Multi-Level Learning Approaches for Medical Image Understanding and Computer-aided Detection and Diagnosis

Yimo Tao

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
Computer Engineering

Jianhua Xuan
Yue Wang
Chang-Tien Lu

Keywords: medical image understanding, computer-aided detection and diagnosis, image segmentation, medical image annotation, object recognition

April 14, 2010
Arlington, Virginia

Multi-Level Learning Approaches for Medical Image Understanding and Computer-aided Detection and Diagnosis

Yimo Tao

Abstract

With the rapid development of computer and information technologies, medical imaging has become one of the major sources of information for therapy and research in medicine, biology and other fields. Along with the advancement of medical imaging techniques, computer-aided detection and diagnosis (CAD/CADx) has recently emerged to become one of the major research subjects within the area of diagnostic radiology and medical image analysis. This thesis presents two multi-level learning-based approaches for medical image understanding with applications of CAD/CADx. The so-called "multi-level learning strategy" relies on that supervised and unsupervised statistical learning techniques are utilized to hierarchically model and analyze the medical image content in a "bottom up" way.

As the first approach, a learning-based algorithm for automatic medical image classification based on sparse aggregation of learned local appearance cues is proposed [1]. The algorithm starts with a number of landmark detectors to collect local appearance cues throughout the image, which are subsequently verified by a group of learned sparse spatial configuration models. In most cases, a decision could already be made at this stage by simply aggregating the verified detections. For the remaining cases, an additional global appearance filtering step is employed to provide complementary information to make the final decision. This approach is evaluated on a large-scale chest radiograph view identification task and a multi-class radiograph annotation task, demonstrating its improved performance in comparison with other state-of-the-art algorithms. It also achieves high accuracy and

[1] This work is done during the author's internship with Siemens Healthcare Inc., USA.

robustness against images with severe diseases, imaging artifacts, occlusion, or missing data.

As the second approach, a learning-based approach for automatic segmentation of ill-defined and spiculated mammographic masses is presented. The algorithm starts with statistical modeling of exemplar-based image patches. Then, the segmentation problem is regarded as a pixel-wise labeling problem on the produced mass class-conditional probability image, where mass candidates and clutters are extracted. A multi-scale steerable ridge detection algorithm is further employed to detect spiculations. Finally, a graph-cuts technique is employed to unify all outputs from previous steps to generate the final segmentation mask. The proposed method specifically tackles the challenge of inclusion of mass margin and associated extension for segmentation, which is considered to be a very difficult task for many conventional methods.

Acknowledgements

First of all, I dedicate this thesis to my dear family for supporting me all these years through this long journey.

Thanks for the academic guidance and financial support provided by my advisors Dr. Shih-Chung Ben Lo and Dr. Jianhua Xuan. Especially, thank Dr. Lo's patience and kindness to guide me toward the maturity and sophistication. Thank Dr. Matthew Freedman, Dr. Yue Wang, for their support during my graduate study. Thank all my colleagues and friends in CBIL at Virginia Tech, and ISIS Center at Georgetown University.

Thanks for the inspiring and dedicated guidance from my mentor Dr. Xiang Sean Zhou in Siemens. Thank Dr. Bing Jian, Dr. Zhigang Peng, Dr. Le Lu, Dr. Jinbo Bi, Dr. Dewan Maneesh, and Dr. Yiqiang Zhan for their help and guidance during my internship at Siemens Healthcare.

Thank you to all my friends for your support throughout my educational endeavor.

Contents

List of Figures

List of Tables

Glossary

AMINDIST Average Minimum Distance

AOR Area Overlapping Ratio

BIRADS Breast Imaging Reporting and Data System

CAD Computer-aided Detection

CADx Computer-aided Diagnosis

CI Confidence Interval

HS Hausdorff Distance

ISLM Image Sub-patch Level Modeling

KNN K-Nearest Neighbor

LAT Lateral

LS Level Set

MAOR Margin Area Overlapping Ratio

MINDIST Minimum Euclidean Distance

ML Maximum Likelihood

MLAF Maximum Likelihood Function Analysis

PA-AP Posteroanterior/Anteroposterior

PACS Picture Archive and Communication System

PM Probability Map

SSC Sparse Spatial Configuration

SVM Support Vector Machine

WI William Index

1

Introduction

With Wilhelm Conrad Roentgen's dramatic and accidental discovery of "X-rays" in 1895, medical imaging and the discipline of diagnostic radiology began and developed quickly over the past century. Associated with the research and development of modern physics and computer technologies, medical imaging technology has become one of the major sources of information for therapy and for research in medicine, biology and in other fields. It has been widely used in the radiology department for diagnosis of diseases, for assessing acute injuries, for assessing disease severity or responses to a particular therapy, for guiding surgical interventions and health screening. Accompanied with the progress of medical imaging techniques, computer-aided detection and diagnosis (CAD/CADx) has recently emerged to become one of the major research subjects within the area of diagnostic radiology and medical image analysis.

In this section, we first briefly review the image modalities used in this study and the background of CAD/CADx. Then, we present the statement of the problems investigated in this work, specifically, medical image classification and mammographic mass segmentation.

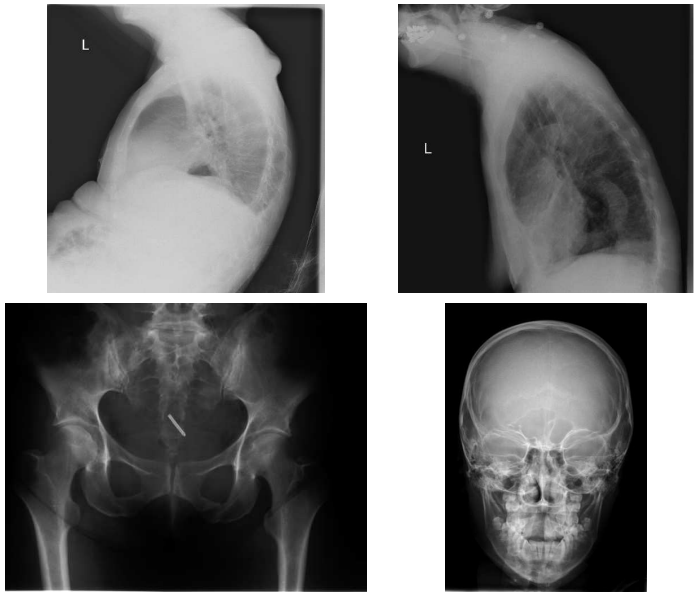


Figure 1.1: Examples of medical radiographs.

1.1 Image Modality

The imaging modality used in this study includes X-ray medical radiographs (e.g. 2D chest radiographs) for the image classification work, and X-ray mammography for the image segmentation work. We briefly reviewed the relevant background here.

(1) Conventional X-ray imaging produces a two-dimensional planar image revealing the internal structures of objects. Since projection images are produced, they have no depth information, and all opaque and semi-opaque structures in the beam are superimposed. X-ray images are formed by X-ray photons interacting with an X-ray detector. During transmission, some of the applied energy of X-ray is absorbed by the body, and some of the energy passes through the body to the detector. The difference between those photons (quanta) that interact and those that do not determines the contrast in the image. An image represents the transmission (or attenuation) distribution of the patient under study. Several example images used in the image classification work is shown in Fig. 1.1.

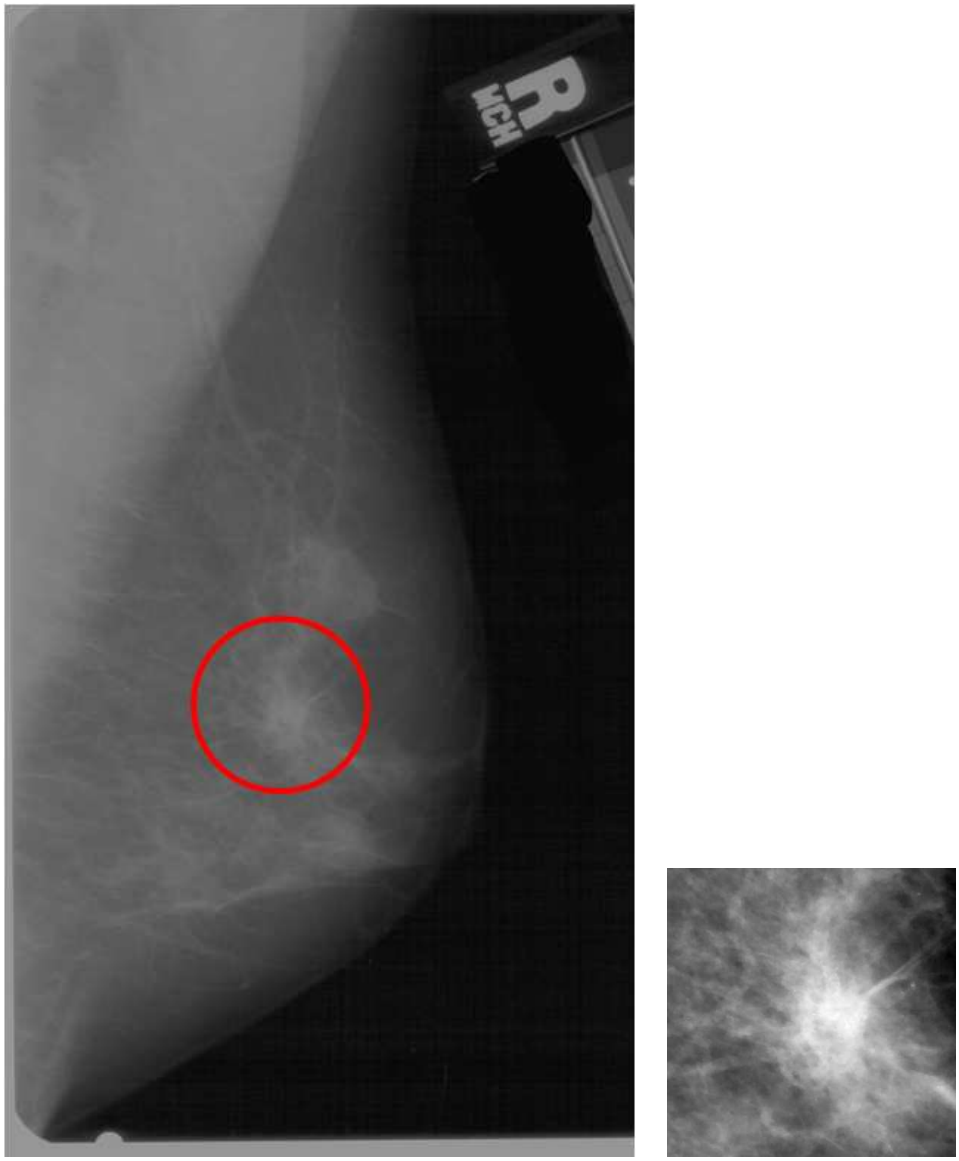
(2) X-ray Mammography is a technique optimized for imaging the breast. A lower

energy X-ray beam is used in order to enhance the image contrast. In addition, mammography film utilizes a single layer emulsion with one intensifying screen and has a higher spatial resolution than conventional radiography. X-ray Mammography has been widely used since the beginning of last century to detect the early signs of breast cancer, such as micro-calcification and small tumors. It has been shown that mammography has the ability to show non-palpable abnormalities in the breast with very high resolution (the equivalent of a 25μm pixel resolution [1]). Early detection of the subtle symptoms of breast cancer increases the probability to cure the cancer. To date, mammography is the best image modality for breast cancer detection and diagnosis with the good combination of sensitivity, specificity, short acquisition, and cost-effectiveness ratio. This leads to their widely usage in the large-scale screening programs throughout the world. An exemplar mammography image is shown in Fig. 1.2.

1.2 Computer-aided Detection and Diagnosis

Inspired from the early concept and preliminary studies of automated diagnosis or automated computer diagnosis [2] in the 1960s, large-scale and systematic research and development of various CAD schemes began from the early 1980s. Along with the development and progress, three organs/anatomies including chest [3], breast [4], and colon [5] have attracted the majority of CAD research interest from the early 2000s. This may be caused by that the detection of cancer in these anatomies has been or is being subjected to screening examinations.

Recently, CAD has become a part of the routine clinical work for detection of breast cancer on mammograms at many screening sites and hospitals in the United States. Prospective studies [6; 7; 8; 9] and results on large numbers of mammographic screenings have been reported. Regarding the effect of CAD on the detection rate of breast cancer, all of these studies indicated an increase in the detection rates of breast cancer with use of CAD. This seems to indicate that CAD/CADx has the potential usefulness to be applied in the real clinical environment. Commercial CAD product for mammogram



(a) (b)

Figure 1.2: Example images of mammography: (a) full size mammography with suspicious mass within red circle, (b) suspicious mass enlarged (window level adjusted to enhance contrast).

[10], lung CT [11] and colon CT [5] have been proved by FDA and are available for pre-clinical/clinical use. Meanwhile, there are many research actively conducted in the detection and differential diagnosis of many different types of abnormalities in medical images (such as brain, liver, and skeletal and vascular systems) obtained from various examinations and imaging modalities. With its current development, it is likely that in the future, CAD/CADx systems together with other image processing software will be integrated with the current Picture Archive Communication System (PACS) [12].

To conclude, computer-aided diagnosis has been integrated as a part of clinical work in the detection of breast cancer by use of mammograms. The whole CAD scheme is still in its preliminary development with potential for many applications of different pathology types from various modalities. In the future, it is likely that CAD schemes will be incorporated into PACS, and that they will be assembled as a package for detection of lesions and also for differential diagnosis. CAD has the potential to be employed as a useful tool for diagnostic examinations in daily clinical work.

1.3 Statement of the Problems

The intention of this research is to develop image content-based computational methods for investigating various medical imaging understanding problems for CAD/CADx applications. In this work, we cover two major topics in the medical image computing domain: (1) automatic image classification, and (2) automatic image segmentation. The developed computational algorithms could be integrated as components or extended features of a CAD/CADx systems (e.g. a mammography CAD system). These algorithms along with their master CAD system may have the potential to improve the efficiency and effectiveness for radiologists in organizing and processing the medical imaging data. Furthermore, the developed systems could assist radiologists in diagnosing diseases in the clinical environment.

Automatic Medical Image Classification: The amount of medical image data produced nowadays is constantly growing. Manually classifying these images is costly

and error-prone. 40% of all radiographs have missing or mislabeled DICOM header information regarding anatomy or imaging orientations. This calls for automatic classification algorithms to perform the task reliably and efficiently. Inspired by the recognition mechanism of human visual system, a learning-based algorithm for medical image understanding is investigated in this work. The main purpose of the system is to automatically recognize the projection view of the chest radiographs. Such algorithm could be integrated with PACS workstation to support optimized image display for improving the PACS radiography workflow. Furthermore, the method could be integrated as a post-processing module for CAD systems for auto-invocation in the background thread after recognizing the anatomic content and orientation of the image. We also demonstrate that the proposed algorithm is generalizable to annotate more image classes on other image modalities.

Mammographic Mass Segmentation: Segmentation of suspicious regions in medical images, e.g. lung nodules in the CT images or breast lesions in the mammograms, is arguably one of the most essential components for a CAD/CADx system. Accurate segmentation is important for the detection and classification in the post processing stage of the system. In this work, we specifically aims at segmenting masses with spiculation and ill-defined boundaries. A multi-level learning-based framework for segmentation mammographic mass is proposed in this work. The so-called "multi-level learning-based" approach lies in the fact that we utilize supervised/unsupervised learning techniques to hierarchically model the image content including the appearance and shape. The proposed method may contribute to mammographic CAD/CADx studies due to its ability to delineate the extended borders of ill-defined masses more robustly and accurately.

2

Robust Learning-based Medical Radiograph Classification

2.1 Introduction

The amount of medical image data produced nowadays is constantly growing, and a fully automatic image content annotation algorithm can significantly improve the image reading workflow, by automatic configuration/optimization of image display protocols, and by off-line invocation of image processing (e.g., denoising or organ segmentation) or computer aided detection (CAD) algorithms. However, such annotation algorithm must perform its tasks in a *very accurate* and *robust* manner, because even "occasional" mistakes can shatter users' confidence in the system, thus reducing its usability in the clinical settings. In the radiographic exam routine, chest radiograph comprise at least one-third of all diagnostic radiographic procedures. Chest radiograph provides sufficient pathological information about cardiac size, pneumonia-shadow, and mass-lesions, with low cost and high reproducibility. However, about 30%-40% of the projection and orientation information of images in the DICOM header are unknown or mislabeled in the picture archive and communication system (PACS) [13]. Given a large number of radiographs to review, the accumulated time and cost can be substantial for manually identifying the projection view and correcting the image orientation for each radiograph.

The goal of this study is to develop a *highly accurate* and *robust* algorithm for automatic annotation of medical radiographs based on the image data, correcting potential errors or missing tags in the DICOM header. Our first focus is to automatically recognize the projection view of chest radiographs into posteroanterior/anteroposterior (PA-AP) and lateral (LAT) views. Such classification could be exploited on a PACS workstation to support optimized image hanging-protocols [14]. Furthermore, if a chest X-ray CAD algorithm is available, it can be invoked automatically on the appropriate view(s), saving users' *manual effort* to invoke such an algorithm and the potential *idle time* while waiting for the CAD outputs. We also demonstrate the algorithm's capability of annotating other radiographs beyond chest X-ray images, in a three-class setting and a multi-class setting. In both cases, our algorithm significantly outperformed existing methods.

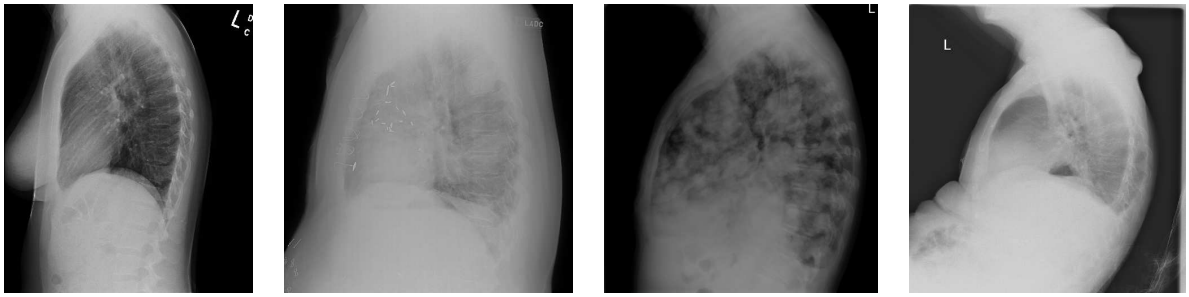
2.1.1 Related Works

A great challenge for automatic medical image annotation is the large visual variability across patients in medical images from the same anatomy category. The variability caused by individual body conditions, patient ages, and diseases or artifacts would fail many seemingly plausible heuristics or methods based on global or local image content descriptors. Fig. 2.1 and Fig. 2.2 show some examples of PA-AP and LAT chest radiographs. Because of obliquity, tilt, differences in projection, and the degree of lung inflation, the same class PA-AP and LAT images may present very high inter patient variability. Fig. 2.3 shows another example of images from the "pelvis" class with considerable visual variation caused by differences in contrast, field of view (FoV), diseases/implants, and imaging artifacts.

Most existing methods (e.g., [15], [16]) for automatic medical image annotation were based on different types of image content descriptors, separately or combined together with different classifiers. Müller et al. [17] proposed a method using weighted combinations of different global and local features to compute the similarity scores between the query image and the reference images in the training database. The annotation

(a) (b) (c) (d)

Figure 2.1: The PA-AP chest images of (a) normal patient, (b) and (c) patients with severe chest disease, and (d) an image with unexposed region on the boundary.



(a) (b) (c) (d)

Figure 2.2: The LAT chest images of (a) normal patient, (b) and (c) patients with severe chest disease, and (d) an image with body rotation.

