Analysis of Features for Blood Cell Recognition

Tomasz Markiewicz, Leszek Moszczyński

Abstract - **The paper will present the method of assessment of features for the recognition of blood cells characteristic for myelogenous leukemia. The features relied on the application of texture, geometry and statistical analysis of the image of the cells are compared to each other and either left or reduced from the set, according to the assessed importance. The results of numerical experiments will be presented and discussed.**

I. INTRODUCTION

The problem of recognition of the blood cells in the bone marrow is a very important subject in the treatment of the patients suffering from myelogenous leukemia. The percentage of blasts is a major factor at defining various subtypes of acute myeloid leukemia [3, 4].

It is known that proper treatment of leukemia patients requires not only recognition of different stages of the development of the blasts but also tedious calculation of their quantity in the aspirated bone marrow.

We can find different cell lines in the bone marrow: the megacaryocytic series, erythrocytic series, monocytic series and granulocytic series. The most known and recognized abnormal cells include monoblasts, promonocytes, monocytes, myeloblasts, promyelocytes, myelocyte, metamyelocytes, proerythroblasts, basophilic erythroblast, polychromatic erythroblast, eosinophilic erythroblast, megacaryoblasts, promegacaryocytes and megacaryocytes [3,4].

The variety of cells occurring in the bone marrow demands a high expertise of the analyst, which is usually verbal one. The reproducibility of the results is sometimes lacking, since the differences among different cell types are very difficult for recognition. For improving the reliability of the analysis and diagnosis computer based digital image processing offers a useful tool. Automatic segmentation of the image of the smear of the bone marrow, leading to the selection of different types of cells, can accelerate the recognition process and at the same time make it more accurate.

This paper will discuss different aspects of the system for automatic generation of the features for the recognition of different cell lines. These features will form the input vector applied to the classifiers, in our case the Support Vector Machines [11]. They will be relied on the application of texture, geometry and statistical analysis of the image of the cells. The results of numerical experiments will be presented and discussed.

II. GENERATION OF FEATURES

The first step in solving the task is extraction of the individual cells from the whole image of the smear of the bone marrow. This was done by applying the morphological operations [1,5], implemented in the form of the watershed algorithm. The individual cells have undergone the signal processing aimed at the extraction of the features well defining the images of the cell. Good features should generate the features close to each other for different representatives of the same class and far distant for the representatives of different classes. In our solution we have relied on three types of features: texture, geometry and pixel intensity histograms [10,11].

A. Texture features

Texture refers to the arrangement of the basic constituents of a material [7]. 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 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 [7,8]. The most common include: the Haralick gray level co-occurrence matrix features (GLCM), Markov random field features, Unser sum and difference histograms, Gabor transformation features, wavelet or fractal descriptions, etc [5,6]. 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 first preprocessing methods.

B. The nuclear geometrical features.

The important information concerning the blast cells is contained in the geometrical shapes and parameters associated with them [2]. It is known from the observation that various cells differ greatly

T. Markiewicz is with the Institute of the Theory of Electrical Engineering, Measurement and Information Systems, Warsaw University of Technology, Warsaw, Poland, e-mail: markiewt@iem.pw.edu.pl, L. Moszczyński is with Institute of Material Technology, University of Technology, Warsaw, Poland

with the size. For example the eosinophilic erythroblasts have the size of 8-10 micrometer, while megacariocyte 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 the 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 and adding one-half of the pixels on the perimeter. It is defined separately for the nuclei and to the total cells. As the features we assume the area of the nucleus and the ratio of the areas of the nucleus to the whole cell.
- The area of convex part of the nucleus
- compactness – measured by the formula: perimeter2/area
- concavity – the severity of concavities or indentations in a cell (the extend to which the actual boundary of a cell lies on the inside of each chord between non-adjacent boundary points)
- concavity points – the number of concavities, irrespective of their amplitudes
- symmetry – the length difference between lines perpendicular to the major axis to the cell boundary in both directions.

This makes together 11 features that should be taken into account at the recognition process.

C. Statistical features of the histogram

The next set of features has been generated on the basis of the pixel intensity distribution of the image. The histograms of such intensity have been determined for all color components RGB. On the basis of this the following features have been generated for the nucleus and the whole cell:

1. mean value of the histogram
2. variance of the histogram
3. skewness of the histogram
4. kurtosis of the histogram
5. mean value of the gradient matrix
6. variance of the gradient matrix
7. skewness of the gradient matrix
8. kurtosis of the gradient matrix.

These features have been calculated for all colors, independently for the nucleus and for the whole cell. This makes together 48 features. All numerical experiments have been implemented on the platform of Matlab, by using the facility provided by the Image Processing Toolbox [9].

