Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.intl.elsevierhealth.com/journals/cobm

Recognition and classification of colon cells applying the ensemble of classifiers

M. Kruk[c], S. Osowski[a,b,*], R. Koktysz[d]

[a]University of Technology, Warsaw, Poland
[b]Military University of Technology, Warsaw, Poland
[c]Warsaw University of Life Sciences, Poland
[d]Military Institute of the Heath Services, Warsaw, Poland

ARTICLE INFO

Article history:
Received 7 May 2008
Accepted 4 December 2008

Keywords:
Inflammatory bowel disease
Colon cell recognition
Classification
Ensemble of classifiers

ABSTRACT

The paper presents the application of an ensemble of classifiers for the recognition of colon cells on the basis of the microscope colon image. The solved task include: segmentation of the 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 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 5% of discrepancy of both results).

1. Introduction

Inflammatory bowel disease is the name of disorders group that causes the intestines to become inflamed (red and swollen). The typical symptoms of this disease are the abdominal cramps and pain, diarrhea, weight loss and bleeding from the intestines. Two kinds of inflammatory bowel disease are Crohn's disease and ulcerative colitis [1,16]. Crohn's disease usually causes ulcers (open sores) along the length of the small and large intestines. Crohn's disease either spares the rectum or causes inflammation or infection with drainage around the rectum. Ulcerative colitis usually causes ulcers in the lower part of the large intestine, often starting at the rectum. The inflammation lasts a long time and usually comes back over and over again leading to the depression of patients [1,16].

The most important evidence that the inflammation has started is the existence of the human defense cells in a stoma of tissue. The most important human defense cells include lymphocytes, eosinophylic granulocytes, neutrophilic granulocytes, plasmocytes. For the purpose of diagnosis the recognition of these cells is essential. This is done on the basis of the image of the biopsy of the tissue stained Haematoxylin&Eosin (HE). Up to now there are no computer programs dedicated to this particular task. The existing general purpose programs for automatic counting of cells, like DiffMasterTM Octavia [14], Eamus [6] or the automatic systems of blood cell recognition presented in [11,15] are not directly applicable since the image of the stoma tissue is of absolutely different morphology and needs special treatment. The important reason is also the difference in the image magnification. In this particular task the usually applied magnification of the microscope is 600×, while in the mentioned above appliances the typical magnification is 1000× or 1500×. Although the general colors of the image remain invariant with respect to magnification there is visible, significant reduction of the details under analysis, that are taken into account by these programs. The specialized programs, suited for this low resolution image recognition, are needed.

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 at the applied magnification of 600× the cells are similar to each other and very difficult to recognize. According to the knowledge of authors no specific automatic cell recognizing system for the images considered in this work is available on the market.

The paper will discuss different aspects of these cells recognition on the basis of the colon image, starting from the extraction of individual cells from the whole image, generation of numerical diagnostic features, selection of the features and finally the recognition and classification of cells using different classifiers arranged in the form of the ensemble network. The results of the numerical experiments of the colon cell recognition using individual classifiers and the ensemble network of these classifiers will be presented and discussed in the paper.

[*] Corresponding author at: University of Technology, Warsaw, Poland. Tel.: +48 222347235.
E-mail address: sto@iem.pw.edu.pl (S. Osowski).

0010-4825/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2008.12.001

2. 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 stained HE, at the magnification equal 600×. 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 this work and should be removed from the image. Our main interest is in the region of white background, out of ducts, containing small particles representing the human defense cells under consideration.

The first step in the whole procedure is to extract these defense cells and form the database of them. This problem was solved by applying the morphological operations and the watershed algorithm [9,13]. After extraction each cell represents the individual image that should undergo further preprocessing in order to be characterized by the numerical values called diagnostic features, representing its image. It is desirable to generate the features that are stable for different representatives of the same family of cells. At the same time they should differ the cells of different families. These features will form the input vector **x** applied to the classifier in the recognition process.

The problem of the recognition of the colon 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 classifiers trained on the database of cells.
- Combining the classifiers into the 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 visible in the viewing field of the image are used by the medical staff 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. The obtained results are the basis for assessing the intensity of inflammation and the advancement of the illness in the human organism.

3. 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 using morphological operations [13]. The applied algorithm of segmentation may be shortly 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 and eliminated from the image. The resulting image is subject to further processing.
- Delete the largest and 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 (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 [2]. As a result we get the local minima of the distance matrix.
- Apply the watershed algorithm [13] 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 considered families.

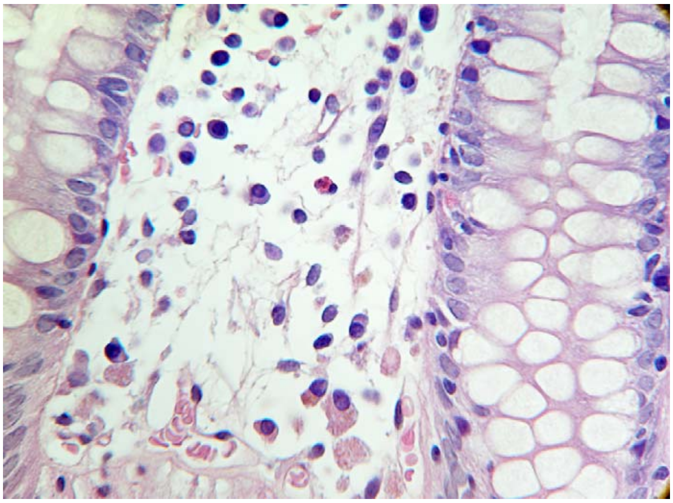


Fig. 1. The microscopic image of the colon tissue.

