Processing of Microscope Colon Images for Supporting the Medical Diagnostics

M. Kruk, S. Osowski, R. Koktysz

***Abstract* – The paper presents the methods of the microscope colon image processing for segmentation and recognition of cells. In the first part we describe the cells segmentation algorithms applying the morphological functions. The second part contains the description of the techniques of the cells recognition using the Support Vector Machine. The SVM classifier recognizes the types of the cells, counts them and determines their percentage contents in a stroma of tissue. The paper describes shortly the technique of SVM, the numerical characterization of the cells and the approach used in training and testing of the classifier system for cells recognition. The results of the numerical experiments will be also presented and discussed.**

I. INTRODUCTION

Inflammatory Bowel Diseases belong to the group of chronic, incurable diseases of gastrointestinal tract. They are characterized by the spontaneous remissions and relapses of their etiology not unexplained up to now. To this Inflammatory Bowel Diseases we may include on one hand the *Colitis Ulcerosa* and Lesniowski-Crohn disease, and on the other one the microscope colitis (collagenous colitis) and also some non-specific inflammations. When the human defense cells are existing in a stoma of tissue we can state that the inflammation starts. The most important human defense cells include lymphocytes, eosinophylic granulocytes, neutrophilic granulocytes, plasmocytes and also some other cells of inflammatory inflitration.

The paper will be concerned with the recognition of these particular cells on the basis of the microscope colon image. The most important stages of this task include: filtration and segmentation of the image, extraction of the individual cells, generation of diagnostic numerical features for each cell and as a final stage the recognition and counting. As a result of such complex process we can certify the intensity of the inflammation, which is very important factor in medical diagnosis.

II. THE PROBLEM DESCRIPTION

The input data in our experiment is a microscopic digitized colon image. It is the image of the biopsy of the tissue of the magnification equal 40. The image is saved as the bitmap file for further processing. The typical image of this kind is depicted in Fig. 1. The large parts of the image are out of our interest. They represent the grandular ducts and should be removed from the image. Our main concern is the region containing the small particles representing the human defense cells.

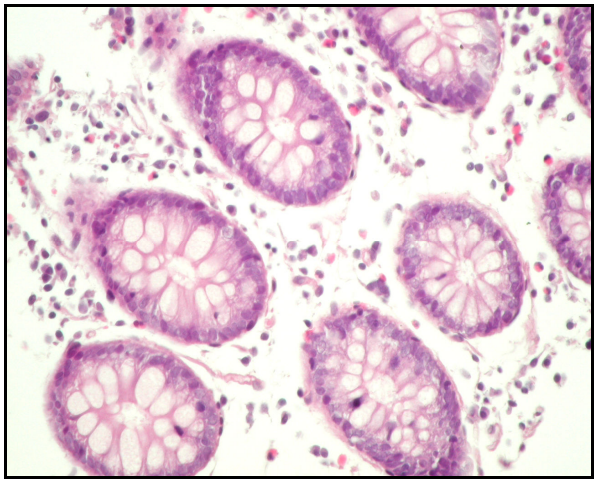


Fig. 1 The typical microscope colon image

The first step of processing is the extraction of the cells trough segmentation. We achieve it using the morphological operations and the watershed algorithm [2]. After extraction each cell represents the individual image which must be preprocessed in order to be characterized by the numerical features. These features should be created in a way to enhance the differences among different types of cells and reduce the diversity existing within the same family. The features characterizing the cells are used by the classifier in the recognition process. As a result we get each cell associated with the particular class. The main steps of the problem solution may be summarized as following:

- extraction of the individual cells from the microscopic colon image to create the database of cells
- generation of the diagnostic features well characterizing different families of cells
- classification of cells using the Support Vector Machine trained on the basis of the generated features
- verification of the trained classifier on the basis of the testing set of cells.

After recognition of all cells we count their total number within each family, their percentage ratio and the total number of all cells existing in the investigated area of the colon. The results given by this automatic

M. Kruk and S. Osowski are with the Institute of Theory of Electrical Engineering, Measurement and Information Systems, Warsaw University of Technology, pl. Politechniki, 00-661 Warsaw, Poland, S. Osowski is also with Military University of Technology, Warsaw. R. Koktysz is with The Institute of Patomorphology, Warsaw Military Hospital.

system are used by the medical doctor to assess the intensity of inflammation and the advancement of the illness in the human organism.

III. CELLS SEGMENTATION

The introductory step of the colon image processing is the extraction of the individual cells from the image. We have done it by applying the filtering and segmentation processes [6]. The applied algorithm consists of the following stages.

