LEARNING-BASED AUTOMATIC BREAST TUMOR DETECTION AND SEGMENTATION IN ULTRASOUND IMAGES

Peng Jiang[1], *Jingliang Peng*[1], *Guoquan Zhang*[2], *Erkang Cheng*[3], *Vasileios Megalooikonomou*[3], *Haibin Ling*[3]

[1]School of Computer Science and Technology, Shandong University,
Shandong Provincial Key Laboratory of Software Engineering, P.R.China
[2] Department of Ultrasound, Shandong Provincial Hospital Affiliated to Shandong University, P.R.China
[3]Center for Data Analytics & Biomedical Informatics, Computer & Information Science Department,
Temple University, Philadelphia, PA, USA

ABSTRACT

Ultrasound (US) images have been widely used in the diagnosis of breast cancer in particular. While experienced doctors may locate the tumor regions in a US image manually, it is highly desirable to develop algorithms that automatically detect the tumor regions in order to assist medical diagnosis. In this paper, we propose a novel algorithm for automatic detection of breast tumors in US images. We formulate the tumor detection as a two step learning problem: tumor localization by bounding box and exact boundary delineation. Specifically, the proposed method uses an AdaBoost classifier on Harr-like features to detect a preliminary set of tumor regions. The preliminarily detected tumor regions are further screened with a support vector machine using quantized intensity features. Finally, the random walk segmentation algorithm is performed on the US image to retrieve the boundary of each detected tumor region. The proposed method has been evaluated on a data set containing 112 breast US images, including histologically confirmed 80 diseased ones and 32 normal ones. The data set contains one image from each patient and the patients are from 31 to 75 years old. Experiments demonstrate that the proposed algorithm can automatically detect breast tumors, with their locations and boundary shapes retrieved with high accuracy.

Index Terms— Ultrasound Image, Breast Tumor Detection, AdaBoost, Support Vector Machine, Random Walks

1. INTRODUCTION

As one of the most common cancers in women worldwide, breast cancer accounts for 16% of all female cancers. For instance, it has been estimated that about 519,000 women died in the year of 2004 due to breast cancer (WHO Global Burden of Disease, 2004). Since the causes of breast cancer remain unknown, early tumor detection is crucial to reduce the death rate. The earlier the cancers are detected, the better treatment can be provided.

Tumor detection in ultrasound (US) images serves as a key step towards automatic or semi-automatic breast cancer diagnosis. However, this task is challenging because of several facts: tumors often have complicate shapes and appearances; their patterns can largely vary from patient to patient; US images often are of low contrast and contain noise, speckles and/or motion artifacts. Fig. 1 gives examples of tumors from different patients, where the third image contains a tumor region with relatively clear boundary and uniform shade, while the first two images contain tumor regions that are not shaped as clearly or shaded as uniformly.

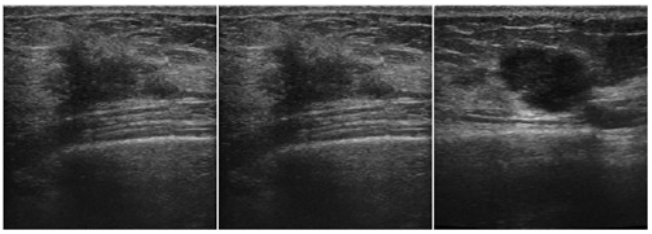

Fig. 1. Examples of breast tumor US images.

In this paper, we propose a two-stage approach for automatic tumor detection and segmentation from US images. The first stage is automatic tumor detection. For this task, we use Adaboost plus Haar feature [7] to locate potential tumor locations and then use a support vector machine (SVM) combined with quantized intensity features for refinement. By this two-level cascading our method inherits both the efficiency from the Adaboost+Haar framework and the discriminative power from the SVM classifier. In the second stage, we adjust the random walk based segmentation algorithm [10] by automatically choosing seeds from the result of the first stage.

