NUCLEI IMAGES ANALYSIS
Technology, Diagnostic Features and Experimental Study

I. Gurevich, D. Murashov
Dorodnicyn Computing Centre of the Russian Academy of Sciences, 40, Vavilov str., Moscow, GSP-1, 119991, Russia

O. Salvetti
Institute of Information Science and Technologies, CNR, 1, Via G. Moruzzi, Pisa, 56124, Italy

H. Niemann
University of Erlangen-Nuremberg, Informatik 5, Martensstraße 3,91058 Erlangen, Germany

Keywords: Medical image mining, image analysis, information technology for automated diagnostics, morphologic analysis of the cytological slides.

Abstract: The information technology for automated morphologic analysis of the cytological slides, taken from patients with the lymphatic system tumours, was developed. The main contributions of the paper are the technology, the set of features for representation of nuclei images in pattern recognition problems (automated diagnostics), and experimental study of the technology and the features informativeness. The main components of the technology are: acquisition of cytological slides, method for segmentation of nuclei in the cytological slides, synthesis of the feature based nuclei description for subsequent classification, nuclei image analysis based on pattern recognition and scale-space techniques. The experiments confirmed efficiency of the developed technology. The discussion of the obtained results is given. The developed technology is implemented in the software system.

1 INTRODUCTION

The technology for morphologic analysis of cytological slides of patient with tumors of the lymphatic system was developed. The technology is intended for computer-aided diagnostic which is based on the properties of tumor cells.

The main contributions of the paper are the technology, the set of features for representation of nuclei images in pattern recognition problems (automated diagnostics), and experimental study of the technology and the features informativeness.

The main components of the technology are described and results of experimental testing are discussed. The task of describing formally the lymphoid tumour cells is challenging. Even high-skilled experts could provide different conclusions on the same specimen. Morphometric approach allows one to enhance precision of the diagnosis by utilizing quantitative characteristics of a specimen. Further advancement could be achieved using technologies based on the information extracted by pattern recognition and other modern mathematical techniques for image analysis.

In this paper the information technology for automated morphologic analysis of the cytological slides taken from patients with the lymphatic system tumors, is presented. The technology includes the set of methods and techniques developed by the authors recently (author's paper). All of the elements of technology are naturally defined by the technology of specimen slides acquiring and specimen image analysis processes.

The main phases of the developed technology are as follows: 1) design of the archive of tumour cell images of patients with high- and low-grade lymphoid tumours, 2) image preprocessing for removing differences in illumination and colour; 3) selection and calculation of features reflecting cell nuclei peculiarities, 4) qualitative and statistical

Gurevich I., Murashov D., Salvetti O. and Niemann H. (2007).
NUCLEI IMAGES ANALYSIS - Technology, Diagnostic Features and Experimental Study.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications*, pages 204-210
DOI: 10.5220/0002071002040210

analysis of features and evaluation of its information weights, 5) composing feature description of a patient based on the results of nuclei cluster analysis; 6) classification (diagnostics) of slides.

For the design of the technology the following problems are solved: a) selection of main features of the nuclei in cytological slides; b) development of techniques for nuclei segmentation in slides; c) statistical analysis of feature description (descriptive statistics, correlation, factor and cluster analysis); d) selecting of diagnostic "mathematical" features; e) experimental testing of efficiency of different classifiers.

2 MORPHOLOGY OF THE SPECIMENS

Primary diagnostics of lymphoid tumors is based on morphological descriptions of cells and relations between different cell populations. The presence of large lymphoid cells points to diffuse large cell lymphoma or high-grade transformation of indolent lymphoma, while predominance of small lymphocyte like cells indicates indolent lymphoma.

Quantitative analysis of cells was based on calculating intensity, textural and structural features of slides that characterized the size and the shape of objects. A very important factor for diagnostic is the visual structure of chromatin in cell nuclei.

