Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

# Recognition of Colon Cells Using Ensemble of Classifiers

M. Kruk, S. Osowski, R. Koktysz

***Abstract*—The paper presents the application of an ensemble of neural classifiers for the recognition of colon cells on the basis of the microscope colon image. The solved task include: the segmentation of individual cells from the image using the morphological operations, the preprocessing stages leading to the extraction of features, selection of the most important features, and the classification stage applying the neural classifiers arranged in the form of ensemble. The paper presents and discusses the results concerning the recognition of four most important colon cell types: eosinophylic granulocyte, neutrophilic granulocyte, lymphocyte and plasmocyte. The proposed system is able to recognize the cells with the accuracy comparable to the human expert (around 4% of discrepancy).**

## I. Introduction

The Inflammatory Bowel Diseases (IBD) refers to two chronic diseases that cause inflammation of the intestines: ulcerative colitis and Lesniowski-Crohn disease [13]. 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, that can be treated as the last mixed class and is not subject for counting.

The paper will be concerned with the recognition of these particular cells on the basis of the microscope colon image. This is quite difficult task, since the cells are similar to each other, especially if we take into account relatively small magnification x40. Although there exist nowadays some semi-automatic systems for recognition of blood cells at magnification factor 1000x [10] no specific automatic cell recognizing system for the images considered in this work has been already developed.

The main concern of the paper is the final stage of recognition and classification of cells using different neural classifiers arranged in the form of the ensemble network. The results of the numerical experiments of the colon cell recognition using individual neural classifiers and the ensemble network of these classifiers will be presented and discussed in the paper.

## II. Problem Formulation

The recognition of the most important defending cells: plasmocytes, lymphocytes and granulocytes existing in the colon tissue will be done on the basis of the microscopic image of the colon tissue taken for the patient in the form of biopsy at the magnification equal 40x. The acquired color image is saved in the form of the bitmap file for further processing. Fig. 1 presents the typical image of the colon tissue. There are visible large structures of the dark background, representing parts of the grandular ducts which are out of interest in our work and should be removed from the image. Our main interest is in the region out of ducts, containing small particles (white background) representing the human defense cells under consideration.

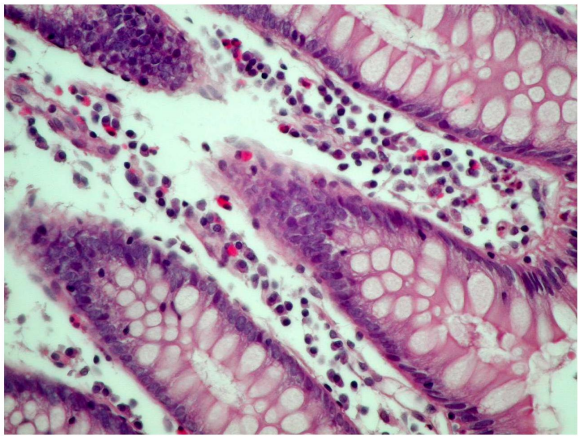


Fig. 1 The microscopic image of the colon tissue

The first step in the procedure is the extraction of the defense cells. We have solved this problem by applying the morphological operations and the watershed algorithm [6,9]. After extraction each cell represents the individual image that should undergo further preprocessing in order to be characterized by the numerical features, representing the image. It is desirable to generate the features that are stable for different representatives of the same family. These features may form the input vector **x** applied to the classifier in the recognition process.

The problem of the recognition of the colon cancer cells may be summarized in the following steps.

- Extraction of the individual cells from the microscopic colon image to create the database of cells.
- Generation of the diagnostic features characterizing the families of cells in a way enhancing the differences among the cells belonging to distinct classes and reducing these differences for the cells of the same class.
- Recognition and classification of cells by using the

M. Kruk is with Warsaw University of Technology, Warsaw, Poland-mail: kruk@iem.pw.edu.pl.

S. Osowski is with Warsaw University of Technology and Military University of Technology, Warsaw, Poland (e-mail: sto@iem.pw.edu.pl).

R. Koktysz is with the Institute of Pathomorphology, Warsaw Military Hospital, Warsaw, Poland

classifiers trained on the database of cells.