Figure 2.3: Images from the IRMA/ImageCLEF2008 database with the IRMA code annotated as: acquisition modality "overview image"; body orientation "AP unspecified"; body part "pelvis"; biological system "musculoskeletal" . Note the very high appearance variability caused by artifacts, diseases/implants, and different FoVs.

strategy was based on the GNU Image Finding Tool image retrieval engine. Güld and Deserno [18] extracted pixel intensities from down-scaled images and other texture features as the image content descriptor. Different distance measures were computed and summed up in a weighted combination form as the final similarity measurement used by the nearest-neighbor decision rule (1NN). Deselaers and Ney [16] used a bag-of-features approach based on local image descriptors. The histograms generated using bags of local image features were classified using discriminative classifiers, such as support vector machine (SVM) or 1NN. Keysers et al. [19] used a nonlinear model considering local image deformations to compare images. The deformation measurement was then used to classify the image using 1NN. Tommasi et al. [20] extracted SIFT [21] features from downscaled images and used the similar bag-of-features approach [16]. A modified SVM integrating the bag-of-features and pixel intensity features was used for classification.

Regarding the task for recognizing the projection view of chest radiographs, Pieka and Huang [22] proposed a method using two projection profiles of images. Kao et al. [23] proposed a method using a linear discriminant classifier (LDA) with two features extracted from horizontal axis projection profile. Aimura et al. [24] proposed a method by computing the cross-correlation coefficient based similarity of an image with manually defined template images. Although high accuracy was reported, manually generation of those template images from a large training image database was time consuming and highly observer dependent. Lehman et al. [25] proposed a method using down-scaled image pixels with four distance measures along with K-nearest neighbor (KNN) classifier. Almost equal accuracy was reported when compared with the method of Aimura et al. [24] on their test set. Boone [13] developed a method using a neural network (NN) classifier working on down-sampled images. Recently, Luo [14] proposed a method containing two major steps including region of interest (ROI) extraction, and then classification by the combination of a Gaussian mixture model classifier and a NN classifier using features extracted from ROI. An accuracy of 98.2% was reported on a large test set of 3100 images. However, it was pointed out by the author that the performance of the method depended heavily on the accuracy of ROIs segmentation.

Inaccurate or inconsistent ROI segmentations would introduce confusing factors to the classification stage. All the aforementioned work regarded the chest view identification task as a two class classification problem, however, we included an additional OTHER class in this work. The reason is that in order to build a fully automatic system to be integrated into CAD/PACS for identification of PA-AP and LAT chest radiographs, the system must filter out radiographs containing anatomy contents other than chest. Our task, therefore, becomes a three-class classification problem, i.e., identifying images of PA-AP, LAT, and OTHER, where "OTHER" are radiographs of head, pelvis, hand, spine, etc.

In the more broad research field of object detection and recognition, many methods based on the use of local features have been proposed. The objects of interest were in many cases face, cars or people [26; 27; 28; 29; 30; 31]. Cristinacce and Cootes [26] combined boosted detector described by Viola and Jones [32] with the statistical shape model described by Dryden et al. [33]. Multiple hypotheses of each local feature were screened using the shape model and the winning hypothesis was determined for each feature. Agawal et al. [27] presented an object detection algorithm for detecting the side view of a car in a cluttered background. It used a "part-based representation" for the object. The global shape constraint was imposed through learning using the Sparse Network of Winnows architecture. Mohan et al. [30] proposed a full-body pedestrian detection scheme. They first used separate SVM classifiers to detect the body parts, such as heads, arms and legs. Then, a second SVM classifier integrating those detected parts was used to make the final detection decision. Leibe et al. [31] proposed a method for robust object detection based on learned codebook of local appearances. To integrate the global shape prior, an implicit shape model was learned to specify the locations, where the codebook entries might occur. Our work was inspired by many ideas from the non-medical domain, but with more suitable models of human anatomy, accommodating the fact that in the medical domain "*abnormality is the norm*".

2.1.2 Proposed Approach

We adopt a hybrid approach based on *robust aggregation of learned local appearance findings*, followed by the exemplar-based global appearance filtering. It combines the use of learned local-feature detectors, sparse and distributed shape prior constraints, and an exemplar-based global appearance check mechanism.

The robustness and advantage of the proposed approach lie in several aspects:

- The algorithm starts with detections of *semantic* local visual cue representations. By *semantic* we mean that these local cues are specified in an anatomically meaningful way instead of, for example, using sub-images on a regular grid. These detectors generates a concise codebook representation which, acting together, also normalizes transformations and geometrical variations in translation, scale, and rotation. Compared with the popular bag-of-features approaches (e.g., [16; 20]), spatial anatomical location is preserved in our model, and this is beneficial in at least two ways: first, each classification task is easier; second, shape priors can be learned and enforced.

- Our shape prior modeling module is based on a group of learned *sparse spatial configuration models*. This step enforces the spatial anatomical consistency of local findings. Because of its *sparse* nature, i.e., it is a collection of spatial relations among many small groups of landmarks, the shape prior constraint could still take effect even with many missed detections. Compared with methods using global shape representations (e.g., [34; 35]), our algorithm can be particularly effective on challenging cases with a large percentage of occlusion, or missing data, such as cases with large tumor or liquids in the lungs.

- Although the components outlined above worked very well (see Section 2.3.2: 98.47% of our method versus 96.18% from the literature), we found out that an additional step of global appearance check of low classification confidence cases, through an exemplar-based KNN filter, further improved the final performance (to 98.81%). It suggests that this additional *fusion* step provides complementary

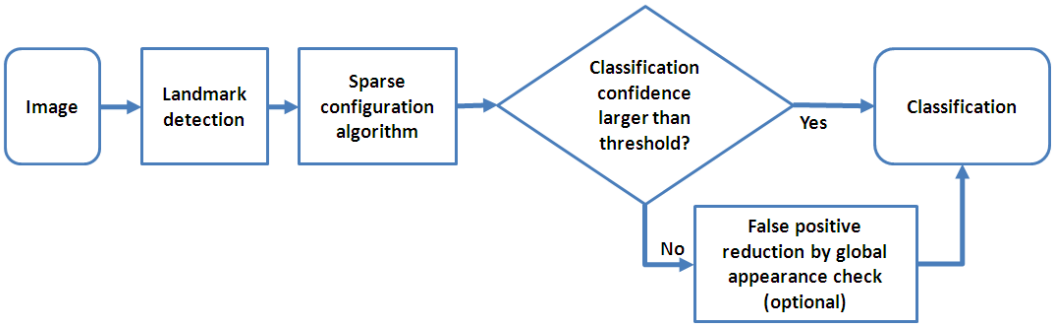


Figure 2.4: The overview of our approach for automatic medical image annotation.

global information that is not fully captured by the integrated local detections. It may seem that the percentage gains in discussion here are not large, however the improvement in users' experiences in the clinical environment is quite dramatic: for a busy clinic, the difference above represents *one error per months* versus *one error several days*.

- Our framework is designed to be generalizable, and we show that it can be applied to other image modalities and applications, such as anatomy/organ ROI prediction and optimized image visualization.

2.2 Methods

Fig. 2.4 shows the overview of the algorithm. Our algorithm is designed to first detect multiple focal anatomical structures within the medical image. This is achieved through a learning-by-example landmark detection algorithm that performs simultaneous feature selection and classification at several scales. A second step is performed to eliminate inconsistent findings through a robust *sparse spatial configuration* (SSC) algorithm, by which consistent and reliable local detections will be retained while outliers will be removed. Finally, a reasoning module assessing the fitered findings, i.e., remaining landmarks, is used to determine the final content/orientation of the image. Depending on the classification task, a post-filtering component using the exemplar-based global appearance check for cases with low classification confidence may also be

included to reduce false positive (FP) identifications.

2.2.1 Landmark Detection

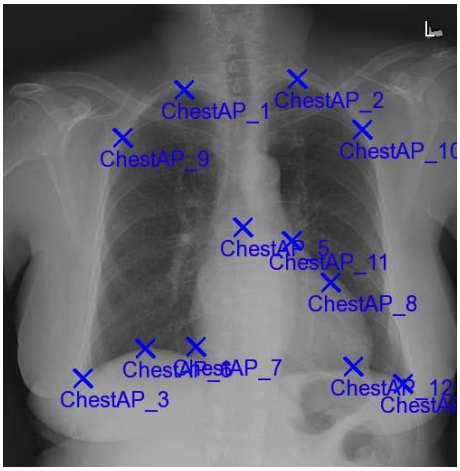


Figure 2.5: Landmark annotation/detectoin examples (shown as crosses) in a PA-AP chest image. The labels under the landmarks specify different anatomic position within the PA-AP chest image.

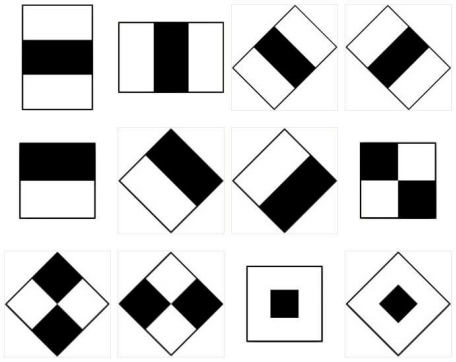


Figure 2.6: Illustration of some 2D Haar features: the sum of pixels which lie within the black rectangle(s) are subtracted from the sum of pixels in the while rectangle(s). These features could be compupted efficiently using integral images.

The landmark detection module in this work was inspired by the work of Viola and Jones [32], but modified to detect points (e.g., the carina of trachea) instead of a fixed region of interest (e.g., a face). We use an adaptive coarse-to-fine implementation in the scale space, and allow for flexible handling of the *effective scale* of anatomical context for each landmark.

Firstly, we collect a number of training images along with a group of anatomic landmarks (as shown in Fig. 2.5) annotated by radiologists according to the literature [36]. Then, to train a landmark detector, image sub-patches centered at the annotated positions are cropped and collected as positive training samples; then, an over-complete set of extended Haar features are computed within patches as shown in Fig. 2.6. The sizes of patches for different landmark detectors vary from 13×13 to 25×25, and they are determined independently based on positions within original (or down-scaled) images.

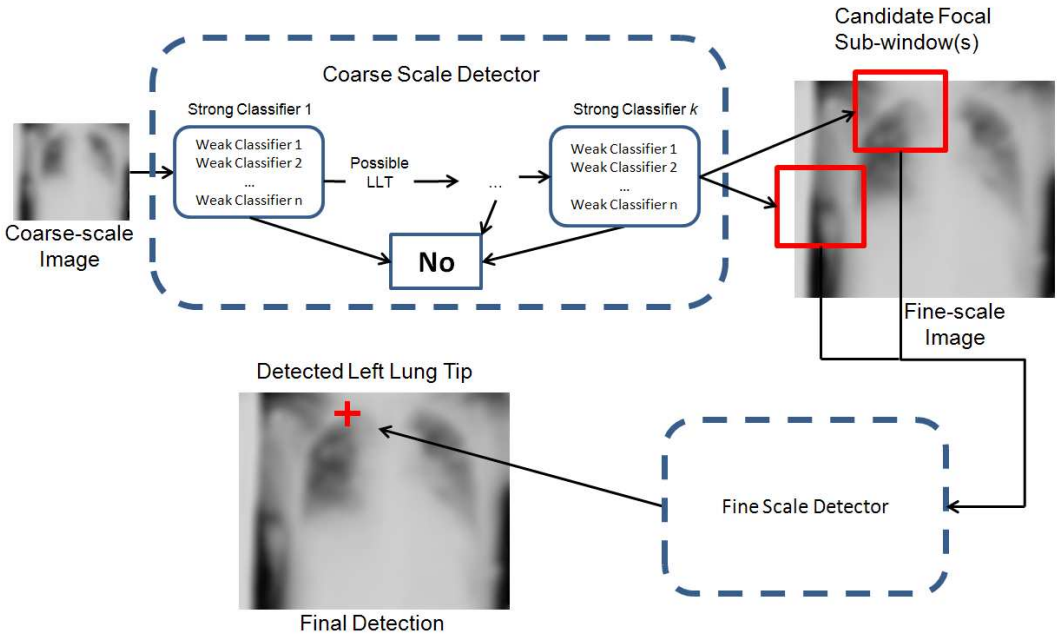


Figure 2.7: Illustration of landmark detection. To detect the position of left lung tip (LLT), the corse scale LLT detector first scans on the re-sampled coarse-scale image and detects candidate focal sub-window(s) (shown as rectangles), where responses are large than a pre-defined threshold. Next, the fine scale LLT detector runs within the sub-window(s) in the image, and detect one final position with the highest response (shown as cross).

The sub-patches are allowed to extend beyond the image border to certain extent, in which case the part of the patch falling outside the image is padded with zeroes. The extended portion for the patch should be smaller than 50% of the total size in order to be used as an effective positive sample. This ensures sufficient and effective context information to be extracted. The size of a landmark detector is thus determined manually to ensure that at least 80% of annotated landmarks are utilized as effective training samples.

Regarding the classifier, we employ the boosted cascade method [32; 37] for simultaneous feature selection and classification, which guarantees great run time efficiency. For each level cascade of a detector, the training criterion is to achieve high recall (99.9%) and moderate false positive deduction rate (0.1%) on the validation training set. For the first level cascade, negative training samples are collected by randomly cropping sub-patches in images belonging to negative classes. For subsequent levels, the

negative training samples are obtained by collecting false positives using the partially trained cascade detector. The whole training process stops when the ratio between the total number of collected negative samples and that of positive ones is less than 0.1. A typical trained cascade detector has 6 to 8 levels. Different landmark detectors are trained independently across several scales within down-scaled images. For the image annotation application, two scales are adopted to balance the computational time and detection accuracy.

During the testing/detection phase, the trained landmark detectors at the coarsest scale are used first to scan on the whole image to determine the candidate position(s), where the response(s)/detection score(s) are larger than a predefined threshold. After that, the landmark detectors at finer scales are scrutinized at previously determined position(s) to locate the local structures more accurately and, thus, to obtain a single detection with the highest response. An illustration of the landmark detection procedure is shown in Fig. 2.7. The final outputs of a landmark detector are the horizontal and vertical (x-y) coordinates in the image along with a response/detection score. Joint detection of multiple landmarks also proves beneficial (see Zhan et al.[38] for detail).

The landmark detection procedure is invariant to image translation since detectors are scanned on the whole image. The rotation and scale robustness is naturally achieved through classifier training procedure. To further improve the trained detectors' robustness against image rotation, we added to the training set an additional number of rotation modified training images ($\pm 15^\circ$ with the gap of 5°).

2.2.2 Sparse Spatial Configuration Algorithm

Knowing that the possible locations of anatomical landmarks are rather limited, we aim to exploit this geometric property to eliminate the possible redundant and erroneous detections from the first step. This geometric property is represented by a spatial constellation model among detected landmarks. The evaluation of consistency between a landmark and the model can be determined by the spatial relationship between the landmark and other landmarks, i.e., how consistent the landmark (as a candidate) is

according to other landmarks. In this work, we propose a local filtering algorithm (Alg. 1) to sequentially remove false detections until no outliers exist.

Algorithm 1 Sparse spatial configuration algorithm

```
for each candidate x_i do
  for each voting group X_v generated from the combinations of X\x_i do
    Compute the vote of x_i from X_v
  end for
  Sort all the votes received by landmark x_i. (The sorted array is defined by γ_{x_i}).
end for
repeat
  x̌ = arg min_{x_i} max γ_{x_i}
  if max γ_{x̌} < V_threshold then
    Remove x̌ and all votes involved with x̌.
  end if
until No more candidate are removed
```

In general, our reasoning strategy "peels away" erroneous detections in a sequential manner. Each candidate x receives a set of votes from other candidates. We denote the ith detected landmark as x_i, which is a two dimensional variable with values corresponding to the detected x-y coordinates in the image. Each candidate x_i receives a set of likelihood scores generated from its spatial relationship with voting groups formed by other landmarks. The likelihood score received by candidate x_i from jth voting group V_{ij} is modeled as multi-variant Gaussian as following:

$$\eta_{ij}(x_i|V_{ij}) = \frac{1}{2\pi\left|\Sigma_{ij}\right|^{1/2}} e^{-(x_i-\nu_{ij})^T\Sigma^{-1}(x_i-\nu_{ij})} \tag{2.1}$$

where Σ_{ij} is the estimated covariance matrix, and the prediction $\nu_{ij} = q_{ij}(x_i|V_{ij})$. Here $q_{ij}(\bullet)$ is defined as:

$$q_{ij}(x_i|V_{ij}) = A_{ij} \times [V_{ij}] \tag{2.2}$$

where A_{ij} is the transformation matrix learned by linear regression from a training set, and $[G_{ij}]$ is the array formed by the x-y coordinates of landmarks from the voting group V_{ij}. A high likelihood score of $\eta_{ij}(x_i|V_{ij})$ means that the candidate x_i is likely to be a

good local feature detection according to its spatial relations with other landmarks in V_j.

Here we briefly illustare the transofmation matrix learning procedure. For simplification, we assume that three groups of annotated landmark sequence (collected from n images) are given by as:

$$\text{Landmark1} : (x_{11}, y_{11}), ..., (x_{1n}, y_{1n})$$

$$\text{Landmark2} : (x_{21}, y_{21}), ..., (x_{2n}, y_{2n})$$

$$\text{Landmark3} : (x_{31}, y_{31}), ..., (x_{3n}, y_{3n})$$

where x_{ij} and y_{ij} stands for the horizontal and vertical positions of the jth sample for the landmark candidate i. Assuming that landmark 2 and landmark 3 form one voting group for landmark 1, we try to predict the horizontal position of landmark 1. Mathematically, their spatial relationship is modeld using linear regression as:

$$x_1 = f(x_2, y_2, x_3, y_3) + \varepsilon = \beta_0 + \beta_1 x_2 + \beta_2 y_2 + \beta_3 x_3 + \beta_4 y_3 + \varepsilon \tag{2.3}$$

where ε is the random noise of $\varepsilon \sim N(0, \delta^2)$ independent of x_2, y_2, x_3, and y_3. Formulating training set variables in a vector format as:

$$B = [\beta_0, \beta_1, \beta_2, \beta_3]$$

$$X_1 = [x_{11}, \ldots, x_{1n}]$$

$$X_v = \begin{bmatrix} 1, x_{21}, y_{21}, x_{31}, y_{31} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 1, x_{2n}, y_{2n}, x_{3n}, y_{3n} \end{bmatrix}$$

It is straightforward to show that the transformation matrix could be obtained using maximum liklihood estimation as:

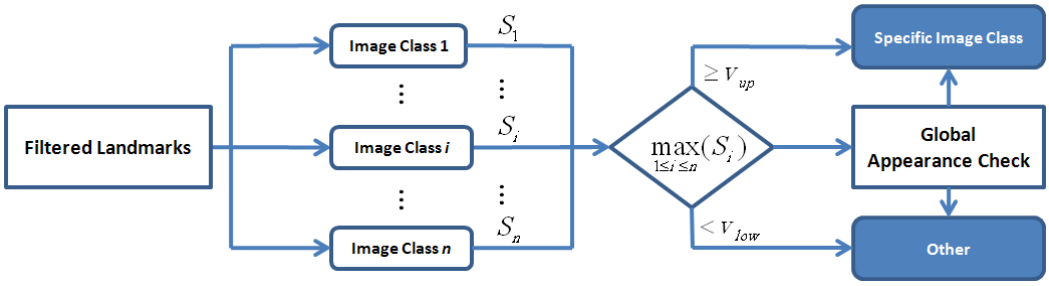


Figure 2.8: The diagram of classification logic.

$$B = (X_v^T X_v)^{-1} X_1 X_v \tag{2.4}$$

For each landmark x_i, multiple voting groups are generated by the combination of different landmarks from the landmark set excluding x_i (denoted as $X \backslash x_i$). The size of each voting group is designed to be small, so that the resutlant sparse nature guarantees that the shape prior constraint could still take effect even with many missed detections, thus leading its robustness in handling challenging cases such as those with a large percentage of occlusion, or missing data. In this work, we set the sizes of the voting groups to be 1 to 3. Therefore, for an image class with 10 landmark detectors, there are a total of $\sum_{i=1}^{3} C_9^i = 129$ voting groups for each landmark candidate.