3 DIGITAL IMAGING

Footprints of lymphoid tissues were Romanovski-Giemsa stained and photographed with digital camera mounted on Leica DMRB microscope using PlanApo 100/1.3 objective. The equivalent size of a pixel was 0,0036 μ2; 24-bit color images were stored in TIFF format, and analyzed as described below.

4 SEGMENTATION SPECIMEN IMAGES

Automated hematological diagnostics using specimen images deals with evaluation of values of a set of diagnostically important features which are used for nuclei classification. The segmentation of nuclei images is necessary for further nuclei image analysis, in particular, for calculation of feature values.

The two-level method for automated segmentation of nuclei images based on the active contour model was developed (Murashov, 2005). The main features of the method are: a) the rough segmentation is achieved using the blue component of the image taking into account the properties of the stain; the precise correction is made using the green component; b) the modified Gaussian filter based on the heat equation with heat source or sink is implemented. The technique is successfully implemented for segmenting cytological specimen images.

5 NUCLEI IMAGES ANALYSIS

The problem of diagnostic-oriented feature description of specimens is very actual for many years (see, for example (Rodenacker, 1995, Rodenacker and Bengtsson, 2003), but it didn't get final solution till now. In order to select and describe diagnostically valuable attributes we have created an archive of pictures of footprints of biopsies from patients with 4 diagnoses: non-tumor (reactive) enlargement of lymph nodes (RL), B-cell chronic lymphatic leukemia (CLL), large cell transformation of B-cell chronic lymphatic leukemia (TRCLL), de novo large-cell lymphoma (LS) (Jaffe, 2001).

Two methods are used for composing nuclei feature description and obtaining diagnostic criteria: 1) selecting a large set of features using image analysis techniques and extracting minimal feature set for classification of nuclei with suitable accuracy (Churakova, Gurevich, et.al., 2003); 2) selecting a set of features used by experts in visual analysis (the features are invariant in some degree to slides quality and parameters of image acquisition process) (Gurevich, Murashov, 2004).

5.1 The Nuclei Descriptions in the Form of a Vector of Numerical Features

The diagnostic features of tumor cells are as follows: a) nuclei size optical density; b) the shape of nuclei (round, elliptic, folded), c) presence of invaginations, d) textural characteristics of the chromatin (condensed/dispersed, the chromatin fibril's diameter, presence of granules of condensed chromatin and size of these granules), e) presence of nucleoli. We provided formal equivalents for some of the above features. Thus, the following 47 features were chosen to describe nuclei morphology (Churakova, Gurevich, et.al., 2003): an area of nucleus in pixels, 4 statistical features calculated on nucleus brightness histogram (average, dispersion, 3rd and 4th central moments); 16 granulometric features of nucleus; 26 features calculated on the Fourier-spectrum of nucleus. Further steps included statistical and factor analysis of features and cluster analysis of nuclei on appropriate sets of features.

Statistical and qualitative analysis included: feature correlations, feature distribution, feature histograms and their moments; feature "robustness" to variations of nuclei geometry. The analysis allowed excluding unstable and high-correlated features. The feature values distribution was estimated by Shapiro-Wilks W-test. It turned out that the distributions for majority of features are not normal once.

After calculation of descriptive nonparametric statistics (medians and quartiles) it is appeared that values of some textural features are grouped within 3 separated areas, i.e. the considered nuclei are divided into 3 types. In case of CLL, cytological slides contain mainly "mature" nuclei. In slides corresponding to TRCLL one can find "mature" nuclei as well as "transformed" nuclei. LS is characterized by a larger percent of transformed" nuclei. Since well-known for experts distribution of "mature" and "transformed" nuclei over diagnosis coincides with its distribution over obtained types, it is possible to conclude that these nuclei types correspond to "mature" and "transformed" nuclei.

