FEATURE GENERATION FOR THE CELL IMAGE RECOGNITION OF MYELOGENOUS LEUKEMIA

Stanislaw Osowski[1,2], Tomasz Markiewicz[1], Bozena Marianska[3], Leszek Moszczyński[1]

[1]Warsaw University of Technology, [2]Military University of Technology, Warsaw, [3]Institute of Haematology, Warsaw, POLAND, Email: sto@iem.pw.edu.pl

ABSTRACT

The paper presents the preprocessing methods of the leukemic blast cells image in order to generate the features well characterizing different types of cells. The solved problems include: the segmentation of the bone marrow aspirate by applying the watershed transformation, selection of individual cells, feature generation on the basis of texture, statistical and geometrical analysis of the cells. These features are used as the input signals applied to the support vector machine used as the classifier. The numerical results of recognition of 12 different cell types are presented and discussed.

1. INTRODUCTION

The acute leukemia is a disease of the leukocytes and their precursors. It is characterized by the appearance of immature, abnormal cells in the bone marrow and peripheral blood. The aspirated marrow is found to be infiltrated by abnormal cells.

The recognition of the blast cells in the bone marrow of the patients suffering from myelogenous leukemia is a very important step in the recognition of the development stage of the illness and proper treatment of the patients [2,3,9]. The percentage of blasts is a major factor at defining various subtypes of acute myeloid leukemia. According to French-American-British (FAB) standard, 8 acute leukemia types are classified on the basis of the ratio of myelo/monoblasts, the number of erythroid precursors or non-erythroid cells as well as megacarioblasts cells. It is known that proper treatment of leukemia requires not only recognition of different stages of the development of the blasts but also calculation of their quantity in the aspirated bone marrow.

We can find many different cell types in the bone marrow. The most known and recognized abnormal cells include monoblasts, promonocytes, monocytes, myeloblasts, promyelocytes, myelocytes, metamyelocytes, proerythro-blasts, basophilic erythroblasts, polychromatic erythroblasts, orthochromatic erythroblasts, lymphocytes, plasmocytes, megacaryoblasts, megacaryocytes, etc [2,3]. The variety of cells occurring in the bone marrow demands a high expertise of the analyst, which is usually verbal one. For improving the reliability of the analysis and diagnosis, computer based digital image processing offers a useful tool.

This paper is dedicated to the task of feature generation for the automatic blast cell recognition. The well-defined features should suppress the differences among the cells belonging to the same class and amplify them for cells belonging to different classes. These features are applied to the support vector machine network (SVM) fulfilling the role of recognizing and classifying system. The presented solution may be treated as the first step in building up an automatic system able to recognize different blood cells.

2. AUTOMATIC SEGMENTATION OF THE IMAGE

Image segmentation is a division of the image into different regions, each having certain properties. In a segmented image, the picture elements are no longer the pixels, but connected set of pixels, all belonging to the same region. We will use the segmentation techniques to separate the individual cells from the set of cells creating the image.

The recognition and separation of individual cells from the image of the blood cells is a very difficult task, since different regions are of little grey level variations and the borders of individual cells are hardly visible. Fig. 1 presents the exemplary image of the blast cells of the bone marrow containing different types of cells.

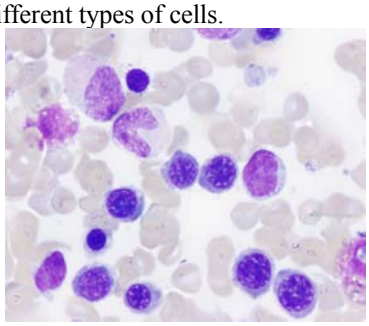

Fig. 1: The exemplary image of the bone marrow smear of the acute leukemia patient containing different blast cells.

The individual cells are close to each other and the borders among them are not well defined. The task of segmentation of the image is focused on the automatic recognition and separation of each cell for further processing, in order to obtain stable features, useful in recognition of the cell. In solving the segmentation task we have used the morphological operations.

The morphological operations aim at extracting relevant structures of the image by probing the image with another set of a known shape called structuring element, chosen as the result of prior knowledge concerning the geometry of the relevant and irrelevant image structures. The most known morphological operations include erosion, dilation, opening

and closing [6]. The morphological approach to image segmentation combines regions growing and edge detection techniques. It groups the pixels around the regional minima of the image. The boundaries of adjacent grouping are precisely located along the crest lines of the gradient image. In our experiments, we have accomplished this through an operation called the watershed transformation. The watershed transformation [4,6] takes its origin from the topographic interpretation of the gray scale image. According to the law of gravitation, the water dropped on such surface will flow down until reaches a minimum. The whole set of points of the surface, whose steepest slope paths reach a given minimum, constitutes the catchment's basis associated with this minimum. The watersheds are the zones dividing adjacent catchment's basins.

