

Histological Image Mining for Exploring Textural Differences in Cancerous Tissue

Vassili Kovalev, Ihar Safonau and Aliaksandr Prus

United Institute of Informatics Problems, Belarus National Academy of Sciences, Gomel, Belarus

Telephone(Fax): +375-232-70-09-39

Email: vassili.kovalev@gmail.com, ihar.safonau@gmail.com, alex_prus@tut.by

Abstract—This paper addresses the problem of histological image mining. The ultimate goal was to explore textural differences between the cancerous tissue images acquired from ovarian and thyroid gland probes. The study group consisted of 46 subjects including 23 patients suffering from ovarian cancer and 23 patients with thyroid cancer. Tissue probes were immunohistochemically processed using D2-40 mesothelial marker for highlighting tumor angiogenesis. A test database containing 4000 color images (35 Gigabytes) of 2048×1536 pixels was acquired with the help of Leica DMD108 microscope, 1000 images for each of four tissue classes: ovary non-tumor, ovary tumor, thyroid non-tumor, thyroid tumor. Both color and grayscale version of extended multidimensional co-occurrence matrices defined on pixel triplets were used as descriptors of spatial image structure. A method for detecting and visualizing non-trivial differences between image classes was suggested and its abilities were demonstrated on pathological and non-pathological regions of tissue probes. In addition, the inter-class differences were assessed using recent statistical and classification procedures (SVM, Random Forests).

I. INTRODUCTION

Motivation of this work stems from two main points: the importance of research on tissue angiogenesis and substantial social consequences associated with the cancer types considered with this study.

Over the last several years numerous factors involved in the development and progression of solid cancers have been identified. Among these, tumor angiogenesis is certainly one of the most important owing to the fact that, in order to grow and to develop, a tumor needs to be supplied with a vascular system following its growth. Without formation of new blood vessels, tumors cannot exceed a size of about 1 mm^3 [1]. It is anticipated that angiogenesis research will probably change the face of medicine in the next decades [2], with more than 500 million people worldwide predicted to benefit from pro- or anti-angiogenesis treatments.

This work is a part of two large angiogenesis research projects. The first one is dedicated to the ovarian cancer. According to the 2008 American Cancer Society statistics, ovarian cancer is and will be distinguished as the most lethal gynecologic cancer. Despite an aggressive treatment, the overall 5-year survival rate is only 30 – 35%. Another project is concentrated on the angiogenesis research of thyroid gland cancer which is proven to be directly related to nuclear accidents and potential problems somehow associated in public opinion with the nuclear power engineering. Thus, exploring the

angiogenesis-related processes based on histopathological images of ovarian and thyroid cancer for discovering any non-trivial, unobvious relationships is an important problem of biomedical image analysis which can be referred to the image mining.

Generally, the image mining can be defined as the process of extracting hidden patterns from images. More specifically, image mining deals with extraction of implicit knowledge, image data relationship or other patterns not explicitly stored in the image database [3], [4]. As more and more image collections are gathered, the image mining becomes an increasingly important tool for transforming these visual data into knowledge. Image mining research has borrowed some ideas from such well developed areas as computer vision, image analysis, image retrieval, machine learning, and artificial intelligence. Typical image mining procedure involves preprocessing, segmentation, feature extraction and discovering significant patterns out of the extracted features accomplished by a semantic interpretation and obtaining the final knowledge.

Similar to the conventional image analysis algorithms, the image mining can be performed at different levels of image representation. Although the segment-based image mining procedures are also under development in framework of the project, this work presents the first results on exploring textural differences of cancer tissue images considered on the pixel-level. To this end, both color and grayscale version of extended multidimensional co-occurrence matrices defined on pixel triplets are used to describe spatial image structure. Once the reliable inter-class differences expressed in terms of certain regularities on matrix element values are found, they mapped back to the original images, corresponding image regions/structures are visualized and their semantic usefulness is evaluated.

II. MATERIALS AND METHODS

A. Patients, probes and histological images

Cases were retrieved from the data of the department of morbid anatomy at the Minsk City Hospital for Oncology. The study group consisted of 46 subjects including 23 patients suffering from ovarian cancer and 23 patients with thyroid cancer. Tissue probes (one slide per patient containing tumor and surrounding tissue samples) were immunohistochemically processed using D2-40 mesothelial marker for highlighting tumor angiogenesis. A test