- Combining the neural classifiers into one ensemble forming the expert system of increased accuracy of recognition.
- Application of the trained classifier expert system in the on-line recognition of the colon cells.

The results of recognition of all cells are used by the medical doctor to count the total number of cells within each family, the percentage ratio of different families and also the total number of all cells existing in the investigated area of the colon. These results given by the automatic system are the basis for assessing the intensity of inflammation and the advancement of the illness in the human organism.

## III. Extraction of the Cells from the Image

The first task is to extract the individual cells from the image and place them in the database. This problem was solved by applying the filtering and segmentation of the image [9]. The applied algorithm of segmentation may be summarized as following.

- Read the input bitmap image $\mathbf{I}(x, y, r, g, b)$ of the cell. In this notation x and y are the coordinates in horizontal and vertical axes, while r, g and b denote the intensity levels of each color component.
- Use the K-means algorithm to divide all pixels into three classes according to their brightness. The brightest pixels are converted to white color. The other two are associated with the nucleus and cytoplasm.
- Delete the largest and the smallest elements of the image. The smallest elements are treated as the noise and largest as the grandular ducts. We have solved it by applying the transformation of the real image to the binary form, labeling the image elements and finally deleting the extreme size elements which are larger than 3000 and smaller than 40 pixels. The mentioned sizes follow from the analysis of the size of many cells under interest.
- Transform the resulting image to the grayscale and then to the binary one (all elements larger than threshold are white, the others are black).
- Filtrate the resulting image using the morphological operation of opening. The structuring element in the form of a disk of the size equal 2 was used in this operation.
- Generate the map of distances from the black pixel to the nearest white pixels of the image and compute the distance matrix using the linear time Euclidean distance transformation algorithm [1]. As a result we get the local minima of the distance matrix.
- Apply the watershed algorithm [9] based on these distances for the final division of the image into catchment's basins, each corresponding to one cell. As a result we get the compact regions representing the cells.
- Extract the regions corresponding to the individual cells and return to the color picture.
- Add the contours to each separated cell. As a result we get the image containing the segmented cells which can be extracted to the individual files, representing the considered families.

All presented above operations have been supported by the functions of Image Processing Toolbox of Matlab [7]. Fig. 2 illustrates the image of the segmented cells corresponding to the colon tissue of Fig. 1. All grandular ducts have been removed and only the cells of interest remained. The next step is to extract each cell and add it to the data base for further processing.

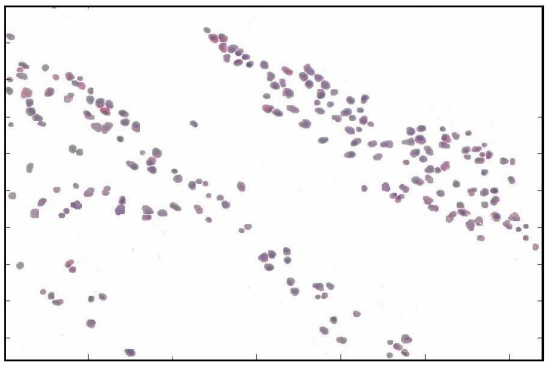
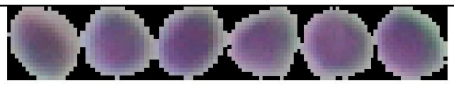


Fig. 2 The view of the segmented cells of the colon image of Fig. 1

Table I presents the examples of 6 cells representing the considered four types of cells: lymphocyte, plasmocyte eosinophilic granulocyte and neutrophylic granulocyte. They have been acquired at the magnification 40x.

TABLE I
THE EXAMPLES OF THE SEGMENTED CELLS

| Cell type | The representative images of 6 cells |
|---|---|
| Lymphocyte |  |
| Plasmocyte |  |
| Eosinophillic granulocyte |  |
| Neutrophylic granulocyte |  |

The lymphocytes contain 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 neutrophilic granulocyte is very similar and close to half.

## IV. Generation of Diagnostic Features

To create the efficient classification system we have to generate the proper set of diagnostic features, forming the input signals to the classifier. They should distinguish different classes and take similar values for the samples belonging to the same class. In the proposed solution we

have applied the features belonging to four groups: the parameters describing the histogram of the image, geometrical features, textural features and the features comparing the color intensities.

