35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Multi-Label Fast Marching and Seeded Watershed Segmentation Methods for Diagnosis of Breast Cancer Cytology

Paweł Filipczuk[1], Marek Kowal[1] and Andrzej Obuchowicz[1]

Abstract— **Digital cytology plays an increasingly important role in breast cancer diagnosis. However, analysis of cytologic images is a very difficult task. Especially, nuclei segmentation is extremely challenging. In our work on fully automated medical diagnosis system we encountered the problem of densely clustered nuclei. We decided to use a segmentation algorithm that is rather rarely found in the literature. Multi-label fast marching was applied and compared to well-known and extensively used seeded watershed algorithm. In both methods, it is critical to determine the appropriate starting points (seeds). The seeds were determined using a combination of adaptive thresholding in grayscale, clustering in color space and conditional erosion. The proposed segmentation procedure was tested for suitability for diagnosis of the cancer. Experiments were conducted on a set of 450 microscopic images of fine needle biopsies obtained from patients of the Regional Hospital in Zielona Góra, Poland. The images were classified as either benign or malignant using 84 features extracted from isolated nuclei. Both methods gave very promising results and showed that our method is effective and can be successfully applied for computer-aided diagnosis system.**

I. INTRODUCTION

Breast cancer is the most common cancer among women. According to the International Agency for Research on Cancer, in 2008 there were 1,384,155 diagnosed cases of breast cancer and 458,503 deaths caused by the disease worldwide [1], [2]. The effectiveness of treatment largely depends on the timely detection of the disease. An important and often used diagnostic method is the so-called triple-test. The triple-test is based on 3 medical examinations. It includes self examination (palpation), mammography or ultrasonography imaging, and fine needle biopsy (FNB) [3]. In FNB, the material is collected directly from the tumor and examined under a microscope. This approach requires extensive knowledge and experience of the cytologist. Computer-aided diagnosis can assist the specialist and help make the results objective. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for detection of diseases, infections as well as many other disorders [4], [5], [6].

Nuclei segmentation is the key functional component of the computer-aided cytology. Many authors have reported problems with a segmentation of clumped and overlapped nuclei [7], [8]. Segmentation errors introduced by the clustered nuclei give a significant distortion in nuclei features. Eventually, this results in low classification accuracy. Seeded watershed (SW) algorithm is often used to handle this problem [9]. We proposed the alternative solution based on the fast marching algorithm. It is usually applied to tissue or organ segmentation, while rather rarely used to segmentation of cytologic images. The classical algorithm is designed to background-foreground segmentation and can not be directly employed to extract nuclei. In order to segment multiple objects we used multi-label fast marching (MLFM) [10]. MLFM and SW were then compared in terms of their suitability for breast cancer diagnosis.

The paper is divided into 4 sections. Section I presents an introduction into breast cancer diagnosis. Section II describes segmentation methods. Section III shows the experimental results. Finally, section IV delivers our conclusions.

II. NUCLEI SEGMENTATION

On cytologic images nuclei often create clusters, overlap each other, their boundaries are not clear and their interiors are not uniform. In our previous works we have already developed methods able to extract nuclei from cytologic images [11], [12], [13], [14]. However, some clustered nuclei were not properly segmented. To cope with this problem, we propose extended segmentation procedure. Firstly, hybrid method based on adaptive thresholding and k-means clustering is used to discover nuclei region. Next, conditional erosion is applied to binary image of nuclei to localize nuclei markers. Finally, nuclei are segmented using MLFM or SW. Fig. 1 presents sample segmentation obtained using both methods.

A. Nuclei Marker Extraction

The procedure starts from converting the original RGB image into the binary image representing nuclei region. This is done using adaptive thresholding and k-means clustering. Adaptive thresholding is applied to distinguish objects (i.e. nuclei, cytoplasm and red blood cells) from the background. Threshold is calculated adaptively for subsequent pixels of the image using averaging filter.