III. ANALYSIS OF FEATURES

Applying methods of feature generation presented above we can generate many features. Some of them are more or less correlated and as such should be avoided. Moreover some features are of very large variance for the same class. This is also not welcome. To distinguish between different classes the centers of such classes should be separated as much as possible. These requirements are the main points in the analysis of the features, on the basis of which we can analyze and compare the separation ability of different sets of features.

The numerical experiments have been performed for 12 classes of blood cells. 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), prolymphocyte (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.

For the purpose of unification all features have been normalized by dividing each column (feature) by the maximum value of it. The next step was to eliminate the features that are correlated with some others or when their variance within the same class exceeded some threshold.

The first step was to analyze the variances of the features for each class. Table 1 presents the results of such analysis, performed for 12 classes of the normalized data and for 20 chosen features. As it is seen the variances of the data forming the classes are changing but remains within the acceptable limits.

The next results are concerned with the distances between the centers of the samples forming the succeeding classes. The larger these distances. the better features and the better chances to get good separation between two different classes. Table 2 presents the results concerning this problem. Taking into account the small variances of the classes (range of 1E-4 up to 1E-2) the distances between centers seem to be very good. However for some features (for example Nr 10) some distances are not high enough. Such feature should be considered for removing from the set.

TABLE 1 THE VARIANCES OF THE BLOOD CELLS FOR 12 CLASSES AND 20 FEATURES

1	2	3	4	5	6	7	8	9	10	11	12
1,37E-2	3,53E-2	6,34E-3	2,16E-2	1,31E-2	5,44E-3	3,98E-2	2,13E-2	9,56E-3	8,98E-3	1,02E-2	8,81E-3
9,98E-3	6,64E-3	1,95E-2	6,25E-3	9,97E-3	5,54E-3	2,15E-2	1,40E-2	5,17E-3	7,86E-3	1,09E-2	4,80E-3
4,42E-3	3,85E-2	3,69E-3	1,99E-2	1,52E-2	3,36E-3	3,79E-2	1,27E-2	6,57E-3	9,14E-3	3,30E-3	7,17E-3
2,86E-3	2,44E-3	1,64E-2	1,45E-3	3,80E-3	1,41E-3	2,75E-2	8,75E-3	1,36E-3	2,88E-3	4,16E-3	1,58E-3
1,41E-2	1,53E-2	1,47E-2	1,45E-2	3,08E-2	1,31E-2	1,87E-2	1,68E-2	1,81E-2	1,00E-2	1,28E-2	5,22E-3
1,24E-2	9,53E-3	1,16E-2	1,19E-2	4,54E-2	1,81E-2	4,87E-3	6,83E-3	2,13E-2	1,70E-2	9,86E-3	5,10E-3
1,44E-2	1,41E-2	2,92E-2	1,44E-2	3,36E-2	1,94E-2	4,38E-2	2,17E-2	3,43E-2	1,68E-2	1,27E-2	1,15E-2
2,13E-2	2,44E-2	4,74E-2	1,99E-2	3,14E-2	2,18E-2	4,08E-2	2,12E-2	3,79E-2	1,93E-2	1,54E-2	1,18E-2
5,08E-3	8,47E-3	5,59E-3	7,36E-3	9,83E-3	1,10E-2	1,21E-2	7,75E-3	8,10E-3	8,16E-3	2,86E-2	1,18E-3
1,08E-2	1,13E-2	6,04E-3	1,27E-2	5,04E-3	1,11E-2	1,73E-2	1,31E-2	9,13E-3	2,74E-2	1,99E-2	7,02E-4
2,25E-2	2,05E-2	2,72E-2	4,47E-2	9,68E-3	1,41E-2	2,76E-2	2,36E-2	1,85E-2	1,56E-2	4,17E-3	2,18E-3
2,13E-2	1,33E-2	2,27E-2	8,58E-3	7,74E-3	6,92E-3	1,51E-2	1,40E-2	1,55E-2	5,51E-3	5,10E-3	1,55E-3
9,93E-3	2,05E-2	1,15E-2	2,50E-2	2,94E-2	2,18E-2	6,49E-3	7,59E-3	2,98E-2	2,12E-2	9,30E-3	5,90E-3
9,54E-3	1,55E-2	7,72E-3	6,62E-3	1,02E-2	8,43E-3	1,35E-2	1,28E-2	1,45E-2	1,23E-2	4,67E-3	3,94E-3
9,75E-3	2,01E-3	1,78E-2	4,22E-3	2,71E-2	2,74E-3	1,10E-2	5,09E-3	1,51E-2	1,84E-3	3,04E-2	6,88E-3
9,15E-3	3,64E-3	2,39E-2	7,42E-3	2,37E-2	5,32E-3	1,24E-2	5,92E-3	1,58E-2	2,50E-3	2,85E-2	8,76E-3
1,05E-3	8,41E-3	1,41E-2	7,31E-3	1,37E-2	2,28E-2	1,68E-2	1,44E-2	1,05E-2	3,59E-2	2,74E-3	6,14E-3
1,95E-3	1,36E-2	2,05E-2	5,35E-3	1,13E-2	1,42E-2	1,79E-2	1,44E-2	5,88E-3	1,72E-2	3,21E-3	1,19E-2
4,23E-3	1,00E-3	2,65E-2	2,09E-3	2,02E-2	2,03E-3	9,52E-3	2,29E-3	1,33E-2	1,63E-3	1,64E-2	9,92E-3
5,54E-3	2,59E-3	5,76E-2	4,30E-3	2,15E-2	7,30E-3	1,55E-2	2,63E-3	1,59E-2	3,10E-3	1,86E-2	1,77E-2