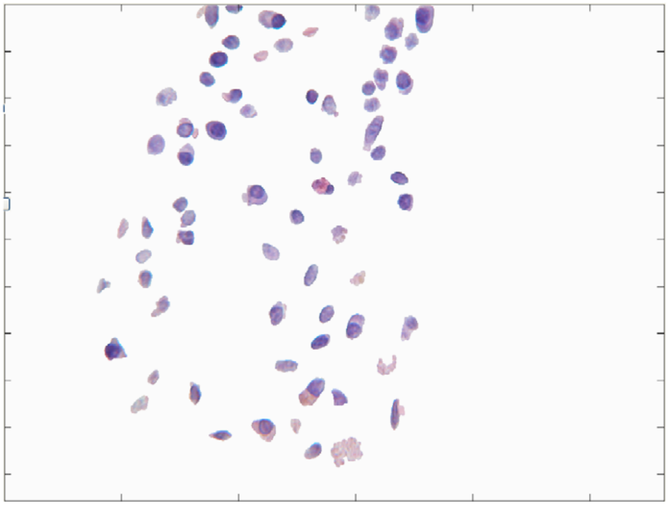
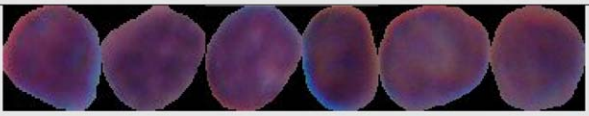
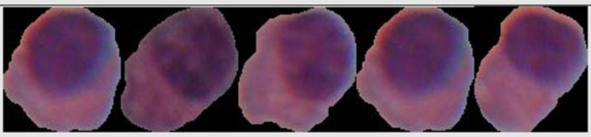
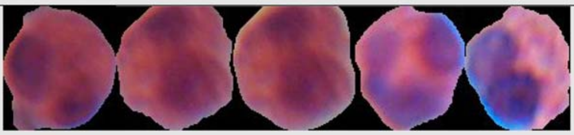
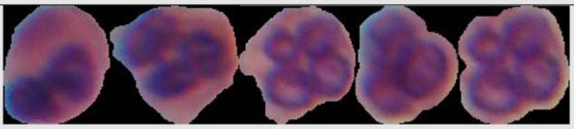


Fig. 2. The view of the segmented cells corresponding to the colon image of Fig. 1.

Table 1
The examples of the segmented cells under consideration.

Cell type	The representative images of defense cells
Lymphocyte	
Plasmocyte	
Eosinophilic granulocyte	
Neutrophylic granulocyte	

All presented above operations have been supported by the functions of Image Processing Toolbox of Matlab [10]. 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 database for further processing.

Part of the database is used for the purpose of learning the classifier. At the stage of learning we need not only the cell but also its class membership defined in advance. The class membership of the data used in learning was determined by the human experts of high experience.

Table 1 presents the magnified examples of cells representing the considered four types of cells: lymphocyte, plasmocyte eosinophilic granulocyte and neutrophylic granulocyte. The lymphocytes contain small amount of cytoplasm (most of the cell body is a nucleus) and this is very important diagnostic factor in their visual recognition. On the other hand the proportion of the area of the nucleus to the total area of the cell for plasmocyte is very similar and close to half. In the case of granulocytes few small nuclei are very characteristic.

4. 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 assume similar values for the cells 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.

4.1. The features based on the histogram

The histogram of the image carries a lot of information of the cell. The histograms can be created for the whole cell and for the nucleus. 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.

Fig. 3 presents the typical histograms of the cell images of the eosinophilic and neutrophilic granulocytes, lymphocyte and plasmocyte corresponding to the red color. They differ by the span, position and the value of maximum point as well as the distribution of bins. Moreover the histograms differ for different colors, so their descriptive parameters should be determined for all three R, G and B colors. 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. They have been defined for the whole cells, nucleus and cytoplasm, separately for each color. Up to 40 features have been defined in this way.

Table 2 presents the average values of the chosen statistical parameters characterizing the histograms of the whole cells. The data have been calculated for three colors: R, G and B (the first number corresponds to R, the second to G and the third one to B color). They represent the mean values of the whole families of cells (few hundred of cells of each family) available in the experiments. The differences among the values of these parameters, characteristic for each family, are evident.

4.2. The geometrical features

Different cell families differ by the size, and to some degree by the 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 radius, perimeter, area of the whole cell, the area of the nucleus and the cytoplasm, the compactness, symmetry, the major and minor axis lengths, the mean distance of pixels to the central pixel of the nucleus, the ratios of different parameters characterizing the nucleus and the whole cell. Up to 23 features have been created in this way. Table 3 presents the mean values and standard deviations (the number after the ± sign) of some chosen geometrical parameters calculated for the whole families of cells gathered in the database.

The significant differences among their values for all four considered cell types can be once again observed. However, notice that the standard deviations of some parameters are quite large, which means that these parameters are not quite stable for different representatives of cells belonging to the same family.

