Computers in Biology and Medicine 63 (2015) 238–250

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Region based stellate features combined with variable selection using AdaBoost learning in mammographic computer-aided detection

Dae Hoe Kim [a], Jae Young Choi [a,b,1], Yong Man Ro [a,*]

[a] Image and Video Systems Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Yuseong-Gu, Daejeon 305-701, Republic of Korea
[b] Department of Biomedical Engineering, Jungwon University, 85 Munmu-Ro Goesan-Eup Goesan-Gun, Chungcheongbuk-Do 367-805, Republic of Korea

ARTICLE INFO

Article history:
Received 27 March 2014
Accepted 16 September 2014

Keywords:
Mammography
Spiculated masses
Region-based stellate features
Variable selection metric
AdaBoost
Computer-aided detection

ABSTRACT

In this paper, a new method is developed for extracting so-called region-based stellate features to correctly differentiate spiculated malignant masses from normal tissues on mammograms. In the proposed method, a given region of interest (ROI) for feature extraction is divided into three individual subregions, namely core, inner, and outer parts. The proposed region-based stellate features are then extracted to encode the different and complementary stellate pattern information by computing the statistical characteristics for each of the three different subregions. To further maximize classification performance, a novel variable selection algorithm based on AdaBoost learning is incorporated for choosing an optimal subset of variables of region-based stellate features. In particular, we develop a new variable selection metric (criteria) that effectively determines variable importance (ranking) within the conventional AdaBoost framework. Extensive and comparative experiments have been performed on the popular benchmark mammogram database (DB). Results show that our region-based stellate features (extracted from automatically segmented ROIs) considerably outperform other state-of-the-art features developed for mammographic spiculated mass detection or classification. Our results also indicate that combining region-based stellate features with the proposed variable selection strategy has an impressive effect on improving spiculated mass classification and detection.

1. Introduction

Currently, the most cost-effective method for early detection of breast cancers is screening mammography [1]. In screening mammography, breast mass is known to be one of the major signs of breast cancers [2]. Especially, breast masses with spiculated margins have a high likelihood of malignancy [2,3]. It has been reported that about 81% of spiculated masses are malignant [3]. Due to the high probability of malignancy, early detection and diagnosis of spiculated masses and architectural distortions may significantly improve the chance of survival for patients with breast cancer.

There has been a limited but increasing amount of work on the development of computer-extracted features for the detection and classification of spiculated masses on mammograms for computer-aided detection (CAD). Kegelmeyer et al. [4] proposed an analysis of local oriented edges (ALOE) features that represent stellate patterns of mammographic masses. The features were used for detecting spiculated lesions. Karssemeijer et al. [5] proposed pixel-wise stellate features for the purpose of detecting malignant masses on mammograms. Mudigonda et al. [6] developed directional gradient strength and coefficient of variation of gradient strength. In [7,8], the authors proposed the fuzziness of mass margin, radial to tangential signature information, and spiculation measure based on relative gradient information. The authors in [9] proposed the so-called texture flow-field analysis based features for the purpose of false positive (FP) reduction. Wei et al. [10] used multi-resolution spatial gray level dependence (SGLD) texture features to characterize the pattern of spiculated masses on mammograms.

While a few feature extraction methods have been suggested so far for automatic classification and detection of spiculated masses in mammography, developing effective and robust stellate features—providing acceptable classification performance—still remains a challenging and open question. In this paper, we present new feature extraction approach to correctly differentiate spiculated masses from normal tissue in mammographic CAD systems. Key technical contributions can be summarized in the following aspects:

- One particularly important characteristic of mammographic spiculated masses is having *stellate patterns* that appear as a number of line structures reside in a padiation pattern [4,5]. In light of this fact, our proposed features, termed *region-based stellate features*,

* Corresponding author. Tel.: +82 42 350 3494.
E-mail address: ymro@ee.kaist.ac.kr (Y.M. Ro).
[1] Present address: Department of Biomedical Engineering, Jungwon University, Chungcheongbuk-do, Republic of Korea.

http://dx.doi.org/10.1016/j.compbiomed.2014.09.006
0010-4825/

have been designed for well encoding the stellate pattern information of spiculated masses. To this end, when extracting the feature form a given region of interest (ROI), the ROI is divided into three *individual subregions* (namely, core, inner, and outer parts) based on a band of pixels surrounding the segmented mass object in the ROI. The statistical characteristics (e.g., mean and standard deviation) associated with stellate (or spiculated) pattern information are then separately extracted from three different subregions. Finally, we combine these complementary sources of information to form the *so-called region-based stellate features*—allowing for characterizing the stellate patterns of spiculated masses at a *regional level*.

