missed and a dense feature map would be generated, we set parameters on the segmentors so that many non-MC objects are found.

There are two types of segmentors used: all-at-once and at-every-pixel. All-at-once segmentors find objects throughout an image, returning a single image with objects indicated. Then for every point that is part of an object, features are computed for the object that contains the point. Since neighboring points are usually part of the same object, they will have equal feature values.

At-every-pixel segmentors find the best object centered at every pixel. No one single image with all the objects is returned since the objects for neighboring points would overlap. Rather, for each pixel a small sub-image is extracted around that pixel and the one best object within that sub-image which contains the center point of the sub-image is found. Features are computed for every point in that object. In this case, neighboring points may be within similar objects so feature values may be similar, but this need not be the case. The segmentation algorithm is given a slightly different sub-image so the resulting objects can be different.

In the at-every-pixel approach, the size of the sub-image (1100×1100 microns) is large enough to contain most calcifications, therefore, any objects that are not completely contained within the sub-image are discarded. Similarly, in the all-at-once approach, if the object is not completely contained with a sub-image centered at a point, no features are computed for that point. Note that other points within the same object may have features computed since the object may be contained with the sub-image centered on the other point.

Since a segmentation is desirable at every point, at-every-pixel methods are attractive. However, some segmentation methods are more naturally applied in an all-at-once fashion. Below is a brief description of the six segmentors used. The first three (MES, Peak and Adaptive Threshold) are at-every-pixel algorithms. The second three (Boundary, Woods and Outlier) are all-at-once approaches.

Minimal Error Segmentor (MES): This is a region growing algorithm where the "error" of merging regions is minimized [6]. First, 4-connected regions with equal gray-level are found in the sub-image. Then regions are incrementally merged with their neighbors in an "optimal" fashion. That is, at each point the merge is chosen which will introduce the least amount of squared pixel error into the image. This continues until an "optimal" stopping point has been reached, one which minimizes complexity while maximizing accuracy. That is, let N be the current number of regions in the image, $MS1$ be the maximum amount of error possible to the image (determined by merging until only one homogeneous region is left) and MSN be the mean squared error in the image if one merges until N regions are left. Then the termination function used is

$$f(N) = \frac{N}{\log(MS1) - \log(MSN)}.$$

Here, the complexity in the image is modeled by N, the number of regions in the image. log (MS1), the natural log of the maximum amount of error possible to the image, is an information theoretic measure of the amount of information in the image. log(MSN) is a measure of the error introduced, or the information lost. Thus their difference is a measure of the information still present, or the accuracy of the image. And so a global minimum in the ratio of complexity to accuracy, as defined above, is the stopping point we seek. Accordingly, the MES algorithm computes $f(N)$ for all N, and then selects the stopping point, the optimal N, to be the one at which $f(N)$ is minimized.

Peak Segmentor: In this approach the sub-image is thresholded at a percentage of the gray-level value of the central pixel, [7]. Then all objects that do not overlay the central pixel are discarded.

Adaptive Threshold: The histogram of the points in the sub-image is found. If an MC is present then there should be two peaks in the histogram, one from the MC and one from the background. This algorithm finds the valley between the peaks and uses this value as a threshold. The sub-image is then thresholded with this value.

Boundary: First an edge detector is applied to the entire image. Small gaps in detected lines are filled as follows. The edge image is dilated and eroded once with the dilation only being performed if 4 or fewer neighboring pixels are edge pixels. Then any pixel which was an edge point in the original edge image but is not in the dilated and eroded image, is converted back to an edge point.

This corrects lines which are broken by the eroding. Finally, an edge thinning algorithm is applied. Every enclosed region is then an object.

Woods (Threshold): This segmentor was developed by Woods [8]. For each point in the entire image, the difference between that pixel value and the mean of its surround is computed. Each pixel in the difference image is then thresholded with the threshold set to one of two values depending on the maximum value around the point in the difference image. If the maximum value is greater than 15 then a threshold of 10 is used. If the maximum value is less than 15 then a threshold of 5 is used. Woods then eliminates all but the highest contrast points, though here we retain all resulting points since a dense feature map is desired.

Outlier Threshold: This algorithm is a combination of the approaches taken in [9, 10]. For every point p in the entire image, a local window w around the point is considered. If the original gray-level of p is greater than a threshold defined as the mean of w plus three times the local standard deviation of w, then the point is considered to be "extreme". The extreme points are then grouped into segmented calcifications on a four-connected basis.

3.3.3 The Initial Feature Set

Calcifications are one of the most studied of breast cancer signs, to the point that papers surveying and listing the first stirrings of a consensus on the useful features were beginning to appear when we started our research[11, 8]. Rather than starting from scratch, we make use of the most popular features; with the exception of edge strength variation along an object boundary, all these features have been used by others. However, we do have to modify some of them, to convert them to a pixel orientation.

That is, a typical calcification feature (some measure of shape or edge strength, for example) is measured over a segmented region which has been indicated by some pre-processing step as a possible calcification. As mentioned in the previous section, here our approach is to measure each feature over the segmented shapes returned by *each* of the segmentors we are considering.

Below are detailed descriptions of each feature. Table 3.1 summarizes the features and lists other researchers who have used the same or a similar feature, and Section 3.3.4 will present the features that survived after they were weeded down to their most potent subset.

Area: The number of pixels in the object.

Shape: Shape is defined as $P^2 - 4\pi A$ where A is the area as defined above and P is the perimeter, defined as the number of pixel sides of the object that touch a background pixel. "$P^2 - 4\pi A \geq 0$" is known as the Isoparametric Inequality [12]. Equality holds if and only if the object is a circle.

Average Edge Strength: Average of $\sqrt{(x')^2 + (y')^2}$ for all points (x, y) on the perimeter, where x' and y' are the x and y gradients, respectively, for a point. The gradients are computed using 3×3 kernels.

Edge Strength Variation: Difference of maximum and minimum edge strength of perimeter points. Edge strength is defined as above.

Table 3.1: The features computed for each object (except for Laws) returned by a segmentor, with a partial list of researchers who have used the feature or one similar.

Feature	Definition	Used by
Area	Number of pixels	[15, 16, 8]
Shape	$P^2 - 4\pi A$	[15, 17, 18, 16]
Ave. Edge Strength	Mean edge strength of perimeter pixels	[17, 16, 8]
Edge Strength Var.	Difference of max and min perimeter pixel edge strengths	
Contrast	Difference of mean gray-levels of object and surround	[10, 19, 20, 15, 8, 21]
Object Std. Dev.	Standard Deviation of object pixels	[15, 8]
Background Std. Dev.	Standard Deviation of surround pixels	[10, 8]
Laws	Convolution at every pixel in the mammogram image	[22, 23]

Contrast: Difference of average gray-level within the object and average gray-level of a two-pixel wide surround of the object.

Object Standard Deviation: Standard deviation of the gray-level values within an object.

Background Standard Deviation: Standard deviation of the gray-level values within two pixels of the object.

Laws Texture: Unlike the other features, these are computed from the mammogram image at every pixel, rather than on an object returned from a segmentor, as they are intended to learn to recognize the range of normal appearances. The Laws features are based on the application of a small set of convolution kernels to the image (each kernel designed to respond to a different local behavior), followed by the measurement of the local sum of absolute values.

In the notation of Pietikainen et. al.[13], the convolution kernels used in this study are L5*E5, E5*S5, L5*S5, and R5*R5. The size of the window for the computation of the sum of absolute values texture energy measure was 15×15, as suggested by Laws[14].

3.3.4 Feature Set Reduction

Having established this set of 46 features (seven features measured across six segmentors, plus the four Laws features), the next step is to winnow these down to their useful subset.

Segmentor Variation

One would expect that the shapes generated by the various segmentations might be basically similar, and that as a result a given feature would be highly correlated across the segmentors. Interestingly,

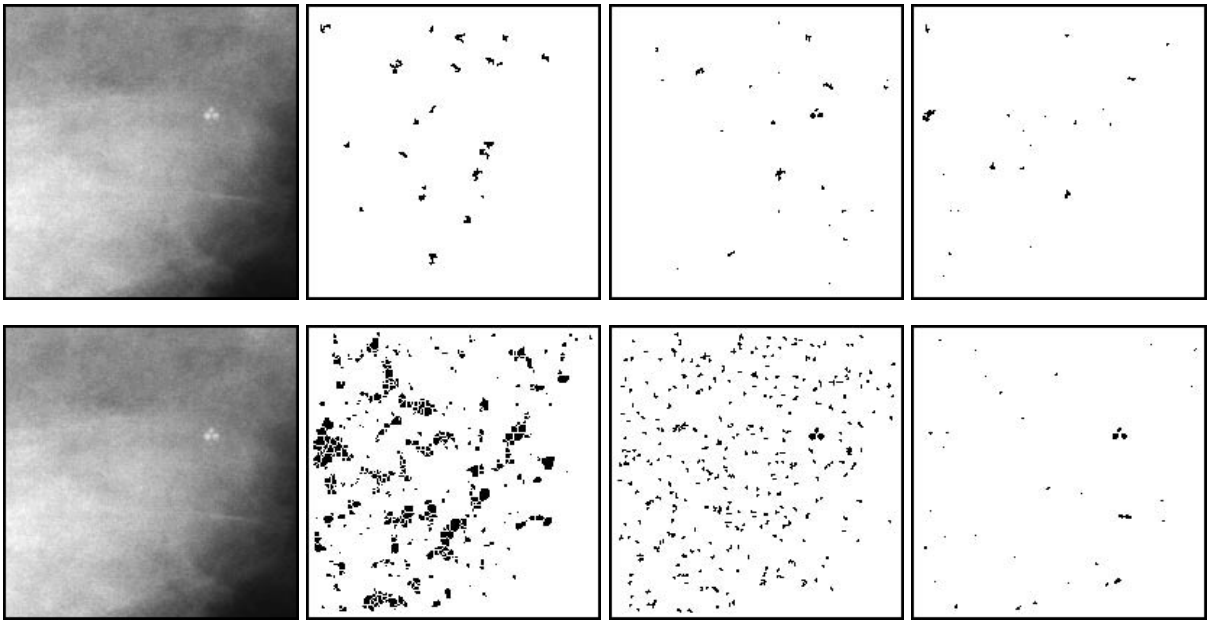


Figure 3.1: The left column shows a portion of a mammogram with three calcifications arranged in a triangle. It is shown twice to facilitate easier viewing along rows. The top row shows the results of the local segmentors (from left to right) MES, Peak, and Adaptive Threshold. The second row shows the results of the global segmentors (from left to right) Boundary, Woods and Outlier.

however, the segmentors varied fairly significantly in their response to the same data.

For instance, Figure 3.1 presents a patch of one of the mammograms with three calcifications arranged in a triangle. This figure illustrates the response of the six segmentors to that same image. Note that Adaptive Threshold and MES failed to segment the calcifications. Outlier and Woods both captured the shape of the calcifications but Woods had more non-calcification objects.

The Reduced Feature Set

The behavior of the features was investigated through growing trial BDTs and noting which features were most prominent, and through regression and correlation analysis. Attention was paid to examining the behavior of feature sets within the context of single segmentors, and segmentor sets within the context of single features. The general trends are as follows:

- The features derived from adaptive threshold segmentations are not useful, due to that algorithm's frequent failure to produce any segmentation.
- The Laws features are heavily utilized by all test BDTs, indicating that they are indeed fulfilling the desired background discrimination role.
- All of the Boundary-derived set of features are pertinent, with the exception of object standard deviation, which was not useful when computed from any of the segmentors.

- The contrast features are useful from almost all of the segmentors, even in combination with one another, and despite the fact that they were indeed fairly correlated.
- The area and shape features were notably useful only from the Woods segmentor, perhaps implying that it managed to recover shapes that best distinguished between calcifications and false alarms.

Taking these trends into account, the list of features used to produce the current performance results are given in Table 3.2.

Table 3.2: The Reduced Set of Selected Features

Boundary	Peak	MES	Outlier	Woods	Laws
Area	Contrast	Contrast	Contrast	Area	L5*E5
Shape	Ave. Edge Str.			Shape	E5*S5
Edge Str. Var.					L5*S5
Contrast					R5*R5
Background SD					

3.4 Results

Performance results were generated from this data using a half/half classification protocol [2, 22]. That is, in order to provide performance results for the entire data set, it was randomly separated into two sets, A and B, of 20 images each. Two separate BDTs were grown from randomly sampled data points from each image in each subset. Then the tree from set A was used to convert the full original images in set B into probability images, and vice versa. Finally, all of these probability images were mean filtered with a 700 by 700 micron window, to create a spatial consensus for the pixel classifications.

To determine the ROC statistics of the smoothed probability images, they were thresholded 50 times, at increments of 0.02, forming binary report images consisting of segmented regions. These were scored by the criteria discussed in Section 3.3.1, generating (P_D, FAR) pairs for each threshold level. The resulting curve is plotted in Figure 3.2, against a result on the same data from an alternative algorithm based on random field models [24, 5]. Note that where the DFM method approaches the P_D of the random field model, it does so at the cost of a higher FAR of about two clusters per image. We hypothesized that this was partially due to presence of non-calcifications in the calcification portion of the training set, as discussed in Section 3.3.1. Impure training data is not a factor in the random field model results, as there the control parameters are fixed during algorithm development, rather than dynamically optimized for the available data.

So we reconsidered the groundtruth data with an eye to wiping out the most obvious errors, in the hopes of creating more accurate "truth". To aid this process, we created mosaic versions of *each* cluster, mosaics which which presented various views of the same raw data, including adaptive

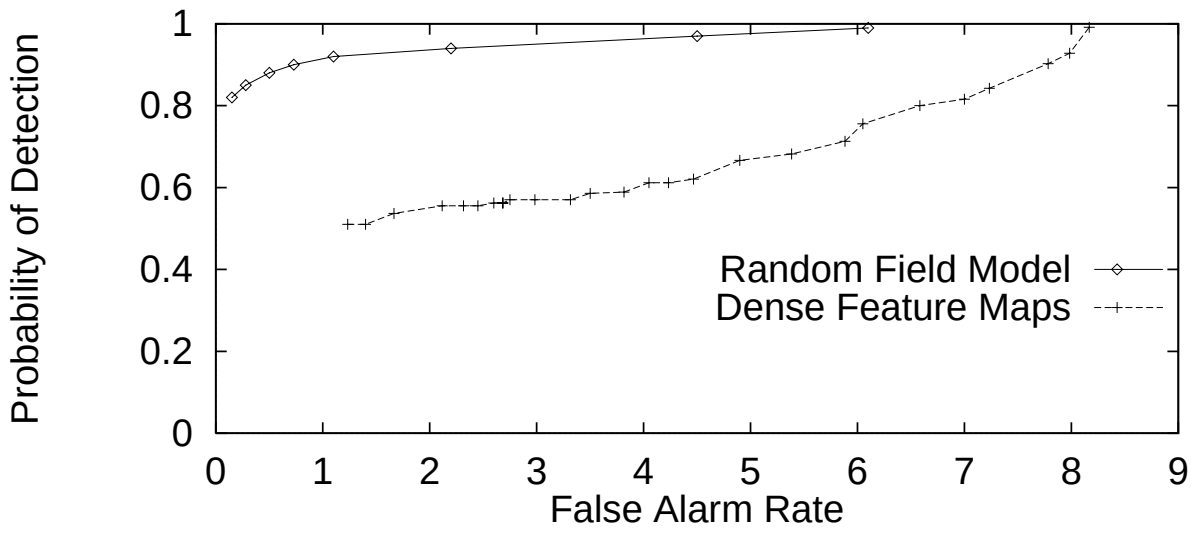


Figure 3.2: Initial ROC Curve Comparing Random Field Model and Dense Feature Map Performance

histogram stretch, adaptive histogram equalization, the original image, the initial segmentation, full image histogram equalization and full-image histogram stretching.

As we are ourselves not trained radiologists, attempting to remove the falsely identified calcification objects was a nerve-racking process, and after a couple of false starts we decided to created two weeded data sets, one in which we would remove only those objects which seemed to be obvious false alarms, and another, much smaller, which we would permit to contain only obvious calcifications. That is, the "borderline" objects were retained in the dataset we called "some" and removed in the case we call "few". The original dataset was called "all".

Retraining on both the old data and the two weeded datasets, results in the trio of ROC curves illustrated in Figure 3.3. In these figures we plot only from 0.5 to 1.0 on the probability of detection axis to better show the differences. Note that the "all" curve, using the initial groundtruth, has improved considerably; this is due only to the feature changes, particularly the revision of the Laws texture choices. Note also that "some" and "few", on the whole, dominate "all", suggesting that the system is more sensitive to the training data than we had first hypothesized.

Finally, we show the "few" curve from the previous figure plotted against the Random Field Model's result. It is still dominated through most all of its length, but it does move up to dominate the competing approach at the far right of the curve. This is important, as the right end of the ROC curve is the end that reflects operation in the sensitive regime that will be required in clinical use.

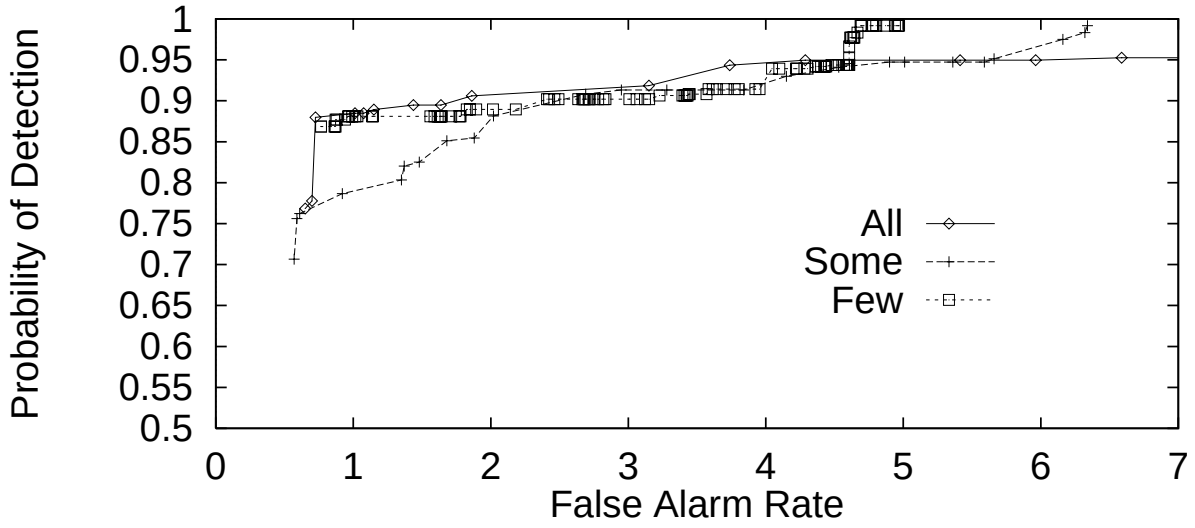


Figure 3.3: Results on the Few, Some, and All Datasets

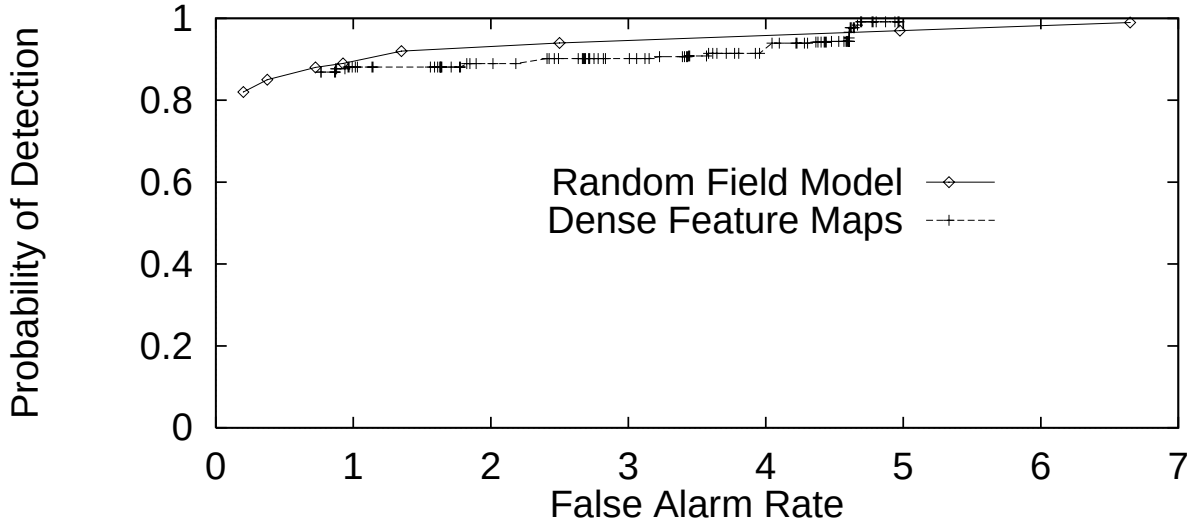


Figure 3.4: Results on the Few, Some, and All Datasets

3.5 Conclusions and Future Work

In this chapter, we have reported on results from the application of the DFM method to the detection of calcifications in mammograms. Our evaluation method has been to share the data and performance criteria of other work on this problem, so as to be able to rigorously compare results. This has served not only to objectively assess the relative weaknesses of the current implementation of the DFM, but also to focus our future attention on the specific problem of false alarm rate.