Please cite this article as: M. Kruk, et al., Recognition and classification of colon cells applying the ensemble of classifiers, Comp. Biol. Med. (2008), doi: 10.1016/j.compbiomed.2008.12.001

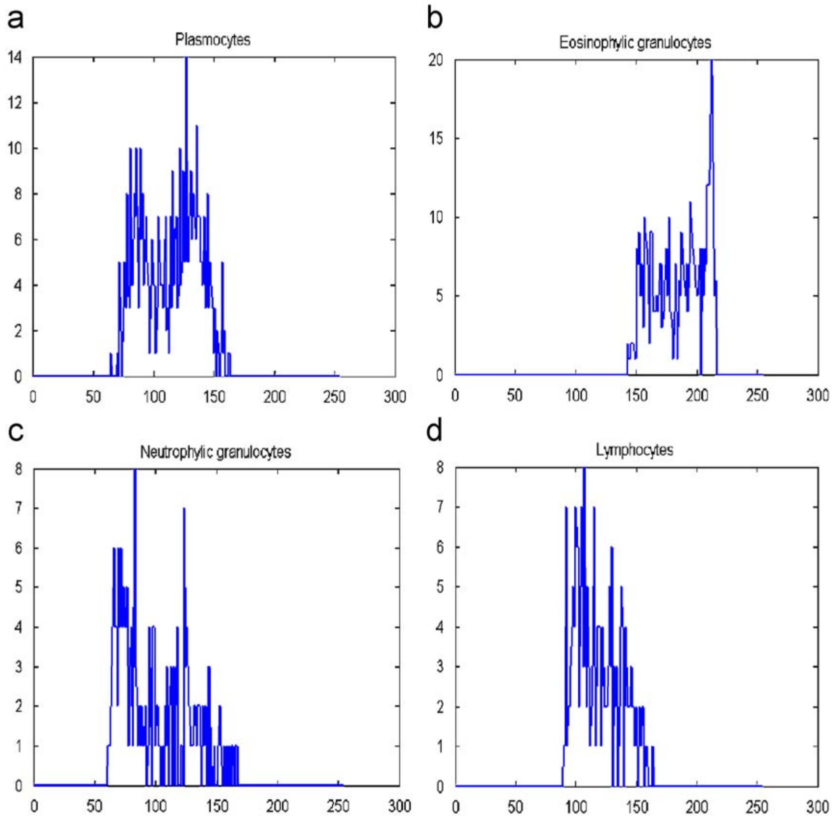


Fig. 3. The histograms of the images of the whole cells of (a) plasmocyte (b) eosinophilic granulocyte, (c) neutrophilic granulocyte, (d) lymphocyte.

Table 2
The average values of the statistical features of the histograms of the whole cells for RGB color representations.

	Lymphocyte			Plasmocyte			Eosinophillic 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

Table 3
The mean values of some geometrical features of the cells.

	Lymphocyte	Plasmocyte	Eosinophylic granulocyte	Neutrophilic granulocyte
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

4.3. 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 the spatial arrangements of the image pixels. They are seen as the changes in the intensity patterns, or the gray tones. For characterization of texture we have applied the Haralick matrix [18] 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 (all for gray representation of colors), separately for nucleus and separately for cytoplasm. They have been used as the textural diagnostic features. Up to 14 features have been defined in this way.

4.4. The colorimetric features

The last family of features has been defined on the basis of the image intensity. 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. As colorimetric features we have used the mean of pixel intensities of the whole cell, the nucleus and cytoplasm image for each color. Additional set of features was created in the form of ratio of the mean of the individual colors (R, G and B) to the mean of all colors. Up to 24 colorimetric features have been created in this way. Totally the features defined on the basis of the histogram, geometry, texture and color characterization form together the set of 101 components.

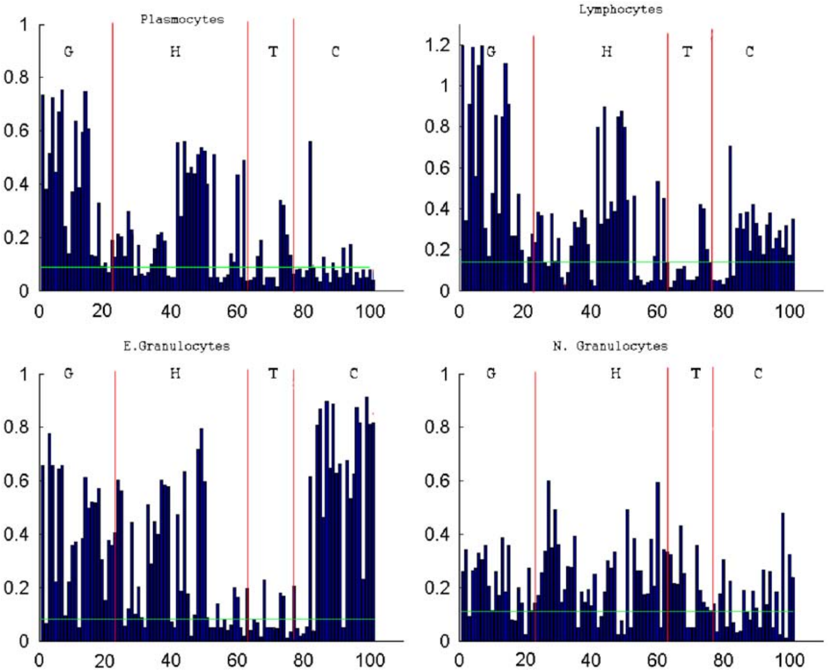


Fig. 4. Values of the discriminating coefficient $S_{AB}(f)$ of all features at the recognition between the recognized cells and the rest.

5. Feature selection

It is well known that features may have different impact on the classification process [3,11]. 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 for the data not taking part in learning.