In numerical implementation of the watershed algorithm the original image is transformed so, as to output an image whose minima mark relevant image objects and whose crest lines correspond to image object boundaries. In this way the image is partitioned into meaningful regions that may correspond for example to the individual blast cells. In our experiments we have used the watershed algorithm implemented using Matlab platform [10]. The applied procedure of the image segmentation and cell separation consists of the following stages:

- Transformation of the original image into gray scale.
- Transformation of the gray image to binary one by applying the biased segmentation.
- Application of closing and erosion operations to smooth the contours and to eliminate the distortions.
- Generation of the map of distances from the black pixel to the nearest white pixels.
- Application of the watershed algorithm for the image segmentation.

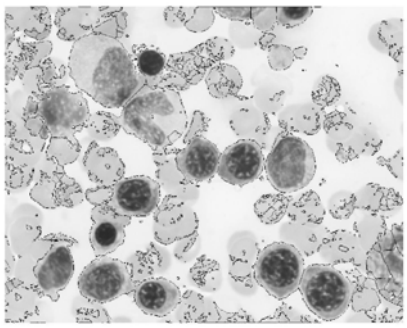

Fig. 2: The segmented image of the bone marrow smear.

The result of such image preprocessing is the image segmented into the regions corresponding to individual cells, with the visible borders between cells.

Fig. 2 presents the result of the segmentation of the exemplary image of the aspirated bone marrow shown in Fig. 1. The dark regions represent different blasts. All of them have been separated as individual regions. Some small inaccuracies are visible, especially at the border of cytoplasm, but the number of such errors is very limited. The blast cells are composed mostly of nuclei. Their shapes are easily visible and have been well reconstructed by the algorithm.

3. FEATURE GENERATION

The efficient recognition of different cells requires the generation of the features of very good discriminative ability. In this work we have developed the features belonging to three main groups of the textural, geometrical and statistical nature.

3.1 Texture feature description

Texture refers to the arrangement of the basic constituents of a material. In digital image the texture is depicted by the interrelationships between spatial arrangements of the image pixels. They are seen as changes in intensity patterns, or the gray tones.

The efficient recognition of the texture images requires the preprocessing of them in order to extract the features, characterizing the image in a way suppressing the differences within the same class and enhancing the differences between textures belonging to different classes. There are many different techniques of texture preprocessing for extraction of such features [8]. Each of the preprocessing methods stresses different features of the texture and only numerical experiments can settle which of them is most suitable. In this work after some experiments we have limited ourselves to only two texture preprocessing methods, due to Unser and Markov [8].

Unser algorithm of feature generation is the simplified version of the Haralick method. The sum and difference histograms of gray levels of the neighboring pixel are created for different directions, for example 0^o, 45^o, 90^o and 135^o. 32 different features on the basis of these histograms can be defined for one color. On the basis of some numerical experiments the following features have been selected: mean value, angular second momentum, contrast and entropy of the intensity of the neighbouring pixels in one chosen direction. The values have been calculated for three colors, independently for nucleus and cytoplasm. They include the differences and sums of the pixels intensity. Taking this into account 42 features have been generated.

In Markov random field method the signal at each pixel location is regarded as a random variable. Each type of texture is characterised by a joint probability distribution of signals that accounts for spatial inter-dependence, or interaction among the signals. The random field texture model is characterised by geometric structure and quantitative strength of interactions among the neighbours. At the assumption that the pixel interactions are translation invariant, the interaction structure is given by a set N of characteristic neighbours of each pixel. The autoregressive model parameters of the probability distribution have been used as the features. For this particular application we have generated 11 Markov features .

Both methods of feature generation produce different sets of features. Some of them represent the textures of different blast cells in less or more distinctive way. After generating the whole feature set, the discriminative quality of features are analyzed and only the best are chosen, while the rest is discarded.