The raw performance of our approach was shown to be comparable to the best competitor on the same dataset, and further, to be better in the most sensitive regime. This is important as it is in this regime that any practical, clinical CAMS system will operate.

Future efforts towards improving the performance of the DFM method for microcalcifications would include

- Development of a segmentor and features customized to address lacks in the initial selection set turned up by our analysis.
- Objective re-examination of the entire feature set, to determine the truly optimal subset of pertinent features.
- Use of training data in which a radiologist has certified the location and extent of each calcification.

Chapter 4

Mass Detection

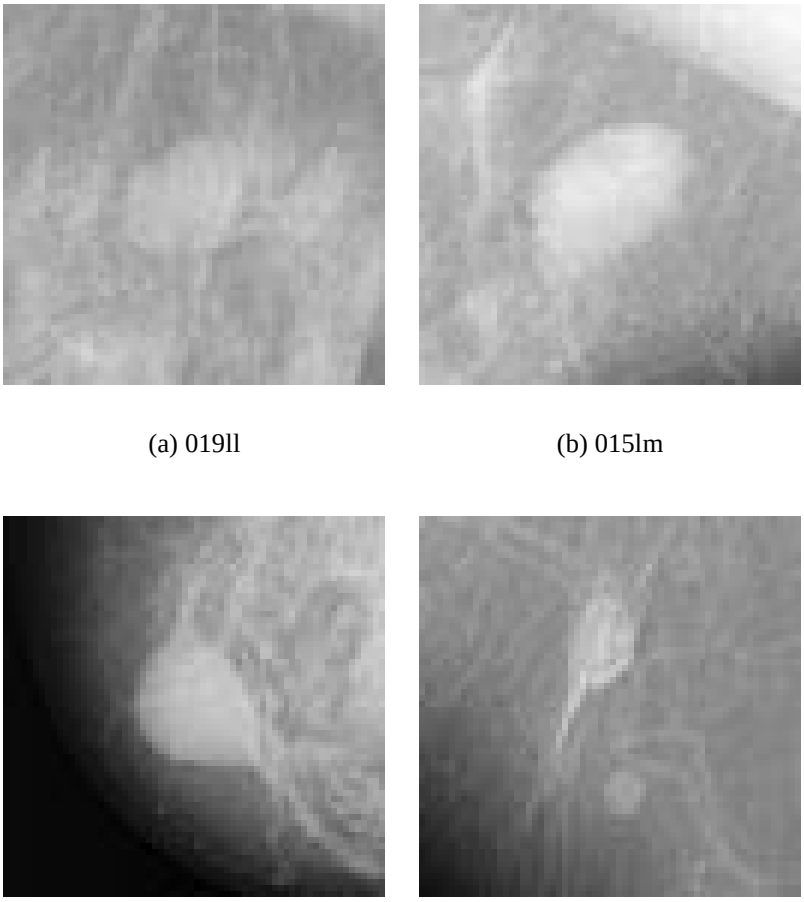
Masses appear in mammograms as compact shapes ranging from millimeters to centimeters in size, with densities lower to greater than that of surrounding tissue. The challenge is to detect these masses consistently in the face of the broad variation of normal tissue types.

4.1 Problem Statement

We address a subset of the image data and possible mass characteristics present in typical 4-view mammogram cases, choosing to analyze single views and modeling masses as approximately circular and denser than surrounding tissue. Our objective is to generate a feature that responds strongly to masses and minimally to normal tissue which is consistent across breast tissue types and is amenable as input to the dense feature map (DFM - see Section 1.2) approach. The context in which the masses are to be detected is one containing a great deal of interfering structure due to the complex appearance of normal parenchyma and wide ranges in brightness due to the presence of fatty and/or glandular tissue. Figure 4.1 illustrates this complexity in shape and background for relatively obvious masses. Accordingly, the situation calls for a detection approach which performs well over wide ranges of scale and contrast, despite partial or distorted shapes and interfering structure.

4.2 Hypothesis

Our detection hypothesis is that a significant subset of masses are reasonably well characterized by a roughly circular disk that is denser than surrounding tissue. This disk is presumed to have a well-defined edge around a substantial portion of the circumference, but is embedded in normal tissue varying widely in density and structure. These considerations motivate the development of a method based on the circular Hough transform.



(a) 019ll (b) 015lm

(c) 021ll (d) 132rx

Figure 4.1: Subimages of selected MIAS masses

4.3 Methods

Our method for detecting circumscribed masses in mammograms is built around the circular Hough transform [25]. The input is a mammogram as described in Sections 4.4.1 and 4.4.2, and the end result is a feature image that correlates with the likelihood of a circumscribed mass at each pixel location in the input mammogram. The intermediate steps of the algorithm we use are outlined in the flowchart in Figure 4.2. We segment the input mammogram, preprocess the image to normalize

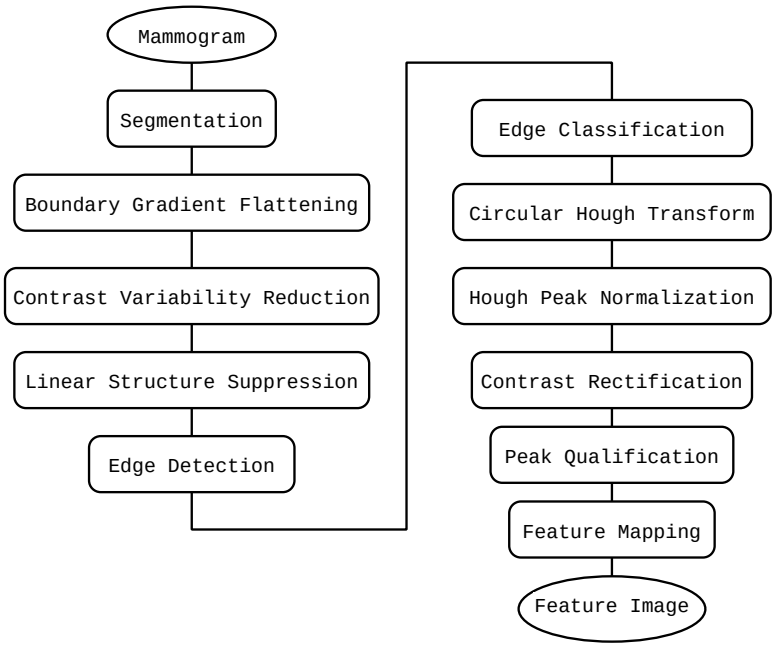


Figure 4.2: Circumscribed mass detection algorithm flowchart

the contrast and suppress linear structures, detect intensity edges, sort and threshold these edges by length and strength, perform the circular Hough transform, filter the Hough accumulator to normalize it with respect to radius, qualify peaks in the accumulator using the contrast of those peaks, and finally map the qualified peaks into a feature image. We describe these algorithm steps in more detail in Sections 4.3.1–4.3.11

4.3.1 Subsampling and Segmentation

At 50-100 microns, the original data is at a resolution finer than practical for initial algorithm investigation. Accordingly, it was subsampled using Gaussian/sinc interpolation [26] to create data with a pixel resolution of 800 microns. At this spatial resolution, the circumscribed masses in the MIAS database have a radius that varies from 4.5 to 49 pixels.

After subsampling, the images were edited by hand to remove artifacts from the background (non-tissue) regions. They were segmented by a process that thresholds the image at a level dynamically chosen to be just finer than the background histogram peak, and then finds and retains

the largest resulting 4-connected object. The rest of the objects are merged into the background, and the background value is set everywhere to zero. The result is an image in which the breast region is the only region that contains non-zero values.

4.3.2 Boundary brightness Gradient Flattening

The brightness roll-off at the breast boundary varies considerably from image to image, due to both the characteristics of the individual breast and the compression for imaging. This roll-off creates a variable background image gradient which masks subtle mass boundaries in the region and introduces spurious edges.

Figure 4.3 illustrates the comparative differences in the brightness histogram as a function of distance from the boundary for two mammograms. We reduce the boundary roll-off by forming a two dimensional cumulative distribution

$$c(d,v) = c(d,v-1) + h(d,v) \tag{4.1}$$

where

$$h(d,v) = \sum_{\forall i \in I(d,i)=v} 1 \tag{4.2}$$

is the histogram of the image $I(d,i)$ containing the number of image points with value v at distance d from the segmentation boundary. $I(d,i)$ is simply the mammogram mapped onto the coordinate system d,i. The cdf is then that of the image brightness along the segmented boundary of the breast tissue as a function of the distance from the boundary, where I is the image and d is the distance from the boundary and i spans the pixels at each d. The cdf is then normalized

$$c_n(d,v) = \frac{c(d,v)}{c(d,vmax)} \tag{4.3}$$

This normalized cdf is then thresholded

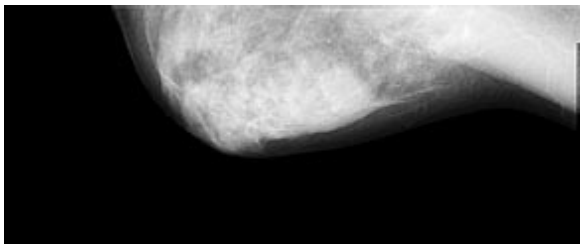
$$t(d) = v_{c_n(d,v)>T} \tag{4.4}$$

where T is typically set to 0.3, so that $t(d)$ captures the general curve of the brightness roll-off.

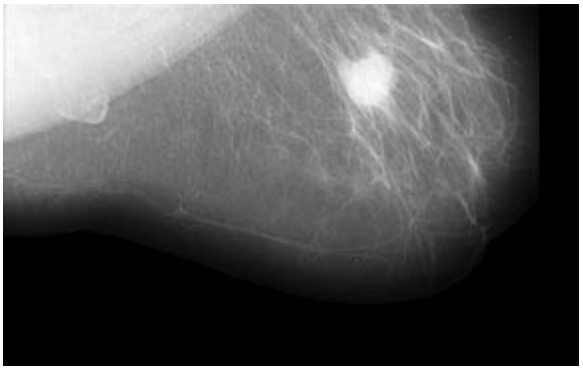
The image values are then mapped through

$$O(d,j) = I(d,j) - t(d)\exp\left(\frac{I(d,j)-B}{S}\right)^5 \tag{4.5}$$

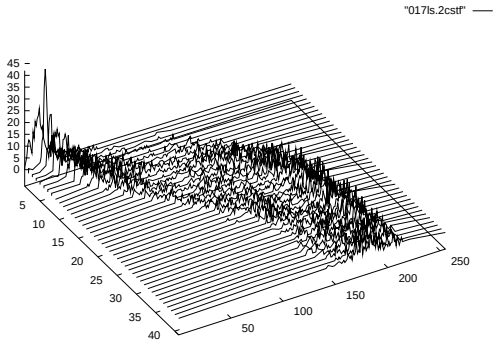
where B is the constant image value of exposed film and S is the value for unexposed film. This maps values near exposed film nearly linearly, while those near the densest tissue are changed little. This mapping flattens the border regions while avoiding over-compensation in areas of dense tissue.



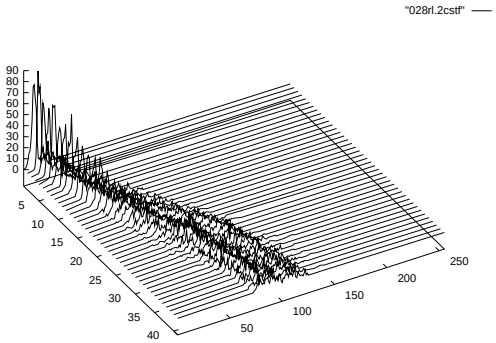
(a) 017ls image



(b) 028rl image

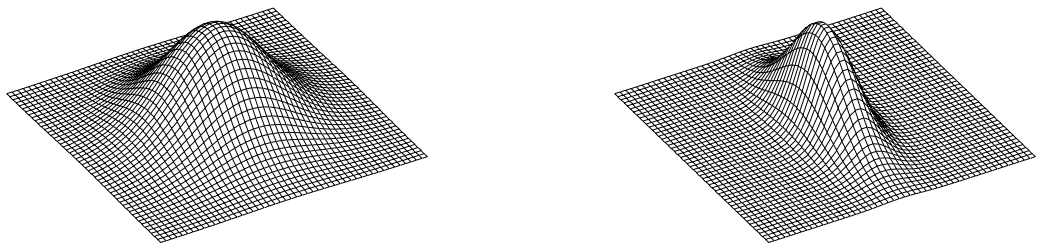


(c) 017ls histogram as a function of distance from boundary



(d) 028rl histogram as a function of distance from boundary

Figure 4.3: Image and boundary histogram examples



(a) Isotropic function (b) Ridge function

Figure 4.4: Linear feature suppression function shapes

4.3.3 Contrast Variability Reduction

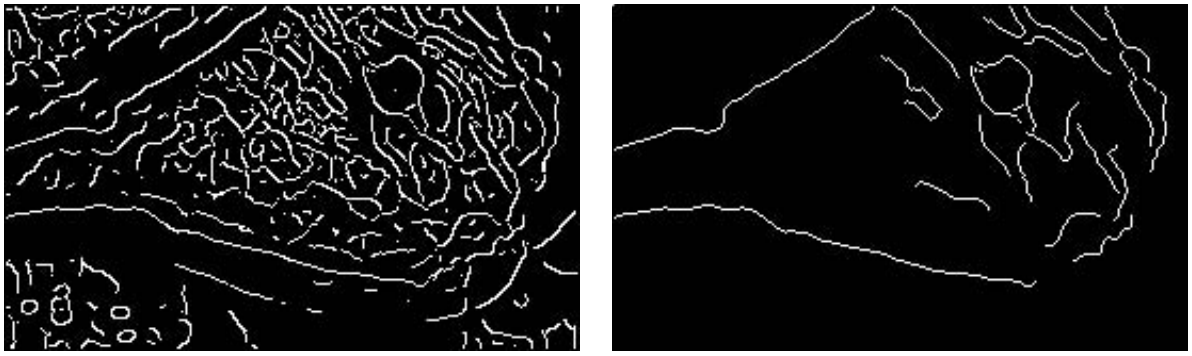
Both the tissue composition of the breast and the compression for imaging contribute to contrast variation across the image, making it difficult to accurately measure the contrast for a candidate mass. We lessen this variation by fitting two bicubic surfaces to the image, one to the breast boundary region and one to the central breast region. These two surfaces reasonably model gross variations over the image while responding minimally to image features on the scale of masses. The two bicubics are then stitched together by choosing points from each along their mutual boundary that minimize the mean squared error with respect to the fit over both surfaces. The mutual boundary is then iteratively adjusted, with the fitting and stitching steps repeated, further minimizing the mutual fit error. Finally, this composite surface is subtracted from the mammogram.

4.3.4 Suppression of linear structures

The linear structures arising from the normal parenchyma and blood vessels in the breast create a large number of edges that tend to mask the often weaker edges created by masses. To improve the detection of mass edges, we characterize each point in the image as best fit either by a smooth isotropic or oriented ridge model similar to those shown in Figure 4.4.

Points fitting an oriented ridge are further qualified by checking for an extrema in the second derivative larger than the first derivative, both at the same orientation and location as the ridge model. Points passing this test are deemed representative of a linear structure in the image and the ridge response, scaled by the second derivative, is subtracted from the image.

This process is repeated over two scales and eight orientations at present, and is similar in concept to that used in the initial detection stages of [27], though the objective is suppression,



(a) Original edge points (b) Edge points above length-strength threshold

Figure 4.5: Edge point qualification

rather than characterization of linear structure.

4.3.5 Edge detection

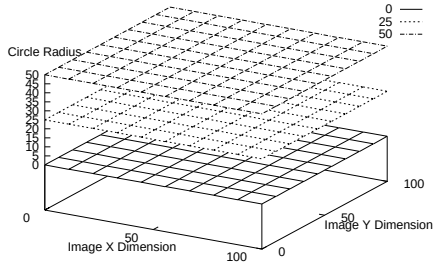
Our objective for edge detection is a sensitive edge indicator that also provides accurate normal directions, as these are required for our implementation of the circular Hough transform. We use a simple algorithm based on that of Canny [28], filtering the input image with a 2-d symmetrical Gaussian ($\sigma_x = \sigma_y = 1.0$), then estimating the image gradient using first differences taken in the image X and Y directions, computing the magnitude and phase angle of the gradient, and finally suppressing all points but those which are a local maxima in the direction of the gradient.

4.3.6 Edge Sorting and Thresholding

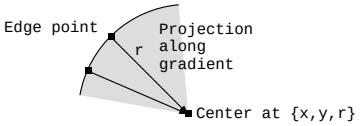
These edges are then sorted and thresholded by length and strength based on the assumption that long, strong edges are more significant that short, weak ones [29], rejecting 50-90% of the original edge points, as can be seen in Figure 4.5.

4.3.7 Circular Hough transform

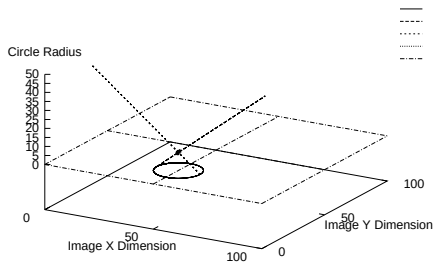
We chose the Hough transform [30, 25, 31, 32] as a detection algorithm for its robust performance over a wide range of scale in the face of noise and background clutter. We parameterize the Hough accumulator for *circles* for two reasons: first, circles are a reasonable first order approximation for circumscribed masses, and second, we wished to limit the dimensionality of the search for peaks in the accumulator. Figure 4.6a schematically illustrates the circular Hough transform accumulator, which is quantized at 1 unit in x, y, and r, where $\{x, y\}$ indicate the origin of a circle and r



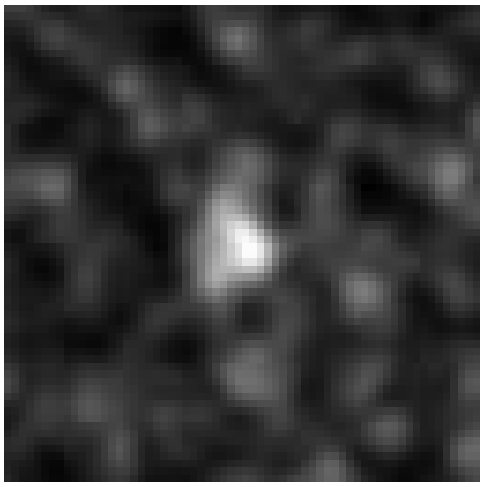
(a) Schematic of Hough accumulator



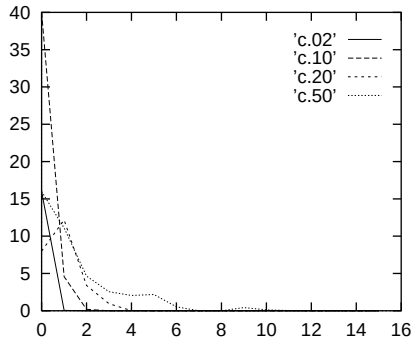
(b) Edge Gradient Projection (XY Plane View)



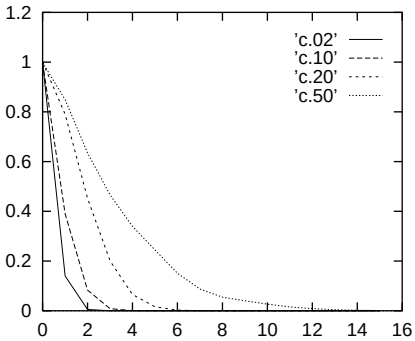
(c) Gradient Projection for circle $\{50, 20, 10\}$



(d) Smoothed Hough accumulator mass subimage at $r \approx r$ mass



(e) Raw accumulator radial profile



(f) Normalized accumulator radial profile

Figure 4.6: Circular Hough Transform Accumulator

indicates its radius. Edge points are projected in the direction normal to the rising edge as shown in Figure 4.6b. Figure 4.6c illustrates the projection of a gradient local maxima at $\{40, 20\}$ with direction $1, 0$ through the Hough accumulator $\{x, y, r\}$, incrementing each point along the line $\{40 + r, 20, r\}$. Similarly, a maxima at $\{60, 20\}$ with direction $\{-1, 0\}$ increments each point in the accumulator along the line $\{60 - r, 20, r\}$. Edge points arranged in a circle will thus result in a cluster of values about the point $50, 20, 10$ in the Hough accumulator where the projected lines converge. The result is a volume of points with higher densities indicating the possibility of circular structure in the original image, as seen in Figure 4.6d.