- Reading the input bitmap image. It is saved in the matrix form $\mathbf{I}(x, y, r, g, b)$, in which x and y are the coordinates in horizontal and vertical axes, while r, g and b are the intensity levels of each color component.
- Deleting the largest and the smallest elements of the image. It consists of the following stages:
  - ✓ transformation to the binary image
  - ✓ labeling the image elements
  - ✓ deleting the extreme size elements which are larger than 3000 and smaller than 40 pixels
- Transformation of the background to the white color: all pixels of intensity larger than threshold value are transformed to white.
- Transformation of the image first to the grayscale and then to the binary one (all elements larger than threshold are white, the other are black). This operation is performed only on the green color.
- Implementation of the morphological operation of opening. We have used the structural element in the form of a disk of the size equal 2.
- Computation of the distance matrix using the linear time Euclidean distance transformation algorithm [1]. As a result we get the local maxima of the distance matrix.
- Application of the watershed algorithm to perform the segmentation of the image [2]. As a result we get the compact regions representing the cells.
- Filtering the resulting image by applying the morphological operations of opening. The smallest elements representing the noise are removed from the image.
- Adding the contours to each separated cell. As a result we get the image containing the segmented cells which can be extracted to the individual file.

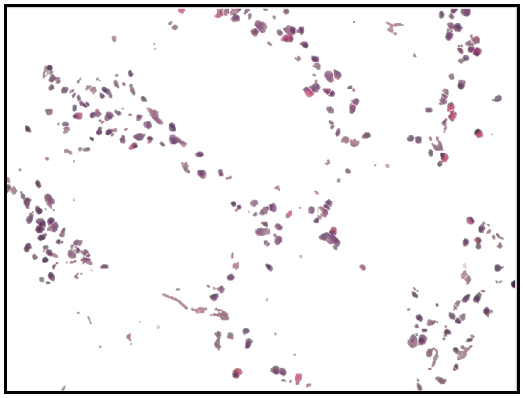
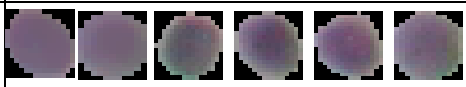
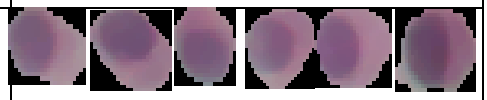
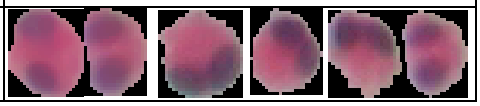


Fig. 2 The view of the segmented cells corresponding to the image of Fig. 1

In our experiments we have processed 10 microscopic images of the colon, extracting the images of three types of cells: eosinophylic granulocyte, lymphocyte and plasmocyte. Table I presents the examples of 6 cells representing the considered types.

TABLE I
THE EXAMPLES OF THE SEGMENTED CELLS

Cell type	The representative images of 6 cells
Lymphocyte	
Plasmocyte	
Granulocyte	

The lymphocytes contains almost no cytoplasm and this is very important diagnostic factor. The proportion of the area of the nucleus to the total area of the cell for plasmocyte and granulocyte is very similar and close to half.

IV. THE RECOGNITION OF THE CELLS

A. *The features generation*

To apply the automatic classifier for the recognition of the cell image we have to characterize the cell by the numerical attributes, called features. Each cell should be described by the individual features. After the study of the subject we have generated the following features used in training and testing of the classifier network.

- The features based on the histogram

The histogram of the image of each cell carries a lot of information. Different measures applied to the histogram may be the source of the whole set of features. The first approach is to approximate the histogram by the Hermite orthogonal functions. The coefficients of the expansion of the histogram into the Hermite functions form the first set of features. The Hermite function of the rank n are defined by the following equation [4]

$$x_n(t) = \frac{1}{\sqrt{2^n n! \sqrt{\pi}}} e^{-\frac{t^2}{2}} H_n(t) \qquad (1)$$

where the Hermite polynomials are determined by the recurrent relation:

$$H_n(t) = 2tH_{n-1}(t) - 2(n-1)H_{n-2}(t) \qquad (2)$$

with the starting values equal with $H_0(t) = 1$ and $H_1(t) = 2t$. Fig. 3 illustrates the histograms of the images of the eosinophylic granulocyte, lymphocyte and plasmocyte (the bars) and their representation by the Hermite expansion (the solid line). Eight order Hermite polynomials have been used in approximation.

It is evident that different cells are characterized by different histograms. The eosinophylic granulocyte

has wide histogram extending from 120 to 200 pixels. The lymphocyte is characterized by very narrow histogram and the histogram of plasmocyte is of the width close to the granulocyte. The Hermite approximation of the histograms deliver essential information of their distribution. The coefficients of Hermite expansion of histogram form the first set of features. The other measures of the histogram shape include: the mean, standard deviation (std), skewness, kurtosis and finally span of the histogram.