3.2 The geometrical features

The important information is contained in the geometrical shapes and parameters [9] associated with them. Various cells differ greatly with the size. For example the

orthochromatic erythroblasts have the size of 8-10 micrometer, while megakcariocyte may be up to 100 micrometer. The shapes of different blasts are either round, oval or kidney-shaped. We have used the following geometrical features of the cells:

- radius –measured by averaging the length of the radial line segments defined by the centroid and border points
- perimeter - the total distance between consecutive points of the border
- the ratio of the perimeter and radius
- area – the number of pixels on the interior of the cell, defined separately for the nuclei and for the whole cell; as the features we assume the area of the nucleus and the ratio of the areas of the nucleus and the whole cell
- the area of convex part of the nucleus
- compactness – given by the formula: perimeter2/area
- concavity – the severity of concavities in a cell
- concavity points – the number of concavities, irrespective of their amplitudes
- symmetry – the difference between lines perpendicular to the major axis to the cell boundary in both directions
- major and minor axis lengths.

11 geometrical features have been generated in this way.

3.3 Statistical features of the image

The next set of features has been generated on the basis of the intensity distribution of the image. The histograms and gradient matrices of such intensity have been determined for three color components R, G and B. On the basis of this the following features have been generated: the mean value and variance of the histogram and the gradient matrix of the image of the nucleus and the whole cell (24 features), skewness and kurtosis of the histogram and gradient matrix of the whole cell (12 features). All these features have been calculated for three colors. 36 features have been generated in this way.

All numerical experiments of feature generation have been implemented on the platform of Matlab [10].

4. THE RESULTS OF NUMERICAL EXPERIMENTS

Fig. 3 presents four different exemplary blasts: neutrophil (cell 1), mono/myelo blast (cell 2), basophilic erythroblast (cell 3) and polychromatic erythroblast (cell 4).

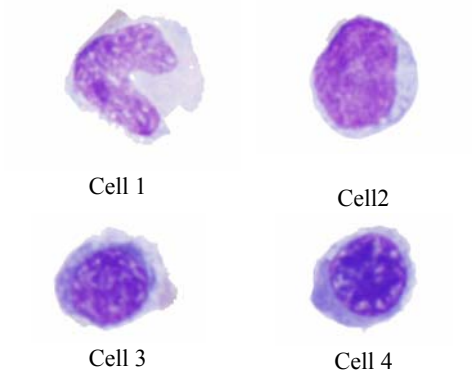

Fig. 3: The blast cells used in numerical experiments.

All of them have been extracted from the image of Fig. 2. They are of different shapes and area. The most prominent part of the cell forms the nucleus (the dark part of the cell) of very special texture and shape. The cytoplasm corresponds to the light colour of the cell at the border of it. The texture features corresponding to different methods have been analyzed. In the case of Markov features the highest discriminative ability had 11th feature. Fig. 4 presents the distribution of samples of this feature for these 4 types of blasts and different cells extracted from bone marrow.

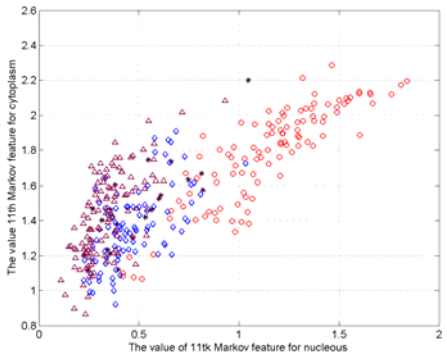

Fig. 4: The distribution of cell locations in plane formed by 11th Markov feature corresponding to cytoplasm and nucleus .

Table 1 depicts differences among chosen geometrical features for the individual representatives of the same set of blast cells shown in Fig. 3.

Fig. 5 shows the distribution of points for these 4 types of cells at different samples of bone marrow on the plane formed by two chosen geometrical features: the ratio of area of nucleus to the area of cell versus the area of cell.

Feature	Cell 1	Cell 2	Cell 3	Cell 4
Radius	89.35	96.42	85.37	81.12
Perimeter	2105	1005	1080	1067
Area	31053	29094	22867	21824
Concavity	945	32	200	211
Symmetry	9014	1717	1632	1578
Compactness	142.70	34.72	50.99	52.17

Table 1: The chosen geometrical features of the nuclei of 4 different blast cells of Fig. 3

The numerical experiments concerning recognition and classification of cells have been performed for 12 different leukemic blast types (classes). They include: basophilic erythroblast (1), poly-chromatophilic erythroblast (2), ortochromatic erythroblast (3), myeloblast/monoblast (4), promyelocyte (5), neutrophilic myelocyte (6), neutrophilic metamyelocyte (7), neutrophil (8), eosinophil (9), prolymphoblast (10), lymphocyte (11) and plasmocyte (12). The numbers in parentheses denote our notation of these particular classes. They represent different cell lines in the bone marrow as well as different stages of development within the same line.