- In this paper, we propose the use of AdaBoost learning [11] to select the best set of variables of region-based stellate features for maximizing the discrimination between spiculated masses and normal tissues. In particular, differing from the existing variable selection algorithms [12,13], the proposed variable selection method has been designed for the targeted classification applications in mammographic CAD systems. In detail, a novel *variable selection criterion* well suited for improving classification of spiculated masses in mammographic CAD is devised to optimally determine variable importance (ranking). To the best of our knowledge, combining the proposed region-based stellate features with variable selection underpinning AdaBoost is a new and novel approach for classifying and detecting spiculated masses in mammographic CAD systems and could have important implications in early detection and diagnosis of spiculated masses with high malignancy.

Comparative and extensive experiments have been conducted to investigate the effectiveness of proposed region-based stellate features in conjunction with our proposed variable selection algorithm. The popular benchmark mammogram database (DB) “Digital Database for Screening Mammography (DDSM)” [14] was used in our experimentation. The results show that the proposed approach for extracting stellate patterns can achieve a high level of classification performance—in the presence of the errors caused by the inaccurate location of the boundaries of masses automatically segmented using computer algorithms (see Fig. 1)—that meets the requirement of real-life clinical applications. In addition, we validate the superiority of our proposed region-based stellate features by comparing other state-of-the-art features recently proposed for representing stellate patterns of mammographic masses. In addition, our results demonstrate that the performance (in terms of classification) can be impressively improved by combining our region-based stellate features with variable selection based on AdaBoost learning.

This paper improves and extends preliminary work presented in [15]. In particular, this paper presents a new approach that combines our region-based stellate features with a novel variable selection algorithm. Further, we report integrated experimental results that are more extensive and rigorous in the following aspects: (1) the comparison of other state-of-the-art spiculated (stellate) features; (2) additional analysis using more classifier models; (3) systematic investigation on the effect of combining stellate patterns each computed from a particular mass subregion.

The rest of this paper is organized as follows. In Section 2, ROI detection and segmentation method used in this study is briefly explained. Section 3 details the proposed region-based stellate feature extraction. In Section 4, the proposed variable selection algorithm on the basis of AdaBoost learning is outlined. In Section 5, we present the experimental results. In Section 6, we discuss issues related to the proposed method. The conclusion is drawn in Section 7.

2. ROI detection and segmentation

Note that a typical CAD system largely consists of three parts [9,16–19]: (1) automated detection and segmentation of mammographic lesions from a given mammogram, which generates ROIs; (2) automated feature extraction for generated ROIs; (3) automated classification of ROIs into true and false positives for a FP reduction purpose. Before explaining the extraction of our region-based stellate features, we briefly describe the ROI detection and segmentation method employed in this study.

In our work, as recommended in [9,20–22] to perform a more realistic assessment of a classification process under the assumption that the incorrect segmentation results would affect the classification performance, the ROIs are automatically detected and segmented from each mammogram by using a fully automated segmentation method. For this purpose, one popular approach of using multi-level thresholding algorithm [9,18,23] is adopted for segmenting masses on mammograms. We chose this segmentation approach because it has been well-documented in previous publications that it can provide “successful” segmentation results. The segmentation algorithm used consists of the following three sequential steps: (1) construction of iso-contour map; (2) formation of inclusion tree; (3) computation of minimum nesting depth. The implementation details of each step have been described in the literature [23].

Fig. 1 shows an example of segmented ROIs generated by segmentation algorithm used. Note that in Fig. 1, the red line is the successfully segmented mass contour with spiculated margins identified by the segmentation algorithm, and the blue line is the mass outline (as ground truth) marked by experienced radiologists. Note that the size of the ROI is chosen such that the computer-marked mass contour and a band of pixels surrounding a segmented object are included in the ROI (a detailed description of determining the width of the band is given in Section 3.1). The mass or normal tissue ROIs are then used as input for region-based stellate feature extraction to be discussed in the following section.

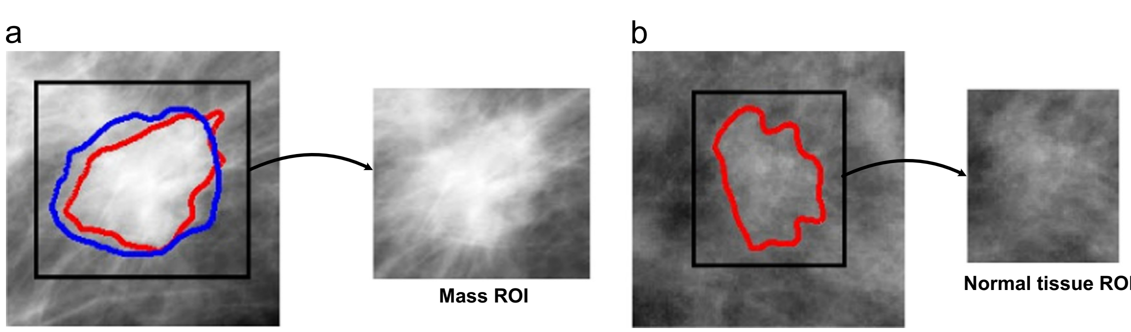

Fig. 1. (a) Mass ROI with spiculated margin. (b) Normal tissue ROI. See text for explanation.