- The features based on the histogram

The histogram of the image carries a lot of information of the cell. Different parameters applied to the description of the histogram may be the source of the whole set of features. The most important fact is that different cell families are characterized by the histograms of various shape. They can be created for the whole cell and for the nucleus. Fig. 3 presents the typical histograms of the cell images of the eosinophilic and neutrophilic granulocytes, lymphocyte and plasmocyte.

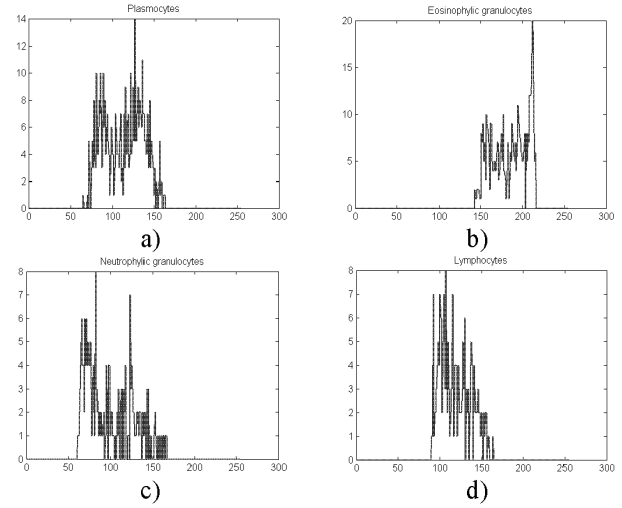



Fig.3 The histograms of the images of a) plasmocyte b) eosinophilic granulocyte, c) neutrophilic granulocyte, d) lymphocyte

It is evident that different cells are characterized by different histograms in RGB colors. They differ by the span, the maximum point and the distribution of bins. To characterize different histograms we have applied the following parameters: the mean, standard deviation (std), skewness, kurtosis, maximum value and the span of the histogram. All these parameters have been determined for all 3 colors. Table II presents the average values of the chosen statistical parameters characterizing the histograms of the considered cells. The data have been calculated for three colors: R, G and B. They have been calculated over the whole families of cells available in the experiments.

TABLE II

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

| | Lymphocyte | | | Plasmocyte | | | Granulocyte | | | Neutrophilic granulocyte | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 8.1 | 7.3 | 9.9 | 14.5 | 13.3 | 15.7 | 20.3 | 22.4 | 23.1 | 34.1 | 36.2 | 33.1 |
| Std | 2.9 | 2.7 | 3.2 | 4.1 | 3.6 | 4.5 | 4.4 | 4.2 | 5 | 1.7 | 1.7 | 2.1 |
| Skewness | 4.7 | 4.9 | 5.2 | 3.8 | 3.4 | 3.9 | 3.1 | 3.1 | 3.3 | 1.9 | 1.9 | 2.4 |
| Kurtosis | 32 | 39 | 38 | 23 | 21 | 24 | 16 | 19 | 18 | 6.5 | 6.3 | 8.4 |
| Span | 20 | 35 | 27 | 56 | 55 | 34 | 52 | 46 | 25 | 60 | 58 | 65 |

- The geometrical features

Different cell families differ by the size and shape. On the basis of it we can define the set of geometrical parameters characterizing the whole cell, the cytoplasm and the nucleus. They include: the perimeter, area of the whole cell, the area of the nucleus and the cytoplasm, the circumference, the ratios of different parameters characterizing the nucleus and the whole cell. Up to 22 features have been created in this way. Table III presents the mean values and standard deviations of some chosen geometrical parameters calculated for the whole families of cells gathered in the data base.

TABLE III

THE MEAN VALUES OF SOME GEOMETRICAL FEATURES OF THE CELLS

| | Lymphocyte | Plasmocyte | E.gran | N.gran |
|---|---|---|---|---|
| Perimeter | 46±6 | 70±8 | 78±9 | 52±4 |
| Nucleus area | 116±19 | 180±31 | 97±27 | 129±20 |
| Area of cell | 220±50 | 461±91 | 596±120 | 258±48 |
| Ratio Nucleus area/cell area | 0.52±0.2 | 0.39±0.12 | 0.16±0.1 | 0.5±0.12 |

The significant differences among their values for all four considered cell types can be observed. At the same time the standard deviations of some parameters are quite large, which means that these parameters are not quite stable.