4.3.8 Hough peak normalization

We desire an indication of circular structure in the input mammogram which is *independent* of radius, as we will be mapping peak heights in the 3-d Hough accumulator down to the 2-d pixel level feature image (Section 4.3.11) to indicate the significance of a detected mass, irrespective of the peak's radius. Unfortunately, the height and density of peaks in the raw Hough accumulator are affected by the number of edge points contributing to a peak and by the accuracy of the edge location and gradient normal. The total number of edge points contributing to an accumulator peak for a circle is $\propto 2\pi r$, but the spatial distribution of the edge points projected along the gradient normal is affected in a complex, deterministic way by the pixel sampling of the circle border, which is both a function of circle radius and location. This results in peak heights and densities that vary considerably with respect to radius, even for images of perfectly circular structures.

As an example, Figure 4.6e is a plot of peak height versus radial distance from the peak center for perfectly circular (digital) disks of radius 2, 10, 20, and 50. We compensate for this behavior by normalizing the Hough accumulator with respect to r using an empirically designed exponential filter matched to the peak shapes created by circular disks. The filtering operation is the convolution

$$g(x, y, r) = f(x, y, r) * h(x, y, r) \tag{4.6}$$

of the raw Hough accumulator $f(x, y, r)$ with the exponential normalizing filter

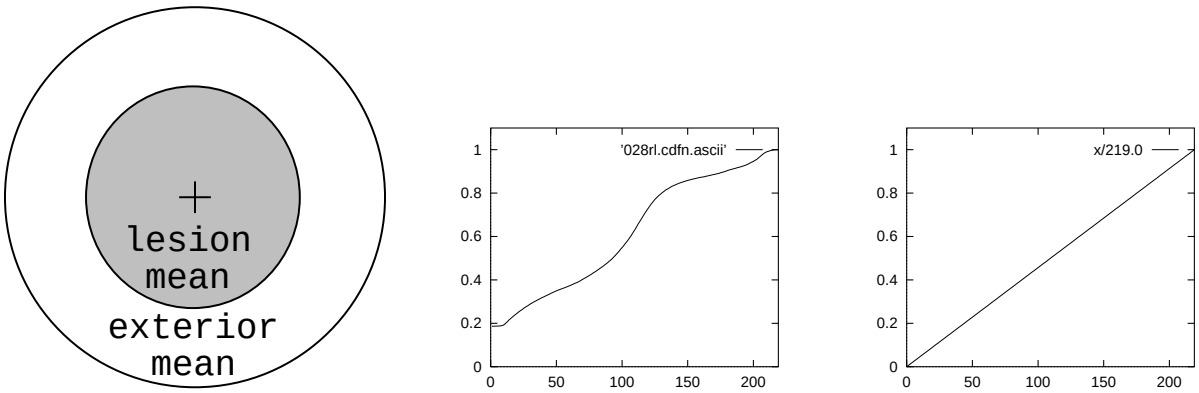
$$h(x, y, r) = c(r) \exp\left(-\frac{\left(x^2 + y^2\right)^{\frac{a(r)}{2}}}{b(r)}\right) \tag{4.7}$$

resulting in the normalized Hough accumulator $g(x, y, r)$. The terms $a(r), b(r)$, and $c(r)$ are weights as a function of the circle radius r, and were empirically determined for each r by summing the Hough accumulator response for a large number of circular disks displaced uniformly over the interval $x_0 \pm 0.5, y_0 \pm 0.5$. Figures 4.6e–f illustrate the effect of normalizing the Hough accumulator for circles of varying radius.

4.3.9 Hough peak qualification

Peaks in the Hough transform accumulator described above indicate circular structures in the mammogram. Unfortunately, many Hough peaks of similar height are created, due to both interfering

structure and to non-circular masses. The Hough transform measures the outline of a structure, as indicated by edge points on its periphery, disregarding the interior. Our approach to improving this performance is to measure the image contrast between the interior and exterior of the circle described by a peak in Hough space by computing the difference between the mean value of the interior of the circular region and the mean of an annulus exterior to the circle as shown in Figure 4.7a. We then use this contrast to scale the Hough peak in a multiplicative fashion, resulting in a measure of the region contrast, as well as the support for a circle.



(a) Mass contrast measurement region

(b) Cumulative distribution mapping

(c) Linear mapping

Figure 4.7: Mass contrast schematic and image mapping functions

4.3.10 Contrast rectification

The significance of any degree of contrast in a given image depends on the range of contrasts possible in that image. A high contrast circular region is much more significant in a fatty mammogram than in a dense one. Accordingly, we address image to image variability in contrast by extracting, for each image, a histogram of the contrast levels seen in that image. Then the specific contrast noted for the disk suggested by a Hough peak is mapped through either a linear or cumulative distribution function, as illustrated in Figure 4.7b and c. As a result, the raw contrast values are converted to a value in the range 0.0–1.0 which reflects the significance of that contrast in that particular image.

This whole process of Hough peak qualification results in a volume of points $\{x, y, r\}$ describing the relative likelihood of a dense circular structure with radius r at a point $\{x, y\}$. To see the effects of contrast rectification on feature separation, consider Figure 4.9. Figures 4.9a and b are plots of the “Peak disk” and “Sum disk” features (see Section 4.3.11) without contrast rectification for all 44 mammograms in the data set. Figures 4.9c and d illustrate the linear map contrast rectification over the set and Figures 4.9e and f illustrate the cumulative distribution map contrast rectification. Note that the rectification does tend to improve spread and reduce variability.

4.3.11 Feature mapping

The last step in the feature extraction algorithm is to create a feature image indicating the relative likelihood of a circumscribed mass at a pixel level. We do this by mapping points in the qualified, rectified Hough image in x, y, r space onto disks of radius r on the image plane at x, y. We experimented with two different methods of collapsing the vector of r possible Hough space values onto the feature image plane:

Peak disk: Setting each feature image pixel to the maximum likelihood encountered over r at the point x, y.

Sum disk: Setting each feature image pixel to the sum of the likelihoods encountered over r at the point x, y.

Figures 4.8 and 4.11 illustrates the "Peak disk" and "Sum disk" features for selected mammograms, as well as the original mammograms.

4.4 Experimental Results

We analyzed two data sets during the course of our investigation. The first was the 22 image subset of MIAS database containing circumscribed masses, and the second was a 129 image subset of the BGSM database. The data set, truth data, performance criteria, and experimental results are described for each data set in turn below.

4.4.1 Results for the MIAS Database

MIAS Test Data Description

The data analyzed in this investigation is a subset of version 1.2 of the digital mammographic database created and distributed by the Mammographic Image Analysis Society (MIAS). (See the WWW site `http://s10d.smb.man.ac.uk//MIAScom.html` for more information.) These are single-view mammograms scanned with the Joyce-Loebl microdensitometer SCANDIG-3, which has a linear response in the optical density range 0–3.2. The resulting pixels have a spatial resolution of 50 microns and a grey-scale resolution of 8 bits.

The MIAS database characterizes each image according to the class of abnormalities present. We shall present results on the 22 ground-truthed images in the database that contain circumscribed masses. To match these images, 22 entirely normal images were selected at random from the set of 204 normal images in the MIAS database, resulting in a data subset containing 44 images.

MIAS Groundtruth data

Version 1.2 of the MIAS database provides groundtruth for its masses in the form of circles which indicate the approximate center and radius of each abnormality. Since circumscribed masses are rarely perfectly circular, and since the MIAS policy was to err on the side of making the groundtruth

circles completely inclusive rather than too small, the groundtruth regions often contain a substantial amount of normal tissue. Therefore, for the subset of masses in which the border was both clearly visible and substantially different from the original groundtruth circle, the truth regions were edited to more closely reflect the actual mass shape[1].

These edited truth regions are used when training classifiers and when applying the performance metrics discussed in the next section. However, the original truth regions are and will be used whenever end-to-end detection rates are computed, in order to permit accurate comparison with results from other institutions making use of the same data.

MIAS Performance Criteria

We use an objective 50% overlap criteria to determine the performance of any given report image. That is, a given true mass is considered to have been detected if there is a reported mass such that the area of their overlap is at least 50% of the larger of their individual areas. If TL and FL are the number of correct detections and false alarms, respectively, across all the test data, NL the number of true masses, and NI the number of test images, then we compute (P_D, FAR) as $(TL/NL, FL/NI)$. To determine the ROC statistics of the smoothed probability images, they were thresholded 50 times, at increments of 0.02, forming binary report images consisting of segmented regions.[2] These were scored by the criteria discussed in the previous section, generating (P_D, FAR) pairs for each threshold level. The resulting curve is shown in Figure 4.10.

MIAS Example Feature Images

The images in Figure 4.8 illustrate the peak and summed disk features generated for the MIAS database.

MIAS Feature Plots

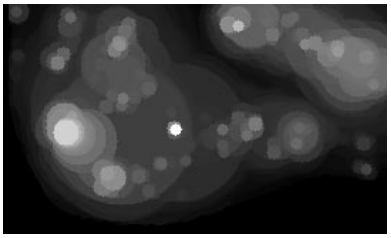
Figures 4.9a and b are plots of the "Peak disk" and "Sum disk" features without contrast rectification (Section 4.3.10) for all 44 mammograms in the data set. Figures 4.9c and d illustrate the linear map contrast rectification over the set and Figures 4.9e and f illustrate the cumulative distribution map contrast rectification.

MIAS ROC Curve Results

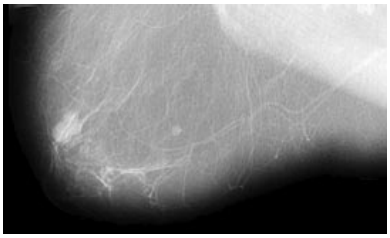
Figure 4.10 illustrates the ROC curve results for the MIAS data set. These results are competitive with published results on other circumscribed mass datasets, particularly in the relatively low FAR.

[1]The edited truth data for the 22 mass images is available via anonymous ftp from `ca.sandia.gov` as `/pub/wpk/stuff/tiwdm/truth.tar.gz`.

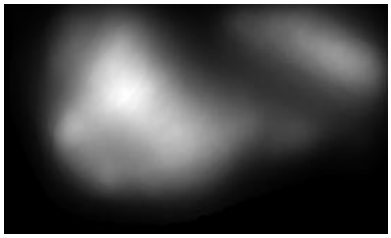
[2]The ROC curve appears to be formed from fewer than 50 points because not all thresholds generate unique report images, and all equivalent report images map into the same performance point.



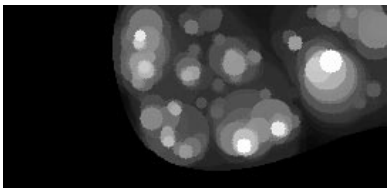
(a) Peak disk feature for “005ll” image



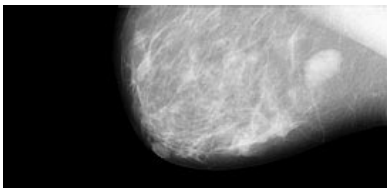
(b) Original “005ll” mammogram



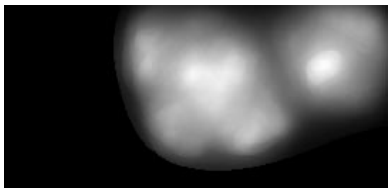
(c) Summed disk feature for “005ll” image



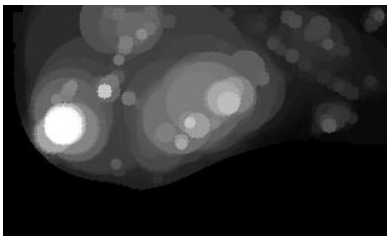
(d) Peak disk feature for “015lm” image



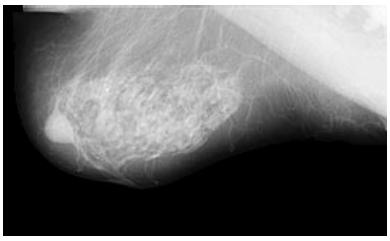
(e) Original “015lm” mammogram



(f) Summed disk feature for “015lm” image

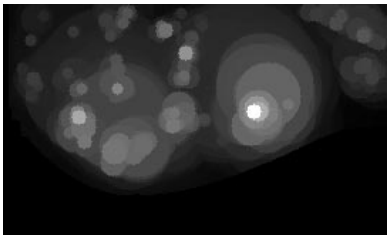


(g) Peak disk feature for “021ll” image

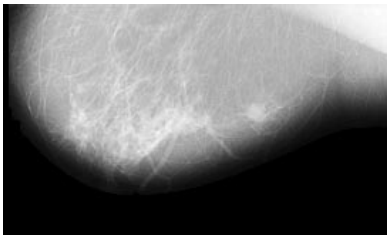


(h) Original “021ll” mammogram

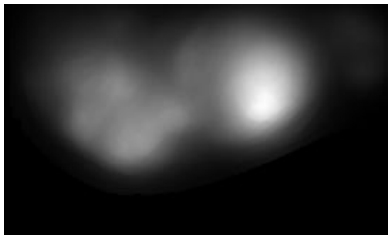
(i) Summed disk feature for “021ll” image



(j) Peak disk feature for “023ll” image

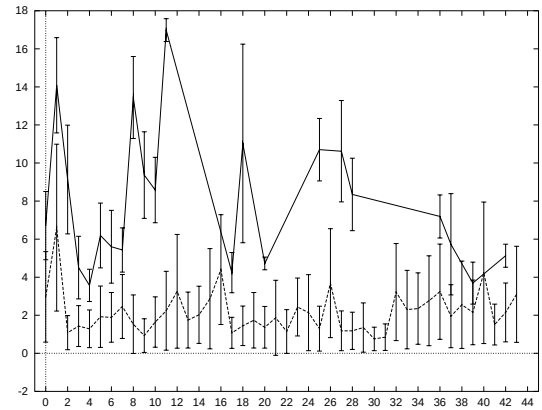


(k) Original “023ll” mammogram

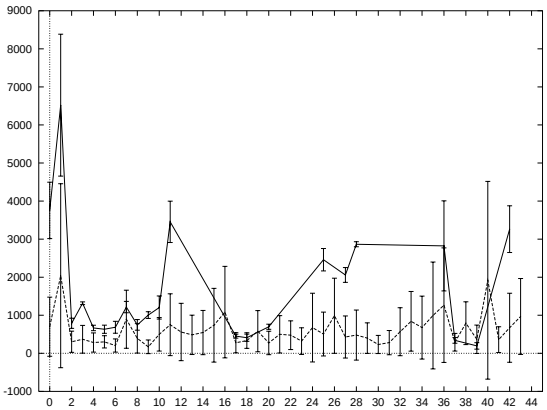


(l) Summed disk feature for “023ll” image

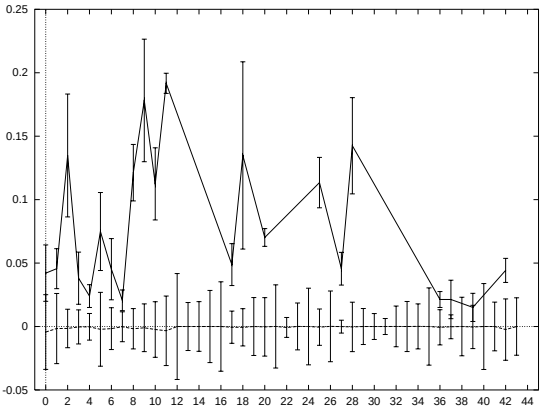
Figure 4.8: Peak, original, and summed disk feature images for selected mammograms from the MIAS database.



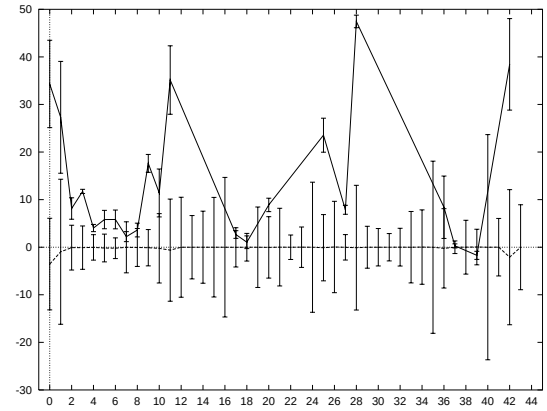
(a) Original peak disk feature plot



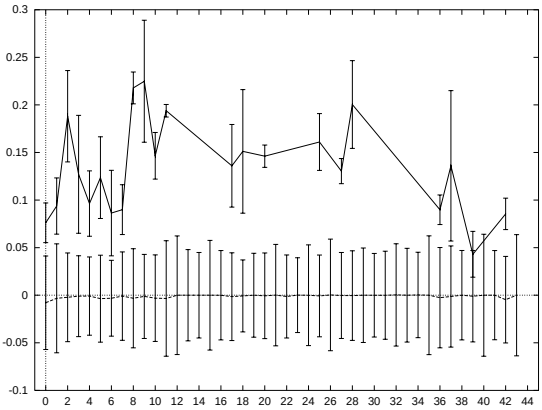
(b) Original summed disk feature plot



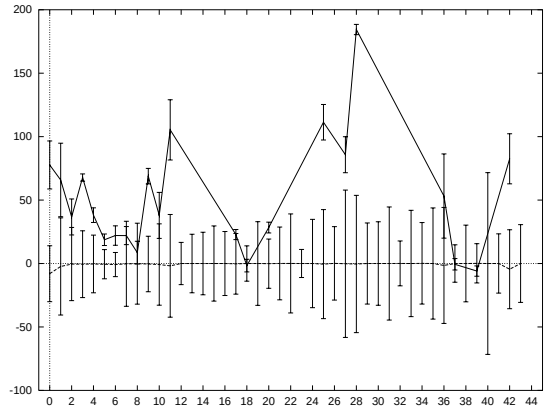
(c) Linear map minus mean peak disk feature plot



(d) Linear map minus mean summed disk feature plot



(e) Cdf map minus mean peak disk feature plot



(f) Cdf map minus mean summed disk feature plot

Figure 4.9: Plots of features across the 44 mammogram MIAS data set.

