image segmentation,
classification,
breast cancer

Marek KOWAL[1], Paweł FILIPCZUK[1], Andrzej OBUCHOWICZ[1], Józef KORBICZ[1]

COMPUTER-AIDED DIAGNOSIS OF BREAST CANCER USING GAUSSIAN MIXTURE CYTOLOGICAL IMAGE SEGMENTATION[2]

This paper presents an automatic computer system to breast cancer diagnosis. System was designed to distinguish benign from malignant tumors based on fine needle biopsy microscope images. Studies conducted focus on two different problems, the first concern the extraction of morphometric and colorimetric parameters of nuclei from cytological images and the other concentrate on breast cancer classification. In order to extract the nuclei features, segmentation procedure that integrates results of adaptive thresholding and Gaussian mixture clustering was implemented. Next, tumors were classified using four different classification methods: k-nearest neighbors, naive Bayes, decision trees and classifiers ensemble. Diagnostic accuracy obtained for conducted experiments varies according to different classification methods and fluctuates up to 98% for quasi optimal subset of features. All computational experiments were carried out using microscope images collected from 25 benign and 25 malignant lesions cases.

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). FNB is collecting nucleus material directly from tumor. Obtained material is examined under a microscope to determine the prevalence of cancer cells [29]. 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 inexperienced 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 [9, 28]. In the literature one can find approaches to breast cancer classification [5,6,10,13,14,16,17,20,21,24,26]. Mentioned approaches are concentrated on classifying FNA (Fine Needle Aspiration) or FNB slides as benign or malignant.

In this paper, we present a decision support system that allows distinguish malignant from the benign breast tumors. The classification of the tumor is based on morphometric examination of cell nuclei [29,30]. In contrast to normal and benign nuclei, which are typically uniform in appearance, cancerous nuclei are characterized by irregular morphology that is reflected in several parameters described in detail further in the article. Features were extracted from segmented images obtained by hybrid segmentation method based on Gaussian mixture clustering and adaptive thresholding.

The quality of segmentation and feature extraction was tested by using the set of classifying algorithms. The measure is based on classification accuracy obtained by leave-on-out cross-validation. In this work four different classification methods were used to rate the feature subsets: k-nearest neighbors, naive Bayes classifier, decision trees and classifiers ensemble.

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.

[1] email: {M.Kowal, P.Filipczuk, A.Obuchowicz, J.Korbicz}@issi.uz.zgora.pl.
[2] This research was partially supported by the National Science Centre in Poland.

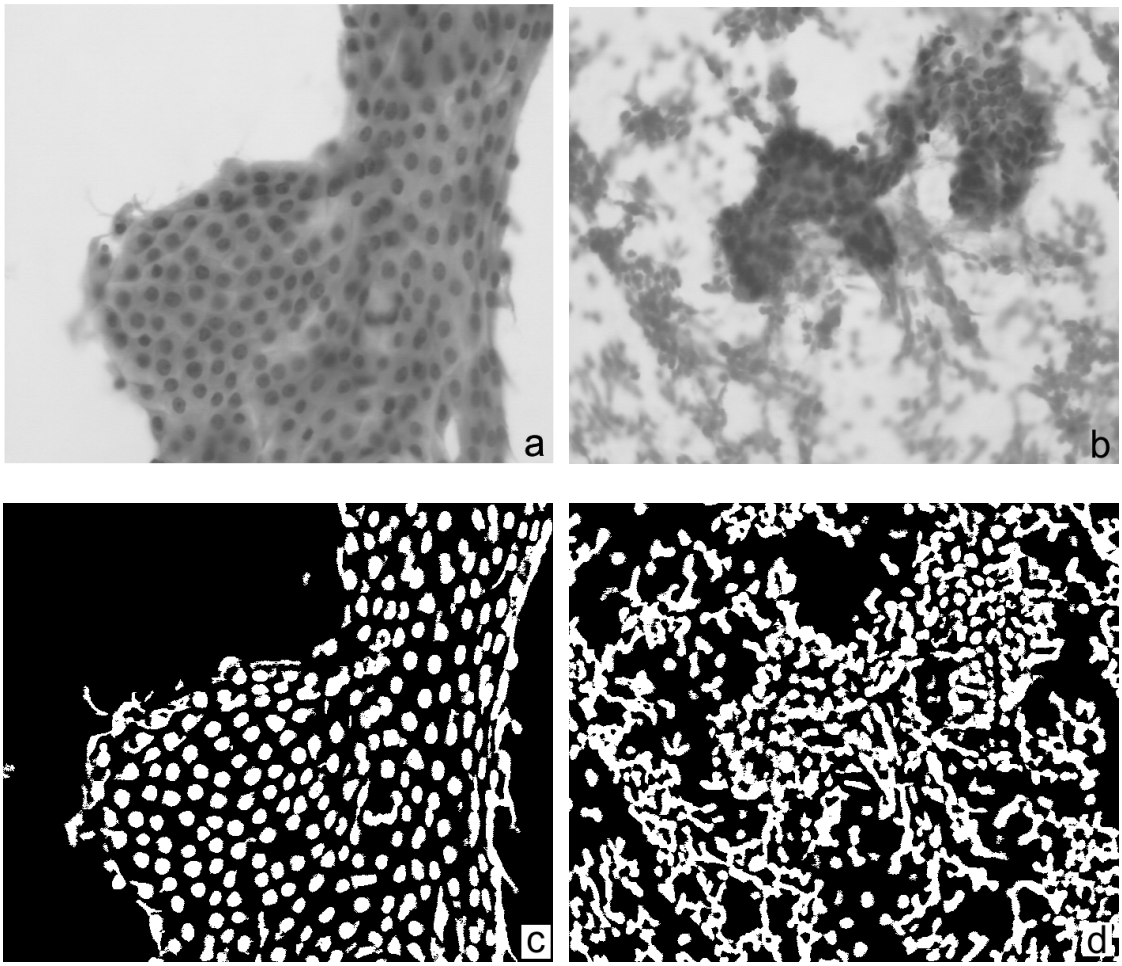
Section 3 deals with segmentation algorithm used to separate cells and extract features. Section 4 shows the experimental results obtained using the proposed approach. The last part of the work includes a conclusions and bibliography.