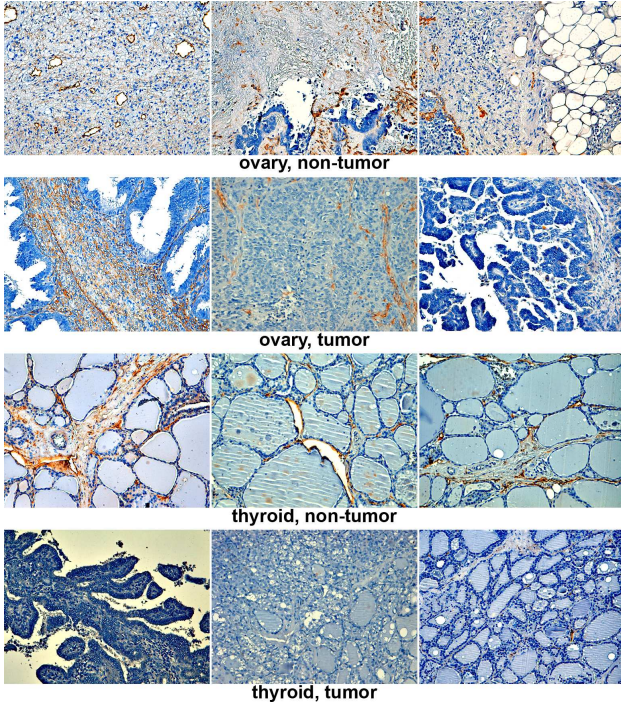


Fig. 1. Examples of histological images of ovarian and thyroid gland tissues (by rows)

database consisting of 4000 color images of 2048×1536 pixels was acquired from the slides with the help of Leica DMD108 optical microscope, 1000 images for each of four tissue classes. The classes (Fig. 1) and their respective conditional names used in the rest part of the paper were as follows: ovary non-tumor (OvaryNorm), ovary tumor (OvaryTum), thyroid non-tumor (ThyrNorm), thyroid tumor (ThyrTum). The total size of image data used in the experiments was amounted to 35 Gigabytes.

B. Methods

Preliminary assessment. At the preliminary stage, the inter-class differences were evaluated using statistical (t -test, PCA, binary logistic regression) and classification (SVM, Random Forests) methods.

Image mining procedure. The essence of the pixel-level image mining procedure suggested with this study can be represented by the following four steps:

- calculating multidimensional co-occurrence matrices defined on pixel triplets and describing spatial image structure,
- merging matrices into a cases-variables (images-matrix elements) data table for given image classes and extracting matrix elements as features distinguishing the classes,
- mapping selected matrix elements back to original images and isolating characteristic image regions of potential interest,
- visual assessment of detected regions together with found quantitative regularities, their semantic interpretation and novelty examination.

Co-occurrence matrices. In this work we promoting a bit further our basic concept of generalized multi-

sort co-occurrence matrices originally suggested in [5] and proven to be fruitful for 2D and 3D shape and texture analysis later on (eg., [6], [7], [8]). This time the extension of matrices is rather straightforward: we still using the single-sort elementary image features such as pixel intensities and pixel colors but we considering pixel triplets as elementary geometrical configurations instead of traditional pixel pairs. Our preliminary experiments performed on different image types suggest that such an extension allows to better describe the spatial image structure. A triangle-based co-occurrence matrix W can be described slightly more formally as follows:

$$W = \|w(V(i), V(j), V(k), d(i, j), d(j, k), d(i, k))\|,$$

where

$i = (x_i, y_i), j = (x_j, y_j), k = (x_k, y_k)$ is a pixel triplet, $d(i, j), d(j, k), d(i, k)$ are the distances between them, $V(i), V(j), V(k)$ are the binned pixel values.

The above matrix is represented by a six-dimensional array with zeros above the leading diagonal in each square-shaped section $V - V$. Also, for reducing dimensionality it is often make sense to consider equilateral triangles only. This case there will be only one dimension corresponding to the triangle side size and the matrix becoming four-dimensional. Some practical advices on intensity binning and computational procedure can be found in [7].

Finally, it should be noted that the use of pixel triplets for calculating co-occurrence matrices automatically means that the third-order spatial statistics are coming into the way. This fact leads to some interesting consequences which are partly detailed in work [9] and further references.

Grayscale vs. color. In the above matrix definition, V stands for either intensity value or color value (color code of suitably paletted-binned input RGB image). Despite the certain advantages associated with a better sensitivity and generally better histological image separation using colorimetric features, it should be pointed out that often colors are quite unstable. More generally, we may say that colors tend to carry certain “fingerprints” imposed by a number of uncontrolled factors such as variability in tissue probe thickness, inhomogeneity of penetration of colorizing agents, specific color distortions associated with given microscope and its imaging sensor and even by specific properties of tissue characteristic of each patient. Thus, considering the potential color instability, it is still worth to consider grayscale version of image descriptors.

Bining pixel values. Our experience suggests that quantification of grayscale values down to 16 bins is a good compromise between the real intensity resolution and the image descriptor size. Given that the color gamut of histological images is very limited, all the input images were binned (paletted) down to 16 colors using an optimal color space reduction performed by commonly known median-cut method. In all the occasions the co-occurrence matrices were calculated using equilateral triangles with side size of 1, 3, and 5 pixels.

