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

Paweł Filipczuk, Marek Kowal and Andrzej Obuchowicz

University of Zielona Góra
Institute of Control & Computation Engineering
Podgórna 50, Zielona Góra, Poland

ABSTRACT

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

Index Terms— Image analysis, image segmentation, breast cancer, fibroadenoma

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. An important and often used diagnostic method is the so-called triple-test. It is based on 3 medical examinations and is used to achieve a high confidence of diagnosis. The triple-test includes self examination (palpation), mammography or ultrasonography imaging and FNB (Fine Needle Biopsy) [1]. FNB is an examination that consists of obtaining cytological material directly from the tumor. Collected material is then examined under a microscope to determine the prevalence of cancer cells. This present approach requires a deep knowledge and experience of the cytologist responsible for the diagnosis. Automatic morphometric diagnosis can make the results objective and assist inexperienced specialists. It also allows screening on a large scale where only difficult and uncertain cases would require additional human attention. 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 [2, 3]. In the literature one can also find approaches to the breast cancer classification [4, 5, 6, 7, 8].

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 a malignant neoplasm. This may confuse the system and cause incorrect automatic diagnosis. In this paper we present the results of the automatic method that allows to distinguish between malignant, benign and fibroadenoma as a special benign case. The approach is based on k-means and adaptive thresholding hybrid segmentation algorithm. All experiments were performed using real case medical data in cooperation with the Regional Hospital in Zielona Góra.

The paper is divided into 4 sections. Section 1 gives an overview of breast cancer diagnosis techniques. Section 2 describes the process of acquisition of images used to test the system. Section 3 deals with the procedure of automatic diagnosis. Section 4 shows the experimental results obtained using the proposed approach. The last part of the work includes conclusions and a 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 Regional Hospital in Zielona Góra. It gives 10 images per case which was the recommended amount by the specialists from the hospital [5, 9] and allows for a 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 a 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 3 seconds. The images were recorded by a SONY CDD IRIS

This research was partially supported by National Science Centre in Poland.

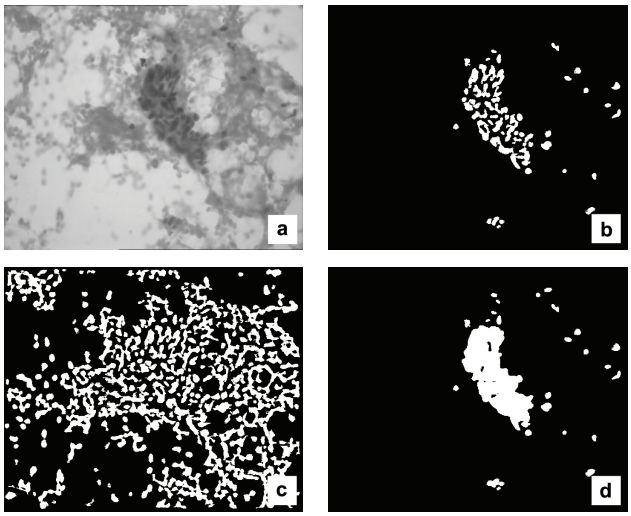


Fig. 1. Sample image (a), final segmentation (b), adaptive thresholding (c) and k-mean clustering (d).

color video camera mounted atop AXIOPHOT microscope. The slides were projected into the camera with 40x and 160x objective and a 2.5x ocular. Each case consisted of 1 overall image generated for enlargement 100x and 9 for enlargement 400x. Images are BMP files, 704x578 pixels, 8 bit/channel RGB. All cancers were histologically confirmed and all patients with a benign tumor were either biopsied or followed for a year.

3. AUTOMATIC DIAGNOSIS PROCEDURE

At first the 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.

The next step is image segmentation. In literature, there have been many different approaches presented to extract cells from microscope images [10, 11, 12, 5, 6, 7, 9, 13]. This task is usually done automatically, using one of the well known methods of image segmentation [14, 15]. Unfortunately, attempts to 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 blur. In order to deal with these problems, we have developed an automatic segmentation procedure that integrates results of image segmentation from two different methods.

First, the RGB microscope image is converted to gray scale and an adaptive thresholding segmentation is employed to separate objects from background. In the considered approach the threshold value is calculated for each pixel and equals the mean intensity of pixels from neighborhood defined by square of size 21x21 pixels. The result is a relatively well isolated objects. Unfortunately, erythrocytes and nuclei have similar luminance and it is very difficult to distinguish between them using this method. K-means clustering is used to tackle this drawback.