2. DATABASE OF FINE NEEDLE BIOPSY MICROSCOPE IMAGES

The database contains 500 images of the cytological material obtained by FNB. The material was collected from 50 patients of Regional Hospital in Zielona Góra. It gives 10 images per case which was recommended amount by specialists from the Regional Hospital in Zielona Góra [17]. This number of images per single case allows correct diagnosis by a pathologist. The set contains 25 benign and 25 malignant lesions 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 images were recorded by SONY CDD IRIS color video camera mounted atop an AXIOPHOT microscope. The slides were projected into the camera with 10 and 160× objective and a 2,5× ocular. One image was generated for enlargement 100× and nine for enlargement 400×. Images are BMP files, 704×578 pixels, 8 bit/channel RGB (Fig. 1a and 1b). All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year.

3. COMPUTER-AIDED 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. For separated objects morphometric and colorimetric features are calculated. Finally, objects are classified, eg. as an benign or malignant case. Our approach follows this scheme.



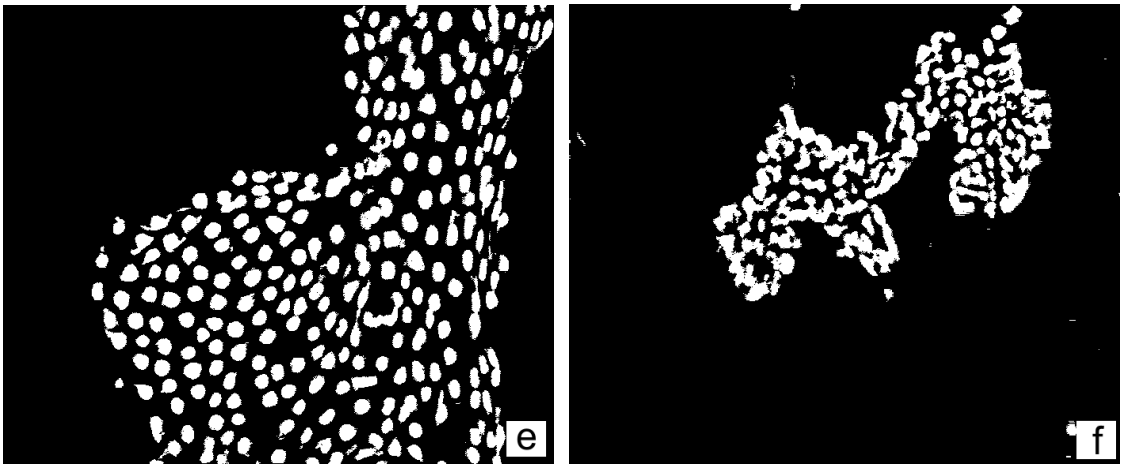


Fig. 1. FNB image segmentation – original images (a, b), adaptive thresholding segmentation (c, d), Gaussian mixture model segmentation (e, f).

