Expert Systems with Applications 42 (2015) 3048–3055

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Automatic detection of abnormal mammograms in mammographic images

Chun-Chu Jen *, Shyr-Shen Yu

*Department of Computer Science and Engineering, National Chung Hsing University, 250 Kuo-Kuang Rd., Taichung 402, Taiwan*

ARTICLE INFO

*Article history:*
Available online 4 December 2014

*Keywords:*
Digital mammography
Preprocessing
Pectoral muscle suppression
Feature extraction
Principal component analysis

ABSTRACT

This paper proposes a detection method for abnormal mammograms in mammographic datasets based on the novel abnormality detection classifier (ADC) by extracting a few of discriminative features, first-order statistical intensities and gradients. As tumorous masses are often indistinguishable from the surrounding parenchyma, automatic mass detection on highly complex breast tissues has been a challenge. However, most tumor detection methods require extraction of a large number of textural features for further multiple computations. The study first investigates image preprocessing techniques for obtaining more accurate breast segmentation prior to mass detection, including global equalization transformation, denoising, binarization, breast orientation determination and the pectoral muscle suppression. After performing gray level quantization on the breast images segmented, the presented feature difference matrices could be created by five features extracted from a suspicious region of interest (ROI); subsequently, principal component analysis (PCA) is applied to aid the determination of feature weights. The experimental results show that applying the algorithm of ADC accompanied with the feature weight adjustments to detect abnormal mammograms has yielded prominent sensitivities of 88% and 86% on the two respective datasets. Comparing other automated mass detection systems, this study proposes a new method for fully developing a high-performance, computer-aided decision (CAD) system that can automatically detect abnormal mammograms in screening programs, especially when an entire database is tested.



## 1. Introduction

Breast cancer is the most common cancer in women worldwide; according to the World Health Organization, claiming the lives of hundreds of thousands of women are threatened each year. With respect to the early detection of breast cancer, mammography has been shown to be the most effective and reliable method (Davies & Dance, 1990; Lau & Bischof, 1991; Siddiqui, Anand, Mehrotr, Sarangi, & Mathur, 2005) for reducing morbidity and mortality. Developing a high-performance CAD system for the detection of breast cancer is crucial to assist radiologists and physicians. However, the presence of artifacts and pectoral muscle can disturb the detection of masses and reduce the rate of accuracy in the computer-aided analysis (CAA). To enhance the image quality of mammograms, filter or segmentation technology is usually the first image processing before application of the detection algorithm to detect suspicious lesions.

Earlier preprocessing methods for breast segmentation, which is a critical stage of breast mass analysis, were generally based on global thresholding (Davies & Dance, 1990; Lau & Bischof, 1991) or gradient analysis (Chandrasekhar & Attikiouzel, 1997; Karssemeijer, 1993). Ojala, Nappi, and Nevalainen (2001) presented an algorithm to segment the breast region from digitized mammograms, but the bright objects outside of the breast region may cause errors in this segmentation. A segmented breast generally consists of the whole breast, pectoral muscles and the nipple extraction. Therefore, the pectoral muscle regions should be removed before detecting the tumor cells so that mass detection can be performed more efficiently. Former work related to pectoral muscle suppression used the Hough transform (Kwok, Chandrasekhar, & Attikiouzel, 2001); Ferrari and Rangayyan (2004) proposed a polynomial modeling of the pectoral muscle. The contrast enhancement technique on mammogram images was used before suppression of pectoral muscle (Maitra, Nag, & Bandyopadhyay, 2011). To make the mass detection more effec-

* Corresponding author.
*E-mail addresses:* phd9717@cs.nchu.edu.tw (C.-C. Jen), pyu@nchu.edu.tw (S.-S. Yu).

http://dx.doi.org/10.1016/j.eswa.2014.11.061
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

tive, an image preprocessing that applies gamma correction equalization (Gonzalez & Woods, 2008) and Otsu's thresholding method (Otsu, 1979) is presented here to segment a suitable breast region with pectoral muscle suppression.

During mass detection, masses are often embedded in and camouflaged by the varying densities of breast tissue structures; moreover, mass shapes can be spiculated, circumscribed and ill-defined. Depending on the experience of the physician, breast cancer detection can be facilitated using computerized feature extraction algorithms. Sameti, Ward, Palcic, and Morgan-Parkes (1997) divided a mammogram into different regions of mass candidates; the discrete texture features were then calculated for the area of each mass candidate. The features were computed based on gray-level co-occurrence matrices (GLCM) that requires high computational loads, and the effectiveness of the textural information possessed by mass regions in comparison with the mass margins was evaluated (Mudigonda, Rangayyan, & Desautels, 2000). Bellotti et al. (2006) characterized regions of interest (ROIs) by means of textural features computed from the GLCM. The separation of normal regions from lesions with masses could be achieved using texture features (Roberto et al., 2006). Mohd Khuzi, Besar, Wan Zaki, and Ahmad (2009) also used GLCM that was constructed at four different directions for each ROI to extract the textural features. Features in the study of Yuan, Giger, Li, and Sennett (2008) were grouped into categories of texture features based on GLCM. Mohamed and Kadah (2007) extracted a large feature set and found that 78 of those features are capable of discriminating between normal and abnormal breast tissues in mammograms with a true positive (TP) rate of 83.3%. In this study, the abnormal ROI from the gray-level quantified ROIs in a mammogram can be distinguished merely by five extracted features containing 3 of first-order statistical intensities and 2 of gradients.