There are many known techniques of feature selection. To the most important belong: the principal component analysis (PCA), analysis of the correlation existing among features, correlation between the features and the classes, application of feature ranking by applying the linear SVM, the analysis of mean and variance of the features belonging to different classes, etc. [3,11]. In our analysis we have chosen the last one, depending on the clusterization of the data and description of the clusters using their means and standard deviations.

It is evident that the variance of the features describing the cells belonging to the same 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 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

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

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 σ_A and σ_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 its small value means that this particular feature is not good for the recognition between classes A and B.

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. Two solutions are possible here. We can specialize the features for recognition of two particular classes or for recognition between the particular class and the mixture of the remaining classes. The choice depends on the applied strategy of classification.

Fig. 4 illustrates the change of the value of the discrimination coefficient $S_{AB}(f)$ for all extracted features at the recognition between the particular cell type (one among plasmocyte, lymphocyte, eosinophilic granulocyte and neutrophilic granulocyte) 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 (close to zero) to high values. All families of features are characterized by both high and low values of this coefficient. We may establish the horizontal line (such as in Fig. 4) denoting the bias, below which the feature may be regarded as insignificant. However, there is no mathematical reason how to choose the value of this bias. Therefore to apply it in practice we have to try different values of bias, train appropriate classifiers and on the basis of their results choose the optimal value of bias. 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 the features providing the highest efficiency of recognition.

6. The applied classifiers

In our cell recognition system we have applied five different classifiers: the multilayer perceptron (MLP), the radial basis function

Please cite this article as: M. Kruk, et al., Recognition and classification of colon cells applying the ensemble of classifiers, Comp. Biol. Med. (2008), doi: 10.1016/j.compbiomed.2008.12.001

(RBF) network, support vector machine (SVM), Fisher linear discriminant (FLD) and k nearest neighbor classifier (KNN). These 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. These are quite important factors when the results of their classification are subject to integration in the ensemble system.

MLP is a multilayer structure of many simple neuron-like processing units of sigmoidal activation function grouped together in layers [4]. The most important point in designing the MLP network structure is the generalization property. The number of weights and the number of hidden neurons 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 cannot be met. In practice on the basis of introductory experiments we have found that five hidden neurons of sigmoidal activation function were optimal and provided good generalization ability of the classifier network.

The RBF classifier is a network structure containing only one hidden layer of radial (Gaussian) neurons acting on the local basis [4], and as many linear output neurons as is the number of classes. The classes were coded in a binary form. The main difference to MLP is the local principle of operation of the RBF neurons. This results in a significant simplification of the learning procedure of RBF network. In practical implementation of this network 25 hidden Gaussian neurons have been applied.

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 [12,17]. 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 separation margin formed in the learning stage according to the assumed value of the regularization constant C provides some immunity of this classifier to the noise, inevitably contained in the testing data. The great advantage of SVM is the unique formulation of the learning problem leading to the quadratic programming with linear constraints, which is very easy to solve. The SVM of the 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. The optimal values of these parameters found in these experiments were as follows: $\gamma = 0.1$ and $C = 300$. To deal with a problem of many classes we have applied one against all strategy [5], 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 $\boldsymbol{y} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ of the norm $\|\mathbf{w}\| = 1$. The measure of separability of two classes denoted by the indices 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 [4]. The learning problem of FLD is transformed to the maximization of the function

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}} \tag{2}$$

where $\mathbf{S}_B$ is the between class scatter matrix and $\mathbf{S}_W$ the within-class scatter matrix.

The KNN classifier makes decision of class membership of the unknown vector $\mathbf{x}$ on the basis of its distances from k known (learning) vectors referred as the prototypes of classes. We assign the unknown vector to the class which appears most frequently in k selected prototypes identified in the previous step [4]. Usually the Euclidean distance of $\mathbf{x}$ to the prototypes is used. The result depends on the value of k and 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. Proper value of k has been selected by parameter optimization using cross-validation. As a result of such experiments we have found in our case $k = 3$ as the optimal one.

7. The ensemble of classifiers

Usually after learning different classifiers we select the best one, rejecting the others. Better approach to get the highest possible accuracy of the classifier system is to combine all of them into an ensemble [8]. In this way we pick up the best peculiarities of each classifier in order to increase the final accuracy of the cell recognition. Different approaches to the integration of the ensemble into one classifying system are used. To find the best possible integration rule we have applied and compared four different strategies: the majority voting, the weighted majority voting, the modified Bayes combination and PCA based two stage approach.

7.1. The majority voting

Suppose we have M classifiers, which were trained on the same data. The committee of these classifiers assigns the pattern to the class that obtains the majority of votes. Each classifier has the same influence on the final score. The majority voting is effective when the probability pr for each classifier to give the correct class label is equal for all input vectors $\mathbf{x}_i$ and at the same time the classifier outputs are independent. However, even in this case we can expect improvement over the individual accuracy pr only when pr is higher than 0.5. In the other case the majority voting integration does not bring any improvement over the individual classifier results [8]. In our case this condition of accuracy was satisfied.

7.2. The weighted majority voting

The weighted voting is a simple technique of establishing the winner taking into account the weighted majority. At M classifiers forming the ensemble, the probability $y_i(\mathbf{x})$ of ith class at the presentation of the input vector $\mathbf{x}$ is determined by the formula

$$y_i(\mathbf{x}) = \sum_{k=1}^{M} w_{ik} z_{ki}(\mathbf{x}) \tag{3}$$

where w_{ik} is the weight from ith class of the kth classifier to the ith output node of ensemble, and z_{ki} is the ith output signal of kth classifier, equal 1 when $\mathbf{x}$ belongs to ith class and zero in other case. The general classification system applying five ($M = 5$) mentioned above individual classifiers combined into one final classifying system is presented in Fig. 5.