The reasoning strategy (Alg. 1) then iteratively determines whether to remove the current "worst" candidate, which is the one with the smallest maximum vote score compared with other candidates. The algorithm will remove the "worst" candidate if its vote score is smaller than a predefined vote threshold $V_{threshold}$. This process will continue until no landmark outlier exists. The bad candidates can be effectively removed by this strategy.

2.2.3 Classification Logic

For an image belonging to certain anatomy class, we assume that there would be a sufficient number of landmarks associated with that class (i.e. true positive detections) to be detected. Therefore, the classification logic (as shown in Fig. 2.8) is determined as

following: the number of landmarks for each image class is divided by the total number of detectors for that class, representing the final classification score (denoted as S_i for the ith image class) . In case that equal classification scores are obtained between several classes, the class with maximum average landmark detection scores are chosen as the final class. Depending on the classification task, a FP reduction module based on the global appearance check may also be used for those images with low classification confidence, i.e. with scores within defined range of (T_{low}, T_{up}). A large portion of these images come from the OTHER class. They have a small number of local detections belonging to the candidate image class, yet their spatial configuration is strong enough to pass the SSC stage. Since the mechanism of local detection integration from previous steps could not provide sufficient discriminative information for classification, we try to integrate a post-filtering component based on the global appearance check to make the final decision. In our experiment for PA-AP/LAT/OTHER separation task, only about 6% of cases go through this stage. To meet the requirement for real-time recognition, an efficient exemplar-based global appearance check method is adopted. Specifically, we use pixel intensities from 16×16 down-scaled image as the feature vector along with 1NN, which uses the Euclidean distance as the similarity measurement. With the fused complementary global appearance information, the FP reduction module could effectively remove FP identified images from the OTHER class, thus leading to the overall performance improvement of the final system (see Section 2.3.2).

2.3 Experiments and Results

2.3.1 Datasets

In this work, we ran our method on four subtasks: PA-AP/LAT chest image view identification task with and without OTHER class, and the multi-class medical image annotation task with and without OTHER class. For the chest image identification task, we used a large-scale in house database, and for the multi-class radiograph annotation

task, we used the IRMA/ImageCLEF2008 database [1].

1) The in-house image database were collected from daily imaging routine from radiology departments in hospitals; it contains a total of 10859 radiographs including 5859 chest radiographs and 5000 other radiographs from a variety of other anatomy classes. The chest images covered a large variety of chest exams and diseases, representing image characteristics from real world PACS. It included upright position from normal patients, supine position of critically ill patient in intensive or emergency care units, and a small number of pediatric images with both suspended and lying position. In addition, the image quality ranged from decent contrast and well-set gray level to low contrast images, or images with severe pathology or implants. In this work, the OTHER class excluded radiographs of certain anatomies (e.g., radiographs of finger and shank) with large difference in the image width to height ratio in comparison to chest radiographs. These radiographs could be easily differentiated from chest radiographs using heuristic rules based on the image width to height ratio. We randomly selected 500 PA-AP, 500 LAT, and 500 OTHER images for training landmark detectors. These training images were also used as the exemplar image database for the post-filtering component. The remaining 9359 images were used as the testing set.

2) For the multi-class medical radiograph annotation task, we used the ImageCLEF2008 database. All images from this database had been labeled with a detailed code that specified acquisition modality, body orientation, body part and biological system. It contained more than 10,000 images from total 197 unique classes. This database had been used for as a part of the ImageCLEF workshop [39] for the medical image annotation task. The distribution of different classes in this database was not uniform. For example, the chest radiographs comprised about 37% of the total images. And the top nine classes comprised about 54% of the total images. In this work, we selected a subset of images (the top nine classes with the most number of images) from this database, including PA-AP chest, LAT chest, PA-AP left hand, PA-AP cranium, PA-AP lumbar spine, PA-AP pelvis, LAT lumbar spine, PA-AP cervical spine, and

[1] http://imageclef.org/2008/medaat

LAT left to right cranium. The remaining images were regarded as one OTHER class. For the PA-AP and LAT chest images, we directly used the detectors trained using the in-house database. 50 PA-AP and 50 LAT chest testing images were randomly selected from the testing set of previous task. For the remaining 7 classes, we randomly selected 200 images for each class. 150 images were used for training landmark detectors, and the remaining 50 images were used for testing. For the OTHER class, we randomly selected 2000 training and 2000 testing images each. All images in the in-house database and the IRMA database were down-scaled to have the longest edge of 512 pixels while preserving the aspect ratio.

2.3.2 Classification Performance

For the chest radiograph annotation task, we compared our method with three other methods described by Boone et al. [13], Lehmann et al. [25], and Kao et al. [23]. For method proposed by Boone et al. [13], we down-sampled the image to the resolution of 16×16 pixels and constructed a five hidden nodes NN. For method proposed by Lehmann et al. [25], a five nearest neighbor (5-NN) classifier using 32×32 down-sampled image with the correlation coefficient distance measurement was used. The same landmark detector training database was used as the reference database for the 5-NN classifier. For method proposed by Kao et al. [23], we found that the projection profile derived features described in the literature were sensitive to the orientation of anatomy and noise in the image. Directly using the smoothed projection profile as the feature along with the LDA classifier provided better performance. Therefore, we used this improved method as our comparison.

For the multi-class radiograph annotation task, we compared our method with the in-house implemented bag-of-features method proposed by Deselaers and Ney [16] (named as PatchBOW+SVM) and the method proposed by Tommasi et al. [20] (named as SIFTBOW+SVM). Regarding PatchBOW+SVM, we used the bag-of-features approach based on randomly cropped image sub-patches. The generated bag-of-features histogram for each image had 2000 bins, which were then classified using a SVM clas-

sifier with a linear kernel. Regarding SIFTBOW+SVM, we implemented the same modified version of SIFT (modSIFT) descriptor and used the same parameters for extracting bag-of-features as those used by Tommasi et al. [20]. We combined the 32×32 pixel intensity features and the modSIFT bag-of-features as the final feature vector, and we used a SVM classifier with a linear kernel for classification. We also tested the benchmark performance of directly using 32×32 pixel intensity from the down-sampled image as the feature vector along with a SVM classifier. To evaluate different methods' performance, we used optimized classification precision. Note that it is possible to carry out the ROC analysis for our method, which may require the parameters of different components in our method (e.g., detection response thresholds, global appearance check thresholds, and etc.) to be modified sequentially.

Table 2.1: PA-AP/LAT/OTHER chest radiographs annotation performance.

ref	Method	PA-AP/LAT	PA-AP/LAT/ OTHER
-	Our method	-	98.81%
-	Our method without FP reduction	99.98%	98.47%
[25]	Lehmann's method	99.04%	96.18%
[13]	Boone's method	98.24%	-
[23]	Improved Projection Profile method	97.60%	-

Table 2.2: Multi-class radiographs annotation performance.

ref	Method	Mutli-class without OTHER	Multi-class with OTHER
-	Our method	99.33%	98.81%
-	Subimage pixel intensity + SVM	97.33%	89.00%
[16]	PatchBOW + SVM	96.89%	94.71%
[20]	SIFTBOW + SVM	98.89%	95.86%

Table 2.1 and Table 2.2 show the recognition rates of our method, along with other methods. It can be seen that our system obtained an almost perfect performance on the PA-AP/LAT separation task. The only one failed case was a pediatric PA-AP

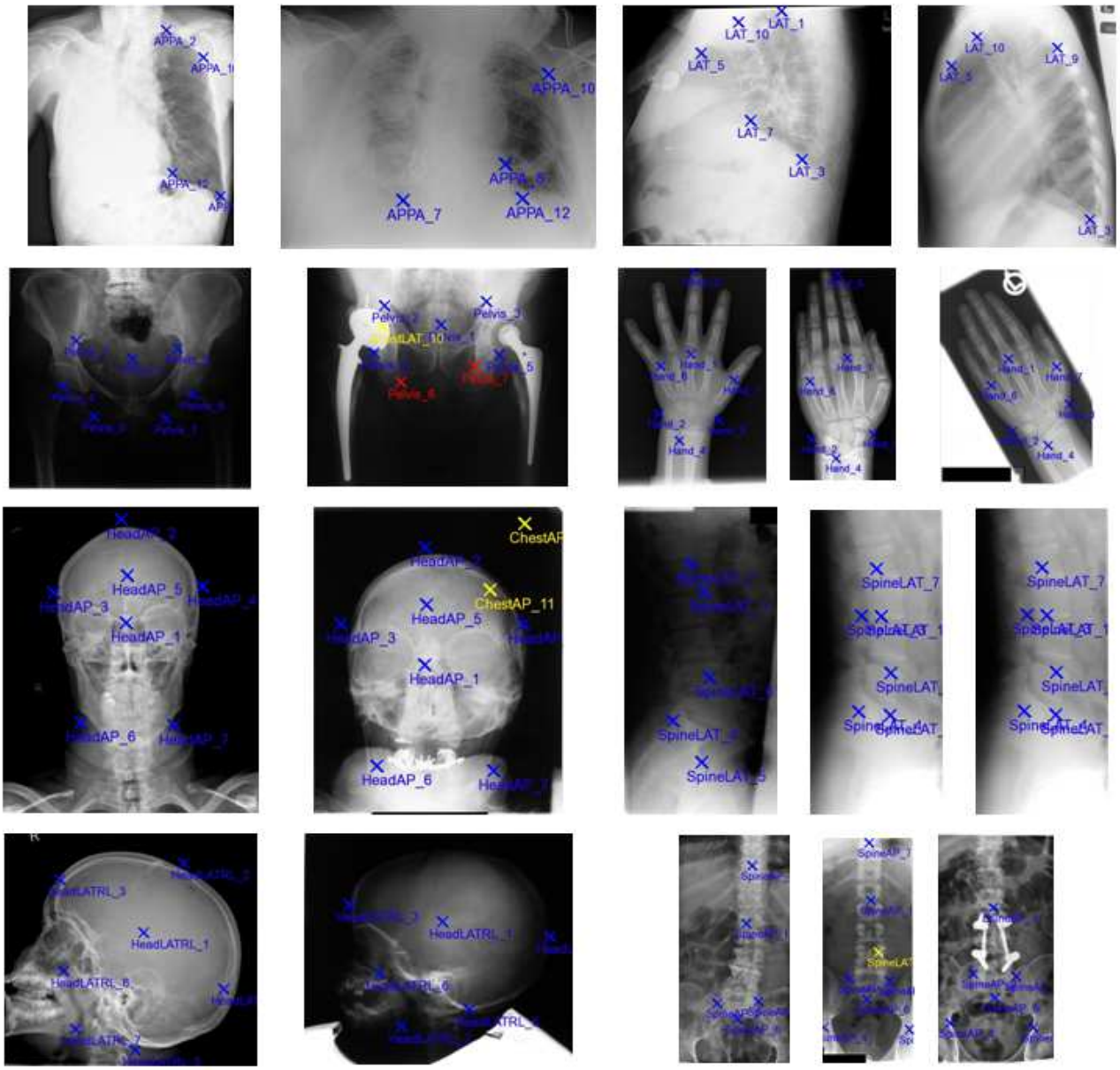


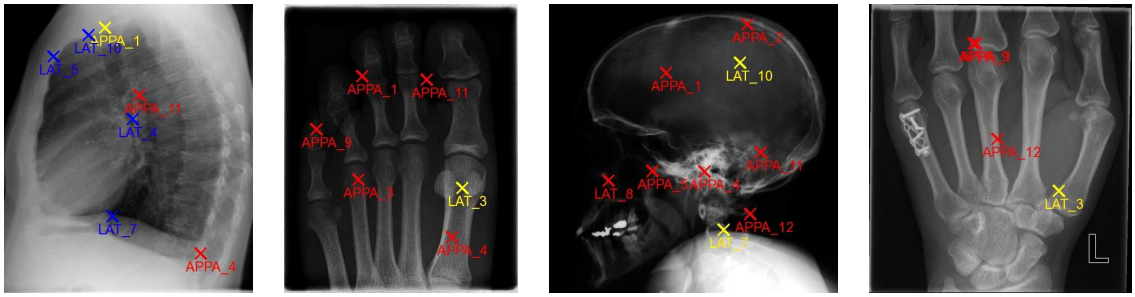
Figure 2.9: Examples of the detected landmarks on different images.

image. Our method also performed the best on the other three tasks. Fig. 2.9 shows the classification result along with the detected landmarks for different classes. It can be seen that our method could robustly recognize challenging cases under the influence of artifacts or diseases. The learned landmark detectors are robust to medium degree of rotation ($\pm 15°$) and scale variance based on the testing result.

2.3.3 Intermediate Results

1) Landmark Detection: We provide here the intermediate results of landmark detectors' performance. In this work, we used 11 landmarks and 12 landmarks for PA-AP and LAT chest images. As for the multi-class radiograph annotation task, we used 7-9 landmarks for other image classes. To test the landmark detectors' performance, we annotated 100 PA-AP and 100 LAT images separately. Since the landmark detectors run on the Gaussian smoothed images, the detected position could deviate from the ground truth position to certain degree, which is allowable for our image annotation application. We determine the detected landmark as true positive detection when the distance between the detected position and the annotated ground truth position is smaller than 30 pixels. Note that the detection performance can be traded off against computational time. Currently in order to achieve real-time performance, we accepted an average sensitivity for the 23 chest landmark detectors at 86.91% ($\pm 9.29\%$), which was good enough to support the aforementioned overall system performance.

2) SSC: For the PA-AP/LAT separation task on the 200 images where ground truth landmarks were annotated, 55 out of 356 false positive landmark detections were filtered by the SSC algorithm, while the true positive detections were unaffected. In addition, the algorithm removed 921 and 475 false positive detections for the PA-AP/LAT/OTHER task and the multi-class task with OTHER class. Fig. 9 shows that the result of the voting algorithm in reducing false positive detections on non-chest image classes. We can conclude that the voting strategy has improved the specificity of the landmark detectors.



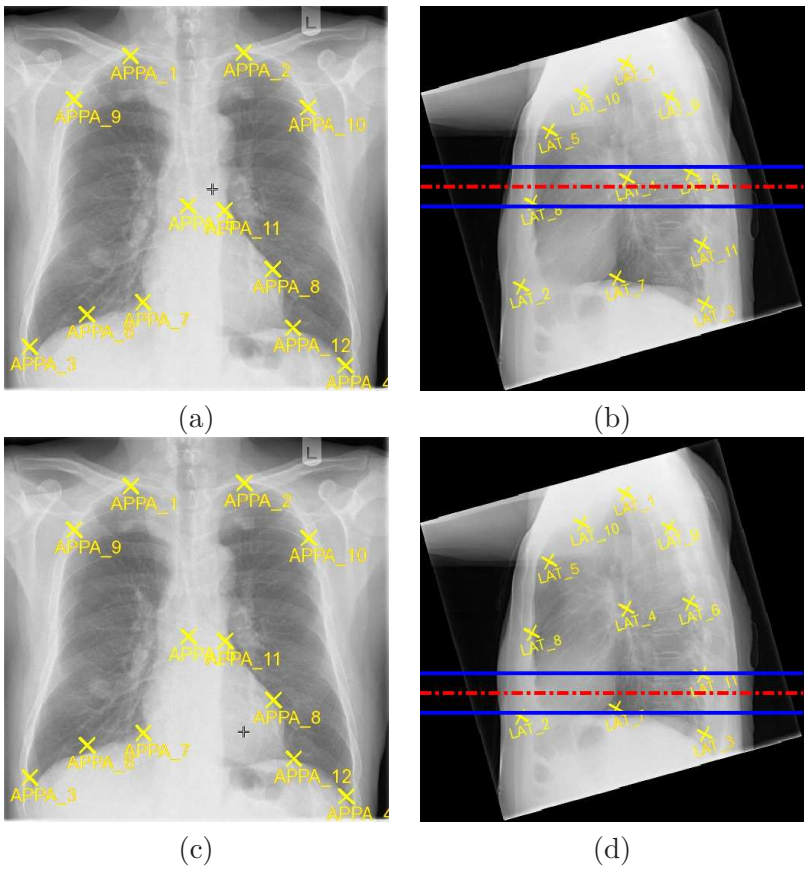
(a) (b) (c) (d)

Figure 2.10: The SSC algorithm performance on different image classes (better viewed in color): (a) LAT chest, (b) foot, (c) cranium, and (d) hand. The blue colored crosses are true positive landmark detections; the yellow colored ones are false positive detections; and the red colored ones are detections filtered by the SSC algorithm. APPA and LAT label under the detected landmarks specify that detections are from PA-AP chest detectors or LAT chest detectors.

2.3.4 System Extension

The proposed algorithm framework is generalizable, and it could be applied to other image modalities and extended for other image parsing applications beyond radiograph annotation. In this work, we further exploit the by-products of the algorithm, more specifically, the detected/filtered landmarks on chest images, for an optimized image visualization application.

The patient may stoop sometimes when taking the LAT chest image, and this may cause the image to present certain degrees of tilt as shown in Fig. 2.2 (c) and (d) compared with Fig. 2.2 (a), which is an image with a standard upright body position. We could explore the detected landmarks on the LAT chest image to perform registration with images with standard upright position. This allows robust online orientation correction for chest radiographs for optimized image visualization. Compared with the system of Luo and Luo [40], where the orientation correction is restricted to images with rotations of only 90°, 180°, 270°, our system is more flexible and robust in estimating degrees of tilt. The orientation corrected images have the potential to help radiologists to view the LAT chest radiograph more conveniently. In addition, with the detected landmarks on PA-AP and LAT chest images, a synchronized view mechanism could



(a) (b)

(c) (d)

Figure 2.11: Optimized image visualization. (a) the PA-AP chest image with pinpointed position shown as cross, (b) and the orientation corrected LAT chest image with the estimated corresponding position/range shown within the blue band. Figure (c) and (d) show that when the pinpointed position on the PA-AP chest image moves, the corresponding position on the LAT chest image moves accordingly.

be modeled to help radiologists to find the corresponding positions between the two images. More specifically, we first compute the relative position of the pinpointed cross on the PA-AP image (as shown in Fig. 2.11(a)) within the frontal lung ROI defined in the PA-AP image. Assuming the relative position is roughly unchanged in the lateral lung ROI in the LAT image, we could estimate the corresponding position/range on the LAT image. Fig. 2.11 shows an example of the synchronized view feature of our visualization workstation system. The blue band area on the LAT image corresponds to the position pinpointed by the cross in the PA-AP image. This optimized display feature has the potential to help radiologists to locate and scrutinize the corresponding findings on PA-AP and LAT chest images simultaneously.

2.4 Discussions

To conclude, a robust and gerneralizable algorithm framework has been proposed for medical radiograph annotation. Extensive experiments were conducted to validate the system performance both in terms of accuracy and speed. Such systems can dramatically improve radiology workflow and save valuable time and cost in the clinical environment.

The computation complexity of our method comes mainly from the landmark detection procedure. Although the cascade classification framework guarantees the run time efficiency for each landmark detector, the computation cost increases linearly with the number of specified landmarks and the number of image classes. To meet the requirement for online recognition, one possible solution is to use a few coarse resolution landmark detectors to first select several candidate image classes, and then the multi-scale detectors from the selected candidate classes are used to determine the final class. In this work, this scheme was adopted for the multi-class medical radiograph annotation task. According to our experiments, based on a multi-thread implementation of the algorithm, the entire process time on average for an image on Intel (R) Xeon (R) 1.86GHz with 3.00GB RAM was about 1s for the PA-AP/LAT/OTHER classifica-

tion task and 2s for the multi-class task. This satisfies our requirement for the online recognition procedure. Due to the generality and scalability of our approach, it has the potential to be extended in several directions, for example, annotation of more classes of radiograph images, extensions to other imaging modalities, and 2D/3D ROI detection tasks, e.g., topogram for CT scan automation [41]. In addition, the same algorithm framework has been extended for 3D medical image applications, e.g., coronary artery detection and tracing in CT image [42].

