Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A new feature extraction framework based on wavelets for breast cancer diagnosis

Q1 Semih Ergin [a,*], Onur Kilinc [b]

[a] Eskisehir Osmangazi University, Department of Electrical and Electronics Engineering, Meselik Campus, 26480 Eskisehir, Turkey
[b] University of Pardubice, Department of Transport Engineering, Koleje Pavilion B, 53009 Pardubice, Czech Republic

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 26 May 2014

Keywords:
Breast cancer
Feature extraction
Digital mammography
Computer aided diagnosis

ABSTRACT

This paper investigates a pattern recognition framework in order to determine and classify breast cancer cases. Initially, a two-class separation study classifying normal and abnormal (cancerous) breast tissues is achieved. The Histogram of Oriented Gradients (HOG), Dense Scale Invariant Feature Transform (DSIFT), and Local Configuration Pattern (LCP) methods are used to extract the rotation- and scale-invariant features for all tissue patches. A classification is made utilizing Support Vector Machine (SVM), k-Nearest Neighborhood (k-NN), Decision Tree, and Fisher Linear Discriminant Analysis (FLDA) via 10-fold cross validation. Then, a three-class study (normal, benign, and malignant cancerous cases) is carried out using similar procedures in a two-class case; however, the attained classification accuracies are not sufficiently satisfied. Therefore, a new feature extraction framework is proposed. The feature vectors are again extracted with this new framework, and more satisfactory results are obtained. Our new framework achieved a remarkable increase on the recognition performance for the three-class study.

1. Introduction

Q2 Breast cancer is a type of cancer that commonly occurs in the milk ducts or the lobules lying within breast tissue [31]. Breast cancer cases represent 30% of all cancer cases. Pal et al. [27] declared that, annually, more than a million women have breast cancer, and 400 thousand of those cases lead to death. This fatal disease affects not only developing countries but also developed countries. It is quite important to diagnose breast cancer in its early phases to prevent death. Cheng et al. [8] state that early diagnosis should not only include breast cancer detection but also specify whether the breast cancer is benign or malignant. It is very complicated to identify lesion (tissue disorder) type via a mammogram image, so even radiologists miss cancerous regions 10–30% of the time. Therefore, Computer-Aided Diagnosis (CAD) applications are necessary to aid specialists in carrying out more accurate diagnoses.

Despite all of the technological advancements in medical electronics, it is still hard to diagnose breast cancer circumstances on digitized mammograms because of the lack of distinction between benign and malignant tissue structures on mammograms. Rabottino et al. [28] mention that if more than one radiologist participates in the diagnosis process, the possibility of diagnosing an abnormal tissue could be increased by up to 10%. Generally, mammographic images have slight information about cancerous regions because there is only a small difference between normal and abnormal tissues due to X-ray permeability. It is especially difficult to understand cancerous regions for young women because of the high density of their tissues. Additionally, the X-ray attenuation over microcalcifications results in a low-contrast region on a mammogram, so especially small lesion areas can be harder to identify.

The classification of mammographic images is negatively affected by the presence of noise. X-ray dose and other medical equipment make the mammograms noisy, and this noise can be modeled as Gaussian additive noise, Poisson noise and multiplicative quantum noise. Naveed et al. [24] state that there could be up to a 21% reduction in mammogram quality due to noise. This decrease in quality reduces microcalcification diagnosis rate from 89% to 67%, and lesion presence detection rate from 93% to 79%. Romualdo et al. [29] used the Wiener filter following the Anscombe Transform to suppress quantum and Poisson noise. In addition, a neuro-fuzzy filter can be utilized to address same noise types, and this filter scheme is an effective way to denoise if the noise presence on a mammogram image is low [3]. However, if the

* Corresponding author. Tel.: +90 222 239 37 50x3265; fax: +90 222 229 05 35.
E-mail address: sergin@ogu.edu.tr (S. Ergin).
URL: http://sergin.ogu.edu.tr/ (S. Ergin).

http://dx.doi.org/10.1016/j.compbiomed.2014.05.008
0010-4825/

Please cite this article as: S. Ergin, O. Kilinc, A new feature extraction framework based on wavelets for breast cancer diagnosis, Comput. Biol. Med. (2014), http://dx.doi.org/10.1016/j.compbiomed.2014.05.008

amount of noise is high, this filter unfortunately discards critical information of the image. Eng and Ma [11] propose a Noise Adaptive Soft Switching Median filter, which is another spatial filter technique to preserve small details of images. Naveed et al. [24] has used the combination of the Frost Filter, Wiener Filter and Non-Local Means (NLM) Filter in order to eliminate noise on mammograms. Continuous Wavelet Transform (CWT) has also been used for the visual enhancement of microcalcifications [17]. Another proposed filtering methodology is an adaptive method comprising the noise matching and Wavelet shrinkage operations [18]. Only horizontal and vertical decompositions of Wavelets have been used, and Wavelet coefficients have been reproduced according to various scale-space constraints by this method. Although it is a flexible method, some parameters need to be determined by the user.

Image restoration, which is a crucial part of classification for mammogram images, includes noise filtering, the reduction of blurred areas and the improvement of the contrast of the image. This subject can be separated into two categories: conventional and blind restoration [20]. The conventional approach requires foreknowledge about the image. The amount of blur presence and distortion can be described by the conventional methods. However, it is not easy to have foreknowledge of images in practice; therefore, the blind restoration techniques are generally preferred. The main purpose of these methods is to save vital information of images by discarding unnecessary information.