At first, images require preprocessing due to their low quality. This phase includes histogram stretching and noise reduction [7]. 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 [1,10,13,16,23,24]. This task is usually done automatically, using one of the well known methods of image segmentation [11,15,22,25,27]. Unfortunately, attempt 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 [11]. Moreover, cells tend to cluster and overlap together and their boundaries are blurred. In order to deal with these problems, we have developed segmentation procedure that integrates results of image segmentation from two different methods.

Table 1. Features extracted from images.

Feature	Description
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 eccentricity of the ellipse that has the same second-moments as the segmented nucleus. The eccentricity is the ratio of the distance between the foci of the ellipse 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 segmented nucleus
minor axis length	the length of the minor axis of the ellipse that has the same normalized second central moments as the segmented nucleus
luminance gradient sum	the sum of luminance gradients in the image of the individual nucleus
luminance mean	the mean of luminance in the image of the individual nucleus
luminance variance	the variance of luminance gradients in the image of the individual nucleus
distance from the centroid	the distance between the geometric center of the selected nucleus and centroid of all nuclei

Firstly, color microscope image is converted to gray scale and adaptive thresholding segmentation is employed to eliminate from further analysis pixels that represent background [22,25]. Thresholding segmentation is based on the idea that pixels with intensity greater than some threshold value are

classified as background pixels and as object pixels otherwise. The key parameter in thresholding segmentation is the choice of the threshold value. In the considered approach threshold is calculated adaptively for individual pixel as mean intensity of pixels from its neighborhood defined by square region of size 21x21 pixels. Finally, a binary segmentation mask is created by assigning each pixel value 0 (background) or 1 (objects). Results of such segmentation can be visualized as black and white image (Fig. 1c and 1d). Adaptive thresholding was able to isolate individual nucleus from nuclei clusters with weak and blurred boundaries. Unfortunately red blood cells, cytoplasm and nuclei have very similar high local contrast so algorithm was unable distinguish them and incorrectly classify all these objects as nuclei (Fig. 1c and 1d). To tackle this drawback, further analysis of pixels labeled by value 1 was required to distinguish the nuclei form the red blood cells and cytoplasm. The problem was solved using Gaussian Mixture clustering algorithm in RGB color space [12,18]. In the first step, Gaussian Mixture Model (GMM) with two normal distribution components $N1(\mu1,\Sigma1)$ and $N2(\mu2,\Sigma2)$ was build using pixels selected from the original color image. Binary segmentation mask obtained during adaptive thresholding was used to select the correct object pixels. $N1(\mu1,\Sigma1)$ and $N2(\mu2,\Sigma2)$ components represent nuclei and other objects (red blood cells and cytoplasm had similar color) respectively. GMM parameters $\theta = [\mu1, \mu2,\Sigma1, \Sigma2]$ where μi represents means and Σi covariances were estimated using an Expectation Maximization (EM) algorithm. The EM is a well known algorithm in statistics and detailed expressions for iterative updating means and covariances can be found in [4,8,31]. Finally, during the clustering step the same data that was used to create GMM were partitioned based on the largest posterior probability computed for each pixel. Based on the clustering results, orignal binary segmentation mask was modified by changing the label of pixels which belongs to second partition described by component $N2(\mu2,\Sigma2)$. The final binary segmentation mask obtained during Gausian Mixture clustering was used to extract the nuclei from the original microscope image. Sample binary segmentation mask after adaptive thresholding and Gaussian Mixture clustering is shown in Fig. 1e and 1f. The scheme which presents whole image processing algorithm designed to extract the nuclei is shown in Fig. 2.

The task of breast cancer diagnosis posed in this work can be boiled down to the classification problem. In order to use well known classification methods, set of features which describes individual nucleus were defined. Morphometric features were calculated based on binary segmentation mask and colorimetric features were calculated based on the original image multiplied by the corresponding binary segmentation mask. Individual nuclei were extracted through the discovery of connected objects in binary segmented mask. Next, the set of features collected in table 1 was calculated for each individual nucleus on the image. The feature extraction procedure was repeated for each image from database described in section 2. Finally, statistics (mean and variance) of features were calculated and normalized for each image.