Cluster sets were obtained by application of FOREL algorithm (Zagoruiko, 1999) to different sets of features. The sets of clusters were evaluated using different criteria: number and size of clusters in each set (large clusters contained more than 400 nucleus), total percent of all nuclei belonging to large clusters, the character of nuclei distribution in large clusters. In the considered problem a taxonomy with a few large clusters accumulating the main part of nuclei is more preferable than a taxonomy with a lot of small clusters where nuclei distribution over clusters is uniform.

A new method to create feature description of a patient was suggested. On the basis of cluster analysis results a patient is described by a new type of features - percentage of a patient's nuclei belonging to large clusters of the taxonomy. Experiments showed that good classification results can be obtained in such feature space.

Factor analysis was conducted for reducing of the feature space. It was applied to several data sets: a) all 8702 available nuclei; b) the sample corresponding to 4 different diagnosis; c) samples for each patient. Three techniques were used: principal factor, centroid and maximum-likelihood. The combination of the Kaizer criterion and the scree-test was used to determine a number of factors, while varimax-rotate strategy was used to calculate factor loadings. Each method yielded the same number of factors for each data set. The mean factor loadings were calculated for three data sets. As a result, the same factors were discovered. Similarity of factors was confirmed by presence of high factor loadings on the same features and, accordingly, by presence of low factor loadings on the rest features. Then, features with high factor loadings (its absolute values exceed some threshold) were determined for each obtained factor.

It is important that selection of the same factors for data sets corresponding to different diagnoses gives the corresponding sets of features with high factor loadings differing substantially. It means that for different diseases considered in our study there are different significant features. It also appeared that there are unique significant features for some diseases that are not significant for the other diseases.

We developed a new diagnostic procedure based on factor patterns designed for considered deceases. The pattern represents high-loadings-feature distribution over factors extracted from the sample corresponding to certain decease. Such distributions are different for different deceases. If the factor pattern of a new patient coincides with the pattern of a particular decease, we could consider that this patient has such decease.

The cluster analysis of nuclei was done on the base of the results of factor analysis. For initial minimization of the feature space Spearman R-statistic was used. Further the features with high loadings were selected from obtained set. It appeared that the taxonomy structure (amount of large clusters and their nuclei proportions) is determined by only 3 features. The values of these features are concentrated into separated areas. The rest of features influences only on total amount and size of clusters, however the nuclei proportions in large clusters are the same (see Table 1, "+" means the presence of explicit partition to malignant and non-malignant groups).

Table 1: Cluster analysis of different sets of features.

Set No	Features	Total amount of clusters	Amount of large clusters (>400 nuclei)	Percentage of all nuclei	Partition to malign. and non-malign.	% of malign. nuclei in non-malig. clusters	% of non-malig. nuclei in malign. clusters
3	19,22,27	23	5	83,5	+	8,5	33,6
4	1,19,22,27	27	5	84,1	+	8,5	33,2
5	1,2,19,22,27	49	5	77,9	+	3,6	35,8
6	1,2,19,22,27,3,6,10	91	5	79,5	+	3,8	34,1
8	18 features	136	5	82	+	5,7	32,7

5.2 Scale-Space Features

Two basic approaches to analysis of a chromatin constitution are known (Rodenacker and Bengtsson, 2003): a) structural - the chromatin distribution is considered as a local arrangement of small objects of varying intensity and the intensity features of dark and bright particles are evaluated; b) textural - based on statistical characteristics of chromatin arrangement and on analysis of regularities of chromatin structure. The second method uses gray level dependency matrices (Weyn, et.al., 1999), co-occurrence, run-length features, rice-field operators, watersheds (topological methods) (Rodenacker, 1993, Rodenacker and Bengtsson, 2003, Weyn, et.al., 1999, Young, et.al., 1986), heterogeneity, clumpiness, margination, radius of particles (Young, et.al., 1986) (the Mayall/Young features), and invariant features (polynomial invariants).

Two methods for nuclei image analysis are developed on the basis of the scale-space approach (Lindeberg, 1994). The approach of Gaussian scale-space is the most relevant for cell nuclei image analysis because it provides the invariance with respect to shift, rotation, scaling, and linear transformations of intensity. It decreases the sensitivity of the analysis to microscope focusing.