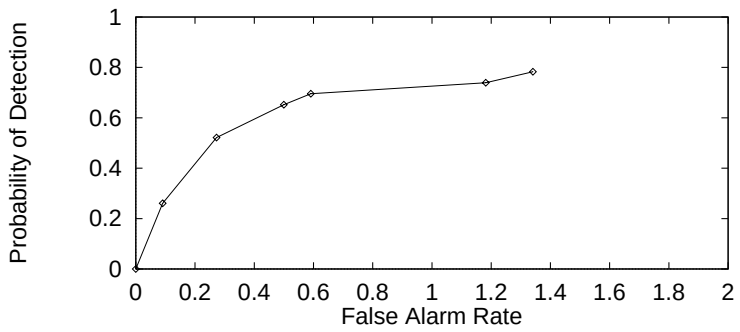


Figure 4.10: Performance Results as a ROC Curve for the MIAS data set.

4.4.2 Results for the BGSM Database

BGSM Test Data Description

The data analyzed in this investigation was drawn from a 117 case data set acquired at the Bowman-Gray School of Medicine. Approximately half of the cases were selected for the presence of a mass, and half selected from those read as normal. The mass cases were confirmed by pathology and the normals by normal followup films for three years. The cases were digitized at 100 microns on a Lumysis 200 film digitizer with a 12 bit resolution over an optical density of 0–3. The 117 4-view cases resulted in 129 images containing masses. Eight films had two masses each, for a total of 137 masses in the data set.

Our results are based on the 129 truthed images that contain masses and the 336 normal images in the data set.

BGSM Groundtruth data

Ground truth for the BGSM data set was obtained by digitizing the mammogram, having a radiologist trace the mass on the film, redigitizing the traced mammogram, registering the two images, tracing the radiologist's outline on the computer, filling in this region, and writing a binary truth image containing 1 inside the mass and 0 elsewhere. These truth regions are used when training classifiers and when applying the performance metrics discussed below

BGSM Performance Criteria

We use an objective 50% overlap criteria to determine the performance of any given report image. That is, a given true mass is considered to have been detected if there is a reported mass such that the area of their overlap is at least 50% of the larger of their individual areas. If TL and FL are the number of correct detections and false alarms, respectively, across all the test data, NL the number of true masses, and NI the number of test images, then we compute (P_D, FAR) as $(TL/NL, FL/NI)$.

To determine the ROC statistics of the smoothed probability images, they were thresholded 50 times, at increments of 0.02, forming binary report images consisting of segmented regions. These were scored by the criteria discussed in the previous section, generating (P_D, FAR) pairs for each threshold level. The resulting curve is shown in Figure 4.13.

BGSM Example Feature Images

The images in Figure 4.11 illustrate the peak and summed disk features generated for the BGSM database.

BGSM Feature Plot

Figures 4.12a–4.12c are plots of the cumulative distribution map contrast rectification, with Figure 4.12a illustrating the features across the images containing malignant masses, Figure 4.12b illustrating part of the normal images containing no masses, and Figure 4.12c illustrating the features across the images containing masses found to be benign.

BGSM Performance Analysis

Tables 4.2– 4.5 indicates the performance we attained for a subset of the algorithm steps on the BGSM database. The table columns represent the major feature algorithm steps from "Visible Without Truth" on the left, which represents the mass in the image, to the "CDF Peak Disk" to the right, which represents the final output for the CDF peak disk feature. The contents of each of these columns are scaled 0 to 4, indicating the visibility of the mass in "VWOT", the number of edge quadrants coincident with the truth for "EdgeQ", and the quarters of area overlap for "HP" and "CDFPD". The rightmost three columns are the "Dense Feature Map Smoothed" and "Dense Feature Map Thresholded" probability images, followed by the "Dense Feature Map Result" indicating the final detection result. The contents of each of these columns are scaled 0 to 4, indicating the quarters of area overlap for "DFMsmth" and "DFMthr", and detection for "DFMres". The table is sorted in order of increasing number of steps feature visibility and detection. There were 60 correct detections and 77 missed detections out of 137 masses. Of the missed detections, 49 coincided with low "CDFPD" feature values. An additional 22 had high "CDFPD" values, but failed to generate a high probability in "DFMsmth", resulting in failed detections. The remaining 6 failed detections had high (≥ 3) feature and probability values which failed to generate detections in the "DFMres" column.

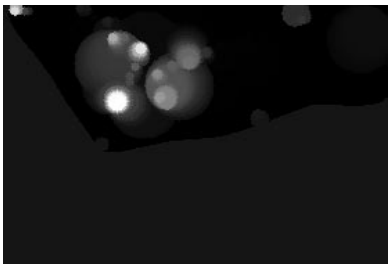
Table 4.1 contains the key to the column titles in Tables 4.2–4.5.

This data was drawn from:

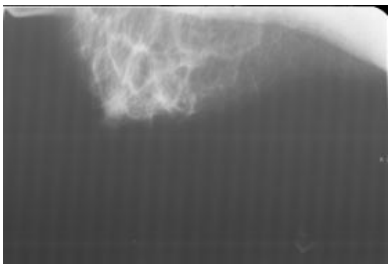
```
gorp:/export/disk3/brg/m/All/Analysis/T.sort.dfmsmth.pd.hp.ed.
```

The full data set may be found in:

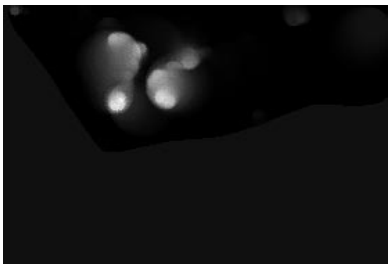
```
gorp:/export/disk3/brg/m/All/Analysis/T.mega.nums
```



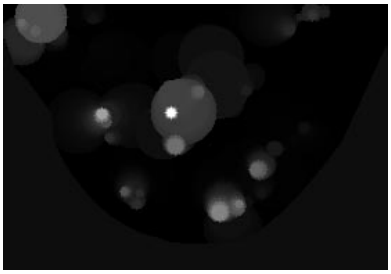
(a) Peak disk feature for "090.m.l.mlo" image



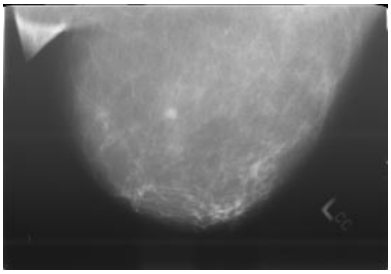
(b) Original "090.m.l.mlo" mammogram



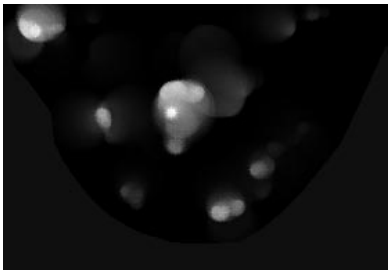
(c) Summed disk feature for "090.m.l.mlo" image



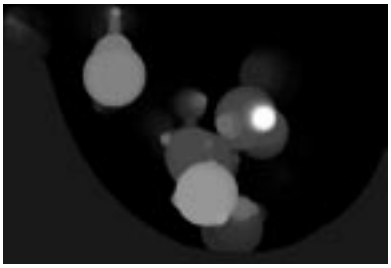
(d) Peak disk feature for "005.m.l.cc" image



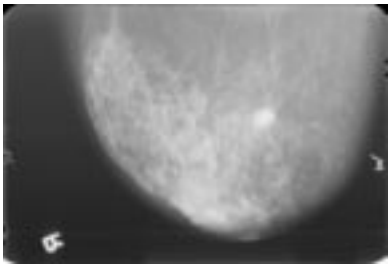
(e) Original "005.m.l.cc" mammogram



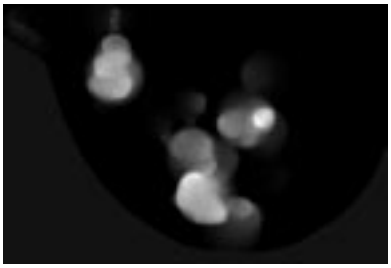
(f) Summed disk feature for "005.m.l.cc" image



(g) Peak disk feature for "024.m.r.cc" image

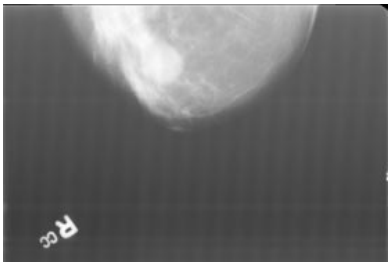


(h) Original "024.m.r.cc" mammogram

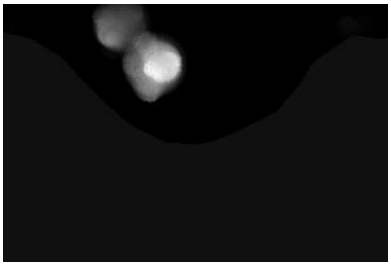


(i) Summed disk feature for "024.m.r.cc" image

(j) Peak disk feature for "080.m.r.cc" image

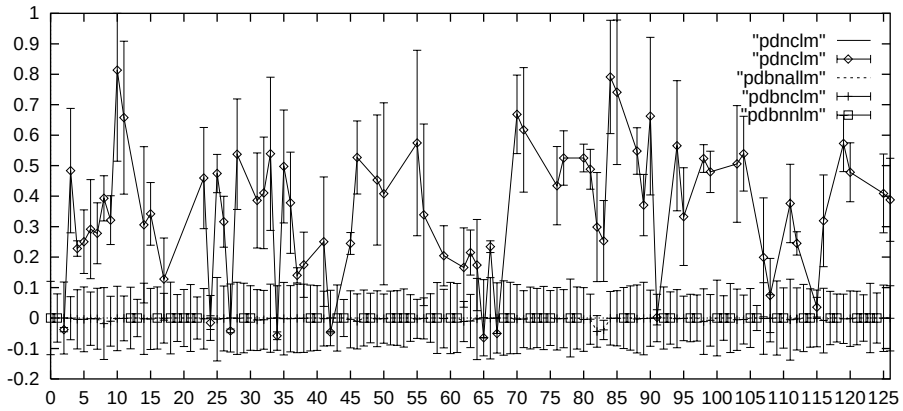


(k) Original "080.m.r.cc" mammogram

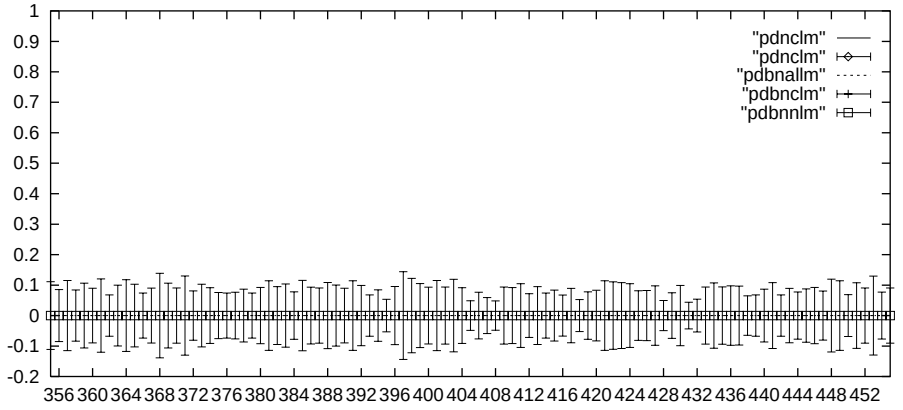


(l) Summed disk feature for "080.m.r.cc" image

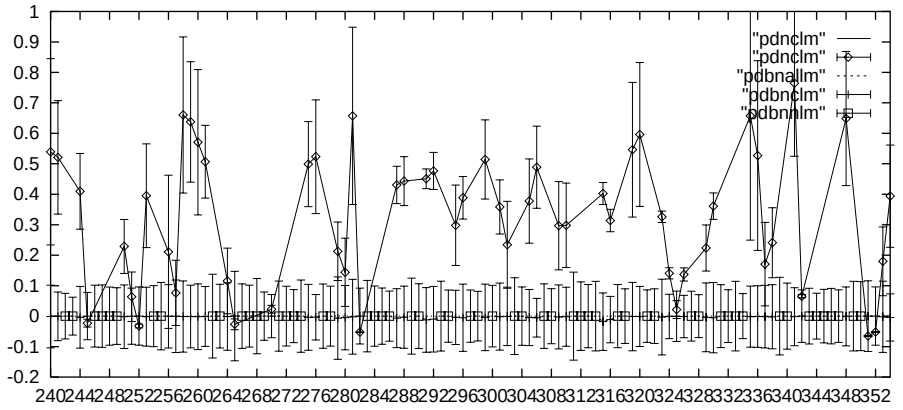
Figure 4.11: Peak, original, and summed disk feature images for selected mammograms from the BGSM mammogram data set.



(a) Plot of cdf-m map of malignant features



(b) Plot of cdf-m map of normal features



(c) Plot of cdf-m map of benign features

Figure 4.12: Plots of features across part of the BGSM mammogram data set.

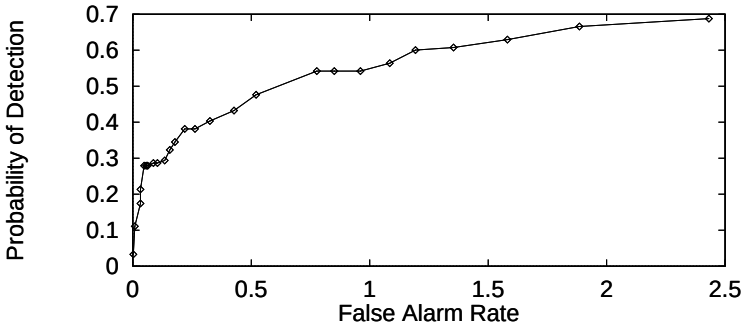


Figure 4.13: Performance Results as a ROC Curve for the BGSM data set.

Caption	Meaning
Filename	Root image filename
Mass	First or second mass in image
VWOT	Visible without truth
EdgeQ	Quadrants of edge coincident with truth
HP	Hough peak
CDFPD	CDF peak disk feature
DFMsmth	Dense feature map smoothed
DFMthr	Dense feature map thresh
DFMres	Dense feature map result
Table Entry	$0 \leq$ feature visibility or detection ≤ 4

Table 4.1: Key to column captions and numerical table entries.

Filename	Mass	VWOT	EdgeQ	HP	CDFPD	DFMsmth	DFMthr	DFMres
008.r.mlo	1	0	0	0	0	0	0	0
009.r.cc	1	0	1	0	0	0	0	0
011.l.mlo	1	0	0	0	0	0	0	0
018.r.mlo	1	0	0	0	0	0	0	0
026.l.mlo	1	0	1	0	0	0	0	0
063.l.mlo	1	0	0	0	0	0	0	0
069.l.mlo	1	0	0	0	0	0	0	0
071.r.mlo	1	0	0	0	0	0	0	0
083.l.mlo	1	0	1	0	0	0	0	0
006.r.cc	2	1	0	0	0	0	0	0
019.l.mlo	1	1	0	0	0	0	0	0
023.l.cc	2	1	1	0	0	0	0	0
072.l.mlo	1	1	0	0	0	0	0	0
023.l.mlo	2	4	0	0	0	0	0	0
065.l.cc	1	4	0	0	0	0	0	0
087.l.mlo	1	4	0	0	0	0	0	0
089.r.mlo	1	4	0	0	0	0	0	0
083.l.cc	1	0	1	1	0	0	0	0
084.l.cc	1	0	2	2	0	0	0	0
089.r.cc	1	0	0	2	0	0	0	0
001.r.cc	1	0	4	4	0	0	0	0
013.l.mlo	1	1	2	4	0	0	0	0
025.l.mlo	1	1	1	4	0	0	0	0
007.l.mlo	1	4	4	4	1	0	0	0
068.l.mlo	1	4	3	4	0	0	0	0
019.l.cc	1	1	2	0	1	0	0	0
079.l.cc	1	4	3	2	1	0	0	0
012.l.cc	1	0	1	2	2	0	0	0
006.r.mlo	2	4	4	2	2	0	2	0
016.r.mlo	1	4	4	2	2	1	1	0
090.l.cc	1	4	2	2	2	0	0	0
033.r.cc	1	1	2	4	2	0	0	0
066.l.mlo	1	1	3	4	2	0	0	0
012.l.mlo	1	4	3	4	2	0	0	0
064.r.mlo	1	4	4	4	2	0	0	0
018.r.cc	1	0	1	1	4	0	0	0
064.r.cc	1	1	1	1	4	0	0	0

Table 4.2: BGSM dataset performance analysis - part 1

Filename	Mass	VWOT	EdgeQ	HP	CDFPD	DFMsmth	DFMthr	DFMres
071.r.mlo	2	4	2	1	4	0	0	0
018.l.mlo	1	0	1	4	4	0	0	0
031.l.cc	1	0	2	4	4	0	0	0
001.r.mlo	1	1	4	4	4	0	0	0
026.l.cc	2	1	4	4	4	0	3	0
033.r.mlo	1	1	4	4	4	0	0	0
067.l.mlo	1	1	4	4	4	0	0	0
023.l.mlo	1	4	4	4	4	0	0	0
068.l.cc	1	4	2	4	4	0	2	0
075.r.mlo	1	4	3	4	4	0	0	0
088.r.cc	1	4	4	4	4	0	0	0
002.r.mlo	1	0	0	1	1	1	1	0
010.r.cc	1	4	2	2	2	1	4	0
008.r.cc	1	0	4	4	2	1	3	0
008.r.mlo	1	0	0	0	0	0	0	0
009.r.cc	1	0	1	0	0	0	0	0
011.l.mlo	1	0	0	0	0	0	0	0
018.r.mlo	1	0	0	0	0	0	0	0
026.l.mlo	1	0	1	0	0	0	0	0
063.l.mlo	1	0	0	0	0	0	0	0
069.l.mlo	1	0	0	0	0	0	0	0
071.r.mlo	1	0	0	0	0	0	0	0
083.l.mlo	1	0	1	0	0	0	0	0
006.r.cc	2	1	0	0	0	0	0	0
019.l.mlo	1	1	0	0	0	0	0	0
023.l.cc	2	1	1	0	0	0	0	0
072.l.mlo	1	1	0	0	0	0	0	0
023.l.mlo	2	4	0	0	0	0	0	0
011.r.mlo	1	4	3	4	2	1	1	4
016.r.cc	1	4	4	4	2	1	4	4
077.l.cc	1	4	4	4	2	1	1	0
006.r.mlo	1	4	3	1	4	1	3	4
080.r.mlo	1	4	1	2	4	1	1	0
010.r.mlo	1	4	3	4	4	1	4	4
015.l.cc	1	4	4	4	4	1	4	4
082.r.mlo	1	0	1	0	1	2	2	0
017.l.mlo	1	1	3	1	1	2	2	0
018.l.cc	1	1	1	1	1	2	2	0
014.l.cc	1	4	2	1	2	2	3	0
002.r.cc	1	0	2	2	2	2	2	0

Table 4.3: BGSM dataset performance analysis - part 2

Filename	Mass	VWOT	EdgeQ	HP	CDFPD	DFMsmth	DFMthr	DFMres
084.l.mlo	1	0	0	2	2	2	2	0
003.l.cc	1	4	1	2	2	2	2	0
086.l.cc	1	4	2	2	2	2	2	0
013.l.cc	1	0	1	4	2	2	4	4
029.l.xcc	1	1	2	4	2	2	2	0
015.l.mlo	1	4	4	4	2	2	4	4
066.l.cc	1	4	2	4	2	2	2	0
071.r.cc	2	4	2	1	4	2	2	0
030.l.cc	1	4	3	2	4	2	4	4
087.l.cc	1	0	4	4	4	2	4	4
066.r.mlo	1	1	3	4	4	2	4	4
003.l.mlo	1	4	4	4	4	2	2	0
004.r.cc	1	4	4	4	4	2	4	4
011.r.cc	1	4	4	4	4	2	4	4
014.l.mlo	1	4	4	4	4	2	4	4
022.r.mlo	1	4	3	4	4	2	4	4
024.r.mlo	1	4	3	4	4	2	2	0
027.l.mlo	1	4	4	4	4	2	3	4
028.l.cco	1	4	4	4	4	2	4	4
031.l.mlo	1	4	3	4	4	2	4	4
032.l.mlo	1	4	4	4	4	2	3	4
065.l.mlo	1	4	4	4	4	2	4	4
072.l.cc	1	4	4	4	4	2	4	4
075.r.cc	1	4	3	4	4	2	2	0
077.l.mlo	1	4	2	4	4	2	4	0
079.l.mlo	1	4	3	4	4	2	4	4
085.r.cc	1	4	4	4	4	2	2	0
085.r.mlo	1	4	4	4	4	2	4	4
086.l.mlo	1	4	3	4	4	2	2	0
029.l.mlo	1	1	3	2	4	3	4	4
004.r.mlo	1	0	2	4	4	3	3	0
070.r.cc	1	0	4	4	4	3	3	0
026.l.cc	1	1	4	4	4	3	3	4
006.r.cc	1	4	4	4	4	3	4	4
022.r.cc	1	4	4	4	4	3	3	4
023.r.cc	1	4	4	4	4	3	4	0
067.l.cc	1	4	4	4	4	3	4	4
070.r.mlo	1	4	4	4	4	3	3	4
073.r.cc	1	4	4	4	4	3	3	4
074.r.cc	1	4	3	4	4	3	3	4