3

Multi-level Learning-based Segmentation of Ill-defined and Spiculated Mammographic Masses

3.1 Introduction

Clinically, the shape and margin characteristics of a mammographic mass are regarded by radiologists as the two most important features for breast cancer diagnosis. While malignant masses usually have ill-defined margins and irregular shapes and/or spiculation, benign masses usually have well-defined margin. More specifically, a spiculated mass consists of a central mass body with "extensions", hence the resulting stellate shape. From medical image analysis perspectives, the image features (e.g., texture and intensity patterns) associated with the extended regions and ill-defined borders are important information for the mass analysis. Therefore, a segmentation algorithm, that is tailored to delineate the mass body and periphery including its irregular details or spiculations, is technically desirable.

Many algorithms have been proposed for automatic mammographic mass segmentation. Region growing is one of the most often used methods. Pohlman et al.[43] developed an adaptive region growing method, whose pixel aggregation criterion was determined from calculations made in 5×5 windows surrounding the pixel of interest. Kupinski and Giger [44] proposed two region growing approaches based on the radial gradient index and a probabilistic model. Kinnard et al. [45] extended the probabilistic model based method by further analyzing the steepest change of cost functions. Their method was found to be able to further include some of the ill-defined mass boundaries. Besides region growing, many other techniques have also been investigated. Te Brake and Karssemeijer [46] proposed a discrete dynamic contour model, which began as a set of vertices connected by edges (initial contour) and grew the subject according to internal and external forces. Li et al. [47] developed a method that employed k-means classification to categorize pixels as belonging to the region of interest (ROI) or background. Sahiner et. al [48] proposed a method consisting of segmentation initialization by pixel-intensity based clustering followed by region growing to improve boundary shape; then the initial result was further augmented by an active contour algorithm for shape refinement. Recently, Shi et al. [49] proposed a level set approach for mass segmentation. The segmented masks were found to be able to improve the performance for discriminating benign and malignant masses in comparison with their previous work. It is worth mentioning that the energy function defined in the level set approach was based on the image gradient information, which would be noisy for masses with ill-defined margins and spiculations. This might reduce the method's reliability and robustness in segmenting ill-defined masses. Domínguez and Nandi [50] proposed a segmentation method based on contour tracing using dynamic programming. Although the method worked well for masses with circumscribed margins, it had difficulties in segmenting masses with less distinct contours. A review of mammographic mass segmentation algorithms could be found in the literature [4].

While encouraging results have been obtained in the aforementioned works, fully automatic segmentation of mammographic masses remains challenging, especially, for

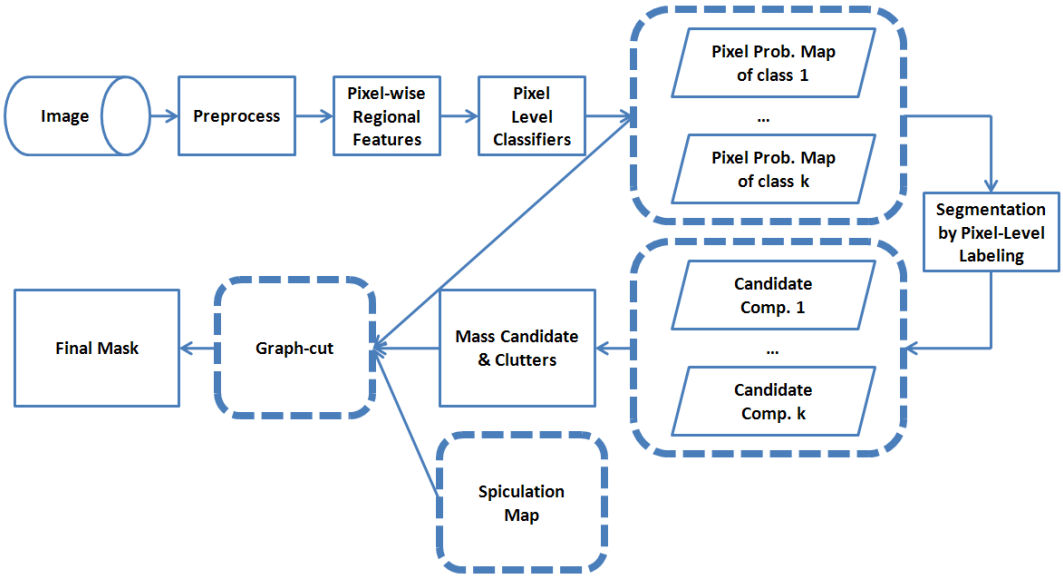


Figure 3.1: Flow chart of the multi-level segmentation approach.

those masses with ill-defined margins and spiculations. In this work, a multi-level learning-based segmentation (MLS) technique is proposed for segmenting various types of masses. The approach specifically tackles the challenge of mass margins and associated extensions, while minimizing the possibility of over-segmentation.

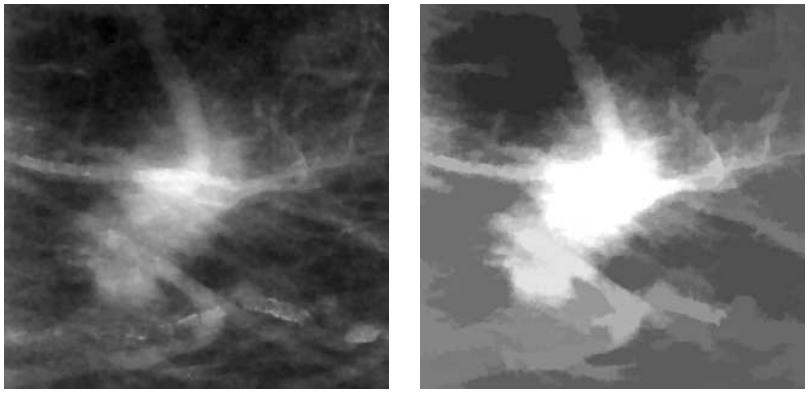
3.2 Method

Our segmentation method is composed of several major components (shown as rounded rectangles) in Fig. 3.1: (1) pixel-wise soft labeling/segmentation, (2) object-scale mass and clutters detection, (3) spiculation detection, (4) segmentation integration by graph-uts. The pixel-level mass and non-mass class labeling and segmentation works as follows: given a region of interest (ROI), the system would label each pixel with its probability associated with mass configuration statistical model trained through supervised learning. A pixel-level probability map (PM) for the whole image is thus obtained. After this, the object-level image classification/detection module takes the PM along with prior information (i.e., shape, size, and spatial distribution) to identify regions of mass and clutters. In order to include irregular shapes and spiculation, a spiculation

detection module based on a multi-scale ridge detection algorithm is then employed to produce a binary image of spicules (names as "spiculation map"). Finally, the graph cuts [51] algorithm is used to integrate the PM and all the object-level findings (i.e., mass region, clutters, and spicules maps) to produce the final segmentation.

3.2.1 Preprocess

We apply the morphological smoothing to remove small intensity peaks and minima in the image. The smoothing operation consists of two steps including image opening by reconstruction, and its complementary operation of image closing by reconstruction. Image opening by reconstruction consists of the application of erosion with a small structuring element, followed by reconstruction of the original image using the eroded version as the maker image. Image closing by reconstruction is carried in a similar way; however, image dilation is used instead of erosion. Image details are suppressed in the processed image as shown in Fig. 3.2. In this work, reconstruction with disk of radius equal to 3 pixels as the structuring element was applied to all images before pixel-level feature extraction step.



(a) (b)

Figure 3.2: Results of morphological smoothing: (a) The original ROI images, (b) The image after morphological smoothing

3.2.2 Pixel-Level Soft Segmentation

3.2.2.1 Pixel-wise Features

In the first step, the segmentation of mammographic mass is addressed with a pixel-level labeling approach, where the probabilistic distribution of the mass are modeled through the image sub patch level. In the large ROI, the system computes a collection of regional features at each pixel position $p(p \subset \Re^2)$ from a sub region of interest (sROI) of size 11 $\times$11 pixel centered at p. In this study, a total of 30 dimensional features (denoted as x_p) including intensity, texture and shape were calculated on the preprocessed image for each scanned sROI. These features are briefly described below.

Gray Level Co-occurrence Matrix (GLCM) [52] is widely used for analyzing texture of 2D image. The co-occurrence matrix stores the co-occurrence frequencies of the pairs of gray levels, which are configured by a distance d and orientation o. The GLCM is constructed in four directions ($\theta = 0°, 45°, 90°, 135°$) and with the pixel distance of 1 pixel. We then extract eight features from the constructed GLCM, including energy, entropy, correlation, inverse difference moment, inertia, cluster shade, cluster prominence, and Haralick correlation [52].

2D Local Binary Patterns (LBP) [53] can be viewed as an intensity- and rotation-invariant generalization of the GLCM. In each pixel position, the LBP operator compares the current pixel's intensity differences with its p neighbors on the circular periphery of radius r. The difference values are binarized and concatenated to form a binary pattern code. The information recorded is intensity-invariant, and rotation-invariance is approximated as p becomes larger. The LBP operator is moved and evaluated throughout the whole texture image, and the binary pattern codes at each pixel position are summed up to form the final LBP histogram.

Wavelets are another important and commonly used feature descriptor for texture analysis, due to their effectiveness in capturing localized spatial and frequency information and multi-resolution characteristics. Here, we extract mean intensities in the decomposed four bands using 2D Haar wavelet.

Vesselness and **Blobness**, computed based on the eigen-analysis of hessian matrix, have also been employed for vascular or blob-like structure detection or enhancement. We implement a 2D multi-scale version of Blobness and Vesselness feature extraction module for only handling both bright objects, which corresponding to mass and line structures. Note that the Wavelets, Vesselness and Blobness depend on their own scales of spatial supporting settings, and the actual neighborhood may be larger or smaller than the size of 11×11. We also extract a group of first order **gray-level statistics features**, including minimum, maximum, mean, standard deviation, skewness and kurtosis.

Since the texture features are computed within a small sROI window at every pixel position, directly texture feature calculation on the image with original intensity values is computationally intensive and sensitive to noise. Besides, the constructed GLCM may be very sparse, causing numerical problem in feature computation. Therefore, we preprocess images using the multi-level thresholding Otsu method [54] to adaptively merge together image regions with similar gray levels. The resulting image is represented by individual texture primitives coded by a smaller gray-level domain. All texture-based features are extracted from this preprocessed image.

3.2.2.2 Segmentation by Pixel-Level Labeling

Based on our ground truth annotation maps of mass and non-mass, feature vectors are split into positives and negatives and fed into an off-line classifier learning process. For the pixel-level classification problem, the size of training samples (as scanned region of size 11×11) can be really large (greater than 100,000). This requires the choosing classifier with good scalability. We choose linear discriminant analysis (LDA) along with Gaussian Mixture Models (GMM) as our classifier, i.e., GMM is used to learn the distribution of the classes in the LDA projected subspace. For each of the binary mass and non-mass class, LDA is first exploited to further project the extracted feature vector x_0 into a vector of x in a lower dimension of d. Then, the Expectation-Minimization (EM) algorithm with multiple random initialization trials is used to determine the

parameter set $\theta = \{\alpha_i, \mu_i, \Sigma_i\}_{i=1}^{k}$ in GMM consisting of k-Gaussian distributions. Here, α_i is the prior, μ_i is the mean, and Σ_i is the covariance matrix of the ith distribution. Here, we briefly describe the basic steps of the EM algorithm. Assume the distribution of random variables $X \in R^d$ is a mixture of k Gaussians given by:

$$f(x|\theta) = \sum_{i=1}^{k} \alpha_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\} \tag{3.1}$$

Given a set of observations $x_1, x_2, ..., x_n$, the maximum liklihood estimation of θ is

$$\theta_{ML} = \arg\max_{\theta} f(x_1, ..., x_n|\theta) \tag{3.2}$$

The EM algorithm iteratively converges from the initial estimation of θ to θ_{ML} according to two steps:

1) Expectation step:

$$w_{ti} = \frac{\alpha_i f(x_t|\mu_i, \Sigma_i)}{\sum_{j=1}^{k} \alpha_i f(x_t|\mu_j, \Sigma_j)} \quad i = 1, ..., k \quad t = 1, ..., n \tag{3.3}$$

$$\hat{\alpha}_i \leftarrow \frac{1}{n} \sum_{t=1}^{n} w_{ti} \tag{3.4}$$

$$\hat{\mu}_i \leftarrow (\sum_{t=1}^{n} w_{ti} x_t)/(\sum_{t=1}^{n} w_{ti}) \tag{3.5}$$

$$\hat{\Sigma}_i \leftarrow (\sum_{t=1}^{n} w_{ti}(x_t - \hat{\mu}_i)(x_t - \hat{\mu}_i)^T)/(\sum_{t=1}^{n} w_{ti}) \tag{3.6}$$

The first step computes the probability that an instance x_t is generated by the ith normal distribution. Then, the maximization step uses the probabilities to update the current value of θ. The process is repeated until the log-likelihood is increased by less than a predefined threshold from one iteration to the next.

As there are many different types of tissues inside the mammogram, such as vessel, glandular tissues, etc., the single-layer LDA classifier may have many false positives

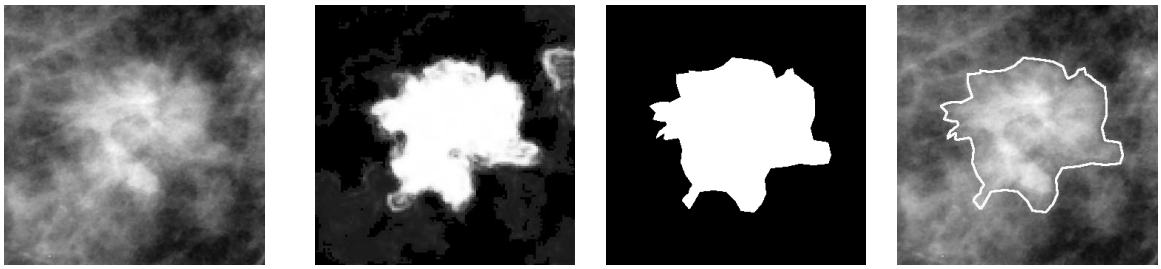
originating from this multi-tissue background. To reduce these mass false positives, a multi-phase classification approach is adopted. It starts with the positive class output PM from single phase, and treats it as a new image. This output image contains for each pixel a probability that it belongs to the structure to be enhanced (mass). Next, another round of pixel-level feature extraction (and selection) and LDA-GMM training process is conducted using both the original image and the output image from the previous phase(s). All these intensity-texture-shape features, in the joint image and PM domain, are used to train a new classifier. This process can be iterated many times, as a simplified "Auto-Context" [55]. The rationale behind this approach is that the structure to be enhanced will be more distinctive in the (intermediate) enhanced image than in the original image. Therefore, adding features from these weighted images will result in potentially more discriminative features between the positive regions and the spurious responses from the previous phase(s). Illustrative examples of generated PMs are shown in Fig. 3.3. It can be clearly seen the generated PM is able to enhance the mass tissue structure, meanwhile suppress the false responses of breast tissues.

3.2.3 Object-level Labeling and Detection

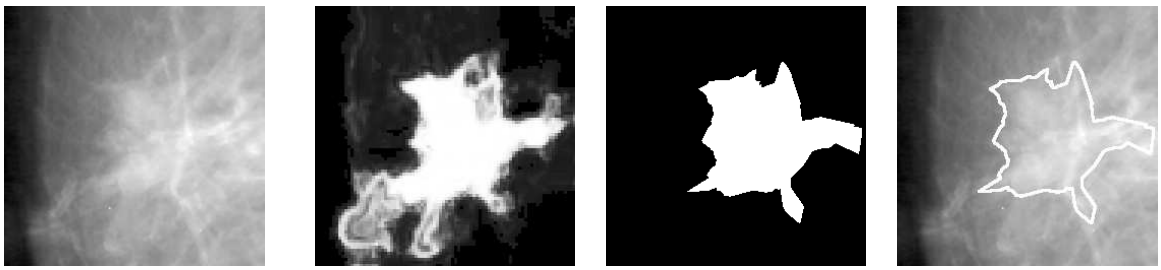
At this stage, the technical objective is to determine the mass region and other clutters in the intermediate PM output (denoted as $prob(p)$). To suppress spurious responses, the multi-scale Blobness filtering, with a larger smoothing kernel (than the pixel-level feature extraction step) was used to capture the mass shape. It was applied on each pixel to obtain a Blobness likelihood map denoted as $blm(p)$. Then, a shape-prior refined probability map $sprob(p)$ was obtained as:

$$sprob(p) = blm(p) \times prob(p) \tag{3.7}$$

Otsu thresholding method [54] was used for discrete quantization of $sprob(p)$ to obtain the potential mass pixels with $sprob(p) > V_{threshold}$. Connected component analysis was then employed to obtain several disjointed regions (DR)s $C_1\{p\}, C_2\{p\}, ..., C_n\{p\}$



(a.1) (b.1) (c.1) (d.1)



(a.2) (b.2) (c.2) (d.2)

Figure 3.3: Results of pixel-level classification: (a.1) & (a.2) the original ROI images, (b.1) & (b.2) the mass probability maps, (c.1) & (c.2) the segmentation ground truth masks generated from multiple radiologists (see section 3.3.1 for detail), (d.1) & (d.2) the ground truth contours (with three radiologists' consensus) superimposed on the originalimages.

on the binarized image. For each DR, we compute a fitness score to determine its likelihood of being a mass as:

$$F_i = \sum\nolimits_{p \in c_i\{p\}} G(p|\mu, \Sigma) \times sprob(p), i =, 1, 2, ..., n \tag{3.8}$$

where $G(p|\mu, \Sigma)$ is a 2D multivariate normal distribution representing a spatial prior that the mass is near the center of ROI. The DR possessing the maximum fitness score is selected as the mass candidate. The remaining DRs are regarded as clutters.