The second step of segmentation is performed in RGB. There are 3 types of objects defined respectively by 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 colored space. Derived clusters represent objects that are characterized by similar color. In the considered case k-means algorithm were 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:

$$J = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{K} \mu_{x,y,k} D^2_{x,y,k}, \tag{1}$$

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, $D^2_{x,y,k}$ is an Euclidean distance measure:

$$D^2_{x,y,k} = (\boldsymbol{c}_{x,y} - \boldsymbol{v_k})^T (\boldsymbol{c}_{x,y} - \boldsymbol{v_k}), \tag{2}$$

where $\boldsymbol{c}_{x,y} \in \mathbb{R}^3$ is a vector of the coordinates of the (x, y)-th pixel in RGB space and $\boldsymbol{v}_k \in \mathbb{R}^3$ 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 [16, 17, 18]. K-means clustering allows to distinguish nuclei from erythrocytes but the objects tend to merge and their edges are inaccurate. In order to combine advantages of both methods the final result is multiplication of obtained segmentation images. Fig.1 presents sample image (a), result of adaptive thresholding (c), result of k-means clustering (d) and final segmentation image (b).

The next step is feature extraction. For each cell the following features were extracted: area (the actual number of pixels of the nucleus), perimeter (the distance between each adjoining pair of pixels around the border of the nucleus), eccentricity (the scalar that specifies the ratio of the distance between the foci of the ellipse that has the same second-moments as the segmented nucleus and its major axis length), major axis length (the length of the major axis of the ellipse that has the same normalized second central moments as the nucleus), minor axis length (the length of the minor axis of the ellipse that has the same normalized second central moments as the nucleus), luminance mean (the mean luminance of pixels of the nucleus), luminance variance (the variance

Classifier	Only benign and malignant cases (2 classes)		Fibroadenoma as a benign case (2 classes)		Fibroadenoma as a separate case (3 classes)	
	Maximum	Average	Maximum	Average	Maximum	Average
KNN	100% (85.78%)	80.52% (70.04%)	90.67% (84.44%)	77.20% (70.63%)	88.00% (73.93%)	61.61% (53.49%)
Naive Bayes	94% (81.33%)	79.05% (71.02%)	90.67% (84.44%)	75.8% (70.26%)	81.33% (70.52%)	61.66% (55.20%)
Decision trees	98% (84.22%)	80.92% (69.17%)	94.67% (82.67%)	76.92% (68.69%)	85.33% (70.52%)	59.41% (51.36%)
Voting procedure	98% (85.33%)	82.02% (71.61%)	90.67% (85.33%)	77.89% (71.76%)	86.67% (75.26%)	63.67% (55.06%)

Table 1. Classification effectiveness for 3 configurations: 2 classes without fibroadenoma cases, 2 classes with fibroadenoma as benign cases and 3 separate classes. Presented values are the recognition rates concerning patients (full set of all 9 images belonging to the patient with voting procedure applied). Values in brackets concerns individual images.

of luminance of pixels of the nucleus) and distance from the centroid (the distance between the geometric center of the nucleus and centroid of all nuclei). For each slide mean and variance of each feature were computed. Finally all input variables were normalized.

The final step is classification. It was decided to use 3 well known classification algorithms: k-nearest neighbor (for k=7), naive Bayes classifier and decision trees [19, 20]. Additionally, we used a voting procedure to obtained the fourth classification result. The procedure involved the use of results from 3 previously mentioned classifiers and provided responses as a result of the majority voting.

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 3 configurations: 3 separate classes (benign, fibroadenoma and malignant), 2 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 medical data. There were 75 patients: 25 malignant, 25 benign and 25 fibroadenoma. Each patient was represented by 10 images. There was 1 overall image and 9 images in maximum enlargement. In the processing only the 9 were used. Since the data set was relatively small the effectiveness was tested using the leave-one-out cross-validation technique [20], where a single case was full set of 9 images representing 1 patient. That means the images belong to the same patient were never at the same time in the training and testing set. The final diagnosis was obtained by a majority voting of the classification of individual images belonging to the patient (e.g. if 5 images were classified as benign and 4 as malignant then the final diagnosis for the patient would be benign).

For each classifier in each configuration an optimal set of features has been determined. To find the set there was applied sequential forward selection, full search for a maximum of 3 features in set and then again forward selection using the result of the full search as a starting point. The sets were used to define the maximum rates. Also, there were calculated average rates using all combinations of 2 and 3 features (the number of available features were 16, mean and variance of 8 features described in section 3, so the total number of combinations was 680). The results are presented in Table 1.