Each classifier is fed by the same input vector $\mathbf{x}$ and produces the response in the form of four signals (one or zero) indicating the membership to one of four classes, representing four colon cell types under recognition. The integration matrix is formed by the weights w_{ij}, with the first index indicating the class and the second—the appropriate classifier. In this solution the weight w_{ij} has been adjusted according to the accuracy of jth classifier at recognition of ith class. Different integration formulas might be applied here. In the case of similar accuracy of each classifier the most reasonable seems to be the simple formula

$$w_{ij} = \frac{\eta_{ij}}{\sum_{k=1}^{M} \eta_{ik}} \tag{4}$$

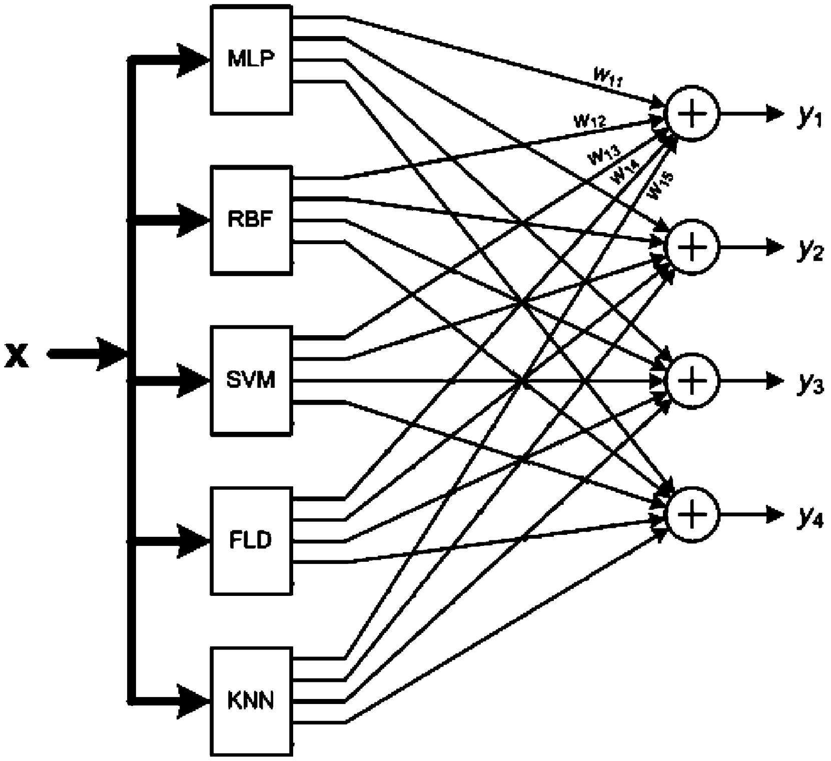


Fig. 5. The ensemble of classifiers for colon cell recognition.

where η_{ik} means the accuracy of kth classifier ($k = 1, 2, \ldots, 5$) at the recognition of ith class ($i = 1, 2, \ldots, 4$) on the learning data (the ratio of the number of proper classifications to the total number of samples belonging to ith class, taking part in learning). When the classifiers are of significantly different accuracy more recommended is the modified form

$$w_{ij} = \frac{\eta_{ij}^m}{\sum_{k=1}^M \eta_{ik}^m} \tag{5}$$

with the exponent m adjusted by the user (the typical value $m = 3$). The final result of classification is determined on the basis of the output signals $y_i(\mathbf{x})$ determined by formula (3) for $i = 1, 2, \ldots, 4$. The highest value of $y_i(\mathbf{x})$ indicates the membership of the applied vector $\mathbf{x}$ to the appropriate class.

7.3. *The modified Bayes combination*

The basic assumption of this method is that the classifiers are mutually independent given a class label. We apply here the modification of the Bayes combination [7,8] since it gives more reliable results at zero estimated probability of some classes. According to this modification the ensemble probability $\mu_j(\mathbf{x})$ supporting the jth class is determined on the basis of the known results of testing the networks on the learning data and is given in the form

$$\mu_j(\mathbf{x}) = \prod_{i=1}^{M} \frac{cm_{js_i}^{(i)} + 1/N}{n_j + 1} \tag{6}$$

where n_j is the total number of elements in training set for class j and $cm_{js_i}^{(i)}$ is the element of the confusion matrix generated for learning data of ith classifier at indication of it to s_ith class as a response to the vector $\mathbf{x}$. The (j, s_i)th entry of the confusion matrix is the number of elements of the learning data set whose true class label was j and were assigned by ith classifier to s_ith class.