3.2.4 Spiculation Detection

It is known that malignant masses in mammograms are often characterized by a radial pattern of linear structures (i.e. spicules) and irregular boundaries [56; 57]. Detecting spiculation is thus essential for further inclusion of mass margins. In this task, a steerable ridge detection approach [58] was employed, and it was further generalized into a multi-scale analysis framework to detect the presence of spiculations. An Mth order ($M = 4$) ridge detectors constructed by Gaussian kernels and their derivatives was employed as:

$$h(x, y) = \sum_{k=1}^{M} \sum_{i=0}^{k} \alpha_{k,i} \underbrace{\left(\frac{\partial^{k-i}}{\partial x^{k-i}} \frac{\partial^{i}}{\partial y^{i}} g(x, y) \right)}_{g_{k,i}(x,y)} \tag{3.9}$$

where $g(x, y)$ is a 2D Gaussian function, and $\alpha_{k,i}$ represents the weight coefficient for the kernel of $g_{k,i}(x, y)$. The ridge detection procedure is formulated as a rotated matched filtering. It involves the computation of inner-products with the shifted and rotated versions of the 2D template $h(x, y)$ at every pixel in the image of $I(\mathrm{x}) : \mathrm{x} = (x, y)$. A high magnitude of the inner-product indicates the presence of the feature and the angle of the corresponding template gives the orientation. Mathematically, the estimation algorithm is formulated as:

$$\theta^*(\mathrm{x}) = \arg\max_{\theta}(I(\mathrm{x}) \cdot h_\theta(\mathrm{x})) \tag{3.10}$$

$$r^*(\mathrm{x}) = I(\mathrm{x}) \cdot h_\theta(\mathrm{x}) \tag{3.11}$$

where $r^*(\mathrm{x})$ is the magnitude of the feature; $\theta^*(\mathrm{x})$ is its orientation at the pixel position x; $h_\theta(\mathrm{x})$ is the rotated template with a degree of θ, and $\cdot$ is the inner product operator. Due to the property of *steerable filters* defined in (3.9), we could cut down on the computational load in (3.10) and (3.11). Specifically, the inner product of a signal $I(\mathrm{x})$ with an $h_\theta(\mathrm{x})$ can be expressed as:

$$I(\mathrm{x}) \cdot h_\theta(\mathrm{x}) = \sum_{k=1}^{M} \sum_{i=0}^{k} b_{k,i}(\theta) I_{k,i}(\mathrm{x}) \tag{3.12}$$

where $b_{k,i}(\theta)$ are orientation-dependent weights computed using trigonometric polynomials of θ (see the literature[58] for details), and the functions $I_{k,i}(\mathrm{x})$ are the inner products of the signal $I(\mathrm{x})$ with un-rotated kernels of $g_{k,i}(\mathrm{x})$:

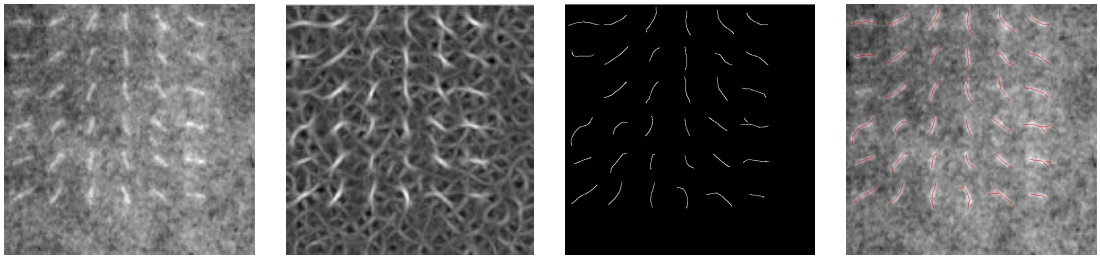
$$I_{k,i}(\mathrm{x}) = I(\mathrm{x}) \cdot g_{k,i}(\mathrm{x}) \tag{3.13}$$

To determine the optimal detector $h(x, y)$ in (3.9), which was equivalent to searching for the optimal weight combinations of $\alpha_{k,i}$, a functional optimization method following the Canny-like criteria was used [58]. The ridge detector of fourth order used in the experiments was determined to be:

$$\begin{aligned} h(x,y) = \quad & -0.392\delta g_{yy} + 0.113\delta g_{xx} + 0.034\delta^3 g_{yyyy} \\ & -0.184\delta^3 g_{xxyy} + 0.025 g_{xxxx} \end{aligned} \tag{3.14}$$

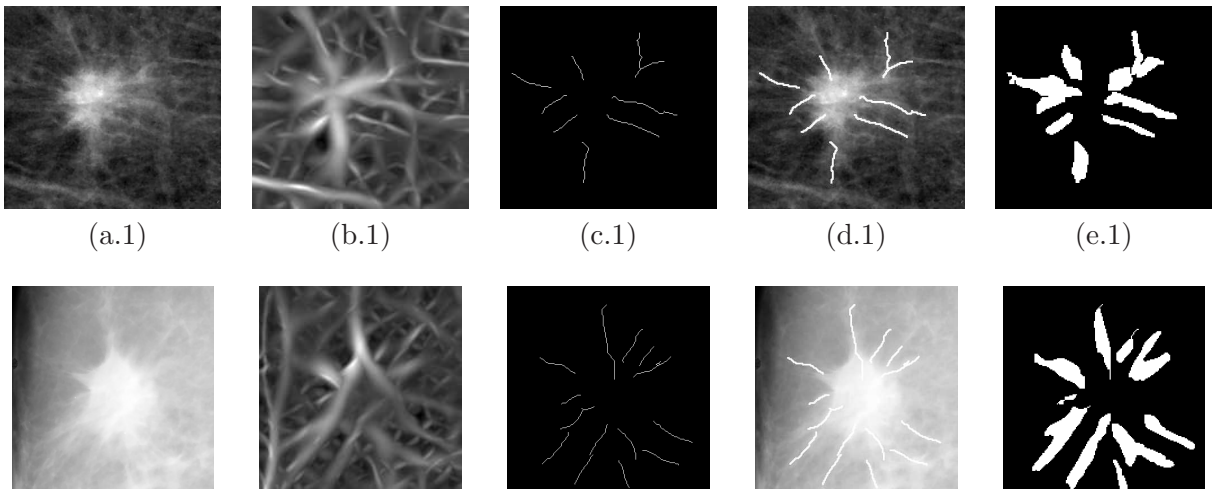
where $g_{xx} = \partial^2 g/\partial x^2$, $g_{yy} = \partial^2 g/\partial y^2$, $g_{xxxx} = \partial^4 g/\partial x^4$, $g_{yyyy} = \partial^4 g/\partial y^4$, $g_{xxyy} = \partial^4 g/\partial x^2 \partial y^2$

Fig. 3.4 shows an example of the ridge detection result on a synthetic image. In order to obtain spicules at different widths, the scale of a ridge detector was progressively increased. An estimate of the ridge scale at each pixel was obtained by normalizing



(a) (b) (c) (d)

Figure 3.4: Example of spiculation detection on synthetic image: (a) synthetic linear structures superimposed on a fatty mammographic background, (b) the ridge detector response, (c) the extracted backbone of the spiculation, and (d) the superimposed backbone on the original image.



(a.1) (b.1) (c.1) (d.1) (e.1)

(a.2) (b.2) (c.2) (d.2) (e.2)

Figure 3.5: Results of multi-scale spiculation detection: (a.1) & (a.2) the original ROI, (b.1) & (b.2) the multi-scale line-strength image, (c.1) & (c.2) the detected spicules, (d.1) & (d.2) the detected spicules superimposed on the ROI, and (e.1) & (e.2) the final spiculation map.

the line-strength obtained at each scale, and choosing the scale that gives the largest response. The line-strength and orientation along with the chosen scale were taken as a representative of the pixel in the image. In this study, three scales ($\sigma = 3, 5, 7$) of detectors were applied on the original image. Based on the line-strength and orientation images, non maximal-suppression, i.e. thresholding with hysteresis, was used to extract the backbone of each structure, followed by thinning to obtain ridges in the image.

To reduce non-spicule false positives generated from other linear structures (e.g., ligaments, ducts, etc.), several post-filtering rules based on geometric relationships (e.g., position and direction) between the spicule candidates and the extracted mass candidate (in section 3.2.3) were applied. A spiculation map was then obtained by applying region growing (using the filtered spicules as seed) on the multi-scale line-strength image. The final output of the spiculation detection module is a binary mask (named as "spiculation map"), where the foreground object(s) represents the detected potential spicule pixels. Fig. 3.5 shows an example of the detected spicules and the spiculation map.

Using the detected spiculation pixels, we compute the ratio between the areas of the detected spiculation to the areas of extracted mass candidate as the spiculation measurement. If the ratio is larger than a pre-defined threshold, we determine the mass as spiculated. This classification decision will be used in the final segmentation integration stage (see section 3.2.5) to determine whether to include these spicules into the graph cuts segmentation algorithm.

3.2.5 Segmentation Integration by Graph Cuts

At the final stage, we employ graph cuts [51] to integrate all the object-level findings, including mass candidate, noisy clutters and spiculation, along with the pixel level PM to generate the final segmentation mask. We construct an undirected graph $G = \langle \upsilon, \varepsilon \rangle$ defined by a set of nodes υ (image pixels) and a set of directed edges ε which connect these nodes. In this graph, there are two distinct nodes (or terminals) s and t, called the source and sink, respectively. The edges connected to the source or sink are called

t-links, such as (s, p) and (p, t).

For the segmentation problem, an energy function E is defined based on the graph G by considering two criteria: (1) the segmentation is guided both by the foreground (i.e. mass) and background (i.e. non-mass) appearance probabilistic statistics; (2) the segmentation is smooth, reflecting a tendency to a solidity of objects. Therefore, we seek a labeling system f, which assigns one label to each pixel in the image, to minimize the energy function E as follows:

$$E = E_{data}(f) + E_{smooth}(f) = \sum_{p \in P} D(p, f_p) + \sum_{(p,q) \in N} V(p, q) T(f_p \neq f_q) \tag{3.15}$$

where E_{smooth} is a piecewise smoothness term, E_{data} is a data error term, P is the set of pixels in the image, $V(p, q)$ is a smoothness penalty function, $D(p, f_p)$ is a data penalty function, N is a four-neighbor system, f_p is the label of a pixel p, and $T(\bullet)$ is 1 if its argument is true and 0 otherwise. In this bipartioning problem, the label f_p is either 0 or 1. If $f_p = 1$, the pixel belong to the mass, otherwise, the pixel is not the mass. We define $V(p, q)$ as follows:

$$V(p, q) = \exp(-\frac{(Pr(I_p|f_p = 1) - Pr(I_q|f_q = 1))}{2\delta^2}) \tag{3.16}$$

where $\Pr(I_p|f_p = 1)$ is the probability of pixel p belonging to class $f_p = 1$. In our case, the $\Pr(I_p|f_p = 1)$ corresponds to the output value in the PM.

Regarding the data penalty term, we use the negative log-likelihoods function as:

$$D(p, f_p) = -\ln \Pr(I_p|f_p = 1) \tag{3.17}$$

A very attractive property of graph cuts is that it can easily incorporate topological constraints into the final segmentation by setting appropriate weights of t-links in the graph. These constraints indicate some image pixels a *priori* known to be a part of the foreground or background. In our scenario, on one hand, we have extracted a mass

candidate along with spiculation map (if the mass is determined to be spiculated), which must be included in the final segmentation of a mass. On the other hand, we regard the noisy clutters as regions must be excluded from the final segmentation. We use the weights of edges defined in (3.18) and (3.19) to completely define the graph G (see [51] for detail derivation).

$$w(p,s) = \begin{cases} \infty & p \in \text{Foreground} \\ 0 & p \in \text{Background} \\ D(p, f_p = 1) & \text{otherwise} \end{cases} \tag{3.18}$$

$$w(p,t) = \begin{cases} 0 & p \in \text{Foreground} \\ \infty & p \in \text{Background} \\ D(p, f_p = 0) & \text{otherwise} \end{cases} \tag{3.19}$$

where $w(p,S)$ and $w(p,T)$ are the weights of edges connecting the pixel p to the source s and the sink t respectively.

By integrating the object level detections along with the mass PM into the graph cuts framework, the optimal segmentation of mammographic mass could be obtained in one single step by finding the minimum cost cut C on the graph G. The cut c, determining a unique labeling function f in the image, can be computed exactly in polynomial time via a standard max-flow algorithm for two terminal graph cuts.

3.3 Experiments and Results

3.3.1 Image Database

In this study, we used a dataset containing a total of 54 (51 malignant and 3 benign) ROIs. The spatial resolution of the image was sampled to $100\mu m \times 100\mu m$. The mass shape, margin, and density were measured by a senior radiologist according to Breast Imaging Reporting and Data System Atlas (BI-RADS) [59] as shown in Fig. 3.6 (a) - (c). Fig. 3.6 (d) shows the size of these masses ranged from 5mm to 50mm. The size

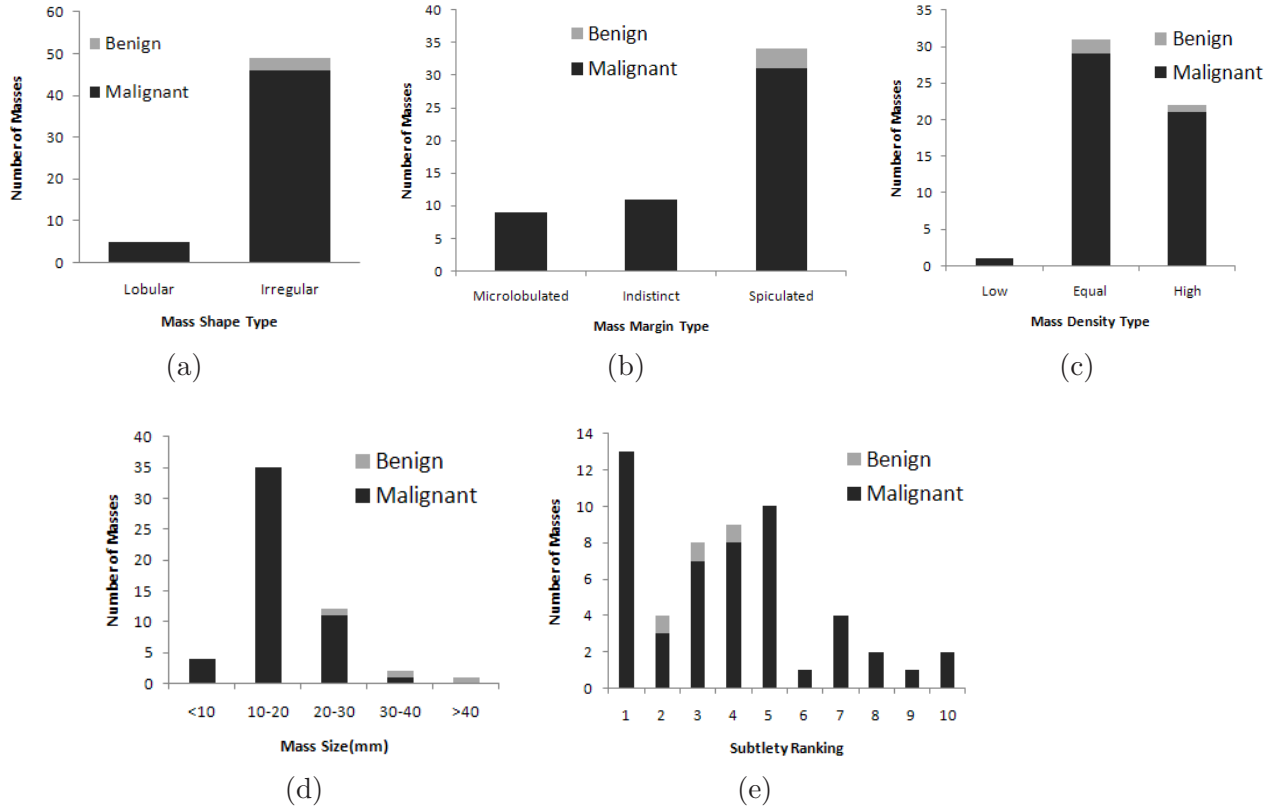


Figure 3.6: The distribution of the mass statistics of shape, margin, density, sizes and subtlety ranking within the database.

of a mass was measured as the longest dimension of the mass. The subtlety of these masses as shown in Fig. 3.6 (e) were ranked on a scale from 1 (the easiest) to 10 (the most difficult) for determining the malignancy of the case. In this study, we collected a total of five radiologists' delineation for each mass in the database independently. With the multi-radiologists' segmentation results, the final ground truth for each mass was then formed by setting the pixel as the foreground (i.e. mass) where at least three radiologists reached censuses. The remaining pixels were regarded as the background (i.e. non-mass).

3.3.2 Pixel-Scale Classification Results

For the pixel level classification, we ran a three-fold cross-validation experiment using the database. A two-phase classification scheme was adopted after experiments, since we found that using more phases did not substantially improve the classification performance. The threshold for the first phase is set to be 0.1 in order to achieve for high recall and moderate FP deduction. We empirically found that the performance is stable with the sROI size in a range of 9 to 15 pixels (note that the default value is 11).

We compare the classification accuracy of using one single layer LDA + GMM classifier with that of the two-phase classifiers. The accuracy with optimized threshold could be improved from 86.15% using the single layer LDA classifier to 87.81% using the multi-phase classifier. The effect of the multiphase approach is illustrated by the enhancement results shown in Fig. 3.7. It is evident that the further reduction of false positive samples, with increasing classification phases, thus improves the overall classification performance.

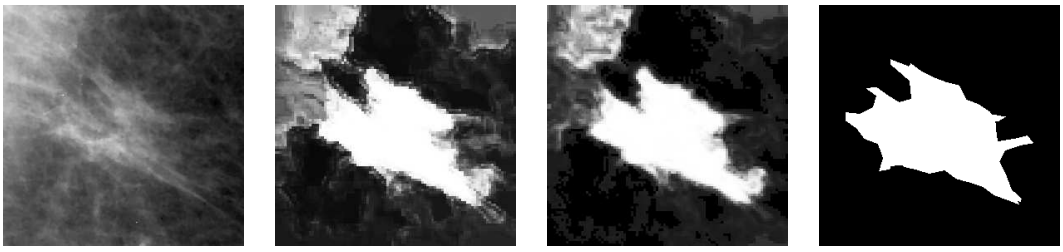
3.3.3 Segmentation Results

To measure the level of agreement between radiologists' delineation and the segmented masses, three measurements including the area overlap measure (AOR), and two contour-based measurements of the average minimum distance ($AMINDIST$), and the Hausdorff distance ($HSDIST$) were adopted.

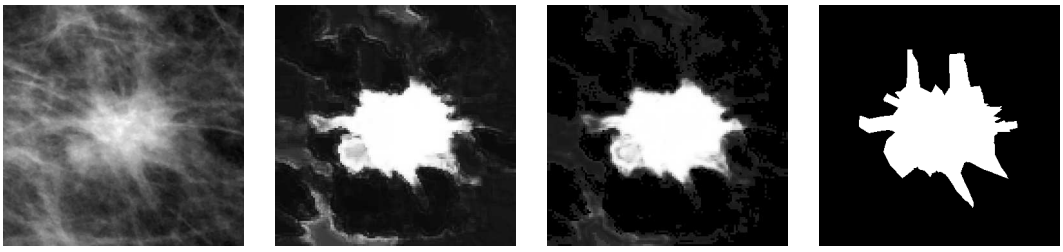
We define the distance measure between a point α and a curve B in terms of the minimum Euclidean distance (MINDIST) in the Cartesian plane. If the curve B is described described in terms of q points $\{b_1, ..., b_q\}$, the $MINDIST(a, B)$ is defined as:

$$MINDIST(a, B) = \min_{i \in \{1,...,q\}} \|a - b_i\| \tag{3.20}$$

The Hausdorff distance [60] as illustrated graphically in Fig. 3.8 is the first distance measure used, and it is defined in terms of the directed Hausdorff distance as:



(a.1) (b.1) (c.1) (d.1)



(a.2) (b.2) (c.2) (d.2)

Figure 3.7: Example outputs of the multi-phase pixel-level classification: (a.1) & (a.2) the original ROIs, (b.1) & (b.2) the PMs generated by the first phase classifier, (c.1) & (c.2) the PMs generated of the two-phase classifiers, and (d.1) & (d.2) the ground truth masks provided by the radiologists. Note that the noisy responses generated from the chest border tissues in the up-right corner of the ROI in (a.1) has been clearly suppressed by the two-phase classifiers.

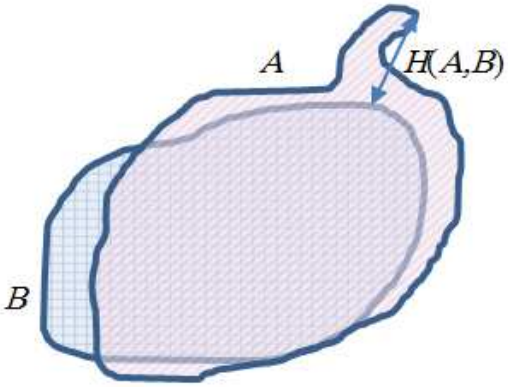


Figure 3.8: The graphical representation of the Hausdorff distance between contours A and B, which can be interpreted in this figure as the maximum of the minimum distances between any point on contour A and contour B. The area overlap measure is defined as the ratio of the hatched area to the area of the cross area.

$$H(A,B) = \max\{h(A,B), h(B,A)\} \tag{3.21}$$

It is well known that the Hausdorff distance is a metric, i.e., it satisfies the identity and symmetry equalities and the triangle inequality, and is nonnegative [60]. It does not require an explicit pairing of points between A and B. One disadvantage of the Hausdorff distance is that it does not measure how much A and B are dissimilar *on the average*. For example, even when the two closed contours are identical at all points except one, the Hausdorff distance can be large. We, therefore, defined a second distance measure, the average MINDIST (AMINDIST), by averaging the AMINDIST of a_i to B and the AMINDIST of b_i to A

$$AMINDIST(A,B) = \frac{\sum_{i=1}^{p} MINDIST(a_i, B)}{2p} + \frac{\sum_{i=1}^{q} MINDIST(b_i, A)}{2q} \tag{3.22}$$

The AOR between two closed contours A and B is defined as:

$$AOR(A,B) = \frac{Area\{A \cap B\}}{Area\{A \cup B\}} \tag{3.23}$$

when A and B are two interior regions. It is easily seen that $AOR(A,B) = 0$ when A and B do not intersect, $AOR(A,B) = 1$ when A and B are identical, and $0 \leq AOR(A,B) \leq 1$ when the similarity of the two closed contours are between these two extremes.