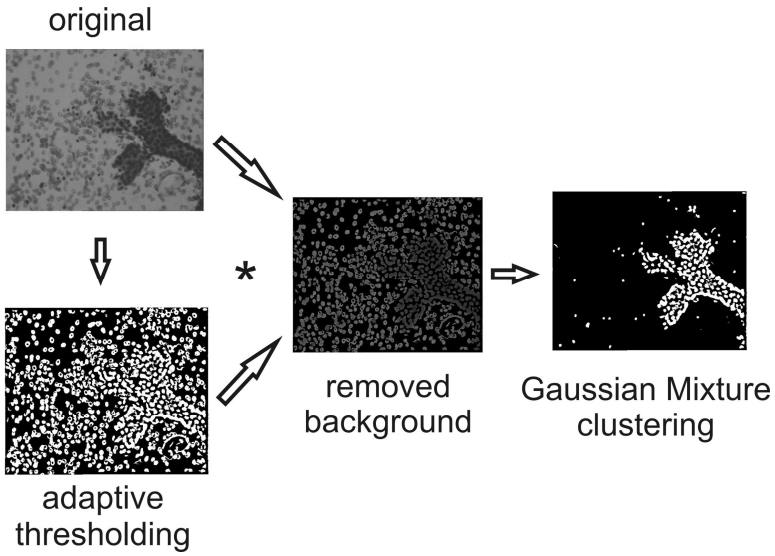


Fig. 2. Nuclei segmentation scheme.

The classification step was realized using four well known classification algorithms: k-nearest neighbors, naive Bayes classifier, decision trees and classifiers ensemble [2,3,19]. 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. All experiments were performed using real case data. There were 50 patients: 25 malignant and 25 benign. Each patient was represented by 10 images. There was one overall image and 9 images in maximum enlargement. In the processing only the 9 were used. The effectiveness was tested using the leave-one-out validation technique, where full set of 9 images representing 1 patient was 1 case.

Table 2. Classification results.

Set of features (statistic)	**Classifier**	**Results**
area (mean), minor axis length (mean), minor axis length (variance), distance to centroid (variance)	k-Nearest Neighbor	98%
eccentricity (mean), perimeter (variance), distance to centroid (variance)	Naive Bayes classifier	92%
perimeter (mean), major axis length (variance), distance to centroid (variance)	Decision Trees	94%
minor axis length (mean), perimeter (mean),	Ensemble Classifier	94%

To find a set of features the best discriminative benign and malignant cases sequential forward selection was applied.

In order to illustrate the effectiveness of proposed computer-aided diagnosis system an experimental results was collected and presented in the table 2. The best classification rate was 98%. Such result is very promising according to the fact that the quality of the segmentation method might be improved by detecting overlapping and joined objects.

5. CONCLUSIONS

Presented segmentation approach is similar to our previously presented k-means, 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 incorrect clusters when nuclei are represented by very few pixels on the image. Despite that, the classification rate is 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 may require new way of analysis.

BIBLIOGRAPHY

[1] AL-KOFAHI Y., LASSOUED W., LEE W., ROYSAM B., Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images, IEEE Trans. on Biomedcial Engineering, Vol. 57, No. 4, 2010, pp. 841-852.
[2] BISHOP C., Pattern recognition and machine learning, Springer, New York, 2006.
[3] BREIMAN L., FRIEDMAN J., STONE C.J., OLSHEN R.A., Classification and Regression Trees, Chapman & Hall, Boca Raton, 1993.