7.4. *The PCA based two step integration*

Consider a data arranged in the matrix $\mathbf{A}$ of the size $p \times (c \cdot M)$, where p is the number of data points, M —number of classifiers and c —number of classes. Each classifier output (a class label) is in the form of a binary vector $\mathbf{z}$ with 1 in the position of the recognized class label and 0s at the other positions. A row of the matrix $\mathbf{A}$ is a composition of the concatenated outputs of M classifiers for the respective data point. The aim of the first step is to map these p objects into a lower-dimensional space. In this way each data point $\mathbf{z}$ (the row of the matrix $\mathbf{A}$) will be represented now by the vector $\mathbf{y}$ of smaller dimension K, containing sufficiently high percentage of the original information. This can be done by using PCA. On the basis of the learning data set we form the PCA matrix $\mathbf{W}$ transforming the $c \cdot M$-dimensional vector $\mathbf{z}$ into K-dimensional vector $\mathbf{y}$, where $\mathbf{y} = \mathbf{W}\mathbf{z}$. The lower dimension vectors $\mathbf{y}$ are used in the second stage as the training data for the next classifier, responsible for the final recognition of the classes. The SVM system working in one-against-one mode is used here as the final classifier.

8. The results of numerical experiments

8.1. *Database*

The numerical experiments of the cell recognition have been performed using the database of cells collected from the microscopic colon images acquired in the Department of Pathomorphology, Military Institute of the Heath Services, Warsaw, Poland. We have used the images of tissues of the magnification equal 600×. Table 4 presents the amount of cells within each family taking part in experiments. These cells have been obtained from 53 patients of

Please cite this article as: M. Kruk, et al., Recognition and classification of colon cells applying the ensemble of classifiers, Comp. Biol. Med. (2008), doi: 10.1016/j.compbiomed.2008.12.001

Table 4
The database of the segmented cells used in numerical experiments.

	Lymphocytes	Plasmocytes	Eosinophilic granulocytes	Neutrophilic granulocytes
Number	511	420	253	187

Table 5
The average misclassification ratio at recognition of four cell families by application of individual classifiers (with reference to the human expert results).

Cell family	RBF (%)	MLP (%)	SVM (%)	FLD (%)	KNN (%)
Lymphocytes	5.6	5.3	3.9	3.5	7.6
Plasmocytes	8.3	8.3	6.4	7.6	9.3
Eosinophillic granulocytes	5.9	5.1	5.9	6.7	7.5
Neutrophillic granulocytes	10.1	10.1	12.3	14.9	15.0
Total	7.17	6.86	6.24	6.89	9.11

different advancement of illness. The number of representatives of each class is significantly different. The granulocytes (especially neutrophilic type) appear only at the higher stage of advancement of illness, and hence are very scarce.

For each cell we have generated the set of features forming the potential input vector **x** applied to the classifiers. From 101 candidate features after application of the selection procedure we have selected different number of the most important features (50 for limphocytes, 79 for plasmocytes, 53 for eosinophilic granulocytes and 48 for neutrophilic granulocytes) according to their experimentally checked importance for the recognition process between the classes of cells.

8.2. The results of experiments

The available data set has been split into five exchangeable parts to enable application of the fivefold 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 trained classifiers. The experiments have been repeated five times, exchanging the contents of the four learning and one testing subsets. The misclassification ratio in either learning or testing mode has been calculated as the mean of all five runs. The same cells have been also recognized by the human expert and his results have been compared to the output of our automatic system. We are aware of the fact that the human score might be not quite accurate and biased because of some imperfections of coloring of the slides, the way of cutting the biopsy, etc. To some degree it may be prevented by asking few experts to do the same job and taking for comparison their average number of cells. However, this problem is still open.

Table 5 presents the statistical results of the average misclassification ratio at recognition of four considered types of cells by the individual classifiers, referring to the testing data not taking part in learning. They have been calculated as the ratio of the misclassified data points to the total number of representatives of the class, averaged over all cross-validation trials. The misclassification is understood as the discrepancy between our results and human expert scores. Additionally the bottom line presents the total errors of recognition of all classes by each classifier. They have been calculated as the ratio of the number of all misclassified cases to the total number of data points (due to different population of all classes it is not equal to the mean of the errors for all classes). It is seen that the classifiers recognize the cell families with different accuracy. The smallest relative error has been obtained at application of SVM and the least accurate classifier was KNN.

The highest misclassification ratio of all individual classifiers corresponds to the neutrophilic granulocyte family. There are two sources of this. First, the number of neutrophilic granulocytes was the smallest one (see Table 6), since they appear only in the highest stage of illness and usually in a small quantity. Therefore even at small number of misclassifications the percentage error might be high (at 1 misclassification only and 10 samples the percentage error is 10%). Secondly the neutrophilic granulocytes are of largely varying shape and color. Some of them resemble the other types of cells.

Table 6
The confusion matrix of the best classifier (SVM) at the testing data.

	L	P	E	N
L	491	15	1	4
P	11	393	7	9
E	4	7	238	4
N	4	12	7	164

Table 7
The confusion matrix of the best ensemble (PCA) of classifiers at the testing data.

	L	P	E	N
L	490	13	4	4
P	11	390	9	10
E	2	4	243	4
N	4	6	5	172

The improvement of the recognition results may be achieved by applying the ensemble of classifiers integrated into one classification system. Four different integration methods have been tried in experiments: the simple majority, weighted majority, modified Bayes and PCA approach. The weighted majority was performed with the weights w_{ij} adjusted by using expression (5) with the value of $m=4$. The best results of PCA approach have been obtained at reduction of the vector dimension from 20 (five classifiers and four classes) to 8.

Simple majority voting applied for integration of their results in the form of ensemble system resulted in a small improvement with respect to the best individual classifier (the total error 6.11% compared to 6.24% of the best SVM classifier). Application of the weighted majority voting has reduced the total error to the value of 5.99%. The modified Bayes integration rule resulted in a similar improvement (total error equal 6.03%). The highest reduction of the total error has been achieved at application of PCA approach. This time the total error was equal only 5.58%. This means 10.5% of relative improvement over the best single (SVM) classifier. Observe that in each case the integration of results of many classifiers resulted in the improvement of the final results.