The first method deals with the analysis of the nuclei image intensity features (Gurevich and Murashov, 2004). A diagnostically important criterion is obtained - a total amount of spatial extrema in scale-space generated by an image of a cell nucleus. The main idea of the analysis is based on the assumption that significant structures in scale-space are likely corresponding to significant structures in the image. The chromatin structure in the nucleus image is represented by a grey-level relief containing connected bright and dark regions (peaks and valleys) corresponding to chromatin filaments, furrows and centers. Each of these regions contains at least one intensity extremum. The developed method is based on the topological properties of iso-intensity manifolds in the neighborhoods of spatial extrema and allows to evaluate approximate amount of chromatin particles represented in the image as grey-level blobs.

The second method is a textural one – an integral feature characterizing the chromatin pattern is involved. The quantitative feature is defined using the second moment matrix which is defined as a covariance matrix of the stochastic value characterizing the nuclei image. The steps of the method are: a) calculation of the module gradient image for emphasizing the chromatin pattern; b) selection of the scale parameter σ corresponding to the most informative image in the scale-space (the most informative image corresponds to the maximal feature value, see (Lindeberg, 1994)); c) calculation features values via 2d moment matrix μ_L (trace $tr\,\mu_L$, determinant $\det\mu_L$, quotient $q_\mu = \det\mu_L / tr\,\mu_L$). The nuclei images of different patient groups prepared for textural analysis are shown in Figure 1.

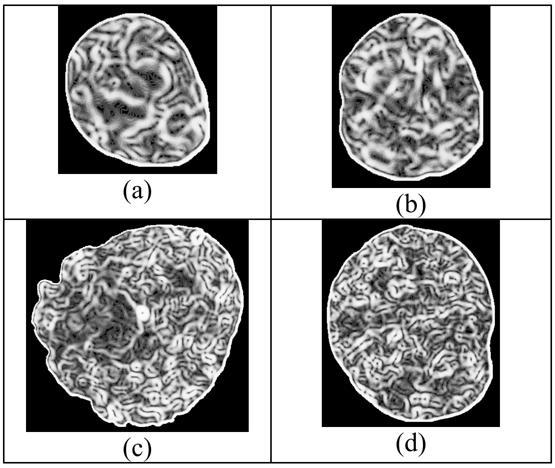

Figure 1: Nuclei images prepared for textural analysis: a) CLL; b) RL, c) TRCLL, and d) LS.

The value of scale parameter σ is obtained experimentally and corresponds to q_μ maximum. For the considered class of images σ is equal to 3,25.

The experimental chart displaying the distribution of cell nuclei is presented in Figure 2, where n is the total amount of spatial extrema, s is an area of a nucleus.

The diagnostic features were tested for: 1) selection of a textural feature providing the best classification result; 2) testing efficiency of the features for diagnosing; 3) testing the diagnostic importance of the nuclei marked by experts and used for training classifiers. In the experiments we used slides from patients with diagnoses CLL, TrCLL, LS, and RL.

The computing experiment includes two steps. At the first step classifiers were trained, and the three tasks pointed out above were carried out. Training control sample included 829 nuclei taken from 25 patients, control sample included 832 nuclei. At this step the diagnostically important nuclei marked by experts in the specimen images were used for training. The following classifiers were applied: k-nearest neighbors, support vector machine (SVM), Fisher liner discriminant, multilayer perceptron, and committee algorithm (Zhuravlev, et. al., 2003). For each nuclei image the features *s*, *n*, $tr\,\mu_L$, $\det\mu_L$ and q_μ were calculated.

The suitable accuracy was obtained by *k*-nearest neighbors and SVM classifiers. Committee classifier increased the reliability of classification. The classification accuracy in case of 2 nuclei classes (non-malignant and malignant) exceeds 90%. In case of 4 different diagnoses the accuracy is comparable with the accuracy of expert estimates. The results of the first step show that the selected feature set can be used for description of nuclei chromatin structure corresponding to the considered diseases.