In the next step, k-means clustering [15] is applied to distinguish nuclei from the rest of the objects. In the considered case, 3 clusters are defined. The clusters correspond to nuclei, red blood cells and cytoplasm. The clustering procedure is carried out in the RGB color space on the subset of pixels provided by adaptive thresholding. The cluster corresponding to the nuclei is determined based on the fact that nuclei are the darkest objects in the image. Next, pixels that belong to

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

[1] P. Filipczuk, M. Kowal and A. Obuchowicz are with Institute of Control & Computation Engineering, University of Zielona Góra, ul. Ogrodowa 3b, 65-246 Zielona Góra, Poland {p.filipczuk, m.kowal, a.obuchowicz}@issi.uz.zgora.pl

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

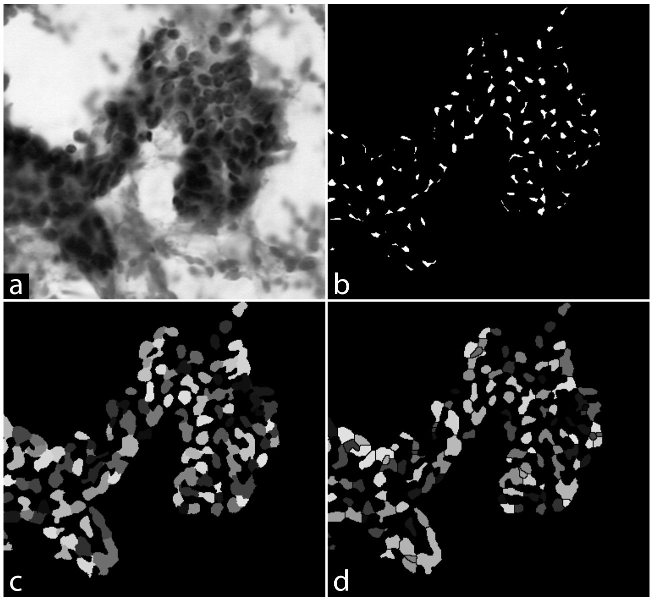

Fig. 1. Input image (a), conditional erosion (b), MLFM (c), SW (d).

the nuclei cluster are used to construct binary image BW. It marks regions in the image where the nuclei are located.

A key stage of nuclei marker extraction is to correctly detect nuclei centers. The method is based on the concept of conditional erosion [9]. Procedure assumes that the erosion is conducted as long as the size of the processed nucleus is large enough. Two masks for erosion operation are designed. They can be referred as fine and coarse erosion structuring elements. The coarse erosion tends to remain the actual shape but reduces the size of clustered nuclei. This can make the nucleus to disappear because of huge reduction in the size. On the other hand, fine erosion structuring element is less likely to make the nucleus disappear, but it will lead to the loss of original shape. Conditional erosion is applied to binary image BW obtained in the previous step. Threshold T_1 for coarse structuring element B_c and threshold T_2 for fine structuring element B_f are chosen experimentally (T_1=350, T_2=50). Next, nuclei are iteratively eroded using coarse element until the size of all objects is smaller than T_1. Finally, erosion with fine element is applied iteratively to the results obtained during coarse processing. Structuring elements B_c and B_f are designed according to the shape of the nuclei which is similar to an ellipse. Objects that have survived the conditional erosion are the markers M used to seed MLFM and SW methods.

B. Multi-Label Fast Marching

Fast marching method is a special case of the level sets approach for monotonically advancing fronts [16]. Algorithm starts with the initial front Γ_0. Next the front Γ evaluates with speed $F(x, y)$ in the normal direction where F is always either positive or negative. Front passes through a point (x, y) at the time $T(x, y)$. Under this formulation the arrival time function $T(x, y)$ satisfies the Eikonal equation:

$$|\nabla T|F = 1. \tag{1}$$

In order to solve the equation, the gradient $|\nabla T|$ is estimated using upwind entropy-satisfying scheme. By limiting our considerations to two-dimensional grid, we must solve the following quadratic equation:

$$\begin{aligned}1/F_{i,j}^2 = \max\left(\max(d_{i,j}^{-x}T,0), -\min(d_{i,j}^{+x}T,0)\right)^2 + \\ \max\left(\max(d_{i,j}^{-y}T,0), -\min(d_{i,j}^{+y}T,0)\right)^2,\end{aligned} \tag{2}$$

where d^- and d^+ are backward and forward difference operators.

The algorithm constructs the narrow band around the initial front and next marches this band forward, freezes the values of existing points and brings new ones into the narrow band. The procedure is repeated until the narrow band is empty. The behavior of the front is driven by the speed function F. It must be designed in a way that the front stops exactly at the boundary of the nuclei. We decided to use speed function based on the image local gradient:

$$F = e^{-\alpha|\nabla(H_\sigma * I)|}, \tag{3}$$

where α is a weighting factor, I is the original image and H_σ is a Gaussian smooth operator.