- The textural features

The texture refers to an arrangement of the basic constituents of the material and in the digital image is depicted by the interrelationships between spatial arrangements of the image pixels. They are seen as the changes in intensity patterns, or the gray tones. For description of texture we have applied the Haralick matrix [12] description defined for directions of 0, 45° and 90°. For this matrix we have defined the contrast, energy, correlation, compactness, entropy, the average sum and variance (for three colors) used as the textural diagnostic features.

- The colorimetric features

The colorimetric features have been defined on the basis of the intensity of pixels of the cell image for each R, G and B component. We have calculated the mean of pixel intensities of the cell image for each color. The ratios of the means of R and G, R and B, G and B and also R and RGB, G and RGB and B and RGB have been calculated and used as the features (6 colorimetric features together).

## V. FEATURE SELECTION

It is well known that features may have different impact on the classification process [2]. Good feature should be characterized by the stable values for samples belonging to the same class and at the same time they should differ significantly for different classes. Thus the main problem in the classification and machine learning is to find out the features of the highest importance for the problem solution. Observe that the elimination of some features leads to the reduction of the dimensionality of the feature space and improvement of performance of the classifier in the testing mode at the data not taking part in learning.

From many known techniques of feature selection [2] like principal component analysis, correlation existing among features, correlation between the features and the classes, application of feature ranging by applying the linear SVM, analysis of mean and variance of the features belonging to

different classes, we have chosen the last one.

The variance of the features corresponding to cells being the members of one class should be as small as possible. On the other hand, to distinguish between different classes, the positions of means of feature values for the data belonging to different classes should be separated as much as possible. We have combined both measures (variance and mean) together to form the discrimination coefficient $S_{AB}(f)$ defined for the feature f at recognition of two cells belonging to different classes A and B.

$$S_{AB}(f) = \frac{|c_A(f) - c_B(f)|}{\sigma_A(f) + \sigma_B(f)} \quad (1)$$

In this definition $c_A$ and $c_B$ are the mean values of the feature f in the class A and B, respectively. The variables $\sigma_A$ and $\sigma_B$ represent the standard deviations determined for both classes. The large value of $S_{AB}(f)$ indicates good potential separation ability of the feature f for these two classes. On the other side small value means that this particular feature is not good for the recognition between classes A and B.

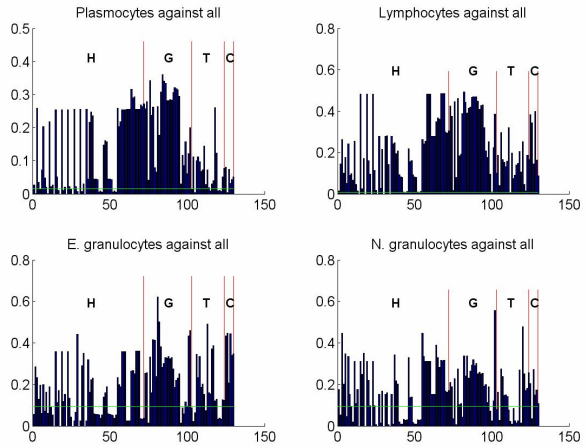



Fig. 4 The values of the discriminating coefficient $S_{AB}(f)$ of all features at the recognition between the recognized cells and the rest

Observe that the particular feature may be good for recognition between two chosen classes and useless for some others. Therefore the class oriented features should be considered to get the optimal choice of features used for the separation of two particular classes. Fig. 4 illustrates the change of the values of the discrimination coefficient $S_{AB}(f)$ for all features at the recognition between particular cell type and the rest of classes. The letters H, G, T and C denote the set of features based on histogram (H), geometry (G), texture (T) and color relations (C). It is evident, that the discrimination coefficient of each feature is different, from very small to high values. We may establish the horizontal line (as in Fig. 4) denoting the bias, below which the feature is regarded as insignificant. In the case of multiclass recognition solved globally in one network (no splitting into 2-class recognition subtasks) each feature is assessed on the basis of the average discrimination value for all 2-class combinations.

