BREAST FIBROADENOMA AUTOMATIC DETECTION USING K-MEANS BASED HYBRID SEGMENTATION METHOD

Pawel Filipczuk¹, Marek Kowal¹ and Andrzej Obuchowicz¹

¹Institute of Control & Computation Engineering, University of Zielona Gora, Podgorna 50, Zielona Gora, Poland

{P.Filipczuk, M.Kowal, A.Obuchowicz}@issi.uz.zgora.pl

This research was partially supported by National Science Centre in Poland

Keywords: Breast cancer, Fibroadenoma, Computer diagnosis, Segmentation

Abstract: Fibroadenoma is a benign tumor having some features similar to malignant one. The aim of this study was to examine the impact of fibroadenoma cases on the results of automatic breast cancer diagnostic system based on quantitative morphometric analysis of fine needle biopsy microscopic images. Database of 50 patients (500 images) of benign and malignant lesions was enriched by an additional 25 patients (250 images) of fibroadenoma cases. Experiments were performed using k-means based hybrid segmentation method. The system was tested on a set of real case medical images with promising results.

1 INTRODUCTION

According to the National Cancer Registry in Poland breast cancer is the most common cancer among women. In 2008, there were 14,576 diagnosed cases of breast cancer in Polish women. Out of these cases, 5362 deaths were the result. There has also been an increase of breast cancer by 3-4% a year since the 1980’s. The effectiveness of treatment largely depends on early detection of the cancer. Important and often used diagnostic method is so-called triple-test. It is based on three medical examinations and allows to achieve high confidence of diagnosis. The triple-test includes self examination (palpation), mammography or ultrasonography imaging and FNB (Fine Needle Biopsy) (Underwood, 1987). FNB is collecting nucleus material directly from tumor. Obtained material is examined under a microscope to determine the prevalence of cancer cells. The present approach requires a deep knowledge and experience of the cytologist responsible for diagnosis. Automatic morphometric diagnosis can make the decision objective and assist unexperienced specialist. It can also allow screening on a large scale where only difficult and uncertain cases would require additional human diagnosis. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for the detection of diseases, infections as well as many other disorders (Gurcan et al., 2009; Śmietanski et al., 2010). In the literature one can find approaches to breast cancer classification (Gil et al., 2002; Filipczuk et al., 2010; Hrebień et al., 2010; Jeleń et al., 2010; Kowal and Korbicz, 2010; Wolberg et al., 1993). Mentioned approaches are concentrated on classifying FNA (Fine Needle Aspiration) or FNB slides as benign or malignant.

Fibroadenoma is a benign tumor of the breast often occurring in women. Despite the fact it is not cancerous it might have some morphometric features similar to malignant neoplasm. This may confuse the system and cause incorrect automatic diagnosis. In this paper we present an automatic method that allows distinguish between malignant, benign and fibroadenoma as a special benign case. The approach is based on k-means and adaptive thresholding hybrid segmentation algorithm. The diagnosis procedure is described in detail in Section 3. All experiments were performed using real case data in cooperation with the Regional Hospital in Zielona Gora.

The paper is divided into four sections. Section 1 gives an overview of breast cancer diagnosis techniques. Section 2 describes the process of acquisition of images used to breast cancer diagnosis. Section 3 deals with segmentation algorithm used to separate cells and extract features. Section 4 shows the experimental results obtained using the proposed approach. The last part of the work includes conclusions and bibliography.
2 MEDICAL IMAGES DATABASE

The database contains 750 images of the cytological material obtained by FNB. The material was collected from 75 patients of the clinic in Zielona Góra. It gives 10 images per case which was recommended amount by specialists from the Regional Hospital in Zielona Góra (Filipczuk et al., 2010; Marciniak et al., 2005). This number of images per single case allows correct diagnosis by a pathologist. The set contains 25 benign, 25 malignant and 25 fibroadenoma cases. Biopsy without aspiration was performed under the control of ultrasonograph with a 0.5 mm diameter needle. Smears from the material were fixed in spray fixative (Cellfix of Shandon company) and dyed with hematoxylin and eosin (h+e). The time between preparation of smears and their preserving in fixative never exceeded three seconds. The images were recorded by SONY CDD IRIS color video camera mounted atop an AXIOPHOT microscope. The slides were projected into the camera with 40x and 160x objective and a 2.5x ocular. One image was generated for enlargement 100x and nine for enlargement 400x. Images are BMP files, 704x578 pixels, 8 bit/channel RGB. All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year.