Standard fast marching is well suited to foreground-background segmentation. Nevertheless, our application must deal with multiple objects. It was realized by using MLFM method [10]. It is initiated by the seeds corresponding to nuclei marker centers. Each seed is associated with the unique label (segment). Propagation speed is the same for all labels. The algorithm maintains single narrow band which contains trial points from all segments. New label for trial point is inherited from the segment that propagates at the current algorithm iteration. In order to prevent leakages of the nuclei segments into background and to reduce computational costs all points classified as background by the adaptive thresholding and k-means are excluded from the fast marching propagation.

C. Seeded Watershed

The classical watershed algorithm used for nuclei segmentation tend to create many micro-segments. Such oversegmentation makes the results of the watershed method completely useless. To deal with this problem we used SW method, a well-known extension of watershed algorithm [9].

Firstly, the topographic surface TS (intensity map) is determined. TS is generated by the Euclidean distance transform of the binary mask of nuclei BW obtained by procedure described in Section II-A. Next, the surface is modified accordingly to found markers using morphological reconstruction. The algorithm impose the minima of surface TS at the locations specified by the markers M. Modified topographic surface TS_m has regional minima preserved wherever M is nonzero. In this way, the markers are incorporated to original topographic surface. It allows to split the clustered nuclei avoiding the oversegmentation.

III. Experimental Results

All methods presented in this work were tested on real medical data. For this purpose, 450 images were collected from 50 patients (25 with benign disease and 25 with malignant) of the Regional Hospital in Zielona Góra, Poland. Each patient is represented by 9 images, which is the number recommended by the specialists from the hospital and allows for correct diagnosis by a pathologist. Fine-needle biopsies without aspiration were performed under the control of ultrasonograph with a 0.5 mm diameter needle. Smears from the material were fixed in spray fixative (Cellfix by Shandon) and dyed with hematoxylin and eosin (H&E). The time between preparation of smears and their preservation in fixative never exceeded three seconds. All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year.

In order to verify the effectiveness of both segmentation methods we performed entire diagnostic procedure and compared the classification accuracy. Firstly, the nuclei were isolated using the methods described in Section II. Then, for each nucleus the following 28 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* (ECC) - 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* (MjAL)- the length of the major axis of the ellipse that has the same normalized second central moments as the nucleus,
- *Minor Axis Length* (MnAL)- the length of the minor axis of the ellipse that has the same normalized second central moments as the nucleus,
- *Distance to Centroid of All Nuclei* (D2A) - the distance between the geometric center of the nucleus and centroid of all nuclei,
- *Distance to c-Nearest Nuclei* (D2cNN) - sum of distances between the geometric center of the nucleus and geometric centers of c-nearest nuclei; after conducting experiments with different values of c, we decided to set this parameter to 1,
- *Gray-Level Co-occurrence Matrix* (GLCM) features - the group of 4 features extracted from GLCM [17] determined for offsets corresponding to 0°, 45°, 90° and 135° using eight gray-levels:
  - *Contrast* (CN),
  - *Correlation* (CR),
  - *Homogeneity* (H),
  - *Energy* (EN),
- *Gray-Level Run-Length Matrix* (GLRLM) features - the group of 11 features extracted from (GLRLM) [18] determined for offsets corresponding to 0°, 45°, 90° and 135° using eight gray-levels:
  - *Short Run Emphasis* (SRE),
  - *Long Run Emphasis* (LRE),
  - *Gray-Level Nonuniformity* (GLN),
  - *Run Length Nonuniformity* (RLN),
  - *Run Percentage* (RP),
  - *Low Gray-Level Run Emphasis* (LGRE),
  - *High Gray-Level Run Emphasis* (HGRE),
  - *Short Run Low Gray-Level Emphasis* (SRLGE),
  - *Short Run High Gray-Level Emphasis* (SRHGE),
  - *Long Run Low Gray-Level Emphasis* (LRLGE),
  - *Long Run High Gray-Level Emphasis* (LRHGE)
- *Mean R, G and B Value* (MRV, MGV, MBV) - the mean value of pixels of the nucleus in channel R, G and B respectively,
- *Variance of R, G, and B Value* (VRV, VGV, VBV) - the variance of pixel values of the nucleus in channel R, G and B, respectively.