The advantage of our method lies in that it is fully automatic. Furthermore, by combining two powerful learning tools, it achieves robustness against image noises commonly existing in US images, and leads to high accuracy of retrieval. We tested the proposed approach on a data set containing 112 US images.The effectiveness of our method is demonstrated

by the experimental results.

2. RELATED WORK

Given a breast US image, it is crucial for a computer-aided diagnosis system to detect and segment the tumor regions for further examination. For this purpose, various algorithms have been proposed during the past years to detect the boundaries of tumor regions in US images. Prevalent algorithms include those following the active contour model (ACM) or the level set model which have been applied to US images of breast [1, 2], cardiopathy [3], prostate [4], thyroid [5] and so forth. These algorithms start from initial contours and deform them in an iterative manner to get as close as possible to the contours of the object contained in the image. However, the results heavily rely on the initial specification of the object contours [6].

It is worth pointing out that the above-mentioned traditional works focus on boundary delineation but not on automatic tumor localization. As such, there are limitations inherent in those techniques: 1) Most of them are semi-automatic methods and manual labeling of a rough tumor position is needed. 2) They are likely to get stuck in complex and noisy US images where tumors may have unclear boundaries.

In the past decade, object detection from visual input has achieved great progress and has been successfully applied to tasks such as face detection [7] using machine learning approaches. Similar approaches have recently been applied to anatomical structure localization tasks as well [8].

3. METHODOLOGY

3.1. Overview

The proposed breast tumor detection algorithm works in two stages: tumor localization and tumor boundary delineation. In the first stage, the AdaBoost classifier using Haar-like features is employed to detect a preliminary set, D, of candidate rectangular boxes each bounding a potential tumor region. Afterwards, an SVM classifier using quantized intensity features is employed to divide the candidate set, D, into two subsets D_t and D_f containing the true and the false tumor regions, respectively. In the second stage, a foreground (background) seed is placed at the center of each bounding box in D_t (D_f) and the random walks algorithm is performed on the US image to obtain the final tumor boundary delineation.

3.2. AdaBoost for Tumor Localization

AdaBoost, first proposed by Yoav Freund and Robert Schapire [9], is the most efficient ensemble learning method and can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost is used not only for predicting in classification tasks, but also for presenting self-rated confidence scores which estimate the reliability of their predictions.

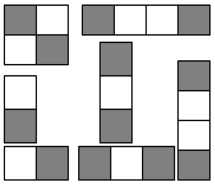

Fig. 2. Feature prototypes of simple Haar-like features.

Let $x \in R^d$ be a d dimensional input feature vector. Here, a feature vector is extracted to represent a training sample. AdaBoost can produce a classifier with better performance which combines a set of weak classifiers. In the beginning of the training stage, all the samples are initialized with equal weights. A distribution of these weights is updated during the training process. AdaBoost selects weak classifiers repeatedly by updating the distribution of these weights. The distribution of weights indicates the importance of samples in the data set. On each training round, the weight of each incorrectly (correctly) classified sample is increased (decreased). Therefore, the new weak classifier in next round will put more focus on those incorrectly classified instances.

In general, a learned AdaBoost classifier is denoted as $h(x) : R^d \rightarrow \{-1, 1\}$ with

$$h(x) = \text{sign}\left(\sum_{i=1\ldots n} c_i h_i(x)\right), \tag{1}$$

where $h_i(x)$ and c_i ($1 \leq i \leq n$) are the weak classifiers and their associated weights, respectively, as obtained from the training process.

In our work, we use Haar-like features with AdaBoost as shown in Fig. 2. Instead of giving a positive or negative label according to Equation 1, we compute a confidence score for each image region under examination, which we define as $f(x) : R^d \rightarrow R$ with

$$f(x) = \sum_{i=1\ldots n} c_i h_i(x). \tag{2}$$

Denoting the maximum confidence score of all the detected image regions as s_M, we only keep those whose confidence scores are within a threshold, τ, from s_M. This way, we form a the preliminary candidate set, denoted as D, of tumor bounding boxes.