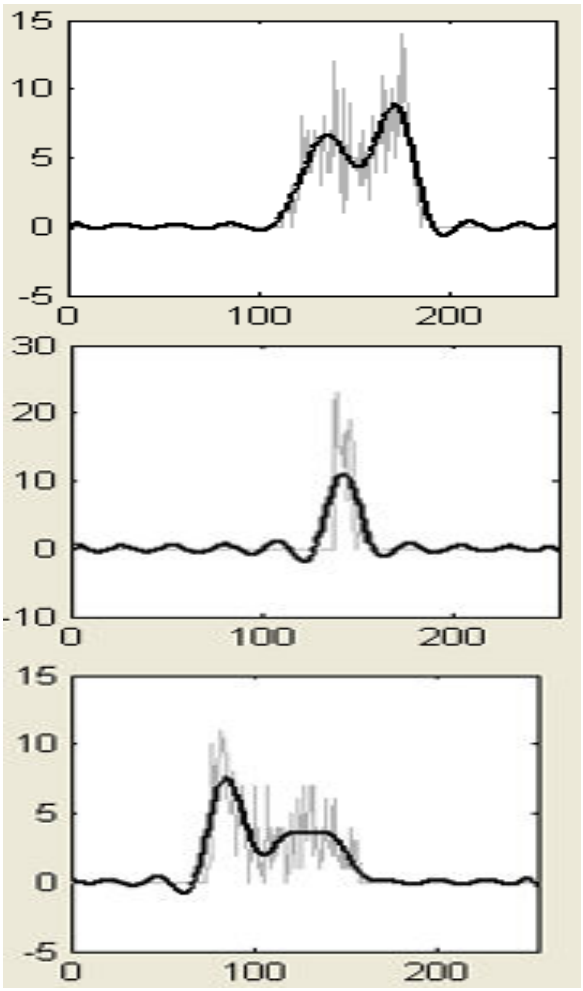


Fig.3 The histograms of the images of a) eosinophylic granulocyte, b) lymphocyte and c) plasmocyte

Table II presents the average values of the statistical parameters characterizing the histograms of the considered cells (the data are calculated for three colors: R, G and B with exception of mean). They have been calculated over the whole families of cells available in experiments.

TABLE II
THE AVERAGE VALUES OF THE STATISTICAL FEATURES OF THE HISTOGRAMS OF THE CELLS

	Lymphocyte			Plasmocyte			Granulocyte		
Mean	0.88			1.57			2.02		
Std	2.9	2.7	3.2	4.1	3.6	4.5	4.4	4.2	5
Skewness	4.7	4.9	5.2	3.8	3.4	3.9	3.1	3.1	3.3
Kurtosis	32	39	38	23	21	24	16	19	18
Span	20	35	27	56	55	34	52	46	25

- The geometrical features

The cells differ also by the size and the proportion between the size of nucleus and cytoplasm. The next set of features has been formed by the geometrical parameters. They include: the perimeter, area of the whole cell, the area of the nucleus and also the ratio nucleus area to the area of the whole cell. Table III presents the mean values of these parameters calculated for the whole families of cells gathered in the data base.

TABLE III
THE MEAN VALUES OF THE GEOMETRICAL FEATURES OF THE CELLS

	Lymphocyte	Plasmocyte	Granulocyte
Perimeter	47	65	74
Nucleus area	210	363	301
Area of cell	229	405	521
Ratio Nucleus area/cell area	0.79	0.65	0.45

We see the essential differences among their values for all three considered cell types.

B. Support Vector Machine classifier

The features generated for the cells form the set of input signals (vector **x**) supplied to the classifier. The recognition of the cells will be done by the Support Vector Machine (SVM) classifier with radial Gaussian kernel [3,4]. The SVM was chosen because of its excellent generalization ability at relatively small number of patterns used in learning. The extensive information regarding SVM can be found elsewhere, for example [3,4]. In all our experiments we have used very efficient Platt algorithm [5].

The important hyperparameters: the spread σ of the Gaussian function and the regularization constant C have been determined in an experimental way using the cross validation approach. After many trials we have found that σ=1 and C=1000 were close to optimal. The SVM was first trained on the learning data and then tested on the next set of data not taking part in learning.

V. THE RESULTS OF NUMERICAL EXPERIMENTS

A. The data base

The numerical experiments of cell recognition have been performed on the data base of cells collected from the microscopic colon images acquired in the Institute of Pathomorphology, Warsaw Military Hospital. We have used the images of tissues of the magnification equal 40. Table IV presents the number of each cell family used in experiments.