TABLE I

CLASSIFICATION ACCURACY FOR ALL 84 INDIVIDUAL FEATURES USING MLFM (LEFT) AND SW (RIGHT) METHODS. NOTE THAT THE FULL NAME OF THE FEATURE IS OBTAINED BY ADDING THE NAME OF THE NUCLEI FEATURE (ROWS) TO THE STATISTICS (COLUMNS). E.G. MEAN AREA DETERMINED USING MLFM IS 63.11%, AND USING SW IS 66.89%.

feature	mean	median	STD
Area	63.11 / 66.89	59.78 / 66.89	58.44 / 56.22
Perimeter	70.22 / 59.56	69.11 / 68.89	47.56 / 44.89
ECC	51.78 / 50.89	48.89 / 62.00	47.78 / 48.44
MjAL	66.44 / 55.33	64.89 / 56.89	45.78 / 53.33
MnAL	62.67 / 66.44	61.56 / 66.44	64.89 / 54.22
D2A	74.89 / 77.56	72.22 / 72.89	69.11 / 72.44
D2cNN	75.11 / 72.00	75.11 / 76.89	70.44 / 69.56
CN	58.00 / 76.22	60.67 / 69.11	52.89 / 71.56
CR	60.44 / 75.56	59.78 / 70.22	61.56 / 73.56
H	54.89 / 66.67	56.89 / 63.78	60.44 / 55.56
EN	55.78 / 49.33	53.78 / 52.89	70.89 / 68.89
SRE	68.67 / 74.44	69.11 / 72.00	76.22 / 75.33
LRE	73.78 / 73.56	69.78 / 72.00	38.67 / 46.22
GLN	57.78 / 49.56	61.78 / 42.89	56.00 / 44.22
RLN	52.22 / 53.78	57.78 / 60.22	47.11 / 55.33
PR	59.33 / 50.44	53.11 / 52.89	54.44 / 51.33
LGRE	53.11 / 74.44	54.22 / 69.11	60.89 / 52.44
HGRE	50.67 / 52.00	57.11 / 58.44	51.56 / 53.78
SRLGE	84.89 / 85.11	81.78 / 82.44	83.11 / 79.78
SRHGE	63.78 / 64.44	54.22 / 50.00	75.11 / 73.56
LRLGE	58.67 / 53.33	58.22 / 53.11	72.89 / 61.78
LRHGE	54.89 / 46.22	56.67 / 58.67	50.00 / 50.67
MRV	59.78 / 56.44	63.33 / 62.00	48.67 / 50.44
MGV	80.00 / 84.00	85.11 / 84.22	50.00 / 51.11
MBV	79.78 / 82.89	82.00 / 86.89	55.33 / 52.67
VRV	53.78 / 50.44	58.22 / 51.11	48.00 / 55.56
VGV	62.44 / 70.89	63.11 / 73.33	58.67 / 63.56
VBV	70.89 / 74.89	66.44 / 70.67	79.78 / 79.33

For each image, the mean, median and standard deviation (STD) were calculated giving a total number of 84 features. The features were then standardized. The approach was tested for the classification accuracy, which is defined as the percentage of successfully recognized cases to the total number of all cases. The classification accuracy was tested using k-nearest neighbor (kNN) classification algorithms and n-fold cross-validation technique. The fold was a set of 9 images representing 1 patient. This means the images belonging to the same patient were never at the same time in the training and testing set.

Tab. I presents classification accuracy for individual features. The approach was also tested using an optimal set of features. The sets for both methods were found using sequential forward selection algorithm. The optimal sets of features determined for MLFM and SW are as follows:

- *Set for MLFM* - MGV (median), area (mean), SRHGE (STD), HGLRE (mean), HGLRE (median), LRE (mean), LRE (median), MRV (STD), perimeter (median), RLN (mean), GLN (median), MjAL (mean), MGV (STD), D2CNN (median), SRHGE (mean), perimeter (STD),
- *Set for SW* - MBV (median), D2CNN (median), SRHGE (mean), MnAL (median), HGLRE (median), LRE (mean), LRE (STD), SRHGE (median), LRLGE (mean), LRLGE (STD).

Surprisingly, both methods gave exactly the same results. The classification accuracy obtained using optimal feature sets was 95.56%, sensitivity 0.97, specificity 0.94, and Matthews correlation coefficient 0.91.

In the considered approach, images were classified individually. However, the diagnostic decision concerns patients, not single images. The final diagnosis obtained by a majority voting of the classification of individual images belonging to the same patient (e.g. for given patient, if 5 images were classified as benign and 4 as malignant then the final diagnosis for the patient would be benign) was for both methods 100%.