3.3. Quantized intensity features with SVM

Due to the challenging nature of tumor detection in US images, the preliminary candidate set, D, obtained with the AdaBoost classifier often contains false positive regions. Observing that true and false positive regions have distinct characteristics in terms of contrast and intensity, we distinguish them based on the probabilistic distribution of pixel intensities. Specifically, we adaptively quantize the US image based on the k-means clustering, compute a feature vector for each

candidate region in D and perform SVM classification to the candidate regions in D accordingly.

For all the pixels, $(p_1, p_2, \ldots, p_n)$, in a US image, we conduct k-means clustering to partition the n pixels into k sets $K_1, K_2, \ldots, K_k$ ($k \leq n$) so as to minimize the approximation error, $\sum_{i=1}^{k} \sum_{p_j \in K_i} \parallel p_j - \mu_i \parallel^2$, where μ_i is the mean value of the pixels in K_i. Based on the k-means clustering result, the US image is quantized to a k intensity level image.

For each candidate region d_i ($d_i \in D$), we define its feature vector v_i as $v_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,k})$ where $t_{i,j}$ ($1 \leq j \leq k$) is the portion of the pixels in d_i with the intensity value of j after quantization. Based on their feature vectors, we use the SVM classifier to divide the candidate regions in D into two subsets, D_t and D_f, containing the true and the false positive regions, respectively, as determined by the algorithm.

3.4. Tumor boundary delineation by Random Walks

We formulate the problem of tumor boundary delineation as the segmentation of a US image into foreground (tumor) and background (non-tumor) regions. As a result, the boundary between the foreground and the background regions gives the closed boundary (boundaries) of the tumor region(s) contained in the US image.

For the purpose of image segmentation, we use the random walk segmentation algorithm [10]. It is based on a graph for the image with each node corresponding to a pixel and each edge corresponding to a neighboring pixel pair and weighted by the similarity between that pixel pair. The algorithm of random walks is given in Algorithm 1 for which we use the centers of the regions in D_t (D_f) as the foreground (background) seeds. It is worthwhile to point out that we perform the random walks segmentation to the US images after quantization as described in Section 3.3.

Algorithm 1 Tumor boundary delineation (adjusted from [10])

1: Let p_{ij} be the probability of walking from node i to node j

$$p_{ij} = \frac{w_{ij}}{\sum_{(i,k)\in E} w_{ik}} = \frac{w_{ij}}{d_i}$$

2: Let x_i be the probability starting at node i, of a random walker reaching the foreground seeds. For foreground seeds : $x_f = 1$, for background seed : $x_b = 0$ and $x_i = \sum_j p_{ij} x_j$ for other nodes.

3: The transition matrix is defined by $P = [p_{ij}]_{n\times n} = D^{-1}W$, $PX = X$ is solved by $X = [x_1, x_2, \ldots, x_n]^T$ except for seed nodes.

4: It can be proved that the random walk solution is equal to minimize following energy[11]

$$\frac{1}{2} X^T L X = \frac{1}{2} \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2$$

$$\text{s.t. } x_f = 1, x_b = 0$$

5: Set $x_f = 0, x_{b=1}$ and repeat 1-4 to get $\overline{X}$, assign pixel point to foreground if $x_i \geq \overline{x_i}$

4. EXPERIMENTAL RESULTS

We experimented using a data set containing 112 breast US images to test the classification accuracy of the proposed method. For each of diseased ones with tumor, we also have an expert-annotated copy to verify the accuracy of the segmentation.

Four-fold cross validation was employed to evaluate the performance of our proposed method. In each combination, we used 84 images with tumor annotations for training and the rest for testing. In order to learn the AdaBoost classifier, we generated 2,352 positive samples (Positive samples were generated by resampling near the annotated tumors, and negative samples were generated randomly by image decomposition.) and used 24 haar features. The learned AdaBoost classifier was obtained with four stages.