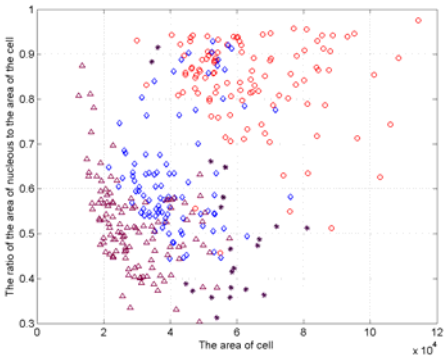

Fig. 5: The distribution of cell locations in the plane formed by 2 geometrical features.

The important problem in assessing the discriminative ability of the generated features for the recognition of the cell is the distribution of these values at many representatives of the same type of the cells. Good features should have similar values for the same type of cells and should differ significantly for different types of cells. After normalization of data the correlation analysis as well as variance and mean value analysis of the clusters have been used to indicate the most and least important features. The least important features have been removed. As a result we have limited the total number of features to 70 and vector x has been also reduced to the same dimension.

As a classifier we have applied the Support Vector Machine (SVM), developed by Vapnik [7]. The learning problem of SVM is formulated as the task of separating the learning vectors xi into two classes of maximal separation margin. All operations in learning and testing modes are done in SVM using so called kernel function, satisfying the Mercer conditions [5,7]. As the kernel we have used the Gaussian function of σ=0.5. At 12 classes we have adopted one-against-one approach. The numbers of available learning and testing data (column 2 and 5 of table 2, respectively) vary from class to class and are dependent on the actual availability of the particular type of cells.

Class	2	3	4	5	6	7
1	50	1	0	46	5	1
2	59	7	0	60	6	1
3	35	3	1	34	5	1
4	54	0	0	54	3	1
5	14	6	0	6	4	0
6	14	0	0	14	3	2
7	10	3	1	11	3	0
8	10	3	0	11	4	2
9	18	0	0	17	2	1
10	10	1	0	11	2	2
11	69	1	1	69	3	1
12	11	0	0	10	0	0
Total	354	25	3	343	40	12

Table 2 : The results of numerical experiments of recognition of 12 classes of cells

Table 2 presents the results of these experiments. The first column presents the notation of the class, column 2 – the number of cells used in learning, column 3 – the number of all misclassifications among learning data, column 5 – the number of cells used in testing only and column 6 – the number of all misclassifications on the testing data set.
In our data set there were cells close to each other in their development stage. Recognition between two neighboring cells is dubious since the transition from one cell to another is continuous and even expert is in trouble in recognizing between them. If we neglect such misclassifications we obtain different (reduced) misclassification rate (column 4 – the learning data, column 7 – the testing data).
As it is seen the recognition accuracy among different cells is changing and depends on the type of cell and on the actual number of data used in learning.

5. CONCLUSIONS

The paper has presented the image processing approach to the recognition and classification of the leukemia cells. The most important points of this approach are: segmentation of the image of bone marrow aspirate using watershed algorithm, the extraction of individual cells from the image, automatic generation of different features of the cell, assessment of the feature quality of the cells using analysis of distribution, correlation and principal component analysis, application of support vector machine for final recognition and classification of cells.

REFERENCES

[1] S. Haykin, "Neural networks, comprehensive foundation", Prentice Hall, New Jersey, 1999

[2] H. Hengen, S. Spoor, M. Pandit, "Analysis of blood & bone marrow smears, SPIE Med. Imag., San Diego, 2002

[3] K. Lewandowski, A. Hellmann, "Haematology atlas", Multimedia Medical Publisher, Gdansk, 2001

[4] O. Lezoray, H. Cardot, "Cooperation of color pixel classification schemes and color watershed", IEEE Trans. Image Processing, vol. 11, pp. 783-789, 2002

[5] O. L. Mangasarian, P. Lagrangian, "Support Vector Machines", Journal of Machine Learning , 161-177, 2001

[6] P. Soile, "Morphological image analysis, principles and applications", Springer, Berlin, 2003

[7] V. Vapnik, "Statistical Learning Theory", Wiley, N.Y., 1998

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

[9] W. Wolberg, W. N. Street, O. L. Mangasarian, "Machine learning to diagnose breast cancer from image-processed features", Rep. of Uni. Wisconsin, 1994

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