TABLE IV
THE DATA BASE OF THE SEGMENTED CELLS

	Lymphocytes	Plasmocytes	Granulocytes
Number	48	103	29

For each cell we have generated the set of features forming the input vector **x** applied to the SVM classifier. The size of the input vector was equal 55. Among input signals there were 4 geometrical features, 5 statistical features generated for each color

(30 altogether) and 21 Hermite coefficients (7 Hermite coefficients for each color). The data sets have been divided into two parts: the learning set (60%) used for training of the SVM network and testing set (40%) applied only for testing purposes of the trained classifier.

B. The results of classification

The feature vectors generated for each cell have been applied to the input of Gaussian kernel SVM classifier. After training the parameters of the network were frozen and the classifier was tested on the testing data set. Among the cells we have to recognize three regular classes of the cells. We have applied one against all approach [4]. In the testing process some cells have been rejected by all classes and they formed the fourth one.

TABLE V
THE COMPARISON OF THE NUMBER OF RECOGNIZED CELLS BY OUR ALGORITHM AND BY THE HUMAN EXPERT

	Our Algorithm	Human expert
Lymphocytes	51	43
Plasmocytes	56	48
Granulocytes	6	5
Other	6	24
Sum of all cells	119	120

Table V present the comparison of the number of cells calculated by the human expert and our automatic recognition system. As it is seen the differences of both scores are within the margin of 15%, acceptable in the hospital practice.

Fig. 4 presents the view of the microscopic image of the colon with the cells annotated by our system. The letters L, P G and I denote different families of cells: L-lymphocytes, P-plasmocytes, G-granulocytes and I-others.

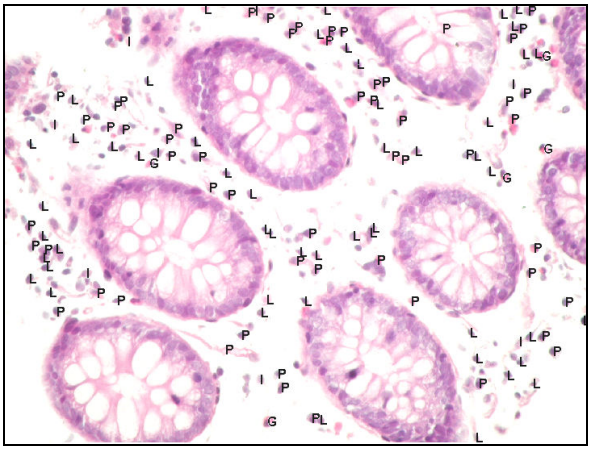


Fig. 4 The output image of the colon with the annotated cells: L-lymphocytes, P-plasmocytes, G-granulocytes and I-others

This image is also of the important result of the work, since it is of great help for the medical doctor in visual inspection of the results of automatic recognition.

VI. CONCLUSIONS

The paper has presented the automatic method of the recognition of different cells existing in the microscopic image of the colon. To the important problems solved in the work belong: the segmentation of the image to extract the individual cells, preprocessing the cell images to generate the numerical diagnostic features, and finally the recognition of the cells using SVM classifier.

The numerical verification of the proposed classification system has been checked on the data base of more than 120 cells, representing the granulocytes, plasmocytes and lymphocytes. The introductory experiments have shown that the mean misclassification rate of the cell recognition is below 15%, acceptable in the medical practice.

Further experiments should be directed to improve the accuracy of recognition. First of all we have to elaborate the new features characterizing the cells and the selection methods able to discover the most important features. The significant task will be also to increase the number of representatives of each class of cells by acquiring more microscopic images of he colon of different patients.

REFERENCES

[1] H. Breu, J. Gil, D. Kirkpatrick, M. Werman, Linear time Euclidean distance transform algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 5, May 1995, pp. 529-533.

[2] V. Luc, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol. 13, No. 6, June 1991, pp. 583-598.

[3] V. Vapnik, Statistical Learning Theory, N.Y.: Wiley, 1998

[4] B. Schölkopf, A. Smola, Learning with Kernels. Cambridge, MA: MIT Press, 2002

[5] L. Platt, L. Fast training of SVM using sequential optimization. (in Scholkopf, B., Burges, B., & Smola, A., Eds. Advances in kernel methods – support vector learning. Cambridge: MIT Press), 1998. pp. 185-208

[6] P. Soille, Morphological Image Analysis, Principles and applications, Springer, 2003

[7] Matlab user manual – Image processing toolbox, MathWorks, Natick, 1999

[8] G. E. Andrews, R. Askey, R. Roy, Hermite Polynomials, §6.1 in Special Functions. Cambridge University Press, Cambridge, England, pp. 278-282, 1999