Table 4.4: BGSM dataset performance analysis - part 3

Filename	Mass	VWOT	EdgeQ	HP	CDFPD	DFMsmth	DFMthr	DFMres
074.r.mlo	1	4	4	4	4	3	4	4
076.r.cc	1	4	4	4	4	3	3	4
078.l.cc	1	4	4	4	4	3	4	4
081.r.mlo	1	4	4	4	4	3	3	4
090.l.mlo	1	4	4	4	4	3	4	4
030.l.mlo	1	4	2	1	1	4	4	4
011.l.cc	1	4	4	4	2	4	4	4
071.r.cc	1	4	4	4	2	4	4	0
082.r.cc	1	0	2	1	4	4	4	4
026.l.mlo	2	0	3	4	4	4	4	4
025.l.cc	1	1	4	4	4	4	4	4
062.l.mlo	1	1	4	4	4	4	4	4
005.l.cc	1	4	4	4	4	4	4	4
005.l.mlo	1	4	4	4	4	4	4	4
009.l.cc	1	4	4	4	4	4	4	4
009.l.mlo	1	4	4	4	4	4	4	4
009.r.mlo	1	4	4	4	4	4	4	4
020.l.cc	1	4	4	4	4	4	4	4
020.l.mlo	1	4	4	4	4	4	4	4
021.r.cc	1	4	3	4	4	4	4	4
021.r.mlo	1	4	2	4	4	4	4	4
023.l.cc	1	4	4	4	4	4	4	4
023.r.mlo	1	4	4	4	4	4	4	0
024.r.cc	1	4	4	4	4	4	4	4
027.l.cc	1	4	4	4	4	4	4	0
028.l.mlo	1	4	4	4	4	4	4	4
032.l.cc	1	4	4	4	4	4	4	4
062.l.cc	1	4	4	4	4	4	4	4
063.l.cc	1	4	4	4	4	4	4	4
066.r.cc	1	4	4	4	4	4	4	4
073.r.mlo	1	4	4	4	4	4	4	4
078.l.mlo	1	4	4	4	4	4	4	4
080.r.cc	1	4	3	4	4	4	4	0
081.r.cc	1	4	4	4	4	4	4	4

Table 4.5: BGSM dataset performance analysis - part 4

4.5 Conclusions

We have succeeded in completing and testing an end-to-end framework for mass detection which is both competitive with those published elsewhere on different data sets and serves to illuminate some of the counter-intuitive areas of success and failure in this extremely complex detection problem. The algorithm performance is illustrated in Figures 4.10 and 4.13, and some of the background material and experiments involved in the algorithm design are described in Appendix C. While this performance is encouraging, much remains to be done prior to introduction in clinical practice. We bring up a number of possible avenues for extension of this work in the next section.

4.6 Discussion

The research and experiments on feature generation described in the last few sections produced an end-to-end detection system with reasonable performance, given the simplistic assumptions made about masses and breast tissue. While the work described herein is a good first step, achieving clinically significant detection and false alarm rates will require more sophisticated methods. One such approach is outlined below:

1. Conduct a study to characterize masses by their size, shape, difficulty of detection by humans, and rate of incidence. In addition, characterize mammograms by breast tissue makeup, difficulty in reading, and density. Analysis of the traced masses in the BGSM database would be a good first step in this direction.

2. Form a mathematical model capable of encompassing the mass shapes and breast tissue types encountered that takes the image formation process into account. Elliptic series [33, 34] are one method capable of describing mass shapes.

3. Determine a computationally feasible detection method employing this mass and image formation model. Simulated annealing techniques, whether stochastic [35, 36] or mean field [37, 38], form one set of methods for approaching problems of this complexity.

Chapter 5

Combination of Multiple Classifiers

5.1 Introduction

The previous two chapters have discussed the extraction of features from mammograms, features suitable for the detection of microcalcifications or circumscribed masses. Equally important to the detection process is the classification method used to analyze those features. In this chapter we report on an approach for improving classification performance through *combining* multiple classifiers, in an intuitive but original fashion.

5.2 Context and Related Work

Many pattern classification techniques exist, each with inherent strengths and weaknesses. No single classifier is suitable for all applications. Similarly, it may be difficult to achieve acceptable performance for complex data distributions using any single classifier. One alternative in the search for greater performance is integrated or adaptive methods which are capable of capitalizing on the strengths of several individual classifiers. Indeed, recent work in handwritten character recognition has shown promising results when several classifiers are combined to reach a decision [39, 40, 41, 42, 43, 44, 45, 46, 47]. This is often referred to as Combination of Multiple Classifiers (CMC). There are two basic approaches a CMC algorithm may take: classifier fusion, and dynamic classifier selection.

In classifier fusion algorithms, individual classifiers are applied in parallel, and their outputs are combined in some manner to achieve a “group consensus". A common characteristic of the fusion-type algorithms is that all classifiers contribute to the final decision. Common types of classifier fusion algorithms are voting methods, such as the majority vote [44, 47], the Borda count [40], unanimous consensus [40, 47], and thresholded voting [47]. The CMC algorithms described in [43] and [45] are polling methods which utilize heuristic decision rules to combine the assignments of two classifiers. The averaged Bayes classifier described in [47] is essentially a voting method that combines classifier decisions at the measurement level. More sophisticated classifier fusion algorithms arrange that better classifiers are weighted more heavily. Ho et al. [40]

describes a method that uses logistic regression to assign weights to the ranks produced by each classifier. Mandler [48] and Xu [47] use the Dempster-Shafer theory to derive weights for each classifier's vote. Methods of multistage classification [41] are basically classifier fusion algorithms in which the outputs of several classifiers are used as the inputs to another classifier. Thus, a generic classifier is used to determine the combination function.

Another class of CMC algorithms, known as Dynamic Classifier Selection (DCS), attempts to predict which single classifier is most likely to be correct for a given sample. First, partitions are defined that categorize input samples with respect to their responses to each classifier. For example, partitions can be defined by the set of individual classifier decisions [42], according to which classifiers agree with each other [40], or even by features of the input samples. Then, the "best" classifier for each partition is determined using training or validation data. For classification, an unknown sample is assigned to a partition, and the output of the best classifier for that partition is used to make the final decision. Thus, the final decision of a DCS algorithm does not directly involve all individual classifiers. Note that the partition assignment will likely utilize all classifiers. Since one or more of the classifier decisions are known to be incorrect, DCS algorithms attempt to ignore those classifiers that are not likely to be correct.

Some other CMC techniques may not be so easily categorized as either classifier fusion or dynamic classifier selection. Ho et al. [40] use a hybrid algorithm which defines partitions, and then uses logistic regression to determine a weighted voting scheme for each partition. This could be considered dynamic model selection since the weights assigned to each classifier's vote are determined after an input pattern has been assigned to a partition. Drucker et al. [39] describe a method of training an ensemble of neural networks, called "boosting". Here, the attention is focused on how the individual classifiers are trained, rather than how the classifiers are combined. Classifiers are trained in stages, and are therefore dependent on the classifiers in previous stages. The classifier decisions are combined using a polling method.

Most of the approaches mentioned treat the training data as a monolithic whole when determining classifier accuracy. However, it seems intuitive that the accuracy might vary with position in feature space. Given an arbitrary test sample and classifiers which may have different feature spaces, it is reasonable to think that a given classifier would perform similarly for other samples near the test sample in its feature space. So this chapter examines a method of dynamic classifier selection that uses estimates of a classifier's accuracy in local regions of feature space. Our proposed CMC algorithm simply selects the classifier which is most accurate for a subset of training samples nearest to the test sample. Note that the particular samples in the subset may vary among classifiers.

5.3 Our Method Compared to Related Approaches

We have selected two previously published algorithms [42, 46] for direct comparison to our proposed algorithm. We also implemented a modified version of one of these algorithms.

5.3.1 The Proposed Approach: DCS-LA

We term our approach to CMC as Dynamic Classifier Selection by Local Accuracy, or DCS-LA. The basic idea is to estimate each classifier's accuracy in local regions of feature space surrounding an unknown test sample, and then use the decision of the most locally accurate classifier. In our implementation "local regions" are defined in terms of the K-nearest neighbors in the training data. We examine two methods for estimating local accuracy. One is simply the percentage of training samples in the region that are correctly classified. We shall refer to this as the overall local accuracy. Another possibility is to estimate local accuracy with respect to some output class. Consider a classifier that assigns a test sample to class C_i. We can determine the percentage of the local training samples assigned to class C_i by this classifier that have been correctly labeled. We shall refer to this as the local class accuracy.

5.3.2 The Behavior-Knowledge Space Approach

The Behavior-Knowledge Space (BKS) algorithm has recently been proposed in connection with an application for recognizing handwritten numerals. Behavior-Knowledge Space is an N-dimensional space where each dimension corresponds to the decision of one classifier. Each classifier can assign a sample to one of M possible classes. Each unit of a BKS represents a particular intersection of individual classifier decisions. Thus, the BKS represents all possible combinations of the individual classifier decisions. Each BKS unit accumulates the number of training samples from each class. For an unknown test sample, the decisions of the individual classifiers index a unit of BKS, and the unknown sample is assigned to the class with the most training samples in that BKS unit.

5.3.3 The Classifier Rank Approach

Sabourin et al. [46] present an algorithm which has some similarities to our DCS-LA approach. One variation of their algorithm selects the classifier that correctly classifies the most consecutive neighboring training samples (relative to the unknown test sample). The selected classifier is said to have the highest "rank". Although they do not associate their algorithm with the concept of local accuracy, their notion of classifier rank certainly has this flavor. We will refer to this algorithm as the Classifier Rank method.

5.3.4 A Modified Classifier Rank Approach

In terms of our work, the Classifier Rank algorithm presented in [46] uses what we would describe as an overall local accuracy estimate. An obvious alternative would be to use local class accuracy. Given a test sample assigned to class C_i by a classifier, local accuracy for the classifier is estimated as the number of consecutive nearest neighbors assigned class C_i which have been correctly labeled. We refer to this algorithm as Modified Classifier Rank.

5.4 Methods

5.4.1 Data Sets

Our experiments focused on the data generated by the microcalcification work in Chapter 3. To review, recall that a data set of 40 digitized mammograms[1] containing some abnormalities was divided into two sets of images, Set A and Set B. From each set of images, pixels from the abnormal and normal class were randomly sampled. Set A has 19,735 samples from the normal class, and 3001 samples from the abnormal class. Set B is made up of 20,028 normal samples and 5159 abnormal samples. For each pixel, 63 features were computed. A more detailed description of this feature data is in Section 3.3.4.

Initially, Set A is used as training data for the individual classifiers and the CMC algorithm, and Set B is used to measure performance. Then the roles of Sets A and B are reversed. Thus, at no time are samples from the same image used for training *and* testing in the same set of experiments. Whenever we talk of training (or test) data, both Set A and Set B have been utilized independently in that capacity. Thus, the results of feature selection, individual classifier performance, and CMC results should be expected to be similar, but not identical, for Sets A and B.

5.4.2 ROC Analysis

Since the pattern recognition task here is to detect abnormalities in mammograms, labeling pixels as either “normal" or “abnormal" tissue, we have a 2 class problem.[2] Further, the accuracy of a classifier in a 2-class problem can be characterized by a plot of the classifier’s true positive detection rate versus its false positive rate, called a receiver operating characteristic (ROC) curve. The Area Under the ROC Curve (AUC) is an accepted way of comparing *overall* classifier performance [49, 50]. Hanley and McNeil [51] describe methods to determine if the observed difference between two AUCs is statistically significant. These standard statistical methods compare AUCs over the full range of TP rates. Our empirical ROC results only cover a portion of the full range, and so AUCs must be expressed as conditional probabilities prior to applying the methods of Hanley and McNeil.

First, the AUCs over the range of interest are estimated using the trapezoid rule for the discrete operating points. The area under a portion of an ROC curve can be expressed as a conditional probability via the following transformation:

$$AUC = \frac{A_p}{TP_2 - TP_1} \tag{5.1}$$

where A_p is the area under the ROC curve computed between TP rates TP_1 and TP_2. The formula

[1] Images were provided by courtesy of the National Expert and Training Centre for Breast Cancer Screening and the Department of Radiology at the University of Nijmegen, the Netherlands.

[2] Though note that the DCS-LA algorithm is also applicable to multi-class problems such as character recognition. Additionally, multi-class problems can be defined in terms of 2 classes by making a binary (yes or no) decision for every class.

for the z statistic is

$$z = \frac{AUC_1 - AUC_2}{\sqrt{{SE_1}^2 + {SE_2}^2}} \tag{5.2}$$

where AUC_1 and AUC_2 are the two estimated AUCs, and SE_1 and SE_2 are the estimated standard errors of each AUC. We use a two-tailed test for statistical significance. The null hypothesis is that the two observed AUCs are the same. The alternate hypothesis is that the two AUCs are different. A critical range of $z > 1.96$ or $z < -1.96$ (a level of significance $\alpha = 0.05$) indicates that the null hypothesis can be rejected.

A conservative estimate of the standard error of an AUC value (from [51]) is:

$$SE(AUC_i) = \sqrt{\frac{\theta(1-\theta) + (n_A - 1)(Q_1 - \theta^2) + (n_N - 1)(Q_2 - \theta^2)}{n_A n_N}} \tag{5.3}$$

where Q_1 and Q_2 are two distribution-specific quantities, θ is the “true" area under the ROC curve, and n_A and n_N are the number of abnormal and normal samples, respectively. The estimate AUC_i is used as an estimate of θ. The quantities Q_1 and Q_2 are expressed as functions of θ:

$$Q_1 = \frac{\theta}{2-\theta} \;\; and \;\; Q_2 = \frac{2\theta^2}{1+\theta} \tag{5.4}$$

Each of the individual classifiers is usually able to generate operating points running from 0% to 100% with fairly small increments between consecutive points. To generate a single operating point for a CMC algorithm, the individual classifiers are all set to approximately the same level of sensitivity, and the CMC is executed. This procedure is repeated with the individual classifiers set to other sensitivity levels, resulting in a series of operating points for each CMC algorithm. The ROC curves generated for each CMC algorithm will not cover the full range of TP rates. Therefore, in a test for statistical significance, two ROC curves are compared *only* over the range of TP rates that are common to both curves.

5.4.3 Individual Classifiers

Six individual classifiers were used in the various CMC algorithms, two parametric and four non-parametric. They are: Linear Bayesian, Quadratic Bayesian, K-Nearest Neighbor (K-NN) with the Euclidean distance metric [52], a fully connected back-propagation artificial neural network (ANN) with sigmoid activation functions [53], the C4.5 decision tree implementation [54], and a CART decision tree classifier [55].

For a CMC approach to be of practical use, it should improve on the best individual classifier, given that the individual classifiers have been reasonably optimized with regards to parameter settings and available feature data. In our work, an earnest effort was made to optimize each individual classifier with respect to selecting “good" values for the parameters which govern its performance. For the K-NN classifier, a value of K must be determined, and the Mahalanobis [52] distance metric is used. For ANNs, the numbers of hidden layers and hidden nodes in a layer must be selected. The parameters for the C4.5 decision tree algorithm are selected based on our

previous experience with this classifier, and the CART parameters are those that were determined to be optimal in the research discussed in Section 3. The Bayesian classifiers do not require any sort of parameter selection or optimization.

If each individual classifier is not given the opportunity to select from all features, then the comparison of CMC algorithms to individual classifiers is biased. So here we performed feature selection for each classifier separately, letting each start from the full set of 63 features and weed down to the subset of those features that maximizes performance for that classifier. After applying these individual feature selection algorithms, each classifier ended up making use of between 5 and 15 features; the number and specific features actually used depends on the individual classifier.

5.4.4 DCS-LA Implementation and Application

The DCS-LA algorithm uses the training data, which may be different for each classifier, and the class assignments made by each classifier. Given an unknown sample, it is first labeled by all the individual classifiers. If all classifiers agree, there is no need to estimate local accuracy. When the individual classifiers disagree, local accuracy is estimated for each classifier, and the decision of the classifier with the highest local accuracy estimate is selected.

Occasionally, two (or more) classifiers with conflicting decisions will have the highest local accuracy estimates. Tie-breaking is handled by choosing the class that is selected most often among the tied classifiers. If a tie still exists, the classifier(s) with the next highest local accuracy will break the tie in the same manner as before. Determining the appropriate size for a local region is part of designing the DCS-LA approach. We ran experiments for various region sizes ranging from $K = 1$ to $K = 51$ using the Mahalanobis distance metric

Also, we would like to investigate the effect of setting the individual classifiers to various sensitivity levels prior to applying CMC. We tested all CMC algorithms with the individual classifiers set to 6 different TP rates: 70%, 75%, 80%, 85%, 90%, and 95%. If a classifier could not be set exactly to the desired TP rate desired, it was set as close as possible. As before, we ran experiments for various region sizes ranging from $K = 1$ to $K = 51$ for each of the 6 levels of individual classifier sensitivity.

5.5 Results

We show only those results obtained when the first half of the data set is used as training data. Nearly identical results were obtained for the experiments which utilized the other half of the data set in the training capacity.

Figure 5.1 shows partial ROC curves plotted for all 6 individual classifiers. The best individual classifier is KNN if the overall AUC is considered. However, there is no single best classifier across all TP rates. As a benchmark for useful CMC performance, we consider a composite ROC curve consisting of the “best" parts of the individual ROC curves. The composite ROC is a lower bound for practical CMC performance. We also plot ROC curves for an Oracle classifier, the theoretical upper bound on CMC performance. The composite and Oracle ROC curves are shown in Figure 1B.