3 AUTOMATIC DIAGNOSIS PROCEDURE

Most automatic diagnostic systems are based on similar configuration of several steps. At the beginning images are adjusted in preprocessing phase. Then objects of interest are extracted from the images in segmentation step, which is the most challenging task. The quality of segmentation determines effectiveness of further operations. For separated objects morphometric features are calculated. Finally, objects are classified, eg. as an benign or malignant case. Our approach follows this scheme.

At first images require preprocessing due to their low quality. This phase includes histogram stretching and noise reduction. Also, the images are affected by vignetting caused by microscope optics. It is removed using a blank slide as a reference.

In literature, there have been presented many different approaches to extract cells from microscope images (Al-Kofahi et al., 2010; Clocksin, 2003; Cloppet and Boucher, 2008; Filipczuk et al., 2010; Herbie et al., 2010; Jelefi et al., 2010; Kowal and Korbicz, 2010; Marciniak et al., 2005; Obuchowicz et al., 2008; Peng et al., 2010). This task is usually done automatically, using one of the well known methods of image segmentation (Gonzalez and Woods, 2001; Sezgin and Sankur, 2003; Suri et al., 2002). Unfortunately, attempt generalize segmentation approaches proposed in literature usually fails because such methods work correctly only for specific images. Slides from various sources may vary significantly depending on the method of smear preparation. Moreover, cells tends to cluster and overlap together and their boundaries are blurred. In order to deal with these problems, we have developed automatic segmentation procedure that integrates results of image segmentation from two different methods. Firstly, proposed algorithm uses adaptive thresholding. The result is relatively finely separated nuclei from background. Unfortunately, any other dark objects, e.g. erythrocytes, are marked as well. Because of that original images are also clustered using k-means algorithm. This task is performed based on color information. It was decided to define three types of objects respectively to their color: nuclei, red blood cells and background. The idea of the image segmentation using clustering algorithms boils down to a search for clusters of pixels in color space. Derived clusters represent objects that are characterized by similar color. In the considered case k-means algorithm was applied to calculate centers for 3 clusters and to determine pixel assignments. The clustering procedure of k-means algorithm is based on minimizing the within-cluster sum of squares for K clusters:

\[ J = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{K} D_{x,y,k}^2, \]  

where \( X \) and \( Y \) defines the size of the analyzed image, \( \mu_{x,y,k} \) is a function specifying whether \((x,y)\)-th pixel belongs to the \(k\)-th cluster, \( D_{x,y,k} \) is an Euclidean distance measure:

\[ D_{x,y,k}^2 = (c_{x,y} - v_k)^T (c_{x,y} - v_k), \]

where \( c_{x,y} \in \mathbb{R}^3 \) is a vector of the coordinates of the \((x,y)\)-th pixel in RGB space and \( v_k \in \mathbb{R}^3 \) is a vector of the coordinates of the \(k\)-th cluster center in RGB space. K-means clustering procedure iteratively change pixel assignments based on distance to nearest mean (cluster center) and updates the cluster centers to match the proper means of the data points that they are responsible for. Detailed expressions for iterative updating the cluster centers and pixel assignments can be found in the following papers (Kanungo et al., 2002; Lloyd, 1982; MacKay, 2003). K-means clustering gives correctly but rather inaccurately separated nuclei from background and other objects. In order to combine advantages of both methods the final result is multiplication of obtained segmentation images.
Table 1: Result classification rate for selected sets of features (m and v in brackets are mean and variance respectively)

<table>
<thead>
<tr>
<th>Set of features</th>
<th>Classifier</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 classes: benign, fibroadenoma, malignant</td>
<td>KNN</td>
<td>88.00%</td>
</tr>
<tr>
<td>area (m), area (v), distance to centroid (m), luminance mean (m), major axis (v), perimeter (v)</td>
<td>Naive Bayes</td>
<td>81.33%</td>
</tr>
<tr>
<td>distance to centroid (m), eccentricity (m), luminance mean (m), luminance variance (m), minor axis (m)</td>
<td>Decision trees</td>
<td>85.33%</td>
</tr>
<tr>
<td>area (m), luminance mean (m), luminance variance (m), major axis (m), minor axis (v), perimeter (m), perimeter (v)</td>
<td>Ensemble classifier</td>
<td>88.00%</td>
</tr>
<tr>
<td>area (v), distance to centroid (m), luminance variance (m), minor axis (m), perimeter (m)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 classes: benign plus fibroadenoma, malignant