At the second step the percentage of the nuclei of different chromatin structure in the specimen images was evaluated. At this step the slides taken from new patients (different from those at step 1) were used. Unlike the previous step all of the nuclei appeared in the slides were analyzed. The feature values (***s***, q_μ, ***n***) were calculated and the nuclei were classified using the algorithms trained at step 1.

The results of the classification presented in Tables 2 and 3 satisfy the existing morphologic criteria mentioned in section 2. The disease is considered as malignant by the experts if the amount of malignant cells exceeds 20% of the total amount of cells in specimens taken from patient. The obtained results (see Tables 2, 3) allow one to make the following conclusions.

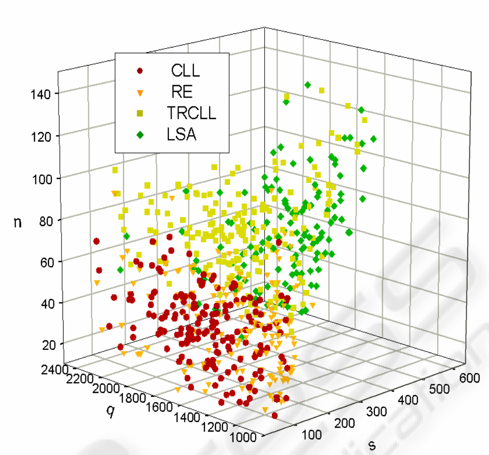

Figure 2: Nuclei distribution in axes s, qμ, and n for 4 groups of diseases.

Table 2: The percentage of nuclei recognized as "malignant" and "non-malignant" in the specimen images corresponding to different diseases.

Diagnosis	Amount of nuclei (%)	
	Non-malignant	Malignant
CLL	90	10
RL	76,5	23,5
TRCLL	62,3	37,7
LS	54,3	45,7

First, a few types of nuclei characterizing the diseases under consideration can be outlined (we deal with "CLL", "RL", "TRCLL", and "LS" types; "CLL" and "RL" form the "non-malignant" group, "TRCLL" and "LS" form "malignant" group). Second, in the slides corresponding to CLL non-malignant nuclei are prevailing. Third, in the slides corresponding to TRCLL the majority of malignant nuclei are classified as TRCLL type. Also, in the slides corresponding to LS the majority of malignant nuclei are classified as LS type (see Table 3). Fourth, in the RL slides nuclei partly (about 20%) are classified as malignant due to enlarged cells, but none of the types TRCLL or LS is dominating (see Table 3). Fifth, the difference in feature values reflecting the peculiarities of different types of chromatin structure corresponding to different diseases will provide the criteria for differential diagnostics.

Table 3: The percentage of different types of nuclei in the specimen images corresponding to different diseases.

Diagnose	Classifier	Amount of nuclei (%)		
		Non-malign.*	TRCLL	LS
CLL	k-nearest neighbors	82	10	4
	SVM	92	4	4
	Committee	92	4	4
RL	k-nearest neighbors	74,5	11,8	13,7
	SVM	90,2	2,0	7,8
	Committee	90,2	2,0	7,8
TRCLL	k-nearest neighbors	47,2	39,6	13,2
	SVM	62,3	35,8	1,9
	Committee	62,3	35,8	1,9
LS	k-nearest neighbors	50,8	4,3	45,7
	SVM	56,5	6,5	37,0
	Committee	56,5	6,5	37,0

* - combined CLL and RL group.

6 CONCLUSIONS

The results of the computational experiments and estimations of hematologists approve the efficiency of the developed technology for automated morphologic analysis of the lymphoid cells. It is confirmed that the objective differentiation of nuclei may be used for patients classifying according to the main classification of lymphatic tumors. The important problems were solved: a) the automated technique for segmenting the lymphoid cell nuclei is developed and implemented; b) a feature set for nuclei description is formed; c) the procedures for nuclei classification is developed; d) the method for estimating the efficiency of the developed procedures is proposed. The technology is implemented in the specialized software system.