The box and whisker plots of the corresponding distributions of these segmentation measurements are shown in Fig. 3.10, with AOR of 0.766 ($\pm$0.144), AMINDIST of 9.79 ($\pm$0.12), and HSDIST of 58.38 ($\pm$31.25). The ROIs shown in Fig. 3.9 demonstrate the effectiveness of the proposed approach. It is seen that the segmented contours are capable of closely delineating mass body contours, and they includes sufficient amount of mass margin portion. The approach is also technically robust in segmenting masses in the presence of ill-defined texture patterns and unsmooth intensity changes inside masses. In addition, the segmented results were seen to include substantial amount of spiculations. An entire gallery showing the segmentation results for all the masses examined in this study using the proposed segmentation algorithm is located in the appendix.

3.3.4 Multi-Observer Agreement

To measure the consistency of our segmentation with multiple radiologists and the consistency within multiple radiologists themselves, we adopted the Williams index (WI) [61]. The WI is a ratio of the agreement of rater $j = i$ to the group versus the overall group agreement and is defined mathematically as follows:

$$WI_i = \left(\frac{n-2}{2}\right) \frac{\sum_{j=1, i\neq j}^{n} s_{ij}}{\sum_{j=1, j\neq i}^{n-1} \sum_{k>j}^{n} s_{jk}} \tag{3.24}$$

where s_{ij} is a measure of similarity or agreement between raters i and j. We used AOR, the reciprocal of $AMINDIST$, and the reciprocal of $HSDIST$ as the similarity measurements. If the upper limit of the confidence interval (CI) of WI is greater than the value one, we can conclude that the measurement data are consistent with the

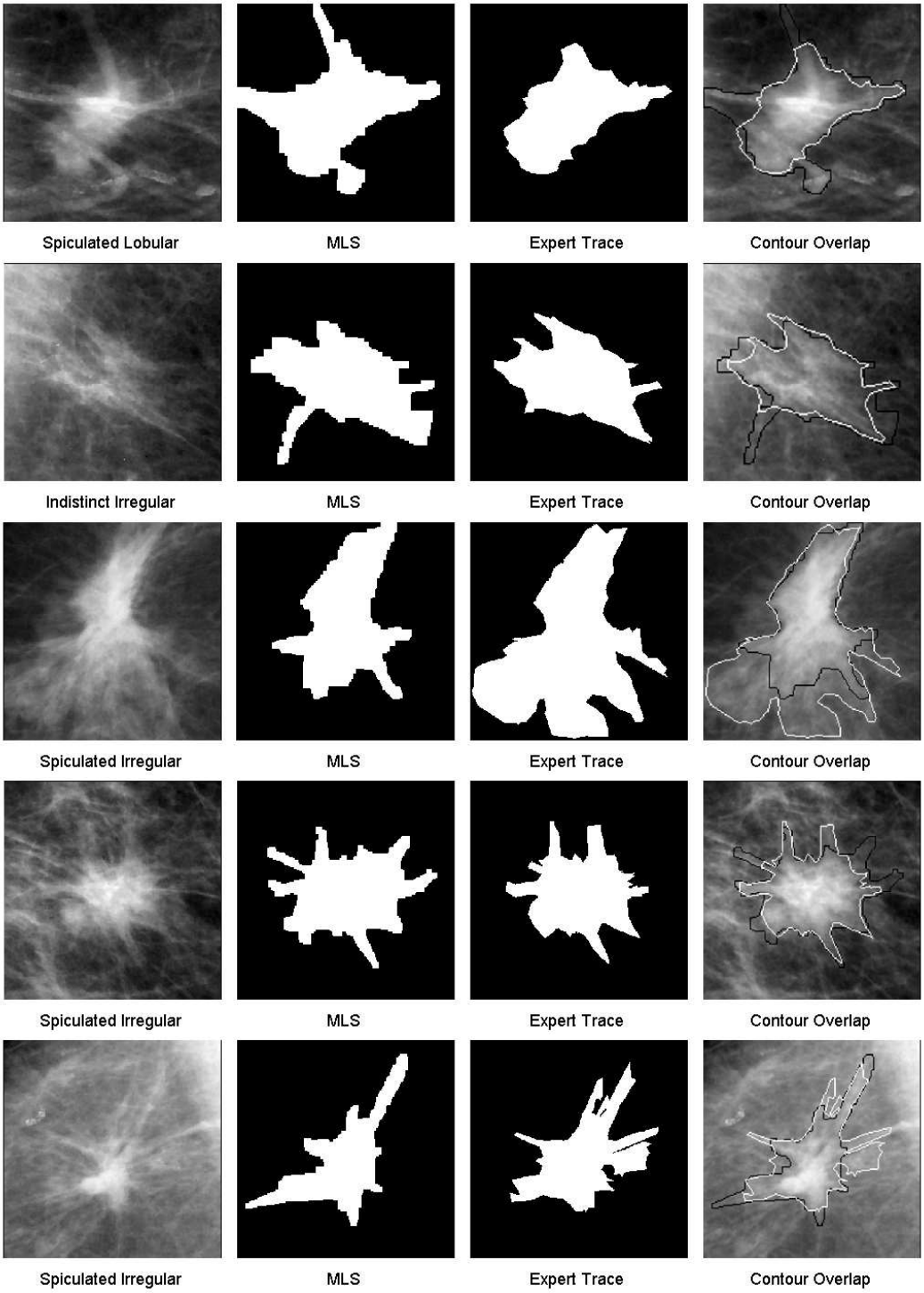


Figure 3.9: Segmentation results: from left to right, they are the original ROI, segmentation result of the MLS approach, manual segmentation, and the contours superimposed on the original image of the MLS approach (black contour) and manual segmentation (white contour). The margin and shape BI-RADS descriptors of a mass are also shown in labels under images in the first column.

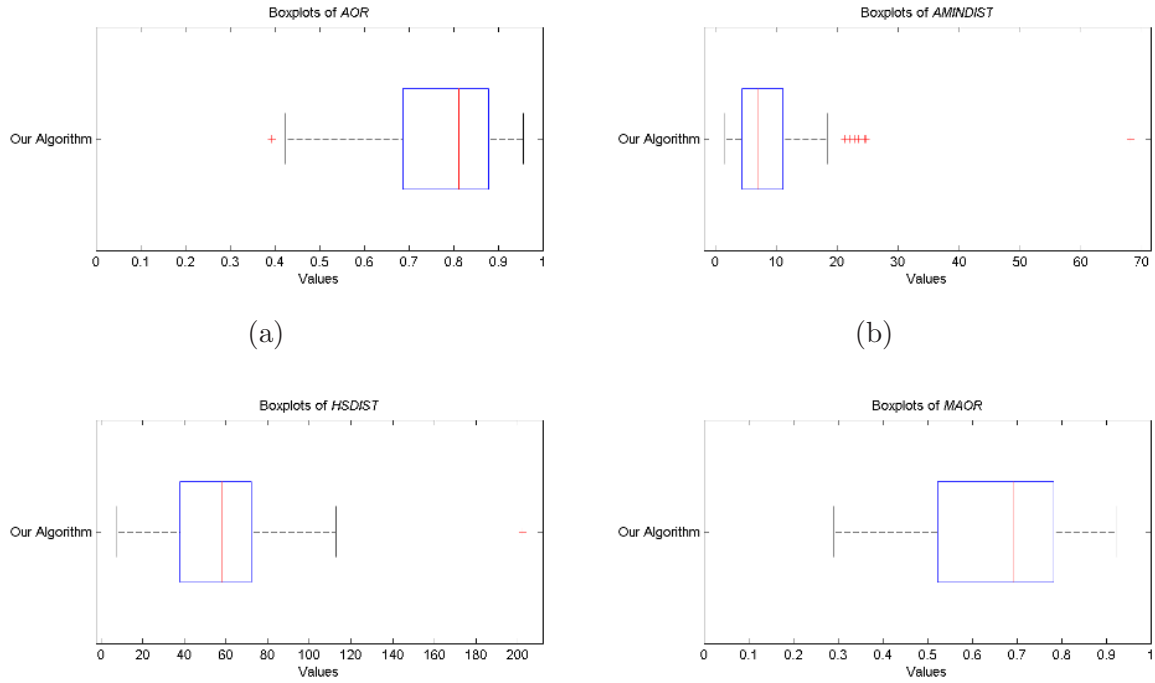


Figure 3.10: The box and whisker plots of the distribution of the segmentation measurements. The vertical lines of the boxes correspond to the lower, median and upper quartile. Outliers are represented by cross.

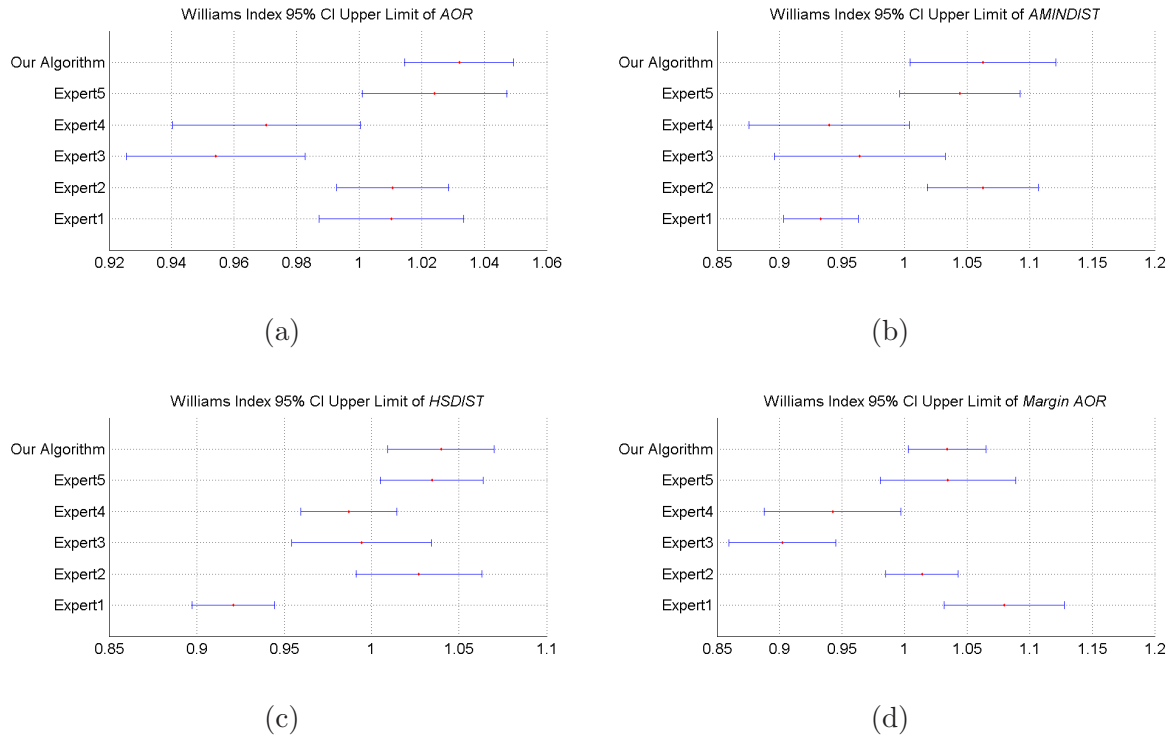


Figure 3.11: The WI for the algorithm and the radiologists.

hypothesis that the individual observer agrees with the group at least as well as the group members agree with each other (i.e., the individual observer is a reliable member of the group). The CI is estimated with a jack knife scheme [61]. The WI of the three segmentation measurements are shown in Fig. 3.11 (a) - (c), with AOR of 1.002 ($\pm$0.010) (CI at 95%), $AMINDIST$ of 0.975 ($\pm$0.047), and $HSDIST$ 0.995 ($\pm$0.029).

3.3.5 Margin Segmentation Results

The performance of the proposed algorithm on segmenting only the margin portion was also evaluated. Here the margin is defined as the remaining foreground pixels by subtracting the mass core region from the complete ground truth. The mass core region was obtained through boundary smoothing via a rotation structure element (ROSE) algorithm [62] followed by morphological erosion. The margin area overlapping ratio ($MAOR$) of the segmented masses with the ground truth margin was then computed.

Box and whisker plot of the distribution of the segmentation results are shown in Fig. 3.10 (d), with $MARO$ of 0.574 ($\pm$0.179). We also measured the WI of $MARO$ between the algorithm and multiple radiologists. The result is shown in Fig. 3.11 (d), with value of 0.9815($\pm$0.021). It is shown that the proposed approach well agreed with multiple radiologists in segmenting mass margin portion.

4

Conclusions and Future Work

4.1 Conclusions and Contributions

We have proposed two learning-based algorithms that can achieve high accuracy and robustness in the tasks of medical radiograph classification and mammographic mass segmentation. The advantage and contribution of our work can be summarized in the following aspects:

(1) Regarding the medical radiograph classification task, we have developed a hybrid learning-based approach that integrates learning-based local appearance detections, the shape prior constraint by a sparse configuration algorithm, and a final filtering stage with the exemplar-based global appearance check. The approach is highly accurate, robust, and fast in identifying images even when altered by diseases, implants, or imaging artifacts. The robustness and efficiency of the algorithm come from: (1) the accurate and fast local appearance detection mechanism with the sparse shape prior constraint, and (2) the complementarity of local appearance detection and global appearance check. The experimental results on a large-scale chest radiograph view position identification task and a multi-class medical radiograph annotation task have demonstrated the effectiveness and efficiency of our method. As a result, minimum manual intervention is required, improving the usability of such systems in the clinical setting. Our algorithm has already been integrated into an advanced image visualization workstation for en-

abling content-sensitive hanging-protocols and auto-invocation of a CAD algorithm on identified PA-AP chest images.

(2) Regarding the work for mammographic mass segmentation, the proposed MLS approach is different from previous approaches, by specifically addressing the technical issue of effective inclusion the ill-defined margins and spiculations for segmentation. The advantages of the MLS approach can be summarized as follows. (1) By image sub-patch level modeling (ISLM), the algorithm can enhance the mass structure to be segmented, while substantially suppressing the influence from clutters and background structures. Traditional region growing based methods [45] may easily "flood" into these unwanted areas, if they are designed to include more mass margin portion. (2) In the probability map (PM) generated by ISLM, image values for the mass are more uniform than the original image patterns (e.g., intensity and texture). In other words, the method is robust in processing masses with the ill-defined margin and appearance variations, which are normalized through ISLM. Therefore, more mass margin can be included relatively easily in comparison with methods directly using the image intensity and gradient information [49; 50].

By integrating spiculation detection, we believe that the MLS approach is a more effective segmentation method for its ability in delineating ill-defined margins, irregular shapes and spiculations. This shall benefit the mass characterization module in many mammographic CAD/CADx systems.

4.2 Future Work

For the future work, many components of the proposed algorithms could be further investigated and tested with more solid experiments:

(1) Regarding the image classification work, a worthwhile study would be to run experiments with more image classes to further investigate the current algorithm's scalability and robustness. With more images of different anatomy classes to be recognized (with high accuracy), the algorithm would have the potential to be integrated with

current PACS systems to benefit organizing medical image data. From the algorithm development point of view, many of the components of the algorithm could be further enhanced. For example, different landmark detection algorithms based on various feature extraction and classification methods could be investigated. Graphic model based outlier detection algorithms could be investigated for the landmark filtering stage. Last but not least, other global appearance filtering mechanisms could be substituted in the framework to see the potential gain of the overall performance.

(2) Regarding the proposed MLS algorithm, it is of value to test on a large database with more types of masses in order to confirm the robustness of the method s in handling different cases in the real clinical environment. Although the method is targeted at ill-defined and spiculated masses in this work, theoretically, it should also be able to handle other types of masses, e.g., masses with circumscribed shape and well-defined margins, given that a suitable training set is provided. It is also interesting to see how to integrate the (learning-based) shape prior constraint into current framework, which would especially benefit for segmenting masses with regular shapes. We have also learned that the ISLM immediate output, i.e., PM, is a desired by-product of the approach, which could be used as the image content descriptors for analyzing the characteristics of mammographic masses. These features may be useful for automatic analysis of malignancy of breast lesions and for retrieving mammograms with similar image patterns in a content-based image retrieval system [63].