IV. CONCLUSIONS

In this paper we compared the suitability of nuclei segmentation using MLFM and SW algorithms for breast cancer classification problem. As a comparative criterion the classification accuracy was used. Tested on real case medical data, both methods gave very similar results. This showed that both methods are a good choice for nuclei segmentation if the initial markers are given. However, the SW algorithm requires less initial parameters than MLFM, which makes it easier to use. Computational complexity of both methods is similar. Result of 95.56% for individual images and 100% for patients shows that proposed approach can provide valuable information for a medical specialist. Experiments also proved that conditional erosion is a very useful tool for detecting nuclei centers even when the nuclei are densely clustered.

ACKNOWLEDGMENT

The authors wish to thank Dr. Roman Monczak from the Regional Hospital in Zielona Góra, Poland for his great help and interesting discussions.

Paweł Filipczuk 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 the European Social Fund and the state budget.

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

REFERENCES

[1] J. Ferlay, H. Shin, B. F., D. Forman, C. Mathers, and D. Parkin, “Globocan 2008 v2.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 10,” http://globocan.iarc.fr, Lyon, France: International Agency for Research on Cancer, 2010, accessed on 30/08/2012.

[2] F. Bray, J. Ren, E. Masuyer, and J. Ferlay, “Estimates of global cancer prevalence for 27 sites in the adult population in 2008,” *International Journal of Cancer*, Jul. 2012, doi: 10.1002/ijc.27711.

[3] J. C. E. Underwood, *Introduction to Biopsy Interpretation and Surgical Pathology*. London: Springer-Verlag, 1987.

[4] 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.

[5] J. Śmietański, R. Tadeusiewicz, and E. Łuczyńska, “Texture analysis in perfusion images of prostate cancer–a case study,” *International Journal of Applied Mathematics and Computer Science*, vol. 20, no. 1, pp. 149–156, 2010.

[6] L. Jeleń, T. Fevens, and A. Krzyżak, “Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies,” *International Journal of Applied Mathematics and Computer Science*, vol. 18, no. 1, pp. 75–83, 2010.

[7] C. Jung, C. Kim, S. W. Chae, and S. Oh, “Unsupervised segmentation of overlapped nuclei using bayesian classification,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2825–2832, 2010.

[8] J. Cheng and J. C. Rajapakse, “Segmentation of clustered nuclei with shape markers and marking function,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 741–748, 2009.

[9] X. Yang, H. Li, and X. Zhou, “Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy,” *IEEE Transactions on Circuits and Systems - I*, vol. 53, no. 11, pp. 2405–2414, 2006.

[10] P. Steć, *Segmentation of Colour Video Sequences using the Fast Marching Method*. Zielona Góra, Poland: University of Zielona Góra Press, 2005.

[11] P. Filipczuk, M. Kowal, and A. Obuchowicz, “Automatic breast cancer diagnostis based on k-means clustering and adaptive thresholding hybrid segmentation,” in *Image processing and communications challenges 3*, ser. Advances in Intelligent and Soft Computing : 102, R. S. Choraś, Ed. Berlin - Heidelberg: Springer-Verlag, 2011, pp. 295–303.

[12] ——, “Fuzzy clustering and adaptive thresholding based segmentation method for breast cancer diagnosis,” in *Computer recognition systems 4*, ser. Advances in Intelligent and Soft Computing: 95, R. Burduk, M. Kurzyński, M. Woźniak, and M. Żołnierek, Eds. Berlin - Heidelberg: Springer-Verlag, 2011, pp. 613–622.

[13] M. Kowal, P. Filipczuk, A. Obuchowicz, and J. Korbicz, “Computer-aided diagnosis of breast cancer using gaussian mixture cytological image segmentation,” *Journal of Medical Informatics & Technologies*, vol. Vol. 17, pp. 257–262, 2011.

[14] M. Kowal, P. Filipczuk, and J. Korbicz, “Hybrid cytological image segmentation method based on competitive neural network and adaptive thresholding,” *Pomiary, Automatyka, Kontrola*, vol. 57, no. 11, pp. 1448–1451, 2011.

[15] J. A. Hartigan, *Clustering Algorithms. (Probability & Mathematical Statistics)*. John Wiley & Sons Inc., 1975.

[16] J. Sethian, “A fast marching level set method for monotonically advancing fronts,” in *Proceedings of the National Academy of Sciences*, 1996, pp. 1591–1595.

[17] R. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[18] X. Tang, “Texture information in run-length matrices,” *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1602–1609, 1998.

HUMAN CAPITAL
HUMAN – BEST INVESTMENT!

Lubuskie
Worth your while

EUROPEAN UNION
EUROPEAN
SOCIAL FUND