[4] DEMPSTER A.P., LAIRD N.M., RUBIN D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, J. the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1, 1977, pp. 1–38.
[5] FILIPCZUK P., KOWAL M., MARCINIAK A., Feature selection for breast cancer malignancy classification problem, J. Medical Informatics & Technologies, Vol. 15, 2010, pp. 193-199.
[6] GIL J., WU H., WANG B.Y., Image analysis and morphometry in the diagnosis of breast cancer, J. Microsc. Res. Tech., Vol. 59, 2002, pp.109-118.
[7] GONZALEZ R.C., WOODS R.E., Digital Image Processing, Prentice Hall, New Jersey, 2001.
[8] GUPTA M.R., CHEN Y., Theory and Use of the EM Algorithm, Foundations and Trends in Signal Processing, Vol. 4, No. 3, 2010, pp. 224-292.
[9] GURCAN M.N., BOUCHERON L.E., CAN A., MADABHUSHI A., RAJPOOT N.M., YENER B., Histopathological Image Analysis: A Review, IEEE Reviews in Biomedical Engineering, Vol. 2, 2009, pp. 147-171.
[10] HREBIEŃ M., STEĆ P., OBUCHOWICZ A., NIECZKOWSKI T., Segmentation of breast cancer fine needle biopsy cytological images, Int. J. Appl. Math and Comp. Sci., Vol. 18, No. 2, 2008, pp. 159–170.
[11] HUANG H.K., PACS and Imaging Informatics: Basic Principles and Applications, John Wiley & Son, New Jersey, 2010.
[12] HUNTER D.R., LANGE K., A Tutorial on MM Algorithms, The American Statistician, Vol. 58, 2004, pp. 30-37.
[13] JELEŃ Ł., FEVENS T., KRZYŻAK A., Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies, Int. J. Appl. Math and Comp. Sci., Vol. 18, No. 1, 2008, pp. 75–83.
[14] JELEŃ Ł., FEVENS T., KRZYŻAK A., JELEŃ M., Discriminatory Power of Cells Grouping Features for Breast Cancer Malignancy Classification, Proc. 4th Int. Conf. on Biomedical Engineering, Kuala Lumpur, 2008, pp. 559-562.
[15] KAWA J., PIĘTKA E., Image Clustering with Median and Myriad Spatial Constraint Enhanced FCM, Proc. 4th Int. Conf. on Computer Recognition Systems CORES' 05, Springer, 2005, pp. 211-218.
[16] KOWAL M., KORBICZ J., Segmentation of breast cancer fine needle biopsy cytological images using fuzzy clustering, in: KORNACKI J., RAŚ Z., WIERZCHOŃ S.T., KACPRZYK J., (eds.), Advances in Machine Learning I, Springer-Verlag, Berlin – Heidelberg, 2010, pp. 405-417.
[17] MARCINIAK A., OBUCHOWICZ A., MONCZAK R., KOŁODZIŃSKI M., Cytomorphometry of Fine Needle Biopsy Material from the Breast Cancer, Proc. 4th Int. Conf. on Computer Recognition Systems CORES' 05, Springer, 2005, pp. 603-609.
[18] MCLACHLAN G., PEEL D., Finite Mixture Models, John Wiley & Sons, 2000.
[19] MITCHELL T.M., Machine Learning, McGraw-Hill, 1997.
[20] NEZAFAT R., TABESH A., AKHAVAN S., LUCAS C., ZIA M., Feature selection and classification for diagnosing breast cancer, Proc. Int. Assoc. of Science and Technology for Development International Conference, Cancun, Mexico, 1998, pp. 310–313.
[21] OBUCHOWICZ A., HREBIEŃ M., NIECZKOWSKI T., MARCINIAK A., Computational intelligence techniques in image segmentation for cytopathology, in: SMOLIŃSKI T.G., MILANOVA M.G., HASSANIEN A.G. (eds.), Computational intelligence in biomedicine and bioinformatics: current trends and applications, Springer-Verlag, Berlin, 2008, pp. 169-199.
[22] OTSU N., A threshold selection method from gray-level histograms, IEEE Trans. Sys. Man. and Cyber. Vol. 9, 1979, pp. 62-66.
[23] PENG Y., PARK M., XU M., LUO S., JIN J.S., CUI Y., WONG F.W.S., SANTOS L.D., Clustering nuclei using machine learning techniques, Proc. Int. IEEE/ICME Conf. on Complex Medical Engineering, 2010, pp. 52-57.
[24] SCHNORRENBERG F., PATTICHIS C., KYRIYRIACOU K., SCHIZAS C., Detection of cell nuclei in breast cancer biopsies using receptive fields, IEEE Proc. Engineering in Medicine and Biology Society, 1994, pp. 649–650.
[25] SEZGIN M., SANKUR B., Survey over image thresholding techniques and quantitative performance evaluation, J. Electronic Imaging Vol. 13, No. 1, 2003, pp. 146–165.
[26] STREET N., Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer, In Jain L. (eds.), Soft Computing Techniques in Breast Cancer Prognosis and Diagnosis, World Scientific Publishing, Singapore, 2000, pp. 297–322.
[27] SURI J.S., SETAREHDAN K., SINGH S., Advanced Algorithmic Approaches to Medical Image Segmentation, Springer-Verlag, London, 2002.
[28] ŚMIETANSKI J., TADEUSIEWICZ R., ŁUCZYŃSKA E., Texture Analysis in Perfusion Images of Prostate Cancer - a Case Study, Int. J. Appl. Math and Comp. Sci., Vol. 20, No. 1, 2010, pp. 149-156.
[29] UNDERWOOD J.C.E., Introduction to biopsy interpretation and surgical pathology. Springer-Verlag, London, 1987.
[30] WOLBERG W.H., STREET W.N., MANGASARIAN O.L., Breast cytology diagnosis via digital image analysis. Analytical and Quantitative Cytology and Histology, Vol. 15, 1993, pp. 396–404.
[31] XU L., JORDAN M.I., On Convergence Properties of the EM Algorithm for Gaussian Mixtures, Neural Computation, Vol. 8, 1996, pp. 129–151.