Over the past few years, many texture feature extraction methods have been proposed that use the calculated GLCM, and there has been an incremental computation cost of detection for large numbers of features; however, their performance is still not very promising. Here, the PCA (Pearson, 1901) technique is used to obtain the feature weights after extraction of a few representative features from the abnormal ROI in the quantized breast object. Furthermore, a novel ADC exploiting the multiple feature weight adjustments is proposed for identifying the abnormal mammograms in a database or a specified dataset. This paper not only proposes effective image preprocessing techniques and the classifier ADC based on a few of extracted features but outperforms other mentioned methods for detecting abnormal mammograms.

The rest of this paper is organized as follows: in Section 2, we specify the abnormal dataset and the databases used; in Section 3, the techniques for image preprocessing, feature extraction and classification are presented; in Section 4, the promising results are shown and compared to other mass-detection methods; and Section 5 presents some concluding remarks.

## 2. Image database

For the development and evaluation of the proposed method, the data used in the experiments was taken from the MIAS database (Suckling et al., 1994) consisting of 322 mammogram images in the mediolateral oblique (MLO) view and the Digital Database for Screening Mammography (DDSM) by Heath et al. (1998). The principal characteristics of the abnormalities were bright regions with irregular shapes and indistinct or spiculated margins. The authors specified set $S_1$, containing 113 abnormal mammograms accurately representing calcification, spiculated masses, circumscribed masses, ill-defined masses, architectural distortion, and asymmetry from the MIAS database, which was reduced to a resolution of 200 μm by pixel and clipped/padded so that every image was 1024 × 1024 pixels. Set $S_2$ consisted of 67 abnormal and 133 normal MLO-view mammograms selected from the DDSM database which has been downsized to 50% using bicubic interpolation and transformed into 256 gray levels with a size of approximately 2500 × 1500 pixels.

## 3. Methods

The mass detector (MD) CAD system for automatic detection of abnormal mammograms mainly consists of image preprocessing, mass detection and the classifier ADC. This image preprocessing of the MD system contains global equalization transformation, image denoising, binarization, breast object extraction, breast orientation determination, and pectoral muscle suppression. An overview of the research methodology is presented in Fig. 1.

### 3.1. Global equalization transformation

$I_{in}$ is the input image of the gray-level digital mammogram. To normalize the range of contrast variation among different images, all gray-level intensities of the pixels within $I_{in}$ will be stretched to the full range of 0–255. $I_{in}(x, y)$ is the intensity value of a pixel located at coordinates $(x, y)$ and the $min_{in}$ and $max_{in}$ are the minimal and maximal gray-level intensities in $I_{in}$, respectively. $I_{in}$ is transformed into $I_s$ by the following formula:

$$I_s(x,y) = 255\left(\frac{I_{in}(x,y) - min_{in}}{max_{in} - min_{in}}\right) \tag{1}$$

Consequently, the intensity range of $I_s$ is normalized to [0, 255].

### 3.2. Image denoising

The noise within a mammographic image could result in imprecise object extraction. Thus, the MD uses the mean filter (Gonzalez & Woods, 2008) to eliminate short-tailed noise such as uniform and Gaussian-type noise from $I_s$, and a neighborhood window size of 11 × 11 is given. $I_m$ is the denoised image obtained by removing the noise from $I_s$.

### 3.3. Binarization

It is necessary to first identify the breast region and remove the non-breast region and information plates to reduce subsequent processing calculations. To determine the breast object, the MD utilizes Otsu's thresholding method to find the optimal adaptive threshold $t_o$ corresponding to the intensity of $I_m$, and the value of $t_o$ is approximately 70–110; it hence specifies a soft-threshold $t_s = \alpha_s t_o$, for binarizing the $I_m$, where $\alpha_s$ is a given constant less than 1. If $I_m(x, y)$ is greater than or equal to $t_s$, then $I_b(x, y)$ is assigned to 1; otherwise, $I_b(x, y)$ is given a 0. As a result, a binary image $I_b$ shown as Fig. 2(a) is obtained with a given $\alpha_s = 0.6$, which can meet the aim of more complete object extraction.

### 3.4. Breast object extraction

The MD excludes the dark background and calculates the area of each disjoined object from $I_b$; it then specifies a flat, disk-shaped structuring element with a radius of 2 pixels that will be used in the study. All probable breast objects in $I_b$ are processed by the basic morphological operations, erosion and dilation (Serra, 1982). Consequently, the largest remaining object is obtained after $\alpha_m$ times erosion then dilates back $\alpha_m$ times: the dilated object is then extracted as the breast object and denoted by $I_o$ as shown in Fig. 2(b). Here, $\alpha_m$ is a given constant of 30. Moreover, the MD