Computers in Biology and Medicine 43 (2013) 1563–1572

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images

Marek Kowal [a,*], Paweł Filipczuk [a], Andrzej Obuchowicz [a], Józef Korbicz [a], Roman Monczak [b]

[a] Institute of Control and Computation Engineering, University of Zielona Góra, ul. Ogrodowa 3b, 65-246 Zielona Góra, Poland
[b] Department of Pathomorphology, Regional Hospital in Zielona Góra, Zielona Góra, Poland

ARTICLE INFO

Article history:
Received 5 July 2012
Accepted 5 August 2013

Keywords:
Computer-aided diagnosis
Breast cancer
Machine learning
Image segmentation

ABSTRACT

Prompt and widely available diagnostics of breast cancer is crucial for the prognosis of patients. One of the diagnostic methods is the analysis of cytological material from the breast. This examination requires extensive knowledge and experience of the cytologist. Computer-aided diagnosis can speed up the diagnostic process and allow for large-scale screening. One of the largest challenges in the automatic analysis of cytological images is the segmentation of nuclei. In this study, four different clustering algorithms are tested and compared in the task of fast nuclei segmentation. K-means, fuzzy C-means, competitive learning neural networks and Gaussian mixture models were incorporated for clustering in the color space along with adaptive thresholding in grayscale. These methods were applied in a medical decision support system for breast cancer diagnosis, where the cases were classified as either benign or malignant. In the segmented nuclei, 42 morphological, topological and texture features were extracted. Then, these features were used in a classification procedure with three different classifiers. The system was tested for classification accuracy by means of microscopic images of fine needle breast biopsies. In cooperation with the Regional Hospital in Zielona Góra, 500 real case medical images from 50 patients were collected. The acquired classification accuracy was approximately 96–100%, which is very promising and shows that the presented method ensures accurate and objective data acquisition that could be used to facilitate breast cancer diagnosis.

1. Introduction

Breast cancer is the most common cancer type among women [1]. According to the National Cancer Registry in Poland, there were 15,752 diagnosed cases in 2009, 5242 of which were terminal cases. Moreover, since the 1980s, an increase in diagnosed cases by 3–4% per year has been observed. Early diagnosis of the disease increases the chances of full recovery substantially. To diagnose breast cancer, the so-called triple-test is often used. This test is based on three medical examinations: palpation, mammography or ultrasonography imaging, and fine needle biopsy (FNB) [2]. In FNB, a cytological sample is obtained directly from the tumor with a fine needle; the sample is then the subject of microscopic examination. This method is painless; therefore, anesthesia is not required. The relatively rapid process of sample extraction results in hospitalization being unnecessary. Unfortunately, the analysis of extracted samples requires a substantial amount of experience and expertise. Computer-aided quantitative analysis can produce objective results, greatly improve the accuracy of diagnosis, and support inexperienced specialists. In recent years, one can notice a significant increase of interest in digital pathology and cytology [3–6].

In previous studies [7–10], an effective hybrid segmentation method for breast biopsy analysis was proposed. This method is based on a combination of adaptive thresholding in grayscale and clustering in color space. In this study, four different clustering approaches are used, and their results are compared. The methods are applied in a medical decision support system. From the nuclei isolated with each of these methods, 42 features were extracted. These features were then inputted as data for classification with three different classifiers. The classification accuracy obtained with all of the incorporated methods is outlined and discussed later in this study. Additionally, a set of candidate features for cytological analysis is proposed, and a comparison of the applied approach with methods proposed by other researchers is presented.

This study is divided into seven sections. Section 1 is an introduction to breast cancer diagnosis. Section 2 is an overview of the system, and also explains the acquisition process for the medical images that are used for the testing. Sections 3, 4 and 5 describe preprocessing, segmentation and classification, respectively, in detail. These actions are consecutive steps of the proposed medical decision support system. Section 6 presents the

* Corresponding author. Tel.: +48 683283217; fax: +48 683284751.
E-mail address: M.Kowal@issi.uz.zgora.pl (M. Kowal).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.08.003

outcomes of applying the proposed approach to the experimental data. The results are discussed in Section 7. The conclusions and bibliography are in the final section.

2. Overview of the system

The presented approach is based on the analysis of microscopic images from FNB samples. The purpose of the system is to diagnose the malignancy of a given tumor. A series of images from a single smear are the input. The images are enhanced in the preprocessing phase, which includes vignetting and noise removal as well as histogram stretching. Next, nuclei identification is conducted with one of the segmentation methods described further in this study. Several morphometric, topological and texture features are extracted from the nuclei. Lastly, these inputted features enable the system to classify the case as benign or malignant. A general scheme of the system is presented in Fig. 1.