The success of restoration and filtering is not enough to make an efficient classification of mammogram images. Feature extraction is the most important step of breast cancer diagnosis over CAD framework. This process can only be performed if the suspicious structures, tumor areas of benign and malignant lesions, are properly described. In addition to a convenient representation, the extracted features must also be affordable in size for feasibility and computational efficiency. Feature extraction of mammographic images needs space transformations to maintain those requirements. The Gabor Wavelet [5], Discrete Wavelet Transform (DWT) [13], Principal Component Analysis (PCA) [4], and Local Binary Pattern (LBP) [32] are some examples of the feature extraction methods of digitized mammograms. [1] have used samples from the Digital Database for Screening Mammography (DDSM) and Overcomplete Wavelet Transform (OWT) with Haar Wavelets to identify masses on mammograms, and they have achieved a 90% classification rate. Martins et al. [23] have suggested the spatial features as contrast, homogeneity, inverse-difference moment, entropy and energy from co-occurrence matrices to discriminate cancerous regions. Without a feature extraction step, the mammogram images have been directly classified with 80% accuracy using a cascaded Support Vector Machine (SVM) classifier on DDSM [7].

In this paper, initially, a two-class classification study separating normal and abnormal (cancerous) breast tissues is conducted. The rotational- and scale-invariant features are extracted by the HOG, DSIFT, and LCP descriptors for all tissue patches. A classification is made utilizing SVM, k-NN, Decision Tree, and FLDA classifiers via 10-fold cross validation. Then, a three-class (normal, benign, and malignant cancerous cases) study is achieved using similar procedures in a two-class case; however, the obtained classification accuracies are not sufficiently satisfied. Therefore, a new framework for the feature extraction phase is proposed. In this framework, an NLM filter [9] is first applied to all of the mammographic patches. Then, the feature vectors are extracted using HOG, DSIFT, and LCP descriptors. Finally, the newly constructed feature vectors are classified with SVM, k-NN, Decision Tree, and FLDA classifiers via 10-fold cross validation, and more satisfactory results are attained. Our new feature extraction framework achieved a noticeable increase the recognition rates. The rest of the paper is organized as follows. All feature extraction techniques and classifiers used in this study are explained in the second and third sections, respectively. Experimental works are included in the fourth section, while all discussions on the obtained experimental results are given in the fifth section. The last section comprises of conclusions.

2. Feature extraction

2.1. Histogram of Oriented Gradients

The Histogram of Oriented Gradients (HOG) descriptor [39] is based on the distribution of a differential intensity histogram of an image. Each image is divided into non-overlapping uniform cells, as shown in Fig. 1. Intensity variations, which belong to differentials for a desired orientation, are calculated for different directions in each cell.

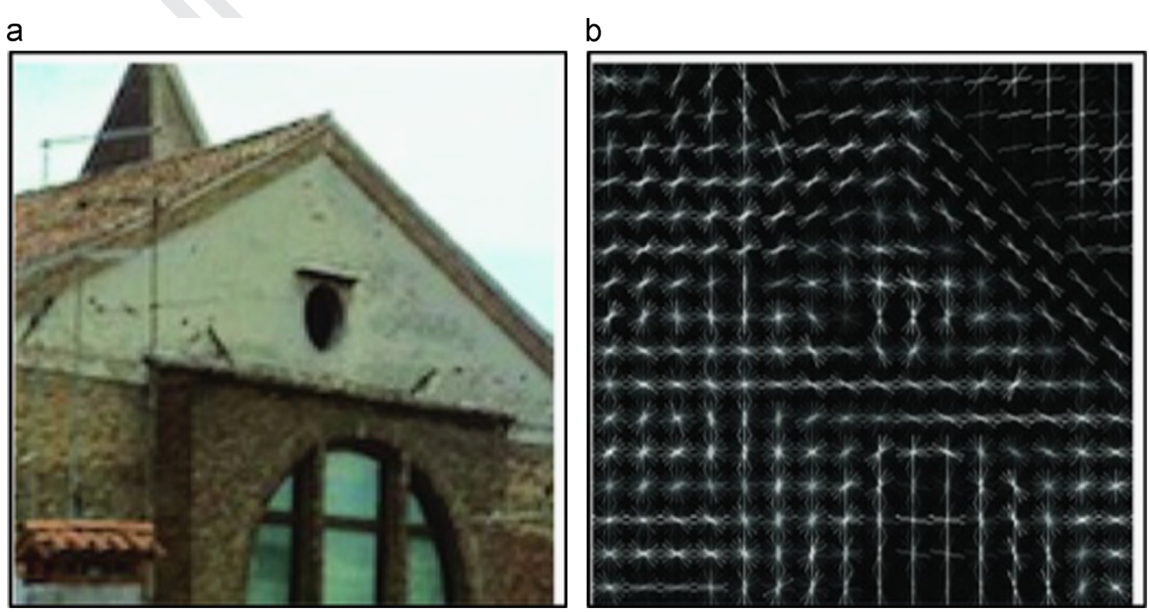

Fig. 1. (a) An original image, and (b) the feature blocks obtained by applying an 8 × 8 cell-sized HOG descriptor on the image given in (a).

Please cite this article as: S. Ergin, O. Kilinc, A new feature extraction framework based on wavelets for breast cancer diagnosis, Comput. Biol. Med. (2014), http://dx.doi.org/10.1016/j.compbiomed.2014.05.008