The determination of the optimal number of the chosen features is a separate problem. We have solved it by trying different number of the most significant features, testing the trained classifier on the validation data set and choosing one of the highest efficiency.

## VI. THE CLASSIFCATION SYSTEM

### A. *The applied classifiers*

In our cell recognition system we have applied five different classifiers [3]: the multilayer perceptron (MLP), the radial basis function network (RBF), support vector machine (SVM), the Fisher linear discriminant (FLD) and k nearest neighbor classifier (KNN). The classifiers differ by the network structure, the definition of the learning principle, the way of taking classification decision, etc. Thanks to such choice each of them looks at the classification problem from different point of view and underlines other aspects of taking decision.

MLP is a multilayer structure of many simple neuron-like processing units of sigmoidal activation function grouped together in layers [3]. The most important point in designing the MLP network structure is the generalization property. The number of weighs of the network should be limited so that the likelihood of correct generalization is increased. But this must be done without reducing the size of the network to the point where the desired target can not be met.

The RBF classifier is a network structure containing one hidden layer of radial (Gaussian) neurons acting on the local basis [3], and as many linear output neurons as is the number of classes. The classes are coded in a binary form. The main difference to MLP is the local principle of operation of the RBF neurons.

The SVM is a feedforward network of one hidden layer (the kernel function layer). It is known as an excellent classifier of good generalization ability [8,11].The learning problem of SVM is formulated as the task of separating the learning vectors into two classes of the destination values either $d_i$=1 (one class) or $d_i$=-1 (the opposite class), with the maximal separation margin. The great advantage of SVM is the formulation of learning problem leading to the quadratic programming with linear constraints

The SVM of Gaussian kernel has been used in our application. The hyperparameters σ of the Gaussian function and the regularization constant C have been adjusted by repeating the learning experiments for the set of their predefined values and choosing the best one at the validation data sets. To deal with a problem of many classes we have combined three different approaches: one against one, one against all and multi-output [4], cooperating together on the basis of majority voting principle.

FLD is a classifier working on the principle of an optimal linear separation of classes. The projection line is defined by $\mathbf{y}=\mathbf{w}^T\mathbf{x}$ of $\|w\|=1$. The measure of separability of two classes recognized by the indexes 1 and 2 is the so called Fisher

discriminant ratio $F(\mathbf{w}) = |m_1 - m_2|^2 / (s_1^2 + s_2^2)$, where $m_1$ and $m_2$ are the means of projections and $s_1^2, s_2^2$ are the scatters [3]. The learning problem of FLD is transformed to the maximization of the function

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \qquad (2)$$

where $\mathbf{S}_B$ is the between class scatter matrix and $\mathbf{S}_w$ the within-class scatter matrix [3].

For KNN classifier we find k nearest neighbors of the unknown vector from the known vectors and then assign the unknown vector to the class which appears most frequently in the vectors identified in the previous step. In the training phase the training samples are mapped into the multidimensional feature space. The space is partitioned into regions by class labels of the training samples. A point in the space is assigned to the appropriate class if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by parameter optimization using, for example, cross-validation.

### B. *The ensemble of classifiers*

We have applied the combination of multiple classifiers by applying the weighted voting principle [5]. Each classifier influences the final decision according to its performance on the training data. The general classification system applying five mentioned above individual classifiers combined into one classifying system is presented in Fig. 5.