In our experiment, the accuracy of our two stages' classifier is $87.5\%(98/112)$, the sensitivity is $88.8\%(71/80)$, the specificity is $84.4\%(27/32)$.

Visual results of running the proposed breast tumor detection and segmentation algorithm on some test images are given in Fig. 3 where we see that the detected tumor boundaries match well with the expert annotations in most cases. The results also show that our algorithm can detect multiple tumors. For the Quantized intensity features, we try different configurations as listed in Table 1.

Table 1. Experimental results. Parameters (k) and (δ) are for k-means and random walks respectively.

AdaBoost type		discrete	real	gentle
Classification Accuracy		93.60%	93.76%	94.12%
	k	δ =50	δ =90	δ =130
Mean overlap rate for k = 2..10	2	70.7%	80.2%	76.7%
	3	74.2%	83.5%	76.1%
	4	66.5%	78.6%	73.4%
	5	72.0%	76.5%	69.5%
	6	54.9%	76.2%	73.2%
	7	71.7%	74.8%	60.4%
	8	71.9%	72.4%	71.4%
	9	20.1%	73.0%	71.4%
	10	40.9%	72.5%	70.0%

5. CONCLUSION

We have investigated using machine learning based methods for breast tumor detection. Specifically, the AdaBoost classifier was employed to localize the tumor candidates, quantized intensity features with SVM were utilized to refine the result set and the random walks algorithm was used for tumor boundary segmentation. The effectiveness of the proposed methods was well demonstrated by the experimental results.