References

[1] M. Kallergi, "Digital mammography: from theory to practice," *Cancer control: Journal of the Moffitt Cancer Center*, vol. 5, no. 1, p. 72, 1998. 3

[2] G. Lodwick, C. Haun, W. Smith, R. Keller, and E. Robertson, "Computer diagnosis of primary bone tumors: a preliminary report," *Radiology*, vol. 80, no. 2, pp. 273–275, 1963. 3

[3] B. van Ginneken, B. M. ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: a survey," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1228–1241, 2001. 3

[4] M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: a review," *Medical Physics*, vol. 36, no. 6, pp. 2052–2068, 2009. 3, 31

[5] L. Bogoni, P. Cathier, M. Dundar, A. Jerebko, S. Lakare, J. Liang, S. Periaswamy, M. Baker, and M. Macari, "Computer-aided detection CAD for CT colonography: a tool to address a growing need," *British Journal of Radiology*, vol. 78, no. 1, pp. S57–S62, 2005. 3, 5

[6] T. Freer and M. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001. 3

[7] D. Gur, J. Sumkin, H. Rockette, M. Ganott, C. Hakim, L. Hardesty, W. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *Journal of the National Cancer Institute*, vol. 96, no. 3, pp. 185–190, 2004. 3

[8] R. Birdwell, P. Bandodkar, and D. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," *Radiology*, vol. 236, no. 2, pp. 451–457, 2005. 3

[9] T. Cupples, J. Cunningham, and J. Reynolds, "Impact of computer-aided detection in a regional screening mammography program," *American Journal of Roentgenology*, vol. 185, no. 4, pp. 944–950, 2005. 3

[10] "R2 technology (2003), image checker: Algorithm." [Online]. Available: http://www.r2tech.com/prd/prd001.html 5

[11] R. Rao, J. Bi, G. Fung, M. Salganicoff, N. Obuchowski, and D. Naidich, "Lungcad: a clinically approved, machine learning system for lung cancer detection," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM New York, NY, USA, 2007, pp. 1033–1037. 5

[12] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 198–211, 2007. 5

[13] J. M. Boone, G. S. Hurlock, J. A. Seibert, and R. L. Kennedy, "Automated recognition of lateral from PA chest radiographs: saving seconds in a PACS environment," *Journal of Digital Imaging*, vol. 16, no. 4, pp. 345–349, 2003. 7, 10, 22, 23

[14] H. Luo, W. Hao, D. Foos, and C. Cornelius, "Automatic image hanging protocol for chest radiographs in PACS," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 302–311, 2006. 8, 10

[15] T. Deselaers, T. M. Deserno, and H. Müller, "Automatic medical image annotation in ImageCLEF 2007: overview, results, and discussion," *Pattern Recognition Letters*, vol. 29, no. 15, pp. 1988–1995, 2008. 8

[16] T. Deselaers and H. Ney, "Deformations, patches, and discriminative models for automatic annotation of medical radiographs," *Pattern Recognition Letters*, vol. 29, no. 15, pp. 2003–2010, 2008. 8, 10, 12, 22, 23

[17] H. Müller, T. Gass, and A. Geissbuhler, "Performing image classification with a frequency-based information retrieval schema for ImageCLEF 2006," in *ImageCLEF 2006*, ser. working notes of the Cross Language Evalutation Forum (CLEF 2006), Alicante, Spain, 2006. 8

[18] M. O. Güld and T. M. Deserno, "Baseline results for the ImageCLEF 2007 medical automatic annotation task using global image features," *Advances in Multilingual and Multimodal Information Retrieval*, vol. 4730, pp. 637–640, 2008. 10

[19] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, "Deformation models for image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1422–1435, 2007. 10

[20] T. Tommasi, F. Orabona, and B. Caputo, "Discriminative cue integration for medical image annotation," *Pattern Recognition Letters*, vol. 29, no. 15, pp. 1996–2002, 2008. 10, 12, 22, 23

[21] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 10

[22] E. Pietka and H. K. Huang, "Orientation correction for chest images," *Journal of Digital Imaging*, vol. 5, no. 3, pp. 185–189, 1992. 10

[23] E. Kao, C. Lee, T. Jaw, J. Hsu, and G. Liu, "Projection profile analysis for identifying different views of chest radiographs," *Academic Radiology*, vol. 13, no. 4, pp. 518–525, 2006. 10, 22, 23

[24] H. Arimura, S. Katsuragawa, T. Ishida, N. Oda, H. Nakata, and K. Doi, "Performance evaluation of an advanced method for automated identification of view positions of chest radiographs by use of a large database," in *Proceeding of SPIE Medical Imaging*, vol. 4684, 2002, pp. 308–315. 10

[25] T. M. Lehmann, O. Güld, D. Keysers, H. Schubert, M. Kohnen, and B. B. Wein, "Determining the view of chest radiographs," *Journal of Digital Imaging*, vol. 16, no. 3, pp. 280–291, 2003. 10, 22, 23

[26] D. Cristinacce and T. Cootes, "Facial feature detection using adaboost with shape constraints," in *British Machine Vision Conference*, 2003, pp. 231–240. 11

[27] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004. 11

[28] K. Yow and R. Cipolla, "Feature-based human face detection," *Image and Vision Computing*, vol. 15, no. 9, pp. 713–735, 1997. 11

[29] T. K. Leung, M. C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labeled graph matching," in *International Conference on Computer Vision (ICCV)*, 1995, pp. 637–644. 11

[30] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, 2001. 11

[31] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 259–289, 2008. 11

[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, 2001, pp. 511–518. 11, 14, 15

[33] I. Dryden and K. V. Mardia., *The statistical analysis of shape*. Wiley, London, 1998. 11

[34] T. Cootes, G. Edwards, C. Taylor *et al.*, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001. 12

[35] M. Leventon, W. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2000, pp. 316 – 323. 12

[36] F. H. Netter, *Atlas of human anatomy, 4th edition*, ser. Netter Basic Science. Elsevier Health Sciences, 2006. 14

[37] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997. 15

[38] Y. Zhan, X. S. Zhou, Z. Peng, and A. Krishnan, "Active scheduling of organ detection and segmentation in whole-body medical images," in *Proc. of 11th MICCAI*, ser. LNCS, D. Metaxas, L. Axel, G. Fichtinger, and G. Szekely, Eds., vol. 5242. Heidelberg: Springer, 2008, pp. 313–321. 16

[39] T. Deselaers and T. Deserno, "Medical image annotation in ImageCLEF 2008," in *CLEF Workshop 2008, Evaluating Systems for Multilingual and Multimodal Information Access*. Aarhus, Denmark: Springer, 2009. 21

[40] H. Luo and J. Luo, "Robust online orientation correction for radiographs in PACS environments," *IEEE Transactions on Medical Imaging*, vol. 25, no. 10, pp. 1370–1379, 2006. 26

[41] Z. Peng, Y. Zhan, X. S. Zhou, and A. Krishnan, "Robust anatomy detection from CT topograms," in *Proceeding of SPIE Medical Imaging*, vol. 7620, 2009, pp. 1–8. 29

[42] L. Lu, J. Bi, S. Yu, Z. Peng, A. Krishnan, and X. S. Zhou, "A hierarchical learning approach for 3D tubular structure parsing in medical imaging," in *International Conference on Computer Vision (ICCV)*, 2009. 29

[43] S. Pohlman, K. Powell, N. Obuchowski, W. Chilcote, and S. Grundfest-Broniatowski, "Quantitative classification of breast tumors in digitized mammograms," *Medical Physics*, vol. 23, no. 8, pp. 1337–1345, 1996. 31

[44] M. Kupinski and M. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Transactions on Medical Imaging*, vol. 17, no. 4, pp. 510–517, 1998. 31

[45] L. Kinnard, S.-C. B. Lo, E. Makariou, T. Osicka, P. Wang, M. F. Chouikha, and M. T. Freedman, "Steepest changes of a probability-based cost function for delineation of mammographic masses: A validation study," *Medical Physics*, vol. 31, no. 10, pp. 2796–2796, 2004. 31, 55, 63

[46] G. Te Brake and N. Karssemeijer, "Segmentation of suspicious densities in digital mammograms," *Medical Physics*, vol. 28, no. 2, pp. 259–266, 2001. 31

[47] H. Li, Y. Wang, K. Liu, S.-C. B. Lo, and M. Freedman, "Computerized radiographic mass detection - part I: lesion site selection by morphological enhancement and contextual segmentation," *IEEE Transactions on Medical Imaging*, vol. 20, no. 4, pp. 289–301, 2001. 31

[48] B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics*, vol. 28, no. 7, pp. 1455–1465, 2001. 31

[49] J. Shi, B. Sahiner, H.-P. Chan, J. Ge, L. M. Hadjiiski, M. A. Helvie, A. Nees, Y.-T. Wu, J. Wei, C. Zhou, Y. Zhang, and J. Cui, "Characterization of mammographic masses based on level set segmentation with new image features and patient information." *Medical Physics*, vol. 35, no. 1, pp. 280–290, 2008. 31, 55, 63

[50] A. Dominguez and A. Nandi, "Improved dynamic-programming-based algorithms for segmentation of masses in mammograms," *Medical Physics*, vol. 34, no. 11, pp. 4256–4269, 2007. 31, 55

[51] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004. 33, 42, 44

[52] R. Haralick, "Statistical and structural approaches to texture," in *Proceedings of the IEEE*, vol. 67, no. 5, 1979, pp. 786–804. 34

[53] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002. 34

[54] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66, 1979. 35, 37

[55] Z. Tu, "Auto-context and its application to high-level vision tasks," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8. 37

[56] C. Vyborny, T. Doi, K. O'Shaughnessy, H. Romsdahl, A. Schneider, and A. Stein, "Breast cancer: importance of spiculation in computer-aided detection," *Radiology*, vol. 215, no. 3, pp. 703–707, 2000. 39

[57] S. Caulkin, S. Astley, A. Mills, and C. Boggis, "Generating realistic spiculated lesions in digital mammograms," in *Proc. 5th International Workshop on Digital Mammography*, 2000, pp. 713–720. 39

[58] M. Jacob and M. Unser, "Design of steerable filters for feature detection using canny-like criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1007–1019, 2004. 39, 40

[59] American College of Radiology, *Breast Imaging Reporting and Data System Atlas, 4th Edition*, 2003. 44

[60] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993. 46, 48

[61] V. Chalana, Y. Kim *et al.*, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997. 49, 52

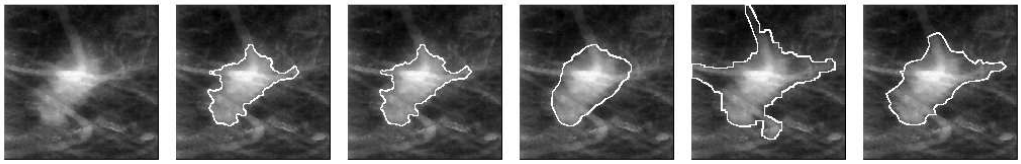
[62] B. D. Thackray and A. C. Nelson, "Semi-automatic segmentation of vascular network images using a rotating structuring element (ROSE) with mathematical morphology anddual feature thresholding," *IEEE Transactions on Medical Imaging*, vol. 12, no. 3, pp. 385–392, 1993. 52

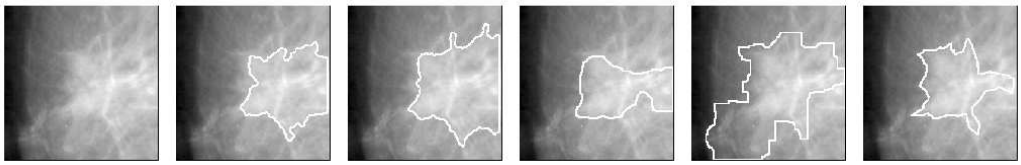
[63] Y. Tao, S.-C. B. Lo, M. T. Freedman, and J. Xuan, "A preliminary study of content-based mammographic masses retrieval," in *Proceedings of SPIE Medical Imaging*, vol. 6514. SPIE, 2007. 56

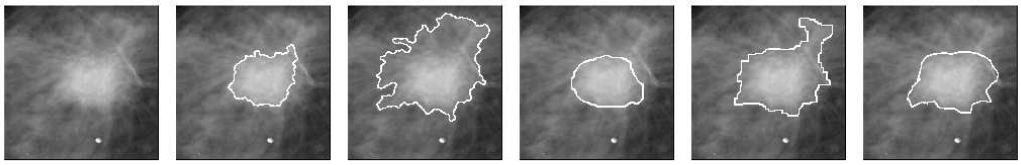
5

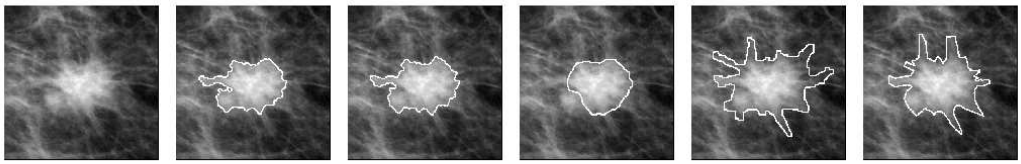
Appendix

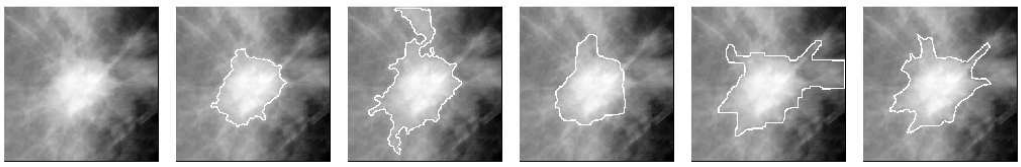
The segmentation results for all masses. For each mass, six images are shown. They are from left to right: original ROI, segmentation result of the maximum likelihood method [45](named as "ML"), segmentation result of the maximum likelihood function analysis method [45](named as "MLFA"), segmentation result of the level set method [49](named as "LS"), segmentation result of our method (i.e., MLS), and manual segmentation. The margin and shape BIRADS descriptors provided by a radiologist are given in the label under the original ROIs.