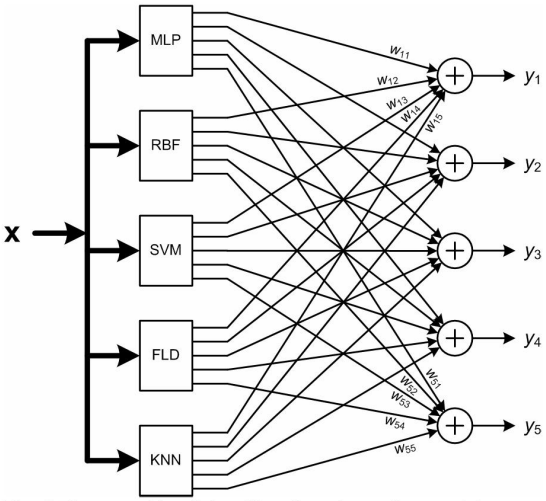



Fig. 5 The ensemble of classifiers for colon cell recognition

Each classifier is fed by the same input vector **x** and produces the response in the form of 5 signals indicating the membership to one of five classes: 4 colon cells under recognition and one class indicating the other cells not belonging to any of the recognized classes (mixed class). The integration matrix is formed by the weights $w_{ij}$, with the first index indicating the class and the second – the appropriate classifier. The weights $w_{ij}$ are adjusted according to the efficiency of jth classifier at recognition of ith class. We have applied the simple formula

$$w_{ij} = \frac{\eta_{ij}}{\sum_{k=1}^{5} \eta_{ik}} \qquad (3)$$

where $\eta_{ik}$ means the efficiency of kth classifier (k=1, 2, ..., 5) at the recognition of ith class (i=1, 2, ...,5) on the learning data (the ratio of the number of proper classifications to the number of samples taking part in learning). The final result of classification is determined on the basis of the output signals $y_i$ (i=1,2,...,5). The highest value of $y_i$ indicates the membership of the applied vector **x** to the appropriate class.

## VII. THE RESULTS OF NUMERICAL EXPERIMENTS

### A. *Data base*

The numerical experiments of the 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 obtained from 53 patients.

TABLE IV
THE DATA BASE OF THE SEGMENTED CELLS

| | Lymphocytes | Plasmocytes | E. gran. | N.gran. | Others |
|---|---|---|---|---|---|
| Number | 324 | 301 | 223 | 287 | 207 |

For each cell we have generated the set of features forming the input vector **x** applied to the classifiers. From 130 candidate features (81 – histogram features, 22 – geometrical, 21 – textural and 6 – colorimetric) after application of the selection procedure we have selected different number of the most important features (from 50 to 110) according to their experimentally checked importance for the recognition process between the classes of cells.

### B. *The results of experiments*

The available data set has been split into five exchangeable parts to enable application of the cross validation procedure. The class representatives have been split equally into all these parts. Four groups have been combined together and used in learning, while the fifth one used only in testing. The experiments have been repeated five times, exchanging the contents of the 4 learning subsets and the testing subset. The misclassification ratio in either learning or testing mode has been calculated as the mean of all 5 runs.

The misclassification ratios on the testing data of all five individual classifiers and the ensemble system are presented in Fig. 6. The misclassification is understood as the discrepancy between our result and human expert score. It is evident that integration of the results of many classifiers, even of unequal efficiency, improves the final score of classification. The best total result of recognition of all

classes has been improved from 4.62% (misclassification rate of the best SVM classifier) to 4.2% (ensemble of classifiers).

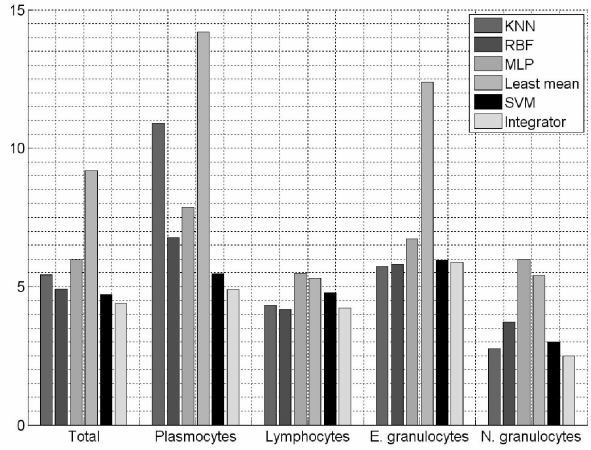



Fig. 6 The comparison of the misclassification ratios of the individual classifiers and the ensemble of classifiers

Table V presents the details of the class recognition errors of the ensemble system in the form of so called confusion matrix, represented as the summed results of the cross validation procedure for the testing data (class L – limphocyte, P – plasmocyte, E – eosinophilic granulocyte, N – neutrophilic granulocyte, O – other cells). The diagonal entries represent the numbers of properly recognized classes. Each entry outside the diagonal means error. The entry in the (i,j)th position of the matrix means false assignment of ith class to the jth one.