Fig. 6 presents the misclassification ratios of all five individual classifiers (SVM, FLD, KNN, RBF and MLP) compared to the ensemble system combined by using different integration strategies (majority, weighted majority, modified Bayes and PCA). It is evident that integration of results of many classifiers, even of unequal efficiency, has improved the final results of classification. The best total result of recognition of all classes has been improved from 6.24% (misclassification rate of the best SVM classifier) to 5.58% (ensemble of five classifiers integrated according to the 2-stage PCA approach).

Analyzing the results of integration of classifiers in an ensemble we can notice quite important difference. The majority and modified Bayes approaches resulted in a similar improvement (around 6% of the final misclassification rate), since all of them rely on a quite similar basis (the efficiency of recognition of samples on the learning data set). The PCA approach applies two stages of classification and its principle of operation is significantly different. In our particular task this system of integration appeared to be the most efficient.

Tables 6 and 7 present the details of the class recognition errors of the best individual classifier (SVM) and the best ensemble system (PCA) in the form of so-called confusion matrix, represented as the

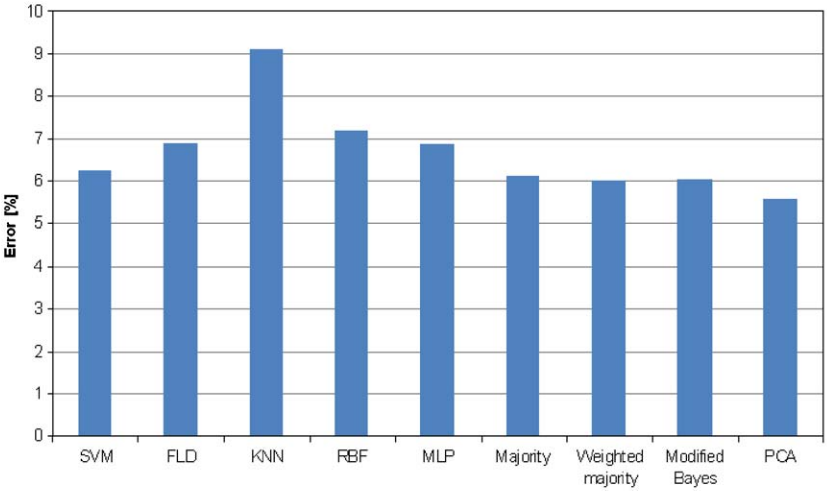


Fig. 6. The comparison of the misclassification ratios of the individual classifiers and the ensemble of classifiers by using different integration formulas.

summed results of the cross-validation procedure for the testing data (class L—limphocyte, P— plasmocyte, E—eosinophilic granulocyte, N—neutrophilic granulocyte). The diagonal entries represent the quantity of the properly recognized cells of the particular class.

Each entry outside the diagonal means error. The entry in the (i,j)th position of the matrix means false assignment of i th class to the j th one. The number of outside diagonal terms (the misclassifications) in the case of ensemble is smaller than for the best individual classifier. The most visible advantage of the application of ensemble is significant reduction of the misclassification cases for the neutrophilic granulocyte cells (the cells recognized with the worst accuracy). The relative error has been dropped from 12.3% of the single classifier to the value of 8.0% in the case of ensemble.

The additional output of the procedure is the annotated image of the colon tissue. After recognition of the image the cells are automatically annotated according to the recognition results. The letters L, P, E and N, used in Tables 6 and 7, have been also applied for denoting different families of cells. Fig. 7 presents the view of two exemplary microscopic images of the colon tissue with the cells annotated automatically by our system. The image of Fig. 7a corresponds to the first stage of illness, when there is no neutrophilic granulocytes in the stoma. The characteristic is a small number of the defense cells in the viewing field of a stoma of tissue. The second image represents the deeper stage of illness, when the neutrophilic cells appear in the stoma. There is also visible significant increase of the total number of defense cells in the viewing field of the image.

The annotated image is an important result of the work and is of great help for the medical staff in the visual inspection of the results of the cell recognition made by an automatic system. If, in the opinion of the human expert, the class membership of some particular cell is not proper, he may introduce immediately his correction to the annotated image using the graphical interface being the integral part of the system. This manual correction automatically results in changing the class membership of the pointed cells and also correcting the size of each cell family and the other medical parameters calculated on this basis. Such interactive cooperation of the system and the human expert can easily compensate the effect of some imperfection of the system performance, caused by the measurement noise or by the nonideality of the chemical processes at the formation of the colon image.