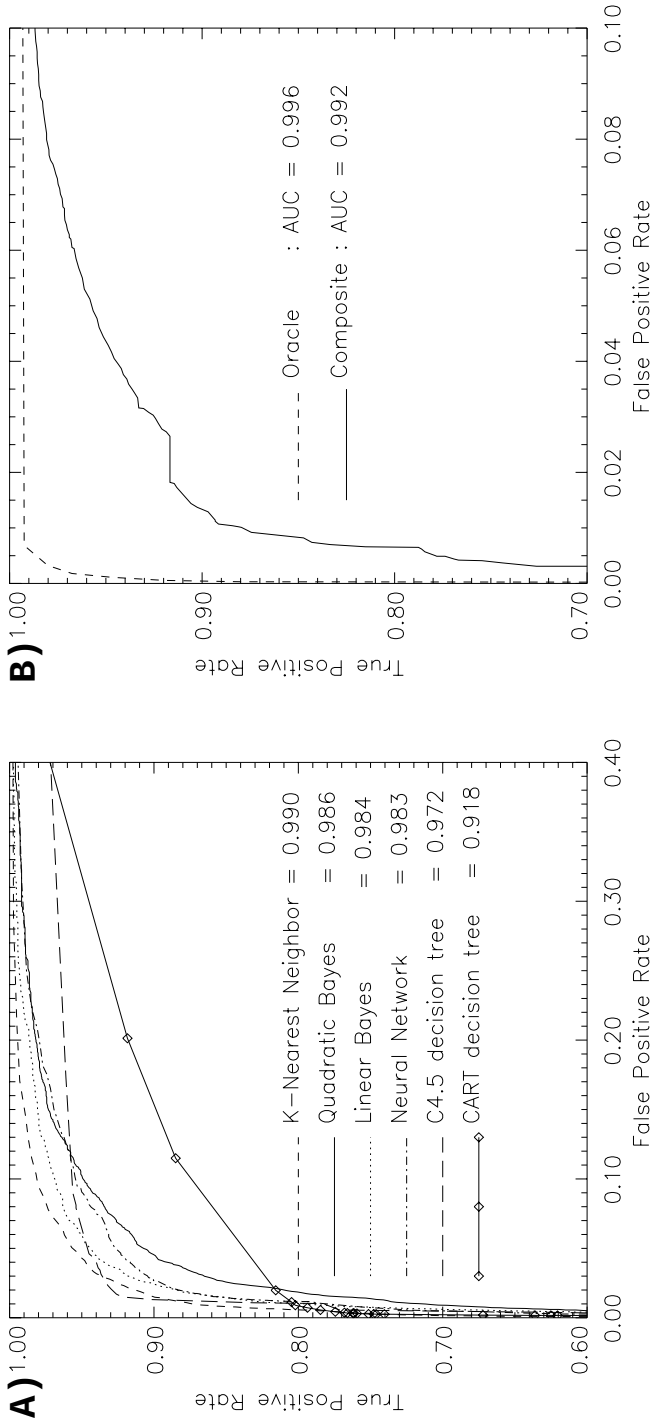


Figure 5.1: A) Partial ROC curves for 6 individual classifiers, and their AUCs. B) Composite ROC curve for the individual classifiers, and the ROC curve for an Oracle classifier.

A comparison of the ROC curves generated by the DCS-LA algorithm using both methods of local accuracy estimation shows that local class accuracy is superior to overall local accuracy. The difference in AUCs, however, is not statistically significant ($z = 1.44$ for TP rates ranging from 78% to 94%). Further ROC analysis of the DCS-LA algorithm with various local region sizes shows that regions defined by $K = 10$ generally seem to result in the best performance for this data set.

Figure 5.2 compares the composite ROC curve with the results for DCS-LA using local class accuracy. We also show the results of the Behavior Knowledge Space, Classifier Rank, and the Modified Classifier Rank algorithms. To be fair, only the best single value of K (10) is used in the plot for the DCS-LA results. Thus, the ROC curves for all four CMC algorithms are composed of 6 operating points each.

The DCS-LA algorithm is better than the best individual classifier at all times. The difference between the AUCs, computed over the range of common TP points (from 82% to 93%), for DCS-LA ROC curve and the Composite ROC curve is statistically significant ($z = 3.51$). The Modified Classifier Rank method performs nearly as well as DCS-LA at lower sensitivities, but less so at higher levels. It is significantly better than the best individual classifier ($z = 2.71$) over the common TP range (82% to 88%). The Classifier Rank method provides improvement, though not statistically significant, at some levels of sensitivity. As with our initial set of experiments, the Behavior Knowledge Space method is not able to improve upon the performance of the optimized individual classifiers. The DCS-LA method performed significantly better than the Behavior Knowledge Space method ($z = 4.91$ for TP rates ranging from 84% to 92%), and the Classifier Rank method ($z = 3.81$ for TP rates ranging from 82% to 91%).

Table 5.1 shows the results of the CMC algorithms when the individual classifiers are set (as close as possible) to a TP rate of 80%. The DCS-LA algorithm uses local class accuracy with $K = 10$. The number of times each individual classifier was selected by the DCS-LA algorithm is also shown. In this example, the DCS-LA algorithm finds operating points with higher TP rates *and* lower FP rates than points obtained by any individual classifier. All classifiers agree on the class assignment for a majority of the test samples (89.5%), and therefore any of the CMC algorithms are actually executed a relatively small percentage of the time. The number of times an individual classifier is selected by the DCS-LA algorithm seems closely correlated to the overall accuracy of the classifier. Results at other sensitivity levels show the same general trends. In general, since the DCS-LA algorithm is attempting to lower the total number of misclassifications, it generates operating points which make the appropriate TP/FP trade-off in order to drive the *overall* error rate down. Consider when all the individual classifiers are set to lower sensitivities (approximately less than 90%). Given the number of test samples per class, it is possible to misclassify fewer total samples by trading off a higher TP rate for a corresponding higher FP rate. By contrast, when we set all classifiers to TP rates of approximately 95%, the DCS-LA algorithm usually generated an operating point with a TP rate lower than 95%. In this situation, trading off the lower TP rate for the corresponding lower FP rate resulted in fewer total classification errors, and therefore an improved overall accuracy.

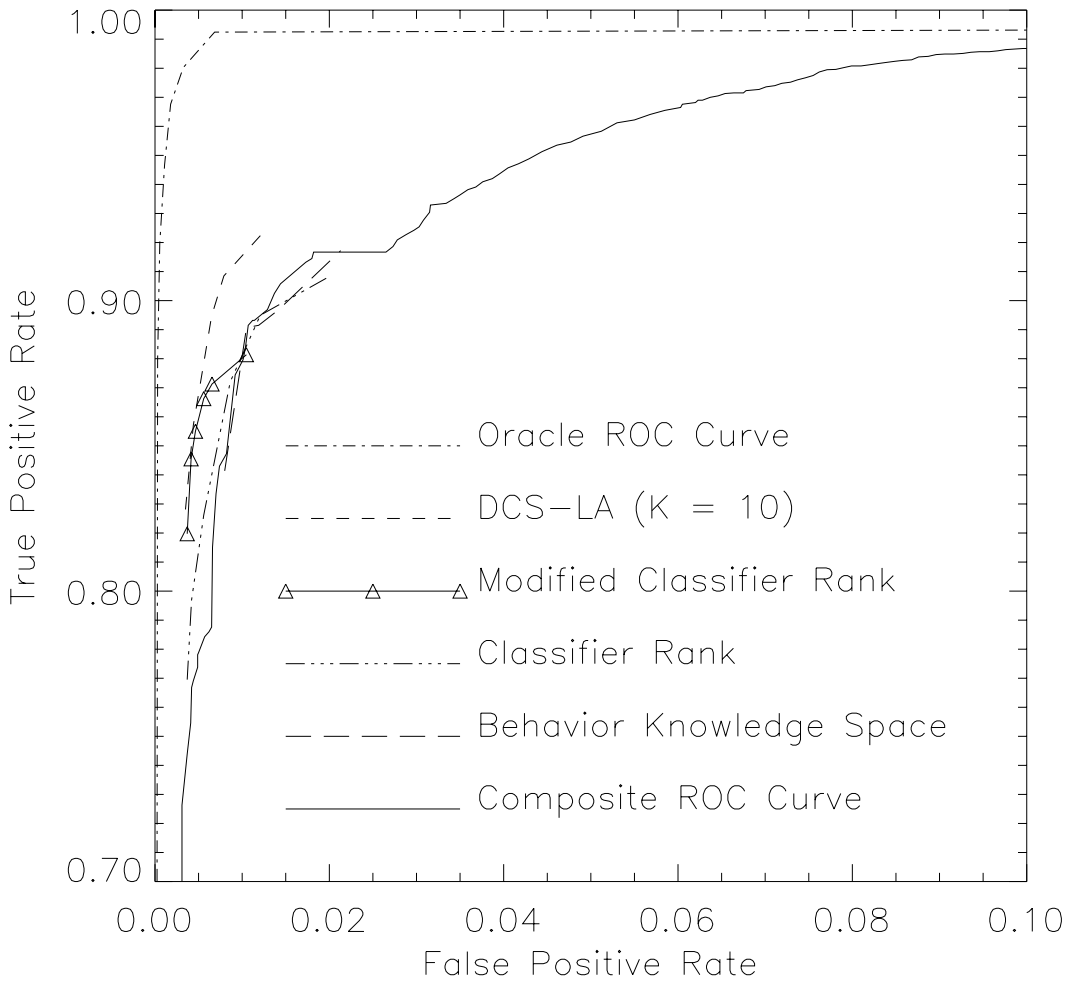


Figure 5.2: The composite and Oracle ROC curves for the 6 individual classifiers compared to the results for the DCS-LA, Behavior Knowledge Space, Classifier Rank, and Modified Classifier Rank methods.

Method of Classification	Set at (TP rate, FP rate)	Overall Accuracy	# times classifier selected by DCS-LA
Neural Network	(80.1, 0.85)	95.3%	1287
K-Nearest Neighbor	(79.8, 0.87)	95.2%	425
CART decision tree	(80.5, 1.11)	95.1%	444
C4.5 decision tree	(78.4, 0.57)	95.1%	287
Quadratic Bayes	(80.0, 1.84)	94.4%	125
Linear Bayes	(80.2, 1.97)	94.4%	67
Oracle	(94.7, 0.11)	98.8%	-
DCS-LA: Local Class Acc.	(87.7, 0.55)	97.0%	-
Behavior Knowledge Space	(89.7, 1.42)	96.8%	-
Classifier Rank Method	(82.6, 0.56)	96.0%	-
Modified Classifier Rank	(85.5, 0.46)	96.7%	-

Table 5.1: CMC results with individual classifiers set to TP rates as close to 80% as possible. All classifiers agree for 22,552 of the samples, or about 89.5% of the time.

5.6 Conclusions

We have shown that even if all the individual classifiers have been optimized, dynamic classifier selection by local accuracy is still capable of improving overall performance significantly. By contrast, simple voting techniques, and even a recently proposed CMC algorithm, were not able to show any significant improvement when the individual classifiers were sufficiently optimized. At times, some of the other CMC algorithms actually hurt performance. The proposed DCS-LA algorithm was *always* capable of improving performance.

In this work, we have attempted to address some issues relevant to the construction of a multiple classifier system which have not previously received attention. First, we have made efforts to optimize the individual classifiers with respect to the available feature data. Certainly it would be preferable to use a single classifier as opposed to a combination of several classifiers if the performance of the two systems is equivalent. Second, we have suggested a systematic procedure for determining if certain classifiers are redundant or detrimental, and could therefore be removed from the mix of individual classifiers prior to CMC. The end result is improved performance, and faster execution time. Finally, we observed the effect of varying the sensitivity of the individual classifiers on the CMC algorithm.

The benefits of a CMC approach may be limited when there is a very small amount of training data, or when the classification accuracy of an individual classifier is sufficiently high. Thus, we believe the greatest potential for CMC algorithms is for large data sets with data distributions that are too complex for most individual classifiers.

Chapter 6

Discussion

6.1 An Outline of Accomplishments

The long-term goal of this LDRD effort was to advance the state of the art in computer-aided mammographic screening. The approach taken was to use the pattern recognition methods that worked so well with spiculated lesions (that is, the use of dense feature maps, binary decision trees, and probability images) and apply them to appropriate features for microcalcifications and circumscribed masses. The aim behind using a common pattern recognition approach was not only to follow a successful lead, but also to investigate the benefits of having a common "malignancy" representation, in the form of probability images. Accordingly, we also looked at methods for combination of classifiers and of representations.

A terse summary of the results of this work is as follows:

- 8 technical journal and conference papers, and numerous conference presentations.
- A demonstration of the improvements gained by use of our microcalcification methods as compared to the prior best known results on a given data set.
- Acquisition of a substantial mammogram database containing a variety of biopsy-proven masses, as well as 3 year followup normals.
- Widely applicable algorithms for flattening contrast near breast tissue boundaries and suppression of linear features in mammograms.
- Completion and testing of an end-to-end mass detection algorithm which performed reasonably well and offered broad insight into the problem area.
- The development of the local-feature-accuracy approach to combination of multiple classifiers, and a statistically rigorous demonstration of the substantial improvements gained by its use on mammographic data.

6.2 Future Work

There are three main research lines or projects that would sensibly continue the current work, outside of the various specific recommendations made in each earlier chapter:

Probability Image Integration: At present we have established image features for detecting spiculated lesions, microcalcifications, and circumscribed masses. In each case, we use the same pattern recognition method, which results in a probability image that suggests, at each pixel, the likelihood that each pixel is located on a lesion. The next step is to work out how best to combine these probability images so as to represent all of the cancer signs in a single image, while further allowing positive signs to reinforce each other.

One approach would be to simply combine the probability images with an appropriate probability calculus; this should be the first line investigated. Much more ambitious, but with a higher payoff, would be to combine the separate features used for each sign, to grow only one binary decision tree, and so only one probability image. This has the potential problem of greatly increasing the dimensionality of the feature space, but the potential benefit of allowing the previously separate features to interact to their mutual benefit.

Dual Image Investigations: Of the three dual-image problems, only asymmetry and sequential change detection have received any attention, and none of it has been entirely successful. All considered approaches have depended on registration of the image pairs. The difficulty here is that the breast is compressed during a mammogram, which seems to so randomize its appearance that registration becomes a possibly insoluble problem.

So an interesting approach would be to address the detailed image registration problem by avoiding it entirely. The core idea is to count on the dense, pixel-level feature representation (in which each pixel becomes associated with a vector of image measures) and the fact of general spatial continuity to provide information on local area trends without the need for precise, pixel to pixel matching.

A Clinical Study: The final step of any such work should be a clinical evaluation of its impact on the sensitivity and specificity of a screening mammographer. The computer's solo performance can be easily measured and reported. However, both the realities of mammographic practice and past experience indicate that the question of primary *clinical* interest is how well the computer/human system works together. Accordingly, any ' integrated detection method should be tested in a final retrospective clinical study. From prior experience, the scale should be such to have screening mammographers examine on the order of 500 four-view mammogram cases with and without computer prompting, to objectively determine the effect of the CAMS technology.

In addition, there remains a larger issue in CAMS research: Computer methods have yet to perform successfully in the clinical setting as a "second reader" or aid to the radiologist for the general mammographic screening process. A useful line of inquiry would be to determine if computer methods can be used to *rule out* cancers with a high degree of accuracy for a subset of the image population, thereby eliminating human readers entirely for that subset. This hypothesis could

be tested in parallel with the usual “second reader” approach by analyzing the images producing correct detections for image properties which would unequivocally identify them as analyzable by computer with high probability of success.

Chapter 7

Bibliography

[1] Robert A. Smith. Epidemiology of breast cancer. In Arthur G. Haus and Martin J. Yaffe, editors, *Syllabus: A Categorical Course in Physics; Technical Aspects of Breast Imaging*, pages 21–34. Radiological Society of North America, second edition, 1993.

[2] W. Philip Kegelmeyer, Jr., Joe M. Pruneda, Philip D. Bourland, Argye Hillis, Mark W. Riggs, and Michael Nipper. Computer-aided mammographic screening for spiculated lesions. *Radiology*, 191(2):331–337, May 1994.

[3] Laszlo Taber and Peter B. Dean. *Teaching Atlas of Mammography*. Georg Thieme Verlag, Thieme Inc., Stuttgart, New York, 2nd revised edition, 1985.

[4] Laurens V. Ackerman and Earl E. Gose. Breast lesion classification by computer and xeroradiograph. *Cancer*, 30:1025–1035, 1972.

[5] Nico Karssemeijer. Adaptive noise equalization and recognition of microcalcification clusters in mammograms. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6), 1993.

[6] W. Philip Kegelmeyer, Jr. A minimal error region merging technique for segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 1983.

[7] Laura N. Mascio, P.W. Verbeek, W-L Kuo, and J.W. Gray. Semi-automated probe mapping using digital imaging microscopy. I. system development. *Cytometry*, 19(1):51–59, 1995.

[8] K.S. Woods, J.L. Solka, C.E. Priebe, W.P. Kegelmeyer, C.C. Doss, and K.W. Bowyer. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, December 1993.

[9] Heang-Ping Chan, Kunio Doi, Simranjit Galhotra, Carl J. Vyborny, Heber MacMahon, and Peter M. Jokich. Image feature analysis and computer-aided diagnosis in digital radiography, 1. automated detection of microcalcifications in mammography. *Medical Physics*, 14(4):538–548, July/August 1987.

[10] Heang-Ping Chan, Kunio Doi, Carl J. Vyborny, Kwok-Leung Lam, and Robert A. Schmidt. Computer-aided detection of microcalcifications in mammograms — methodology and preliminary clinical study. *Investigations in Radiology*, 23(9):664–671, September 1988.

[11] Laura N. Mascio, Jose M. Hernandez, and Clinton M. Logan. Automated analysis for microcalcifications in high resolution digital mammograms. In Murray H. Lowe, editor, *Medical Imaging 1993: Image Processing, Proc. of SPIE 1898*, volume 1905-78, pages 472–479, February 1993.

[12] M. DoCarmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, 1976.

[13] Matti Pietikainen, Azriel Rosenfeld, and Larry S. Davis. Texture classification using averages of local pattern matches. In *Proceedings of the 6th International Conference on Pattern Recognition*, volume 1, pages 301–303. IEEE Press, New York, NY, USA, 1982.

[14] K. I. Laws. *Textured Image Segmentation*. PhD thesis, University of Southern California, 1980.

[15] William G. Wee, Myron Moskowitz, Nai-Ching Chang, Yoeung-Ching Ting, and Seshagirirao Pemmeraju. Evaluation of mammographic calcifications using a computer program. *Radiology*, 116:717–720, September 1975.

[16] D.H. Davies, D.R. Dance, and C.H. Jones. Automatic detection of clusters of classifications in digital mammograms. In *Medical Imaging IV: Image Processing*, volume 1233, pages 185–191. SPIE, 1990.

[17] D.H. Davies and D.R. Dance. Automatic computer detection of clustered calcifications in digital mammograms. *Phys. Med. Biol.*, 35(8):1111–1118, 1990.

[18] Nico Karssemeijer. A stochastic model for automated detection of calcifications in digital mammograms. In *Proceedings of the 12th International Conference on Information Processing in Medical Imaging.*, pages 227–238, 1991.

[19] Wolfgang Spiesberger. Mammogram inspection by computer. *IEEE Transactions on Biomedical Engineering*, 26(4):213–219, April 1979.

[20] Bahaa W. Fam, Sherry L. Olson, Peter F. Winter, and Francis J. Scholz. Algorithm for the detection of fine clustered calcifications on film mammograms. *Radiology*, 169:333–337, 1988.

[21] Heang-Ping Chan, Kunio Doi, Carl J. Vyborny, Robert A. Schmidt, Charles E. Metz, Kwok Leung Lam, Toshihiro Ogura, Yuzheng Wu, and Heber MacMahon. Improvement in radiologist's detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. *Investigations In Radiology*, 25(10):1102–1110, October 1990.

[22] W. Philip Kegelmeyer, Jr. Evaluation of stellate lesion detection in a standard mammogram data set. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6):1477–1492, December 1993.

[23] Peter Miller and Sue Astley. Classification of breast tissue by texture analysis. *Image and Vision Computing*, 10(5):277–282, June 1992.

[24] Nico Karssemeijer. Stochastic model for automated detection of calcifications in digital mammograms. *Image and Vision Computing*, 10(6):369–375, August 1992.

[25] C. Kimme, D. Ballard, and J. Sklansky. Finding circles by an array of accumulators. *Communications of the ACM*, 18(2):120–122, 1975.

[26] E. Karabassis and M. E. Spetsakis. An analysis of image interpolation, differentiation, and reduction using local polynomial fits. *Graphical Models and Image Processing*, 57(1):183–195, May 1995.

[27] Nick Cerneaz and Mike Brady. Enriching digital mammogram image analysis with a description of the curvi-linear structures. In A. G. Gale et al., editors, *Digital Mammography: Proceedings of the Second International Workshop on Digital Mammography, York, England*, International Congress Series No. 1069, pages 297–306, Amsterdam, July 1994. Elsevier. ISBN 0 444 81916 9.

[28] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–697, November 1986.

[29] Svetha Venkatesh and Paul L. Rosin. Dynamic threshold determination by local and global edge evaluation. *Graphical Models And Image Processing*, 57(2):1246–160, March 1995.