TABLE V
THE CONFUSION MATRIX OF THE ENSEMBLE OF CLASSIFIERS FOR THE TESTING DATA

| | L | P | E | N | O |
|---|---|---|---|---|---|
| L | 310 | 4 | 0 | 7 | 3 |
| P | 6 | 286 | 4 | 3 | 2 |
| E | 0 | 6 | 269 | 8 | 4 |
| N | 2 | 2 | 0 | 218 | 1 |

After recognition of the cells on the image of the colon tissue, all cells are automatically annotated according to the recognition results.

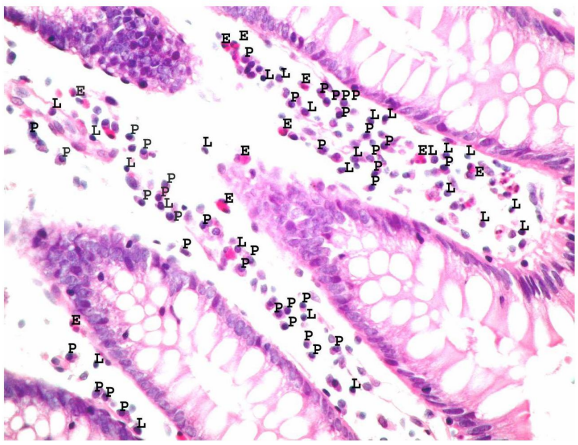


Fig. 7 The output image of the colon with the annotated cells: L-lymphocytes, P-plasmocytes, G-granulocytes and I-others

Fig. 7 presents the view of the microscopic image of the colon tissue with the cells annotated by our system. The letters L, P, E and N denote different families of cells. This image is also the important result of the work, since it is of great help for the medical doctor in the visual inspection of the results of recognition of an automatic system. Moreover there is a grapihical interface enabling the human expert to correct the annotations prepared by the automatic system.

## VIII. CONCLUSION

The paper has presented the automatic method of the recognition of different cells existing in the microscopic image of the colon. To the most 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, selection of the most important features and finally the recognition of the cells using the ensemble of many classifiers.

The numerical verification of the proposed classification system has been checked on the data base of more than 1000 cells, representing the eosinophilic and neutrophilic granulocytes, plasmocytes and lymphocytes. The numerical experiments have shown that the mean discrepancy rate to the human expert score of all cells is below 4% and this accuracy is acceptable in the medical practice. The system may be used for supporting the medical diagnosis simplifying and greatly accelerating the process of cell counting.

## REFERENCES


[1] H. Breu, J. Gil, D. Kirkpatrick, M. Werman, Linear time Euclidean distance transform algorithms, *IEEE Transactions PAMI,* Vol. 17, No. 5, May 1995, pp. 529-533.

[2] I Guyon,.A. Elisseeff, An introduction to variable and feature selection. *Journal of Machine Learning Res.*, 3, 1158 – 1182, 2003

[3] S. Haykin, *Neural Networks, Comprehensive Foundation*, Prentice Hall, 1999, New Jersey

[4] C.W Hsu., C.J Lin, A comparison methods for multi class support vector machines, *IEEE Trans. Neural Networks*, 13, 415-425, 2002

[5] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley, N. J., 2004

[6] V. Luc, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, *IEEE Transactions PAMI,* Vol. 13, No. 6, June 1991, pp. 583-598.

[7] Matlab user manual – *Image Processing Toolbox*, MathWorks, Natick, 1999

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

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

[10] B. Swolin, P. Simonsson, S. Backman, I. Lofqvist, I. Bredin, M. Johnsson, Differential counting of blood leukocytes using automated microscopy and decision support system based on ANN – evaluation of DiffMaster$^{TM}$ Octavia, Clin. Lab. Haem., 2003, vol. 25, pp. 139-147

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

[12] T. Wagner,. Texture analysis. ( in Jahne, B., Haussecker, H., & Geisser, P., Eds. *Handbook of Computer Vision and Application*, Boston: Academic Press), 275-309, 1999

[13] K. W. Zieliński Computerized analysis of medical image, Wydawnictwa Naukowe, PWN Warszawa 2002