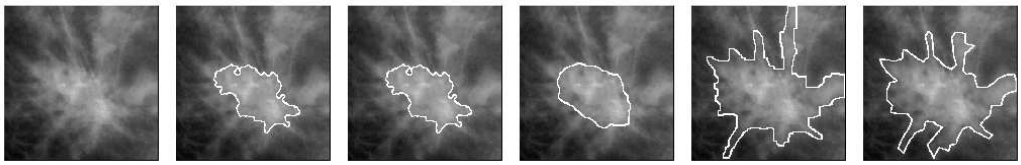

Spiculated Lobular
ML
MLFA
LS
MLS
Expert Trace

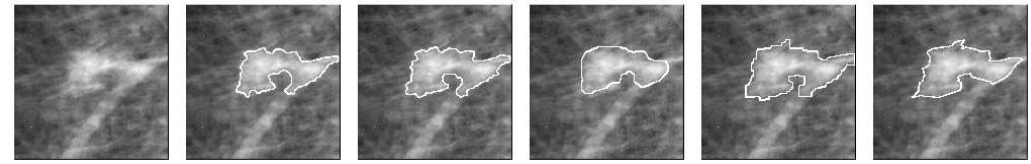

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

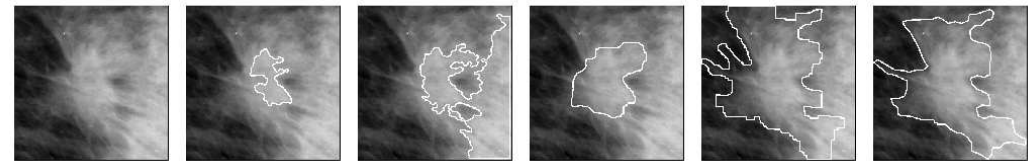

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

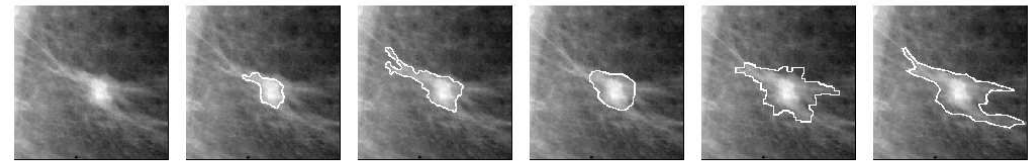

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

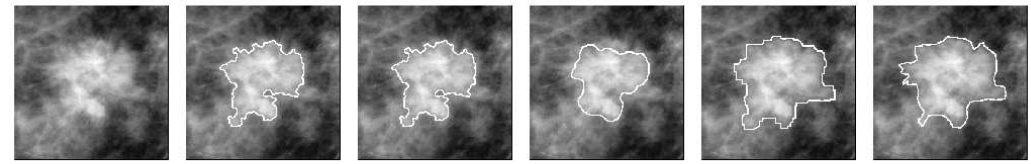

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

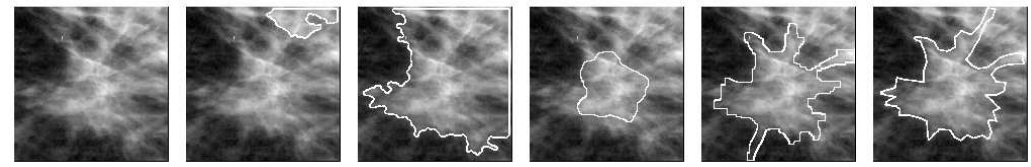

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

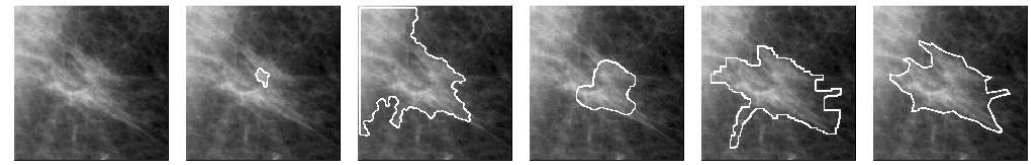

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

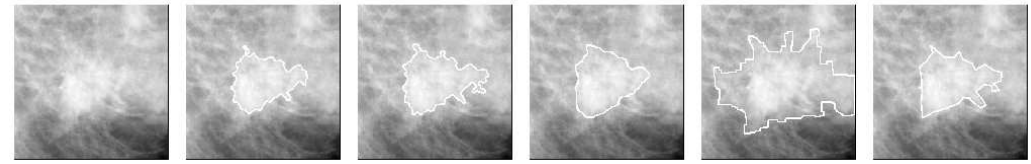

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

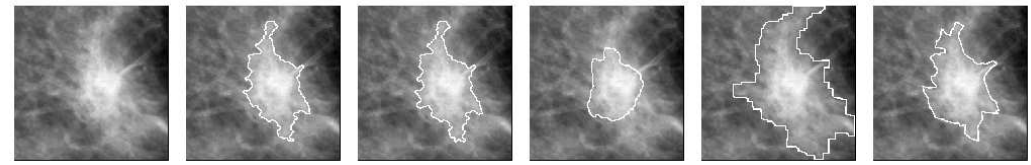

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

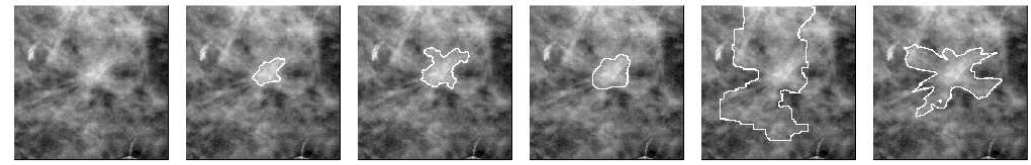

Indistinct Lobular
ML
MLFA
LS
MLS
Expert Trace

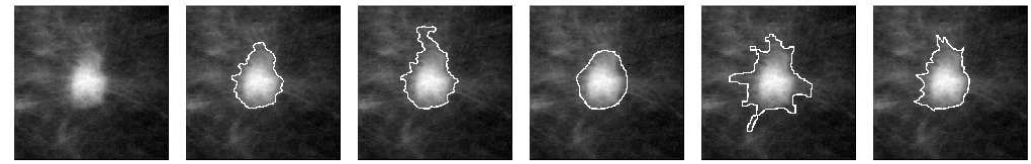

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

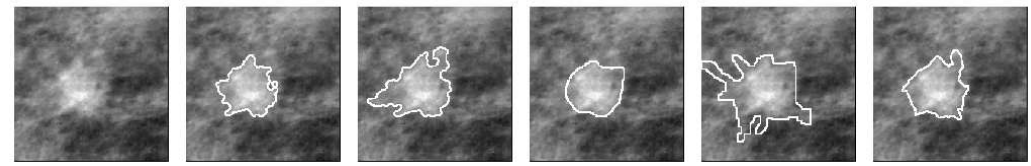

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

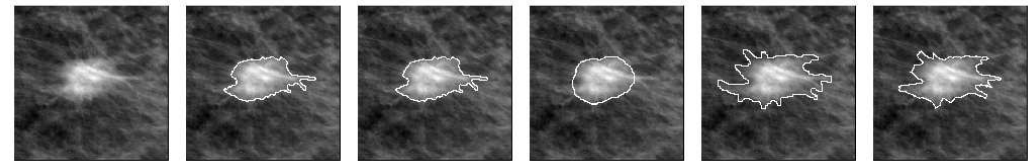

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

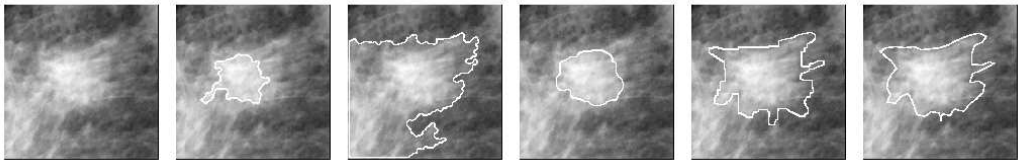

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

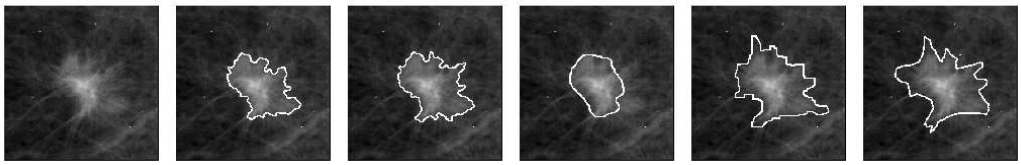

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

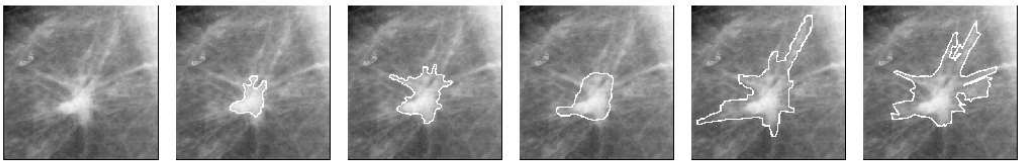

Microlobular Lobular
ML
MLFA
LS
MLS
Expert Trace

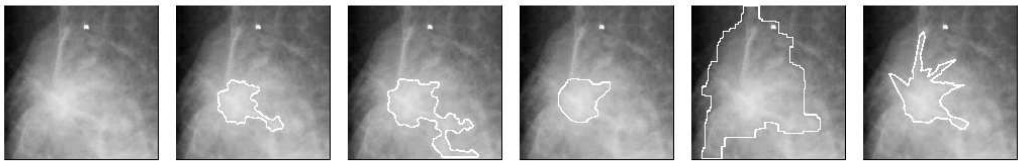

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

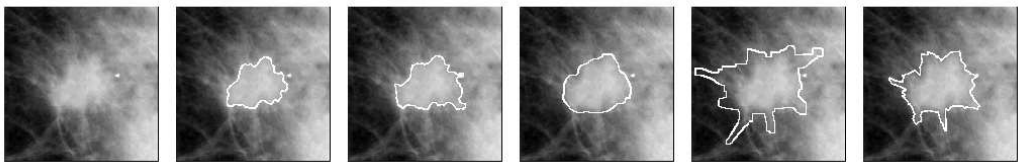

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

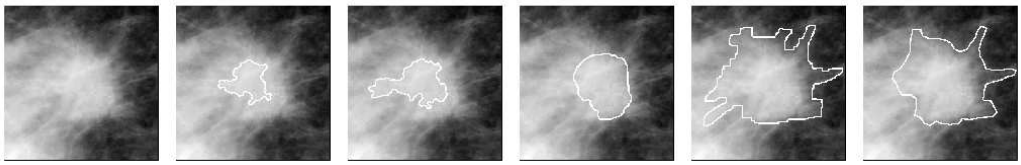

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

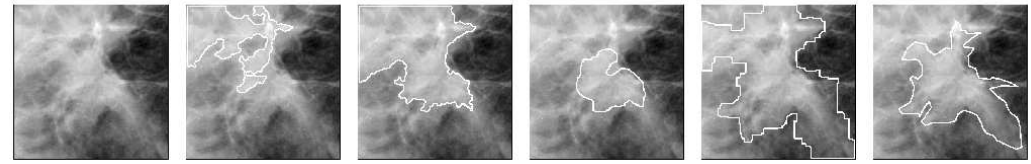

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

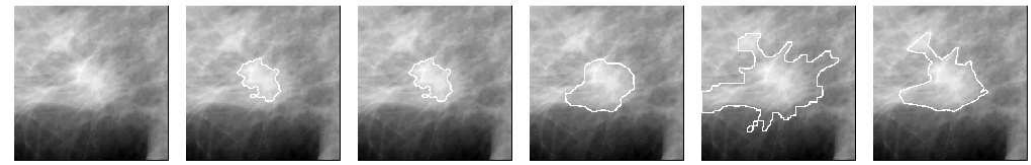

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

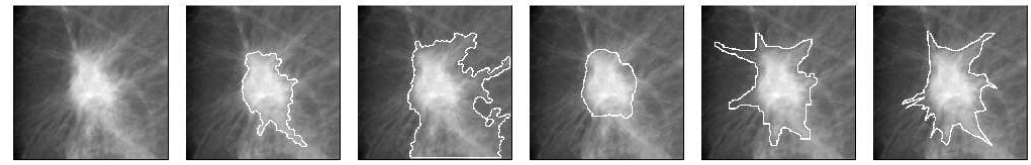

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

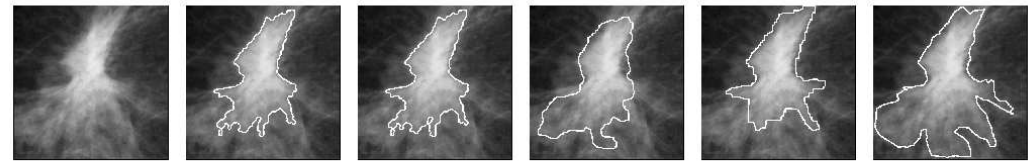

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

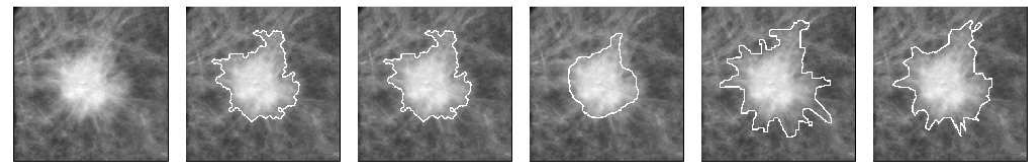

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

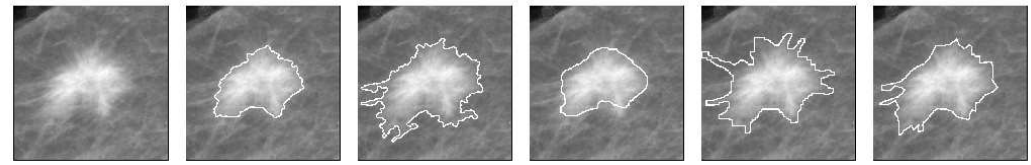

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

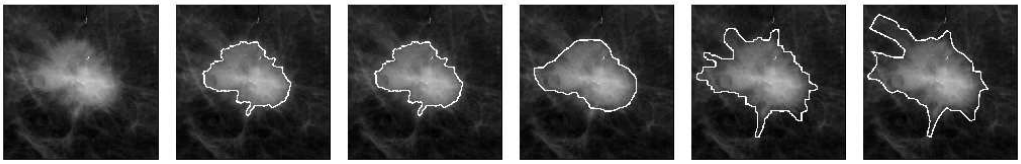

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

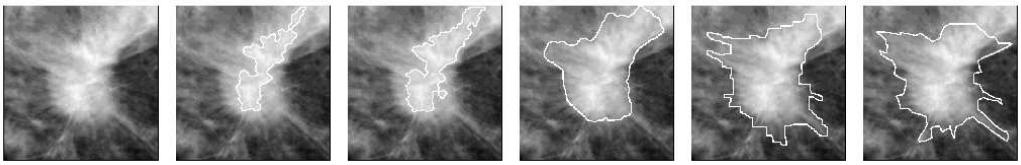

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

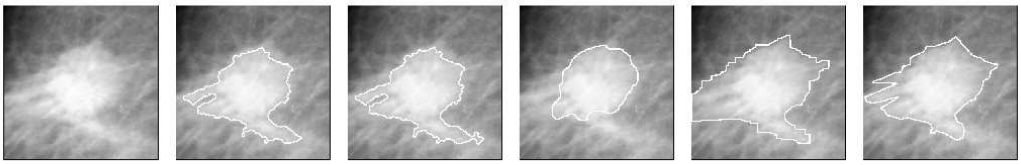

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

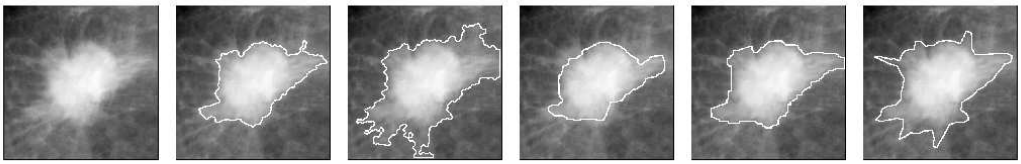

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

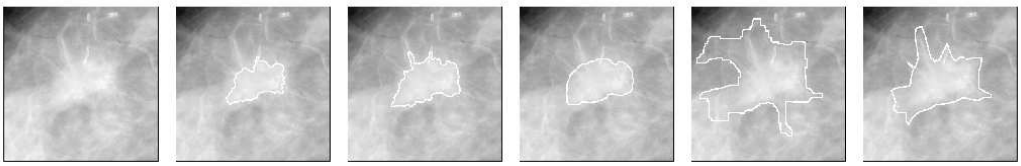

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

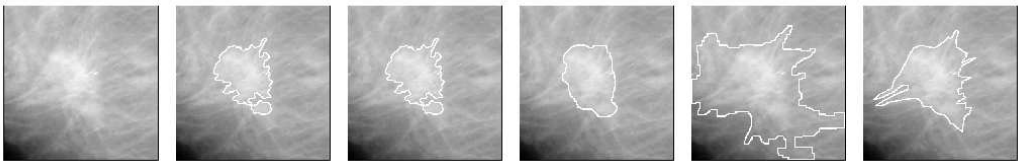

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

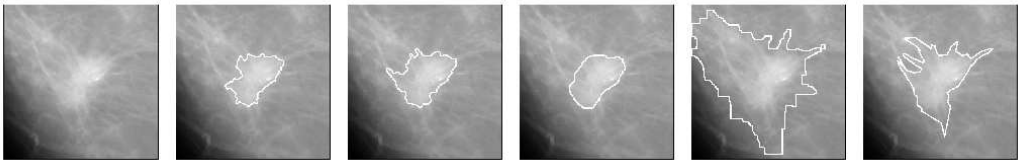

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

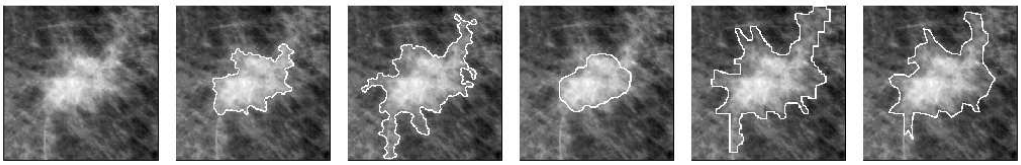

Microlobular Lobular
ML
MLFA
LS
MLS
Expert Trace

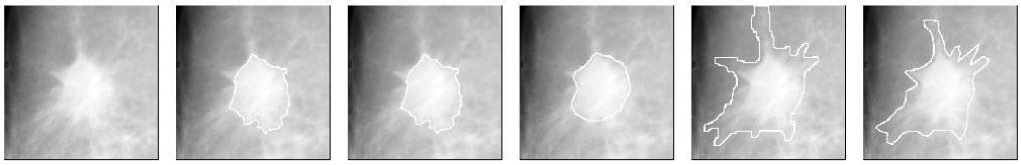

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

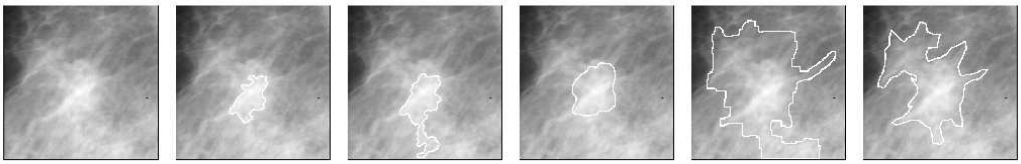

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

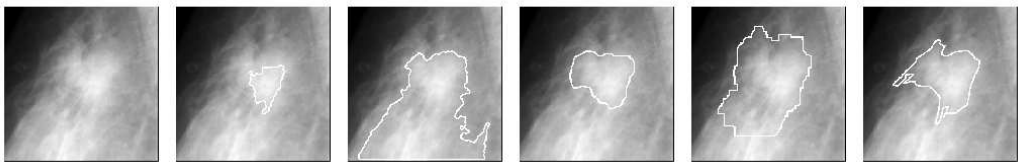

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

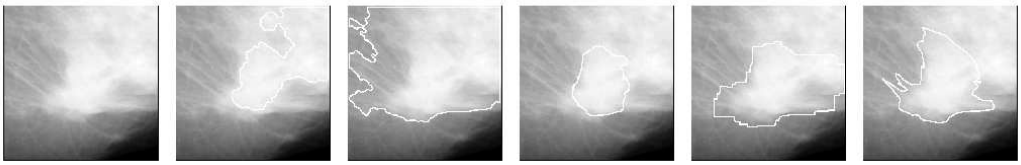

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

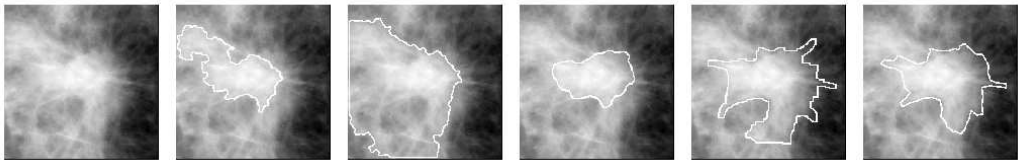

Microlobular Irregular
ML
MLFA
LS
MLS
Expert Trace

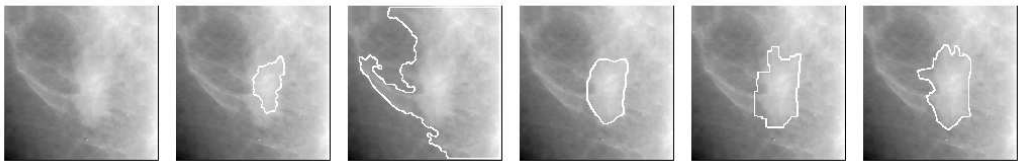

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

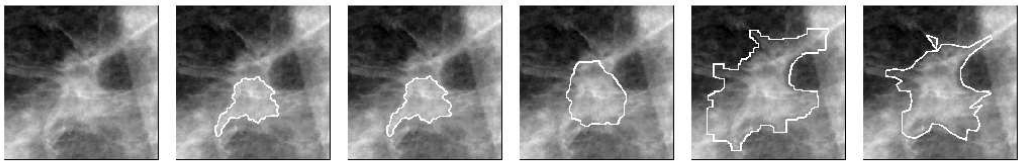

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

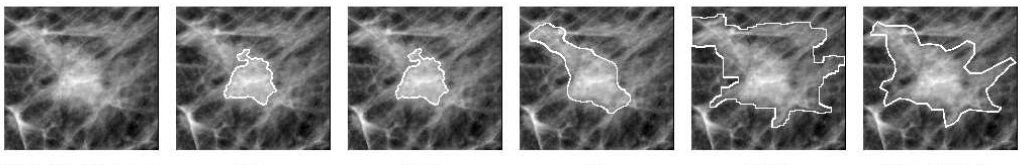

Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace

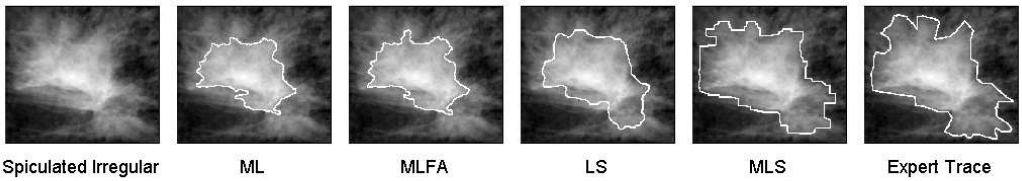

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

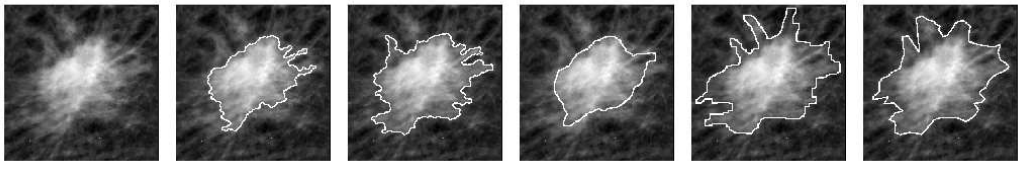

Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace

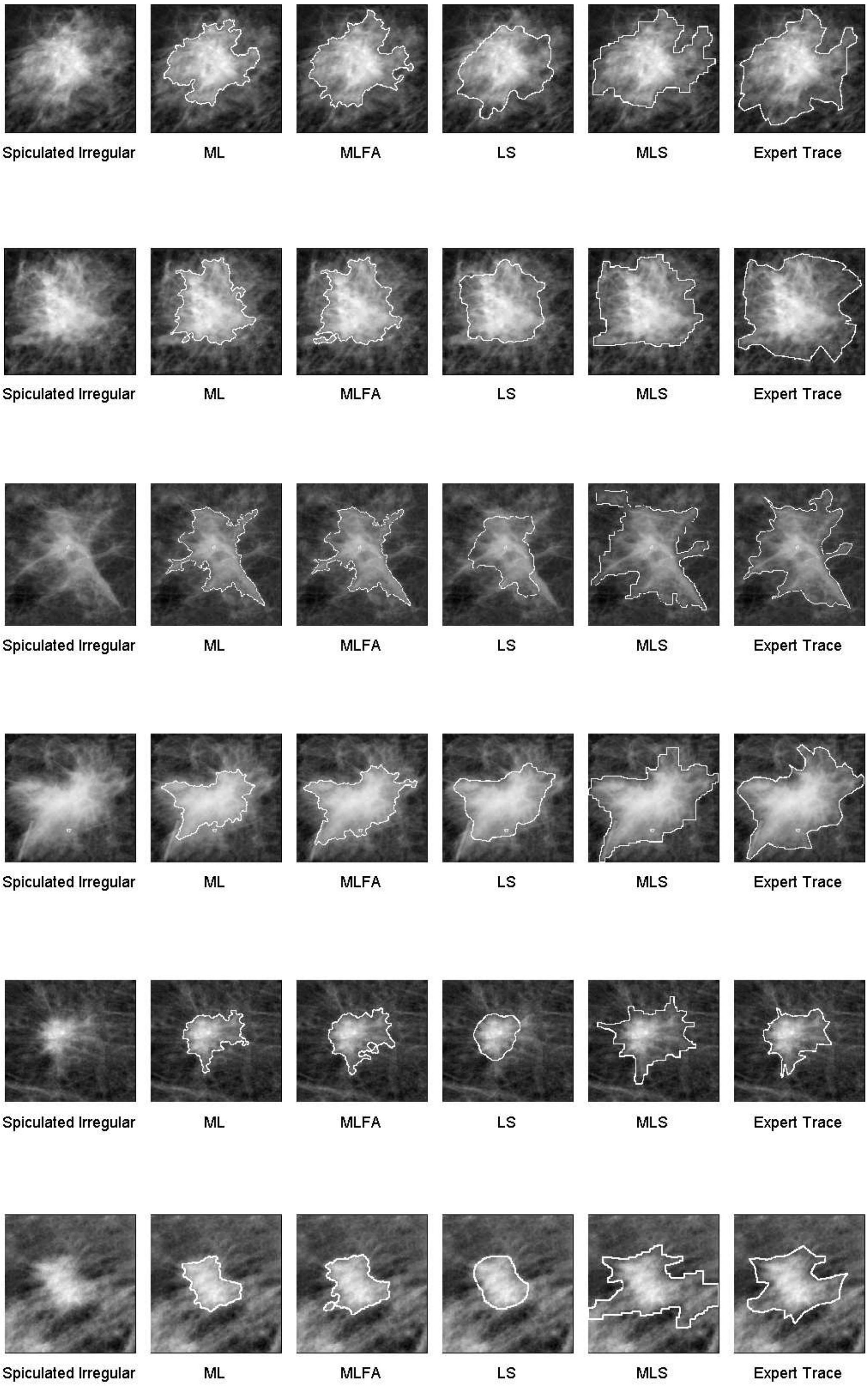

Microlobular Lobular
ML
MLFA
LS
MLS
Expert Trace


Indistinct Irregular
ML
MLFA
LS
MLS
Expert Trace


Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace


Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace


Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace


Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace
Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace
Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace
Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace
Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace
Spiculated Irregular
ML
MLFA
LS
MLS
Expert Trace