[30] R. O. Duda and P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, January 1972.

[31] D. C. W. Pao, Hon F. Li, and R. Jayakumar. Shapes recognition using the straight line Hough transform: Theory and generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1076–1089, November 1992.

[32] V. F. Leavers. Which Hough transform? *Computer Vision, Graphics, and Image Processing: Image Understanding*, 58(2):250–264, September 1993.

[33] Bennett R. Groshong, L. A. Spero, and J. T. Cusma. Maximum a posteriori optimization - a method for calculation of dynamic changes in ventricular contours from angiographic image sequences. In *Proceedings of Computers in Cardiology 1992*. IEEE Computer Society Press, October 1992.

[34] Bennett R. Groshong. Estimating simple closed contours in images. In *Proceedings of the Fifth Annual IEEE Computer Based Medical Systems Symposium*. IEEE Computer Society Press, June 1992.

[35] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[36] E.H.L. Aarts and P.J.M. van Laarhoven. *Simulated Annealing: Theory and Applications*. D. Reidel Publishing Company, Dordrecht, Holland, 1987.

[37] Griff L. Bilbro and Wesley E. Snyder. Applying mean field annealing to image noise removal. *Journal of Neural Network Computing*, pages 5–17, Fall 1990.

[38] Griff L. Bilbro, Wesley E. Snyder, and Reinhold C. Mann. Mean-field approximation minimizes relative entropy. *josa-a*, 8(2):290–294, feb 1991.

[39] H Drucker, R Schapire, and P Simard. Boosting performance in neural networks. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 7(4):705–719, 1993.

[40] T K Ho, J J Hull, and S N Srihari. Decision combination in multiple classifier systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(4):66–75, January 1994.

[41] Y S Huang and C Y Suen. A method of combining multiple classifiers - a neural network approach. In *Proceedings of the 12th International Conference on Pattern Recogntion and Computer Vision*, pages 473–475, Jerusalem, Israel, 1994.

[42] Y S Huang and C Y Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(1):90–94, 1995.

[43] F Kimura and M Shridar. Handwritten numerical recognition based on multiple algorithms. *Pattern Recognition*, 24(10):969–983, 1991.

[44] L Lam and C Y Suen. A theoretical analysis of the application of majority voting to pattern recognition. In *Proceedings of the 12th International Conference on Pattern Recogntion and Computer Vision*, pages 418–420, Jerusalem, Israel, 1994.

[45] C Nadal, R Legault, and C Y Suen. Complementary algorithms for the recognition of totally unconstrained handwritten numerals. In *Proceedings of the 10th International Conference on Pattern Recogntion*, pages 443–449, Atlantic City, NJ, 1990.

[46] M Sabourin, A Mitiche, D Thomas, and G Nagy. Classifier combination for hand-printed digit recognition. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, pages 163–166, Tsukuba Saenie City, Japan, 20-22 Oct. 1993.

[47] L Xu, A Krzyzak, and C Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435, 1992.

[48] E Mandler and J Schurmann. Combining the classification results of independent classifiers based on the dempster-shafer theory of evidence. In *Pattern Recognition and Artificial Intelligence*, pages 381–393, North-Holland, 1988. Elsevier Science Publishers B.V.

[49] C E Metz. ROC methodology in radiologic imaging. *Investigative Radiology*, 21:720–733, 1986.

[50] J A Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14:109–121, 1979.

[51] J A Hanley and B J McNeil. The meaning and use of the area under a reciever operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[52] R O Duda and P E Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[53] K Knight. Connectionist ideas and algorithms. *Commun. of the ACM*, 33(11):59–74, 1990.

[54] J R Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1992.

[55] L Breiman, J H Friedman, R A Olsen, and C J Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[56] Harold L. Kundel, Calvin F. Nodine, and Elizabeth A. Krumpinski. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. *Investigative Radiology*, 25(8):890–896, August 1990.

Appendix A

General Notes on the Appendices

The following appendices are most likely of interest only to someone who is interested in resuming one or more of the lines of research initiated in this project. They contain accumulated notes on ideas that were not investigated thoroughly enough to be included in the main report, various partial results, "to do" lists of issues to investigate, bugs to fix, code to write, and other general loose ends.

In particular, the comments in these appendices are not intended to be complete, polished, or even self-contained. They assume that the interested reader will follow up with searches through the relevant on-line work notes. As of when this SAND report was published, these notes are available as follows:

Philip Kegelmeyer: CA/RAN network, machine `ananda`, directory `~wpk/sandia/`. The work notes are in LaTeX format, and are organized by year and by month within the year. See the Perl script `when` in `wpk/util/bin/` for an easy way to do date-ordered regular expression searches on the work notes.

See also the `wpk/cams/` directory, particularly its `Docs/` subdirectory.

Bennett Groshong: CA/RAN network, machine `gorp`, directory `~brg/text/jarchive/`. The work notes are in LaTeX format, and are organized by year and by month within the year. See also the `bgsm/, khoros_book, ldrd96/, mammo/, mi96/, sand97/, tiwdm96/` sub-directories of `brg/text/`.

Mark Allmen: CA/RAN network, machine `gorp`, directory `~allmen/sandia/notes/daily`. (Though note that the bulk of these notes concern a project unrelated to this computer aided mammography effort). See also `allmen/sandia/applications/cams/`.

Appendix B

Microcalcification Issues

There is one significant outstanding issue in the MC analysis. In the wake of the successful use of quasi-sequential feature (QSFS) analysis to select the best subset of the Laws features for MC detection, that analysis was expanded to cover *all* 63 of the features investigated. The processing and analysis is documented in Philip Kegelmeyer's work notes for the last few months of 1994, and is summarized by the entries on November 23, 1994.

Those investigations turned up an oddity that was never completely explained. The anomaly is that portions of the QSFS analysis did not well match the empirical analysis, and that end-to-end performance testing of some of the QSFS-suggested features resulted in lower scores than the original feature set. This was surprising in that the QSFS analysis was expected to entirely raise scores.

Discussion and proposed explanation are in the cited notes, but the primary possibility is the way the QSFS runs were scored. As an estimate of end-to-end performance (which would have been very time consuming to compute), the resubstitution error rate (RER) of each tree was used. It now seems likely that this was too crude an estimate, in particular since it can be driven artificially to zero simply by reducing the degree to which the tree was pruned. An additional possibility is the way the RER value is currently normalized across the entire set of trees, in order to remove the influence of the simple number of internal nodes.

There are, then, three initial projects that would follow up this concern:

- One should be a smaller scale QSFS experiment in which end-to-end performance *is* considered, to see at what point these results diverge from the RER-generated results.
- It might also be useful to look at *all* 2 case pairs. Though note that even with using the faster RER criteria, to check the best five from Level 1 vs exhaustive matching from Level1 would take about 21 days of parallel processing on HEAT.
- Lastly, consider the exact set of "optimal" features that was selected as part of the analysis in Section 5.4.3. This is yet another opinion as to which of the MC features is most useful, and may provide a useful perspective.

Appendix C

Mass Feature Issues

C.1 Notes and Experiments

The following items are from Bennett Groshong's notes on investigations and experiments pertaining to mass feature algorithm development during the course of the project. They contain background material that supports the current algorithm choices and material that would be relevant to future investigations in this area.

Gauss filtering, edge detection accuracy: We experimented with a larger Gaussian ($\sigma_x = \sigma_y = 2.0$), which increased the accuracy of the gradient phase angle dramatically for smaller radii circles, but failed to improve end-to-end detection performance noticeably. We also experimented with a directional Canny [28] using 8 oriented filters ($\sigma_x = 1.0, \sigma_y = 2.0$) with similar results. We attribute this lack of improved detection performance to the quantity of background structure cluttering in the mammograms. This experiment was conducted prior to adding linear feature suppression and Venkatesh edge classification and thresholding to the algorithm.

Hough transform: Experiments with the Hough transform and htccmax (Hough parameter space peak detection) aren't hugely encouraging. It is too sensitive, detecting peaks that have little apparent relevance in the image - eg. slightly curved vessels. Signal to noise ratio is definitely a problem.

Hough back projection: Talked to WPK about idea of correlation of nearby points. He brought up several ideas - one was "back-projection" of peaks. The idea is to take Hough peaks and then project back into image domain to see whether the peaks are plausible. This is way of relating peaks in the Hough domain to the image domain. It could be a used to find "good" peaks in Hough domain, rather than just big ones.

I mentioned spatial density and clustering in Hough domain. My point was that nearby feature points in model space are correlated, nearby points in image space are correlated, how can we use this to improve performance of Hough? Philip's idea was to back-project a

Circle radius	Hough peak radius
10	1
20	3
30	5

Table C.1: Circle and Hough accumulator peak radius correlation

circle, expanding into a ring - uncertainty due to cloud of points in Hough domain and then do correlation to determine the quality of match.

Hough transform accumulator: Peaks aren't focused, even for exact circles with no noise. Experimentally, the larger the circle, the more spread the accumulator (HTA) values. The correspondence between circle radius and Hough accumulator peak radius is There is an uncertainty in r of about $+/-1$ pixel due to sampling and edge detection, and an angular uncertainty in projecting the edge point along r to the supposed circle center for the same reasons. Since the edge location error doesn't scale with radius, the peak spread change must be due to the angular uncertainty. The uncertainty will be greater for mammograms, with their added noise and interfering structure.

Hourglass Hough accumulator filter: Operate on the Hough accumulator with an hourglass or double cone local maxima approach. This will capture some of the correlation in the Hough accumulator. It should outperform cylindrical filter, but will be horrendously computationally complex.

Cylindrical Hough accumulator filter: Idea for a cylinder maxima filter: determine the maximum over cylinder in r, then determine the peak over r at each point, then output a disk on this to obtain the absolute maximum. This would be reasonably efficient, but not as good as the cone filter above.

Inside-outside means: Post process the cylinder filter (htccmax) data, evaluating the mean for points inside a disk of radius R, and compare this response with the mean of an $R <= r <= 2R$ disk outside. If the inside mean is $> x$, then mark as a candidate. This performs well where the overall background image gradient is fairly low, with performance falling off as the gradient increases.

Hough transform peak characterization: We performed an experiment to measure Hough peak behavior radially out from center for circles of radii $2 <= r <= 30$. The behavior appears to reasonably follow either $1 - exp(1 - x^2.8)/(1 - e)$ or $(1 - x)^3.3$. The tail of the observed data is a little low as compared to the analytic functions.

Hough transform peak normalization: Three empirically determined filtering approaches to normalization of the Hough accumulator with respect to radius were tried:

	No noise											
	Raw HTA			Gaussian			Exponential			Integral		
radius	P	OP	R	P	OP	R	P	OP	R	P	OP	R
06	32	5	6.4	1.37	0.88	1.56	1.0	0.19	5.26	1.0	0.32	3.13
10	40	4	10	2.39	0.3	7.97	1.0	0.1	10	1	0.4	2.5
16	32	7	4.57	3.60	1.15	3.13	1.0	0.23	4.35	1	1.13	0.88
20	16	5	3.20	4.04	0.9	4.49	1.0	0.25	4	1	0.79	1.27
	White Gaussian Noise Added ($\sigma = 50$)											
	Raw HTA			Gaussian			Exponential			Integral		
radius	P	OP	R	P	OP	R	P	OP	R	P	OP	R
06	6	7	0.86	1.01	1.07	0.93	0.22	0.24	0.94	0.5	0.5	1
10	6	6	1	1.51	0.85	1.76	0.25	0.14	1.83	0.9	0.57	1.58
16	6	6	1	1.43	1.19	1.2	0.24	0.18	1.36	0.77	0.71	1.08
20	7	4	1.75	1.79	0.88	2.03	0.38	0.15	2.50	0.95	0.61	1.56

Table C.2: Comparison of peaks for clean and noisy circle images.

Integral: We designed a simple integral or box filter to give exact peak heights for synthetic circles.

Gaussian: We designed a simple Gaussian filter to give exact peak heights for synthetic circles.

Exponential: We designed an exponential filter fit to the peak heights and shapes at each radius, approximating a matched filter for the peak in the Hough accumulator generated by a circle of a given radius.

Experiments with Gauss filter, exponential filter, and integrated sum for noiseless and noisy circles with radii between $6 <= r <= 20$ are shown in Table C.2. One might argue that using the exponential alone makes as much sense as anything else with added noise. The min-max peak ratio for the exponential filter is 1.7 and for the integrated sum is 1.9.

Edge accuracy experiments: We coded and recorded edge accuracy for a radius 30 circle. Edge radial accuracy is described as the distance between the measured radius determined by the edge location and the circle center point and the specified radius, used in constructing the circle. ($sqrt((e_x - c_x)^2 + (e_y - c_y)^2) - r$ where $e(x, y)$ is the edge point, $c(x, y)$ is the circle center, and r is the circle radius.) Edge angular accuracy is the difference between the angle described by the edge pixel location and the circle center ($tan^-1 e_y - c_y, e_x - c_x$), and the angle described by the gradient direction. ($tan^-1(dy, dx)$)

Figure C.1 is a plot of the radial accuracy, generated by

```
gnuplot> plot 'c.30.gglout.sort' using 4:5 with lines
```

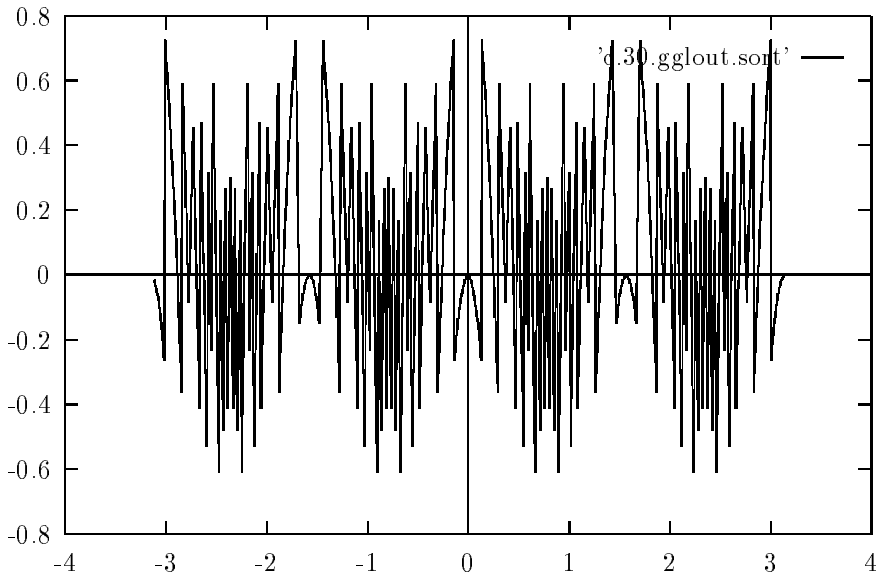


Figure C.1: Plot of radial edge location error vs angle around circle for a radius of 30 pixels.

where 'c.30.gglout.sort' is in ~brg/khoros/k2/exp/circ and describes the behavior of the edge accuracy for each edge point around the radius 30 circle. Columns $4-5$ describe the angle from the edge point to the circle center and the radial error for that point. Edges were computed by the gf3d Gaussian low pass filter with an $x, y\sigma = 1.0$, then computing the gradient, then the gradient local maxima.

Figure C.2 is a plot of the angular accuracy, generated by

```
gnuplot> plot 'c.30.gglout.sort' using 4:6 with lines
```

where 'c.30.gglout.sort' is as above. Columns $4-6$ describe the angle from the edge point to the circle center and the angular error for that point. Figure C.3 is a plot comparing the angular accuracy of the "Canny" and "glmax" edge detectors, generated by:

```
gnuplot> plot 'c.30.gglout.sort' using 4:6 with lines,
      'c.30.cangglout.sort' using 4:6 with lines
```

where 'c.30.gglout.sort' is as above, and '...can...' is the canny output.

Thought: How does angle and location accuracy vary with Gaussian blurring? This should produce a major improvement in angular error.

Figure C.4 is a plot comparing the angular accuracy of the "glmax" edge detector using an input image filtered with a Gaussian of 1 and 2, generated using

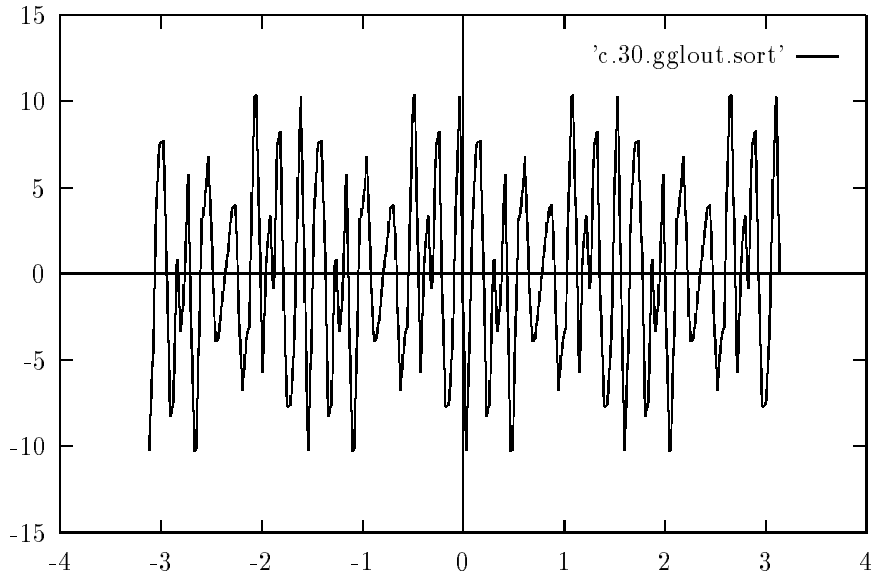


Figure C.2: Plot of radial edge angle error vs angle around circle for a radius of 30 pixels.

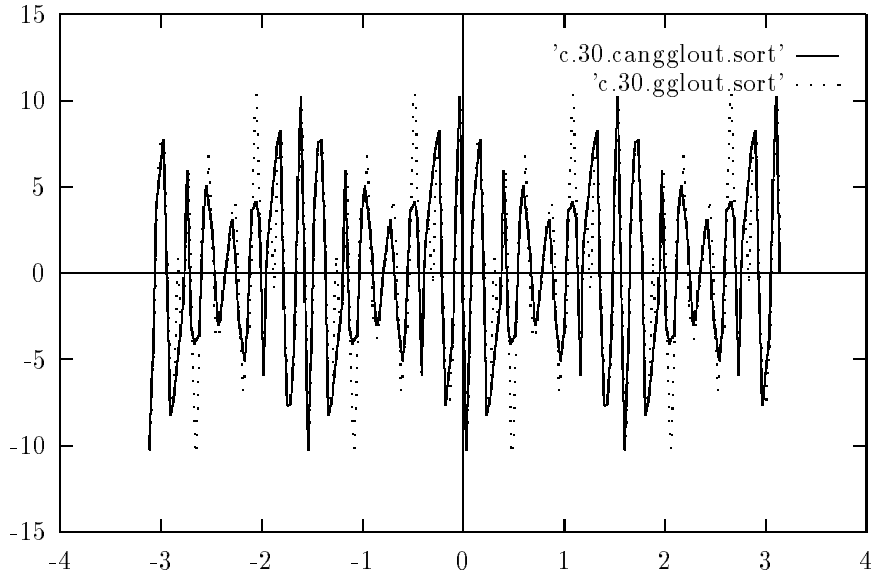


Figure C.3: Plot comparing radial edge angle error for the “Canny” and “glmax” edge detectors.