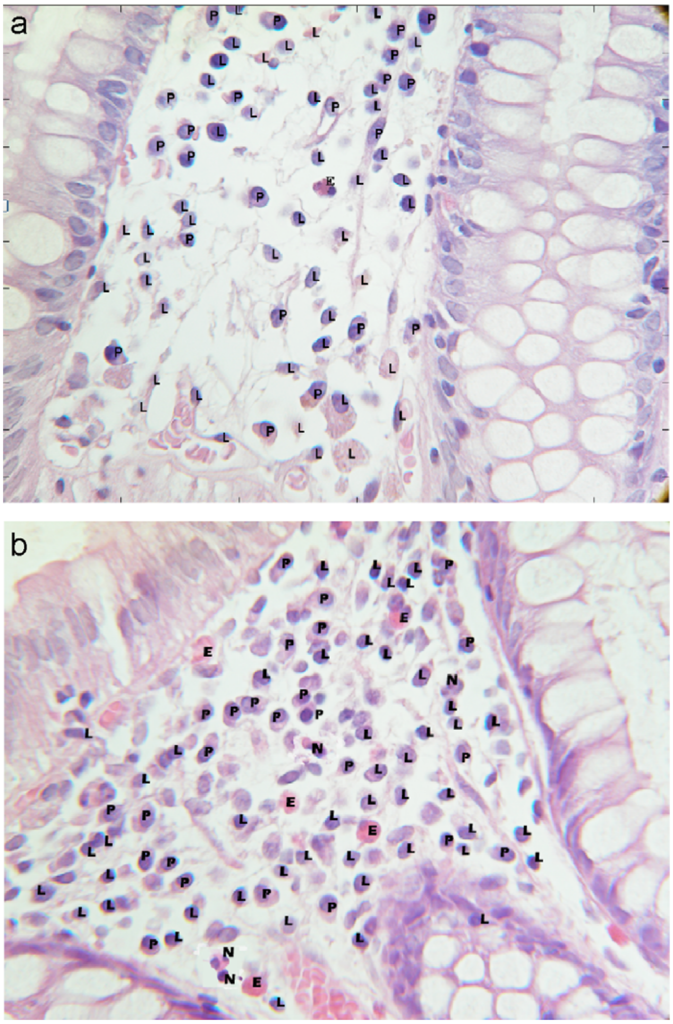


Fig. 7. The output images of the colon with the annotated cells:L—lymphocytes, P—plasmocytes, E—eosinophilic granulocytes, N—neutrophilic granulocytes: (a) the patient of the first stage of illness, (b) the patient of more advanced illness.

9. Conclusion

The paper has presented the automatic method of the recognition of different cells existing in the microscopic image of the colon. The most important problems solved in the work include: the segmentation of the image to extract the individual cells, preprocessing of the cell images to generate the numerical diagnostic features, selection of the most important features and finally recognition of the cells using the ensemble of many classifiers.

The numerical verifications of the proposed classification system have been done on the database of more than 1000 cells, representing the eosinophilic and neutrophilic granulocytes, plasmocytes and lymphocytes. The results of numerical experiments have shown that the mean discrepancy rate between the score of our automatic system and the human expert results for all cells is below 6% and this accuracy is acceptable in the medical practice (the acceptable differences between the scores of different experts are up to 15%). Thus the system has the potential ability for supporting the medical diagnosis simplifying and greatly accelerating the process of cell counting.

The additional advantage of the proposed automatic system is the annotated image of the colon tissue produced by the program. This image might be of great help for the medical staff in the visual inspection of the results prepared by an automatic system. The human expert is equipped with the graphical tools for the possible immediate correction of this annotated image. These tools enable automatic changes of the class membership of the individual cell as well as the population of each cell family following the corrections made by the expert.

Conflict of interest statement

None declared.

References

[1] V.A. Botoman, G.F. Bonner, D.A. Botoman, Management of inflammatory bowel disease, American Family Physician, January 1, 1998 ⟨http://www.aafp.org/afp/980101ap/botoman.html⟩.

[2] H. Breu, J. Gil, D. Kirkpatrick, M. Werman, Linear time Euclidean distance transform algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (5) (1995) 529–533.

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

[4] S. Haykin, Neural Networks, Comprehensive Foundation, Prentice-Hall, Englewood Cliffs, NJ, 1999.

[5] C.W. Hsu, C.J. Lin, A comparison methods for multi class support vector machines, IEEE Transactions on Neural Networks 13 (2002) 415–425.

[6] G. Kayser, D. Radziszowski, P. Bzdyl, M. Werner, K. Kayser, Eamus—internet based platform for automated quantitative measurements in immunohistochemistry, in: International Conference on Society for Cellular Oncology (ISCO), Belfast, 2005.

[7] M. Kruk, S. Osowski, R. Koktysz, Recognition of colon cells using ensemble of classifiers, in: IJCNN, Orlando, 2007.

[8] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley, New York, 2004.

[9] V. Luc, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (6) (1991) 583–598.

[10] Matlab User Manual—Image Processing Toolbox, MathWorks, Natick, 2003.

[11] S. Osowski, T. Markiewicz, Support vector machine for recognition of white blood cells in leukemia, in: G. Camps-Valls, J.L. Rojo-Alvarez, M. Martinez-Ramon (Eds.), Kernel Methods in Bioengineering, Signal and Image Processing, Idea Group Publishing, London, 2007, pp. 93–123.

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

[13] P. Soille, Morphological Image Analysis, Principles and Applications, Springer, Berlin, 2003.

[14] 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, Clinical and Laboratory Haematology 25 (2003) 139–147.

[15] N. Theera-Umpon, P. Gader, System-level training of neural networks for counting white blood cells, IEEE Transactions on SMS 32 (2002) 48–53.

[16] E.M. Thompson, A.B. Price, D.G. Altman, C. Sowter, G. Slavin, Quantitation in inflammatory bowel disease using computerized interactive image analysis, Journal of Clinical Pathology 38 (1985) 631–638.

[17] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

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

Please cite this article as: M. Kruk, et al., Recognition and classification of colon cells applying the ensemble of classifiers, Comp. Biol. Med. (2008), doi: 10.1016/j.compbiomed.2008.12.001