TABLE 2 THE DISTANCES BETWEEN CENTERS OF SUCCEEDING CLASSES

	1	2	3	4	5	6	7	8	9	10	11	12
1	0,000	1,113	1,010	0,813	1,789	0,687	1,154	0,728	0,870	0,766	0,768	1,300
2	1,113	0,000	1,803	0,427	2,420	1,166	1,025	0,831	1,594	0,978	1,415	2,010
3	1,010	1,803	0,000	1,479	1,384	1,062	1,602	1,444	0,873	1,205	1,101	0,525
4	0,813	0,427	1,479	0,000	2,095	0,781	1,020	0,736	1,234	0,601	1,067	1,672
5	1,789	2,420	1,384	2,095	0,000	1,420	2,461	2,307	1,113	1,605	1,276	1,464
6	0,687	1,166	1,062	0,781	1,420	0,000	1,440	1,135	0,525	0,299	0,489	1,267
7	1,154	1,025	1,602	1,020	2,461	1,440	0,000	0,529	1,756	1,320	1,566	1,814
8	0,728	0,831	1,444	0,736	2,307	1,135	0,529	0,000	1,464	1,043	1,267	1,694
9	0,870	1,594	0,873	1,234	1,113	0,525	1,756	1,464	0,000	0,755	0,552	1,120
10	0,766	0,978	1,205	0,601	1,605	0,299	1,320	1,043	0,755	0,000	0,629	1,411
11	0,768	1,415	1,101	1,067	1,276	0,489	1,566	1,267	0,552	0,629	0,000	1,369
12	1,300	2,010	0,525	1,672	1,464	1,267	1,814	1,694	1,120	1,411	1,369	0,000