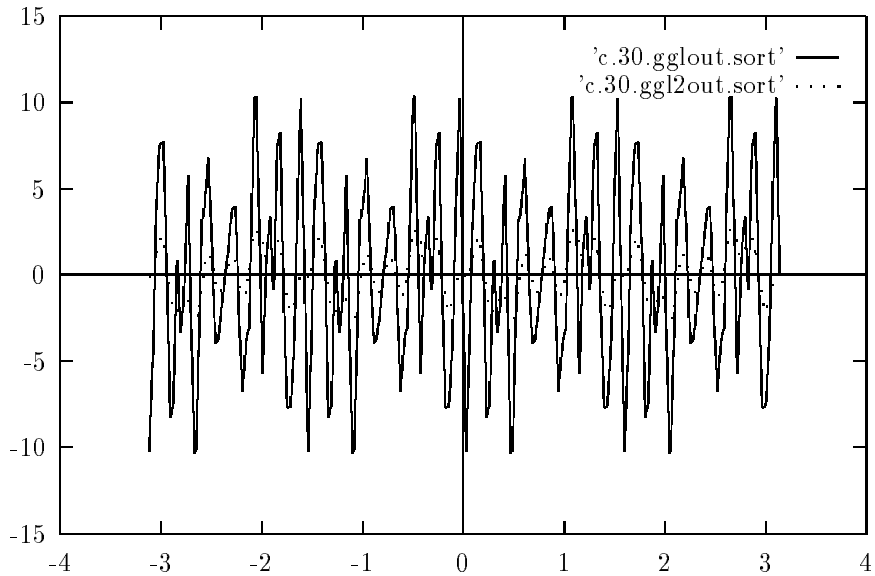


Figure C.4: Plot comparing radial edge angle error for the "glmax" edge detector operating from a Gaussian blur of 1 and 2 for a circle radius of 30

```
gnuplot> plot 'c.30.gglout.sort' using 4:6 with lines,
'c.30.ggl2out.sort' using 4:6 with lines
```

where 'c.30.gglout.sort' is as above resulting from a Gaussian filter $\sigma = 1$, and ...ggl2out... results from a Gaussian blur with $\sigma = 2$.

Now let's compare these errors with those for circles of radius 10 and 60 for a Gaussian lowpass filter of 1.0.

Figure C.5 is a plot comparing the angular accuracy of the "glmax" edge detector using an input image filtered with a Gaussian of 1 for a circle radius of 10, 30, and 60. It was generated using

```
gnuplot> plot 'c.30.gglout.sort' using 4:6 with lines,
'c.10.gglout.sort' using 4:6 with lines,
'c.60.gglout.sort' using 4:6 with lines
```

Improvement for Gaussian filtering with $\sigma = 2.0$ is about 4 : 1 for radii of 10, 30, and 60.

The statistics in Table C.3 indicate the differences in pixel error in the Hough transform accumulator projection using a Gaussian lowpass $\sigma = 1$, compared to that of $\sigma = 2$.

This resulted in a major compaction of the Hough accumulator peak. now we can re-evaluate Hough peak shapes using images filtered with $sigma = 2$.

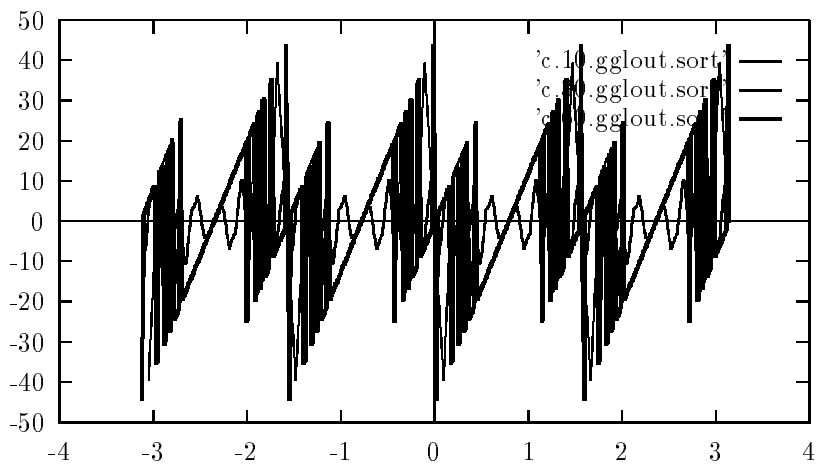


Figure C.5: Plot comparing radial edge angle error for the "glmax" edge detector operating from a Gaussian blur of 1 for circle radii of 10, 30, and 60.

statistic	radius = 10		radius = 30		radius = 60	
Gaussian σ	1	2	1	2	1	2
n	72	72	228	224	474	448
min	0	0	0	0	0	0
max	1.25626	0.814008	5.45835	1.44132	12.668	4.09466
mean	0.59825	0.452372	2.41006	0.741779	4.6941	1.9474
var	0.153853	0.072003	2.31143	0.14218	9.51329	1.54465
sd	0.392241	0.268334	1.52034	0.377068	3.08436	1.24284

Table C.3: Edge accuracy statistics for images filtered with a Gaussian $\sigma = 1, 2$

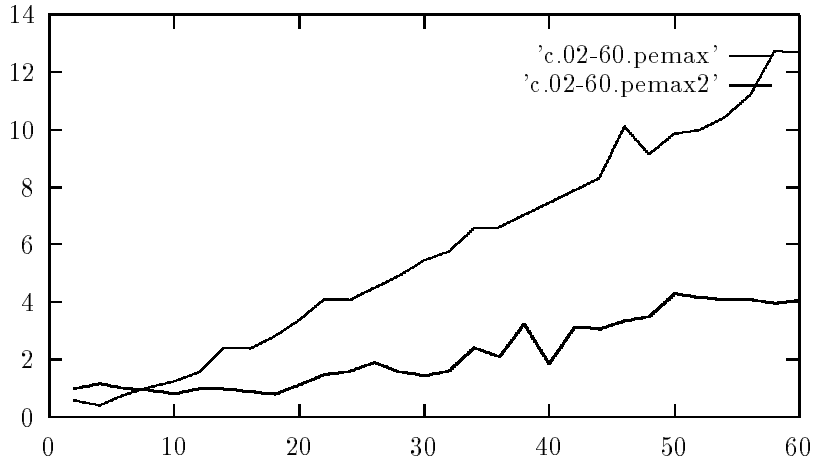


Figure C.6: Maximum pixel error wrt circle center for radii 2-60 and Gauss filter sigma = 1 and 2

Edge accuracy experiments: One of the next moves is to implement a Canny with extended directional shapes. We will need to add orthogonal pairs with same footprint for angle detection. Must add SNR detection when we go to multiple detectors. Maximum response is no longer sufficient.

Figure C.6 is a plot of the maximum error in projected pixel location wrt the circle center for images filtered with a Gaussian $\sigma = 1, 2$.

Figure C.7 is a plot of the angular error for radii of 10, 30, and 60 over $1/8$ of the circle with respect to the circle center for images filtered with a Gaussian $\sigma = 1$.

Figure C.8 is a plot of the angular error for radii of 10, 30, and 60 over $1/8$ of the circle with respect to the circle center for images filtered with a Gaussian $\sigma = 2$.

CAMS Idea: Why not operate directly on the gradient image rather than edge detecting? The idea is to create a polar map or histogram of the gradient pointing at point. Better yet, map a linear function that mimics a mass at each radius —flat out to edge, then Gaussian derivative. We could build this from a filtered image for the center region, plus the derivative at mass edge. It is not clear there is an advantage. I think it is clear that there is information about the mass center we are not using. We want to keep the Hough transform nonlinear summation of contributors versus a linear filter, which will be adversely affected by other structure overlapping the mass.

We want our function then to integrate data from hypothetical mass center, perform differentiation at the mass boundary, ignore anything past the boundary, and be relatively insensitive

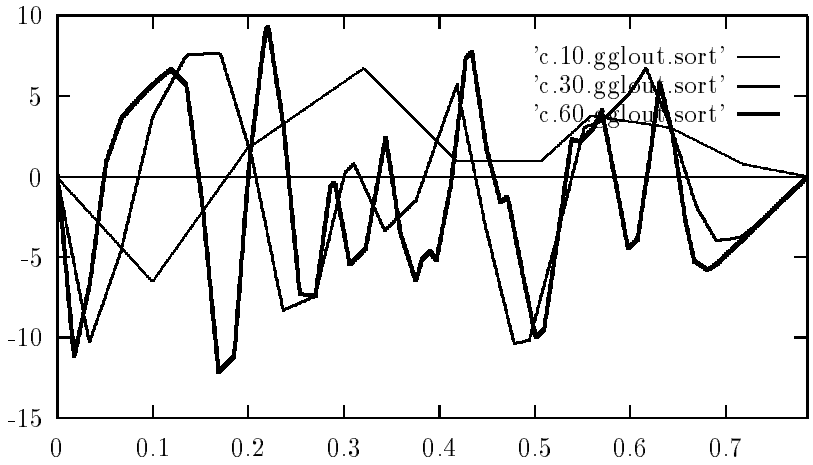


Figure C.7: angle error wrt circle center for radii 10,30,60 and Gauss filter sigma = 1

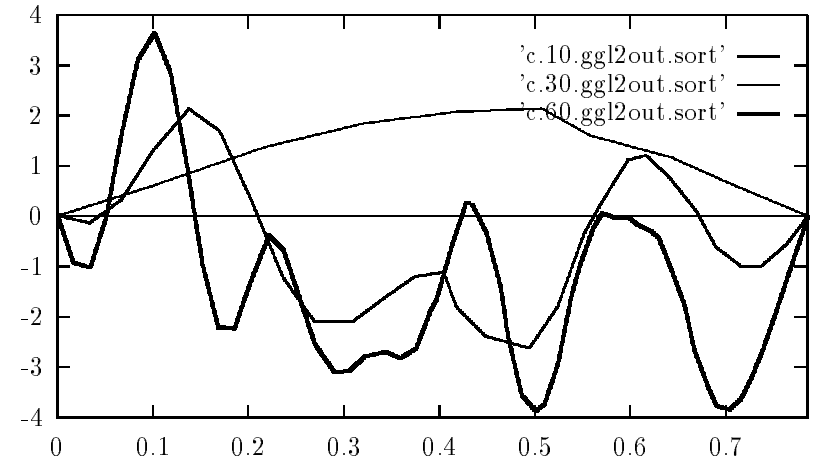


Figure C.8: angle error wrt circle center for radii 10,30,60 and Gauss filter sigma = 2

to structure overlapping the mass. In addition, of course, we want the best noise rejection possible. Furthermore, we want it to be parameterizable as with the Hough.

This is beginning to sound like the elliptic series stuff [34] I did. we've got a cross section. now, instead of a linear derivative calculation, can we come up with an approach similar to the Hough transform.

Once again, it would be wonderful to be able to use some sort of SNR measure on the data. this is basically determining confidence limits or measures on model parameters, where the model is, for instance, the edge detector.

Next, we might measure the response of said function over the radius (distance, parameter) range in question and determine the maximum response at each angle. These votes for maximum response could then be tabulated in a histogram, indicating the most likely parameter setting at that image point. So, for circles, evaluate the 1-d CL detection function at each radius out from the center, determine the maximum, and increment the appropriate r value for the point. repeat over 0 - 2 pi. Filter accumulator in r direction and pick maximum as best representation at x,y. One could also simply output the filtered r value as an indication of the likelihood of mass vs r at that point. Also note that points near in x,y should have similar behavior in r.

What we want ideally is an absolute measure of mass likelihood at each point. The approach outlined above can give the relative likelihood of a mass of radius r over r at each point by computing the statistics over the entire image. A more absolute measure would require statistics over the entire data set. It seems obvious that the measure should be normalized for the image, then these normalized values normalized over the data set.

What if we take a simple first derivative edge detector and operate on it with the radial histogramming indicated above? The same sort of response should occur, albeit with a lower SNR. apply this transform to the image, then analyze the result along r at each point. the r trace should be smooth, with low value out to the radius of the mass, then peak at this radius, then become noisy due to other structure. all we are doing is converting an xy gradient to a radial gradient about a point, and then detecting on this. BUT the one difference is instead of linear summing of the gradient around the circle, only positive contributions are considered. This avoids negation by other structure, as with the Hough. In addition, we can limit contribution to some angle dotted with the vector out from the circle center.

Now, what about non circular objects? - go to an ellipse representation.

Venkatesh edge processing: I've set up the Venkatesh [29] code to allow me to clip edge lengths as well, tshort <= edge lengths <= tlong.

Performance results, as measured by the gnuplots of statistics, are negative. The results appear uniformly worse when the edges are clipped for length in addition to the normal venk. length-strength linear cutoff. All venk. processing appears worse than clipping edge points based on the edge magnitude cdf.

I wonder if Venk. could be modified to use the CDF of the edge length and strength. Obviously it could, the question is whether or not it would result in any improvement.

The premise behind Venk. is that long, weak edges have perceptual significance. For circumscribed masses, the shape formed by the edge points is important.

When I compare the statistics by normalizing with the std dev, I get different results.

C.2 Unexplored Avenues

A number of areas involving mass features remain to be explored:

Wavelet multiscale representation and analysis: I use this term rather loosely. We're currently operating at a single resolution for our mass analysis. It seems obvious that a resolution hierarchy, and analysis thereon, would be an important improvement.

Fourier signature of mass outline: An analysis of mass shape at a range of resolutions (corresponding to image resolutions) would indicate what shapes and shape characteristics characterize masses. I'm unaware of the existence this information at this point.

Radiograph is a transmission image: hence a region can have several labels - each pixel is the sum of several tissue types. It would appear fruitful to explore masses as densities summed on top of normal tissue.

Laws - Gabor filter correspondence: It is apparent that Laws filters correspond exactly to certain Gabor filters. I am unaware of any writeup of this.

C.3 Feature Analysis Methods and Experiments

Feature Separation Plots: Our usual method for taking a quick but crude look at the quality of a given feature was to work up a "feature separation plot", such as those illustrated in Figure 4.9. Those plots illustrate the mean of the foreground and background values for a particular feature, with error bars to indicate a single standard deviation.

A potentially richer, but unexplored, way of looking at the same information would be to use the "boxplot" facilities of the S-Plus data analysis package, which also indicates quartiles, the complete range of the data, and outliers.

Tree-Growing Parameter Analysis: The usual idiom for determining the right tree growing parameters (size of terminal nodes, number of cross-validation trees, degree and proportion of weeding of the training data) was to grow a number of trees in which these parameters were varied, determine which parameters gave the best performance, and use those parameters for all future trees.

This method may not have been very robust, as the *shape* of the performance curve as a function of these parameters was never examined, and so it is possible that the peak was

narrow, or even accidental. In future work, it would be best to examine the parameter sets and their resultant performance as a whole, perhaps via N-dimensional scatter plots, in order to pick a parameter set as robustly optimal as possible.

Peak vs. Sum Features: As discussed in Section 4.3.11, we compute both "peak" and "sum" versions of the overlay of the features. Early in our experiments, these approaches seemed to have complimentary strengths and weaknesses, and so including them both was worthwhile. Once the quality of the features was improved, however, this issue was not revisited, even though it is possible that one may now be shadowed by the other.

Accordingly, it would be useful to run a comparison experiment in which only the peak and Laws features are used in one run, and only the sum and Laws features in the other, to determine whether they are both making roughly equal contributions. In parallel, the vfeatrank tools should be used to examine the current trees, with the same end in mind.

Hough vs. Laws Features: Similarly, it would be interesting to run a comparison experiment where only the peak and sum features are used in one run, and only the Laws features used in the other. Comparing "peak and sum" to the current "peak, sum, and Laws" results should indicate to what degree the current performance depends on being able to recognize and reject normal tissue.

Bare Minimum Performance Characterization: Lastly, it would be interesting to compare the current performance of the current system with the performance that would be observed if one or more of the best Hough features were thresholded at their single best threshold (both as judged by the separation plots). The latter performance is the bare minimum one would expect to achieve; whatever increased performance can be observed in the full system would thus be due to the more sophisticated classification provided by the BDT approach, separate from the issue of dense feature map representation.

Appendix D

Miscellaneous Issues

D.1 Classification Coding

The two main software tools used in the binary decision tree growing and use in this research are varbor (for growing trees) and vbotany (for applying them, to generate output probability or classification images). These are programs that were written at the University of New Mexico under contract from Sandia, and subsequently thoroughly modified. The source code is located at `~wpk/work/tmp/src/`. There are a number of code modifications and fixes that should be considered if this code is to be further used, particular in follow-on efforts to this project:

- There are a variety of versions of the BDT code, spawned by wanting to develop various features in parallel with each other. Now that the features are stable, the BDT versions should be consolidated; see Kegelmeyer's work notes of 9/20/95 and on. This would be usefully tackled before the following list of enhancements.
- Implement dynamic allocation of the array, in version 1.3.1, that allows only a subset of the input feature bands to be investigated. (The "weak link array".)
- It would be good to run tree growing timing tests, with and without masks, and with and without use of the "-fl" flag, to confirm that optimized compilation is happening properly.
- varbor should be upgraded to not require "-fl".
- `DataInput()` in varbor should be upgraded to complain if "-s" is provided as a flag but the mask file does not exist. Once implemented, confirm that it is doing the right thing, as this has been tricky in the past.
- varbor should be upgraded to have a "node specific probabilities" mode as well.
- varbor *and* vbotany should have a "-p 1/0" argument to select or deselect per-node probabilities.
- Debug the vbotany problem documents in Kegelmeyer's 9/4/92 notes.
- Work up better man pages for both, in particular documenting the use and implications of per-node RER computation.

D.2 Analysis Tools

There are a handful of small projects that would be useful but which were never sufficiently high priority during the life of the program:

- To ease examination of large trees, write a single script or alias using rer2ps to extract only the terminal node probabilities.
- Currently, lesions and clusters are reported by drawing a tight boundary around the perimeter of the finding. This is entirely appropriate objective performance evaluation of the computer detection schemes, but it may not be so appropriate for creating output images to be viewed by a human. Work by Krupinski and others[56] suggests that a tight boundary may actually interfere with the human perceptual process, and that an indicator circle, sized slightly larger than the longest object radii, might be a better match to human vision.

  This could be implemented as a conversion tool, "blob to circ" or something like that, which finds blobs, fits a circle to them (by noting center of mass and greatest radial extent), and then expands the circle by the desired number of pixels. A place to start would be to look at labeling software in the Mmach toolbox in Khoros 2.0. .
- A useful tool would be one that takes in a feature image and a BDT, and then paints the pixels of the output image with the *node label* of the node into which that pixel's feature vector fell. This would permit one to note image spatial coherence (or its lack) in the partitioning of the feature space. See Kegelmeyer's 6/21/95 notes for some further design considerations.
- The tool vnico_scr, which scores detected MC clusters according to Nico Karssemeijer's detection criteria, needs interior and man page documentation. It also needs to be considerably sped up; its current slowness is the main reason it could not be used as the performance criteria in the QSFS analysis discussed in Chapter B.
- The S-Plus data analysis package also has some facilities for growing BDTs. Though it lacks good tools for automatically pruning trees, and was designed for considerably smaller data sets, it would nonetheless be instructive to grow trees on some of the same feature data already analyzed. This would serve as an additional verification and validation step for our home-grown BDT codes.
- As part of the earlier, spiculated lesion efforts, Roger Tilley did a fair amount of work to try to wean the various CAMS Khoros codes off of their reliance on RIPPLE, a now-moribund image analysis package. These converted codes should be validated against the RIPPLE-based codes and, once it is confirmed that they generate the same results, moved in to replace them.

Distribution

1 Mark Allmen
Lockheed Martin Astronautics
MS S4360, PO Box 179
Denver, CO 80201

1 Kevin Woods
Dept. of Computer Science & Engineering
University of South Florida
4202 E. Fowler Ave., ENB 118
Tampa, FL 33620-5399

1 Wes Snyder
Dept. of Electrical & Computer Engineering
North Carolina State University
P.O. Box 7911
Raleigh, NC 27695-7911

1 Pete Santago
Medical Engineering
Medical Center Blvd
Winston-Salem, NC 27157-1022

1	MS	0188	Donna Chavez, for LDRD Office, 4523
1		9004	M. E. John, 8100
1		9214	L.M. Napolitano, 8117
20		9214	W.P. Kegelmeyer, Jr., 8117
15		9214	B.R. Groshong, 8117
10		9021	Technical Communications Department, 8815, for OSTI
1		9021	Technical Communications Department, 8815/Technical Library, MS 0899, 4414
4		0899	Technical Library, 4414
3		9018	Central Technical Files, 8940-2