The next step of the research will be devoted to adaptation of the developed technology for analysis of histological slides: 1) the general analysis of tissue structure (quasi-textural analysis of the slides obtained under the low magnification of the microscope); 2) the analysis of single cell images obtained at maximal resolution.

ACKNOWLEDGEMENTS

This work was partially supported by the Russian Foundation for Basic Research Grants 05-07-08000, 06-01-81009, 06-07-89203, by the project of the Program of the Presidium of the Russian Academy of Sciences "Fundamental Problems of Computer Science and Information Technologies", by INTAS Grant № 04-77-7067, and by Cooperative grant “Image Analysis and Synthesis: Theoretical Foundations and Prototypical Applications in Medical Imaging” within agreement between Italian National Research Council and Russian Academy of Sciences.

REFERENCES

Churakova, Zh.V., Gurevich, I.B., Jernova, I.A., et al., 2003. Selection of Diagnostically Valuable Features for Morphological Analysis of Blood Cells. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, Vol. 13, N2, 381-383.*

Gurevich, I., Murashov, D., 2004. Scale-Space Diagnostic Criterion for Microscopic Image Analysis. In *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis. ECCV 2004 Workshops CVAMIA and MMBIA, Prague, Czech Republic, May 15, 2004, Revised Selected Papers / Milan Sonka, Ioannis A. Kakadiaris, Jan Kybic (Eds.): LNCS Vol. 3117.* Springer-Verlag Berlin Heidelberg, pp. 408-416.

Jaffe E. S , Harris N. L., Stein H., et al., 2001. *Pathology and Genetics of Tumors of Haematopoietic and Lymphoid Tissues*. Lyon: IARC Press.

Lindeberg, T., 1994. *Scale-space Theory in Computer Vision*. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers.

Murashov, D., 2005. Method for Segmentation of Low Contrast Cytological Images Based on the Active Contour Model, In: *Yu.I. Shokin, O.I. Potaturkin Eds. Automation, Control, and Information Technology (Proc. of ACIT-SIP 2005, June 20-24, 2005, Novosibirsk, Russia)*, ACTA Preess, pp. 44-49

Rodenacker, K., 1993. Applications of topology for evaluating pictorial structures. In *Reinhard Klette and Walter G. Kropatsch, (eds), Theoretical Foundations of Computer Vision*, Akademie-Verlag, Berlin, pp. 35–46

Rodenacker, K., 1995. Quantitative microscope image analysis for improved diagnosis and prognosis of tumours in pathology. In *Creaso Info Medical Imaging*, *Vol. 22*. Creaso GmbH, Gilching.

Rodenacker, K., and Bengtsson, E. 2003. A feature set for cytometry on digitized microscopic images. *Anal Cell Pathol*, *Vol. 25. N 1, pp. 1–36.*

Weyn B., Van de Wouwer G., Koprowski M., et.al., 1999. Value of morphometry, texture analysis, densitometry and histometry in the differential diagnosis and prognosis of malignant mesothelioma, *Journal of Pathology*, *Vol. 4., N 189, pp. 581-589.*

Young, I.T., Verbeek, P., and Mayall, B.H., 1986. Characterization of chromatin distribution in cell nuclei, *Cytometry. Vol. 7. N 5, pp. 467–474.*

Zagoruiko, N.G., 1999. Applied Analysis of Data and Knowledge, Institute of Mathematics, Siberian Division, Russian Academy of Sciences, Novosibirsk.

Zhuravlev Yu.I., Ryazanov V.V., Senko O.V., et. al., 2003. The Program System for Data Analysis Recognition (Loreg). *In: Proc. of the 6th German-Russian Workshop "Pattern Recognition and Image Understanding" (OGRW-6-2003, Aug, 25-30, 2003)*, Novosibirsk, pp. 255-258.