5. CONCLUSIONS

The aim of the experiment was to test whether fibroadenoma cases would alter the recognition rate for malignant tumors. Although experts from the hospital claimed that fibroadenoma has some features similar to malignant cases this was not reflected in our research. Tests have shown that deterioration of the results is relatively small. Considering the best results achieved for each configuration the difference between the 2 cases, which consisted of 2 classes, with and without fibroadenoma, was 5.83% in average for all classifiers. The difference between average results was 3.68%. Of course, the configuration with 3 classes gave results that was much worse (12.17% worse considering the best results achieved and 19.04% worse considering average results), but mainly because of the additional class.

Presented segmentation approach is based on similar ideas to our previously presented fuzzy c-means and competitive neural networks methods. The classification rates are very promising and show that using relatively simple methods one might achieve satisfactory results. However, the algorithm still meets with some problems. It tends to generate incorrect clusters when nuclei are represented by very few pixels on the image. Also, the nuclei merge sometimes, and for relatively large cells their centers are partially removed. That might be improved by detecting and splitting overlapped cells in addition to more sophisticated features not tested dur-

ing current research. Another challenge will be applying the system to virtual slides generated by virtual scopes which are able to produce images with extremely high resolutions reaching 9 gigapixels and more. Such huge slides will require a completely new way of analysis.

6. ACKNOWLEDGMENTS

The co-author is a scholar within Sub-measure 8.2.2 Regional Innovation Strategies, Measure 8.2 Transfer of knowledge, Priority VIII Regional human resources for the economy Human Capital Operational Programme co-financed by European Social Fund and state budget.



7. REFERENCES

[1] J. C. E. Underwood, *Introduction to biopsy interpretation and surgical pathology*, Springer-Verlag, London, 1987.

[2] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[3] J. Śmietański, R. Tadeusiewicz, and E. Luczyńska, "Texture analysis in perfusion images of prostate cancer - a case study," *Int. J. Appl. Math and Comp. Sci.*, vol. 20, no. 1, pp. 149–156, 2010.

[4] J. Gil, H. Wu, and B. Y. Wang, "Image analysis and morphometry in the diagnosis," *J. Microsc. Res. Tech.*, vol. 59, pp. 109–118, 2002.

[5] P. Filipczuk, M. Kowal, and A. Marciniak, "Feature selection for breast cancer malignancy classification problem," *J. Medical Informatics & Technologies*, vol. 15, pp. 193–199, 2010.

[6] M. Hrebień, P. Steć, A. Obuchowicz, and T. Nieczkowski, "Segmentation of breast cancer fine needle biopsy cytological images," *Int. J. Appl. Math and Comp. Sci.*, vol. 18, no. 2, pp. 159–170, 2010.

[7] L. Jeleń, T. Fevens, and A. Krzyżak, "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," *Int. J. Appl. Math and Comp. Sci.*, vol. 18, no. 1, pp. 75–83, 2010.

[8] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cytology diagnosis via digital image analysis," *Analytical and Quantitative Cytology and Histology*, vol. 15, pp. 396–404, 1993.

[9] A. Marciniak, A. Obuchowicz, A. Monczak, and M. Kołodziński, "Cytomorphometry of fine needle biopsy material from the breast cancer," in *Proc. 4th Int. Conf. on Computer Recognition Systems CORES' 05*, 2005, pp. 603–609.

[10] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. on Biomedcial Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[11] W. F. Clocksin, "Automatic segmentation of overlapping nuclei with high background variation using robust estimation and flexible contour models," in *Proc 12th Int. Conf. Image Analysis and Processing, ICIAP03*, 2003, pp. 682–687.

[12] F. Cloppet and A. Boucher, "Segmentation of overlapping/aggregating nuclei cells in biological images," in *Proc. ICPR, IEEE*, 2008, pp. 1–4.

[13] Y. Peng, M. Park, M. Xu, S. Luo, J. S. Jin, Y. Cui, F. W. S. Wong, and L. D. Santos, "Clustering nuclei using machine learning techniques," in *Proc. Int. IEEE/ICME Conf. on Complex Medical Engineering*, 2010, pp. 52–57.

[14] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, New Jersey, 2001.

[15] J. S. Suri, K. Setarhdan, and S. Singh, *Advanced Algorithmic Approaches to Medical Image Segmentation*, Springer-Verlag, London, 2002.

[16] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y., "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 881–892, 2002.

[17] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[18] D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[19] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, Chapman & Hall, Boca Raton, 1993.

[20] T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.