<table>
<thead>
<tr>
<th>Set of features</th>
<th>Classifier</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>luminance mean (v), luminance variance (v), perimeter (v)</td>
<td>KNN</td>
<td>90.67%</td>
</tr>
<tr>
<td>luminance mean (m), minor axis (m), perimeter (v)</td>
<td>Naive Bayes</td>
<td>90.67%</td>
</tr>
<tr>
<td>area (m), distance to centroid (m), major axis (v), perimeter (v)</td>
<td>Decision trees</td>
<td>94.67%</td>
</tr>
<tr>
<td>area (m), area (v), perimeter (m)</td>
<td>Ensemble classifier</td>
<td>90.67%</td>
</tr>
</tbody>
</table>

2 classes: benign, malignant

<table>
<thead>
<tr>
<th>Set of features</th>
<th>Classifier</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>area (m), distance to centroid (v), minor axis (m), minor axis (v)</td>
<td>KNN</td>
<td>100.0%</td>
</tr>
<tr>
<td>distance to centroid (m), eccentricity (v), perimeter (v)</td>
<td>Naive Bayes</td>
<td>94.00%</td>
</tr>
<tr>
<td>distance to centroid (v), major axis (v), perimeter (m)</td>
<td>Decision trees</td>
<td>98.00%</td>
</tr>
<tr>
<td>minor axis (m), perimeter (m)</td>
<td>Ensemble classifier</td>
<td>98.00%</td>
</tr>
</tbody>
</table>

The next step is feature extraction. For each cell following features were extracted: area, perimeter, eccentricity, major axis length, minor axis length, luminance mean, luminance variance and distance from the centroid of all nuclei on the image. For each slide mean and variance of certain features were computed. Finally all input variables were normalized.

It was decided to use four well known classification algorithms: k-nearest neighbor, naïve Bayes classifier, decision trees and classifiers ensemble (Breiman et al., 1993; Mitchell, 1997). However, it must be mentioned that ensemble of classifiers is not a separate classification technique and its classification procedure is based on the results of other classifiers used in the experiments. The answer of classifiers ensemble is determined by voting procedure.

4 Experimental Results

The system was tested for the recognition rate, which is defined as a percentage of successfully recognized cases to the total number of all cases. The rate was calculated for three configurations: 3 separate classes (benign, fibroadenoma and malignant), 2 separate classes (benign with fibroadenoma and malignant) as fibroadenoma is benign tumor and 2 classes (malignant and benign) without fibroadenoma for comparison purpose.

All experiments were performed using real case data. There was 75 patients: 25 malignant, 25 benign and 25 fibroadenoma. Each patient was represented by 10 images. There was one overall image and nine images in maximum enlargement. In the processing only the nine were used. The effectiveness was tested using the leave-one-out validation technique (Mitchell, 1997), where full set of 9 images representing 1 patient was 1 case.

To find the optimal set of features is was applied sequential forward selection, full search for a maximum of 3 features in set and then again forward selection using the result of full search as a starting point. Selected results are presented in Table 1.

5 Conclusions

The aim of the experiment was to test whether fibroadenoma cases will significantly worse the recognition rate. Tests showed that deterioration of the results is relatively small. The rate in case of 3 classes is about 15% worse in average, but mainly because of
additional class. In case where fibroadenoma is in benign class the rate is only 2-4% worse and amounts above 90%. It must be mentioned that the results concerns patients, not single images. Average rates for single images are almost the same for case with fibroadenoma (2 classes case) and without it. Presented segmentation approach is similar to our previously presented fuzzy c-means and competitive neural networks methods and meets the same problems. Nuclei are often joined. For relatively large cells their centers happen to be removed. Also, the algorithm tends to generate uncorrect clusters when nuclei are represented by very few pixels on the image. Despite that, the classification rates are very promising and shows that using relatively simple methods one might achieve satisfactory results. However, the rate still might be improved by detecting and splitting overlapped cells and addition of more sophisticated features not tested during current investigations. Another challenge will be applying the system for virtual slides generated by virtual scopes which are able to produce images with extremely high resolutions reaching 9 gigapixels and more. Such huge slides require completely new way of analysis.

REFERENCES