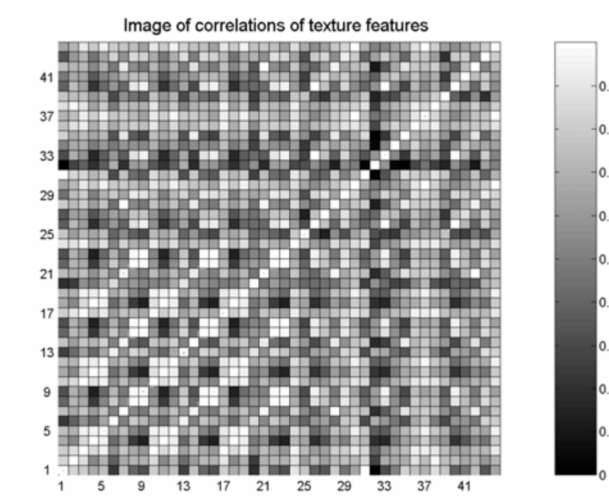

Fig. 1. Image of correlations of texture features

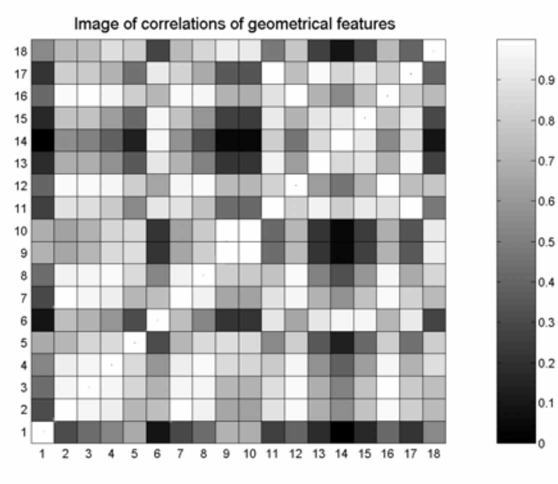

Fig. 2. Image of correlations of geometrical features

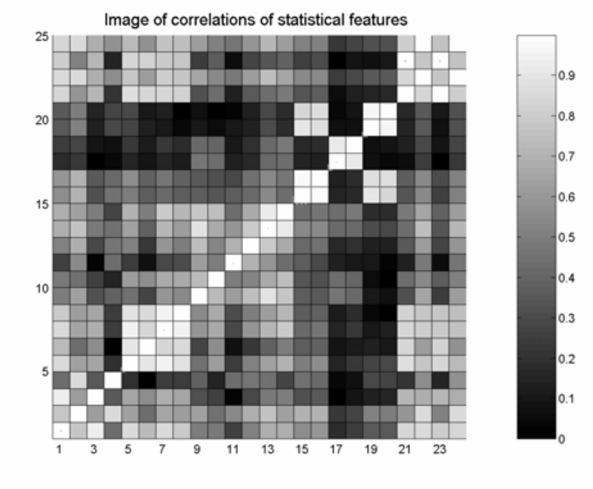

Fig. 3. Image of correlations of statistical features

The aim of our next experiments is to find the correlation among the features. The features strongly correlated with each other should be eliminated from the set. Because of very large size of the whole feature set we have divided them into 3 groups: the texture, geometrical and statistical set of features. Fig. 1 (texture), 2 (geometrical features) and 3 (statistical features) depict the graphical results of experiments. The highest correlations are observed among some geometrical features. The highest differences among the values can be observed in the distribution of the texture features.

The other problem in recognition of the blood cells on the basis of the image is its background. This is important problem, since it is almost impossible to provide exactly the same conditions of acquisition of the cell images in the hospital laboratories. Therefore it is quite important to find the way to avoid the influence of the background on the results. We have introduced so called "scaling" of the image that is subtraction of the bias corresponding to the measured background intensity.

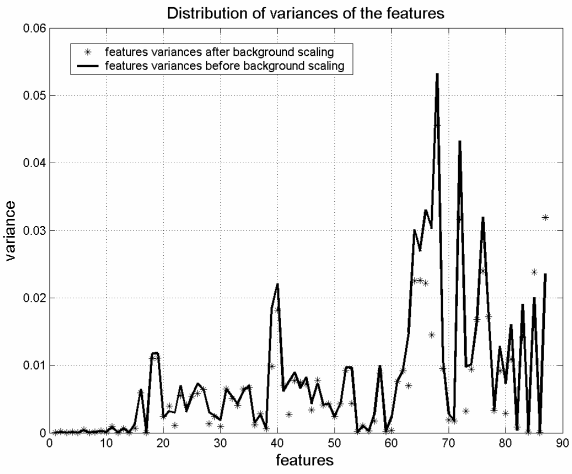

Fig. 4. Features variances modification using background algorithm for polychromatic erythroblast

Fig. 4 presents the distribution of the variances of the features corresponding to the original and to the transformed data. The features have been arranged according to their predicted importance in the recognition process. The best features are at the beginning of the diagram and these of least importance are at the end. As it is seen the scaling generally improves the distribution of the features (smaller variances of individual features).

IV. CONCLUSIONS

The paper has presented the analysis of the features of the images of the blood cells in the myelogenous leukemia. We have taken into account the textural, geometrical and statistical features. Our investigations have been aimed on the comparison of different preprocessing techniques leading to different variances of the data within the same class, the differences of the means of individual classes and correlation among the features. Such analysis enables to compare the features and arrange them into more or less influential order. This is very important step in the generation of the most optimal features, for future recognition of the blood cells.

REFERENCES

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

[2] W. Wolberg, W. N. Street, O. L. Mangasarian, Machine learning to diagnose breast cancer from image-processed nuclear features of fine needle aspirates, 1994, Internal report of University of Wisconsin

[3] K. Lewandowski, A. Hellmann, Atlas hematologiczny, Medyczne Wydawnictwo Multimedialne, Gdańsk, 2001

[4] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, C. Sultan, Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group, Br J Haematol., 1976, vol. 33, pp. 451-458

[5] O. Lezoray, H. Cardot, Cooperation of color pixel classification schemes and color watershed: a study for microscopic images, IEEE Trans. Image Processing, 2002, vol. 11, pp. 783-789

[6] H. Hengen, S. Spoor, M. Pandit, Analysis of blood and bone marrow smears using digital image processing, SPIE Medical Imaging, San Diego, 2002

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

[8] T. Reed, J. Buf, A review of recent texture segmentation and feature extraction techniques, CVGIP: Image Understanding, 1993, vol. 57, pp. 359-372

[9] Matlab user manual, Natick, 1999

[10] T. Markiewicz, S. Osowski, L. Moszczyński, R. Sałat Myelogenous Leukemia Cell Image Preprocessing for Feature Generation, V CPEE, Jazleevets, 2003, pp. 70-73

[11] S.Osowski, T. Markiewicz, B. Marianska, L. Moszczyński, Feature generation for the cell image recognition of myelogenous leukemia, IEEE Int. Conf. EUSIPCO, Vienna, 2004

ACKNOWLEDGMENT

The work has been supported by Polish Ministry of Science (KBN) as the research project.