During the development phase of a medical decision support system, it is important to test the system on real medical data. In this research, a database of 500 images of cytological samples that were obtained from FNB was used. The samples were extracted from 50 patients of the Regional Hospital in Zielona Góra. There are 10 images per case, which was the number recommended by specialists from the hospital [11,12] and which allows a pathologist to conduct a reliable diagnosis. The set contained 25 benign and 25 malignant cases. Biopsy without aspiration was performed under the control of an ultrasound scanner with a 0.5 mm diameter needle. Smears in the samples were fixed with spray fixative (Cellfix of Shandon Company) and dyed with hematoxylin and eosin (h+e). The amount of time between the smear preparations and their preservation in the fixative never exceeded 3 s. The images were recorded with a SONY CCD IRIS color video camera mounted on top of the AXIOPHOT microscope. The slides were projected into the camera with 40× and 160× lenses and a 2.5× ocular lens. Each case contained 1 overall image generated at a 100× enlargement and 9 images generated at a 400× enlargement. The captured areas were selected by a pathologist and contained at least 10 nuclei each. The images were BMP files, 704×578 pixels, 8 bit/channel RGB. All of the cancers were histologically confirmed, and all of the patients who carried benign disease were either biopsied or followed for a year. Fig. 2 shows samples of typical benign and malignant images.

3. Image preprocessing

The first step of the diagnostic procedure is preprocessing. The microscopic images must be enhanced due to their low quality. First, the noise caused by the capturing camera is reduced. This reduction is performed with a Gaussian filter:

$$h_1 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{1}$$

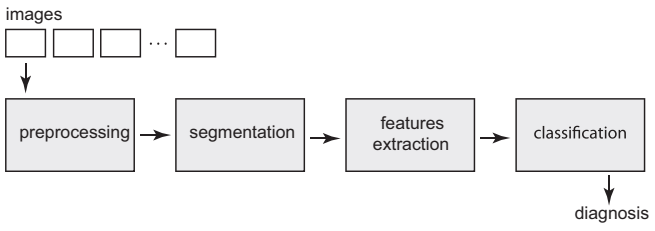

Fig. 1. General scheme of the system.

to blur the images; then, the images are sharpened back with the following sharpening filter:

$$h_2 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{2}$$

The next step is related to handling the vignetting caused by the microscope optics; this distortion is removed with a blank slide as a reference. Lastly, the histogram is stretched to enhance the contrast, and the images are cropped to the size 697×569 to remove the frame and other artifacts that might be observed on the edges of the images. Fig. 3 shows an example image before and after the preprocessing.

4. Methods of nuclei segmentation

Tumor classification as benign or malignant requires nuclei identification in the cytological image. This task is challenging because the images usually contain many clustered and overlapping objects. Several researchers already attempted to address this problem and reported difficulty in the segmentation of clumped nuclei structures [11,13–22].

Furthermore, cytological images from various sources can vary significantly depending on the imaging method or smear preparation. Consequently, attempts to generalize the segmentation methods proposed in the literature usually fail because certain approaches are feasible only for specific cytological images.

Errors that arise from inadequate segmentation of nuclei introduce significant distortion in the nuclei features. Eventually, this distortion results in low classification accuracy. To solve this problem, a two-stage segmentation procedure was proposed (see Fig. 4). First, adaptive thresholding is used for background–foreground segmentation. Second, a clustering algorithm is applied to identify nuclei in the remaining foreground objects (red blood cells and others). Four clustering methods used for nuclei segmentation are compared in terms of their accuracy in tumor classification.

The previous paragraph provided general guidelines of the proposed segmentation approach; next, some implementation details of this approach will be explained. The complete procedure starts from the image preprocessing step described in Section 3. Then, an RGB color image E is converted into a grayscale image G. At this point, a background–foreground segmentation procedure is initiated. The image is segmented into two subsets by means of the thresholding procedure. The key problem of thresholding segmentation is the selection of the correct threshold value t. There are multiple methods for computing the correct threshold value [23–26]. Most of these methods are based on histogram analysis and can be divided into global and local methods. A global threshold is calculated once and is then applied to the whole image. The most popular global methods are based on the Otsu criterion. Unfortunately, this method cannot be applied to detect subtle brighter borders between clustered nuclei, and frequently, it produces very large spurious nuclei that are composed of several actual nuclei (see Fig. 5). To overcome this problem, the threshold t is calculated adaptively for subsequent pixels of the image with the averaging filter. Each pixel has its own threshold $t_{i,j}$, which is calculated as follows:

$$t_{i,j} = \frac{1}{m^2} \sum_{k=-n}^{n} \sum_{l=-n}^{n} g_{i+k,j+l}, \tag{3}$$

where $g_{i,j}$ is the pixel luminance value, $n=(m-1)/2$ and m is the size of the filter window (an odd integer). It is assumed that, for non-existent pixels that are outside the bounds of a given image, the values of the nearest boundary pixels are taken. Lastly, a binary image is obtained after the thresholds to the corresponding pixels have been applied by means of the adaptive thresholding, which