In the future, we plan to advance the study by including

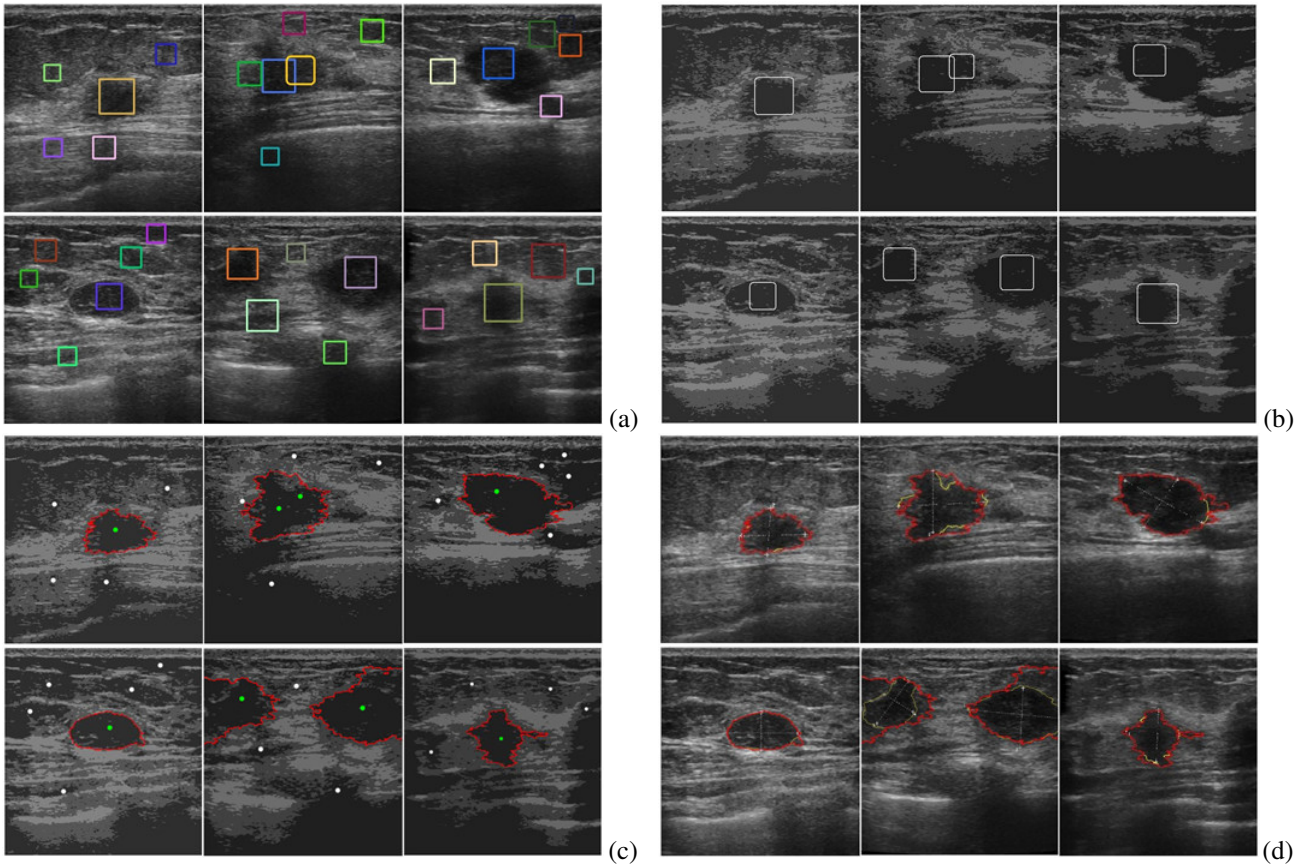

Fig. 3. Steps of tumor detection and segmentation: (a) candidate tumor regions detected by the AdaBoost classifier, (b) k-means-based image quantization and result set refinement with SVM, (c) random walks segmentation of the quantized image with green (white) points at the centers of the true (false) positive regions, (d) comparison with expert annotations.

more data and conducting a more thorough study. In addition, we will explore more advanced classification techniques and boundary segmentation strategies.

6. ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (Grants No. 61070103 and No. U1035004), in part by Program for New Century Excellent Talents in University (NCET) in China and in part by the US National Science Foundation (IIS-0916624 and IIS-1049032).

7. REFERENCES

[1] X.J. Zhu, P.F. Zhang, J.H. Shao, Y.Z. Cheng, Y. Zhang, J. Bai, "A snake-based method for segmentation of intravascular ultrasound images and its in vivo validation," *Ultrasonics*, 51(2):181–189, 2011.

[2] J.A. Noble, D. Boukerroui, "Ultrasound image segmentation: a survey," *IEEE Trans. Med. Imag*, 25(8):987–1010, 2006.

[3] K.Y.E. Leung, M.V. Stralen, G.V. Burken, N.D. Jong, J.G. Bosch, "Automatic active appearance model segmentation of 3D echocardiograms," *ISBI*, 320–323, 2010.

[4] R. Medina, A. Bravo, P. Windyga, J. Toro, P. Yan, G. Onik, "active appearance model for prostate segmentation in ultrasound images," *IEEE Eng. in Medicine and Biology*, 3363–3366, 2005.

[5] D.E. Maroulis, M.A. Savelonas, D.K. Iakovidis, S.A. Karkanis, N. Dimitropoulos, "Variable background active contour model for computer-aided delineation of nodules in thyroid ultrasound images," *IEEE Trans. Inf. Technol. Biomed*, 11(5):537–543, 2007.

[6] Y.L. Huang, D.R. Chen, "Automatic contouring for breast tumors in 2-D sonography," *IEEE Eng. in Medicine and Biology*, 3225–3228, 2005.

[7] P. Viola and M Jones, "Robust real-time face detection," *Int'l J. of Computer Vision*, 57(2):137–154, 2004.

[8] H. Ling, M. Barnathan, V. Megalooikonomou, P. Bakic, and A. Maidment. "Probabilistic Branching Node Detection using Hybrid Local Features," *ISBI*, 233–236, 2009.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci*, vol. 55(1), pp. 119–139, 1997.

[10] L. Grady, "Random walks for image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28, 2006.

[11] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17(4), 2007.