TABLE I
CLASSIFICATION ACCURACY FOR GRAYSCALE DESCRIPTORS, %.

Data	Classes	SVM	Random Forests
Original	OvaryNorm–ThyrNorm	98.40	97.40
	OvaryTum–ThyrTum	93.15	89.90
PCA	OvaryNorm–ThyrNorm	98.35	96.75
	OvaryTum–ThyrTum	89.80	91.55
BLR–sel.	OvaryNorm–ThyrNorm	98.30	97.75
	OvaryTum–ThyrTum	92.55	92.45

TABLE II
EXAMPLES OF CONFUSION MATRICES FOR CLASSIFICATION BY ORIGINAL GRAYSCALE DESCRIPTORS.

Classes “OvaryNorm–ThyrNorm”					
SVM			Random Forests		
	ON	TN		ON	TN
ON	992	8	ON	991	9
TN	24	976	TN	43	957

Classes “OvaryTum–ThyrTum”					
SVM			Random Forests		
	OT	TT		OT	TT
OT	963	37	OT	914	86
TT	100	900	TT	116	884

III. RESULTS

A. Assessing the general inter-class differences

Simple variable-wise statistical differences. Preliminary statistical assessment of inter-class differences was performed by way of applying unpaired Student’s t -test separately to every variable (matrix element). In case of color descriptors, $N = 456$ out of 500 variables were found to be significantly different at $p < 0.05$ when comparing OvaryNorm–ThyrNorm classes and $N = 447$ variables for the pair of tumor classes OvaryTum–ThyrTum. It was revealed that normal tissue of the two organs is more dissimilar (mean $t = 13.4$, $STD = 6.8$) than their respective cancerous regions (mean $t = 9.94$, $STD = 7.3$). The use of grayscale descriptors leads to the essentially same but notably more distinct results: $N = 559$, mean $t = 30.5$, $STD = 19.2$ for the pair of normal tissue classes OvaryNorm–ThyrNorm and again much lower mean $t = 11.3$, $STD = 7.7$ for tumor regions. Some more sophisticated multivariate statistical methods were also tried but they provide results well compared with the above straightforward t -tests.

Feature selection. Recently, there is a large arsenal of feature selection methods available for pattern recognition community. In this particular study we were primarily interesting in procedures that allow identifying some “important” variables (matrix elements) to be directly mapped into original images afterwards. In order to investigate the problem, we have examined and compared different feature selection techniques. In simplest case, the above procedure of variable-wise assessment of statistical significance was employed as a feature selection tools. The only problem remains is to manually exclude variables

TABLE III
CLASSIFICATION ACCURACY FOR COLOR DESCRIPTORS, %.

Data	Classes	SVM	Random Forests
Original	OvaryNorm–ThyrNorm	97.60	96.40
	OvaryTum–ThyrTum	93.85	93.30
PCA	OvaryNorm–ThyrNorm	96.15	94.00
	OvaryTum–ThyrTum	94.20	94.20
BLR–sel.	OvaryNorm–ThyrNorm	98.00	97.25
	OvaryTum–ThyrTum	94.05	94.15

with high t -values which somehow duplicate other ones due to a natural mutual correlations. Another technique we tried was based on the principal component analysis (PCA) method. As a result, 26 components were selected by the criteria of covering 95% of variance from color co-occurrence descriptors and 12 components in case of grayscale descriptors. Results of classification experiments (below) confirm the expected approximate equivalence of the variables-components substitution. Once the principal components are found, the important variables were selected using the known procedure of projecting original variable set onto the components and selecting the most collinear ones. On the top, we also examined popular procedure of binary logistic regression (BLR) with the forward strategy of variable dropping as well as the results of ranking variable importance produced by Random Forests classifier as a byproduct. Finally, we concluded that all the approaches produce similar results and ended up with the binary logistic regression which selects variables in a most automated way.

Classification quality. A number of experiments on pair-wise classification of normal and pathological tissue images were performed using different kinds of input features and different classifiers (SVM with 10-fold cross-validation and Random Forests) in order to get a clear picture of general inter-class differences. Results are summarized in Tables I, II and III. Note that in Table II the class name “OvaryNorm” is shortened down to “ON”, etc.

As it can be easily seen from the Tables, all the classification results are reasonably uniform with non-tumor tissue classes (conditionally called here as “norm”) separated better than pathological ones. This is in agreement with the above statistical tests.

B. Visualizing textural differences

Implementation and testing the image mining procedure described in the above sections suggests that relatively independent variables correspond to certain parts of different morphological image structures (sort of “seeds”) while their correlates refers to the near neighboring image regions. Let us call the seed features together with their correlates by feature families for brevity. Mapping distinct feature families back to original images has demonstrated that about 7–10 types of different key regions holding different characteristics (eg., areas) in different classes can be revealed. They include already known (or easily

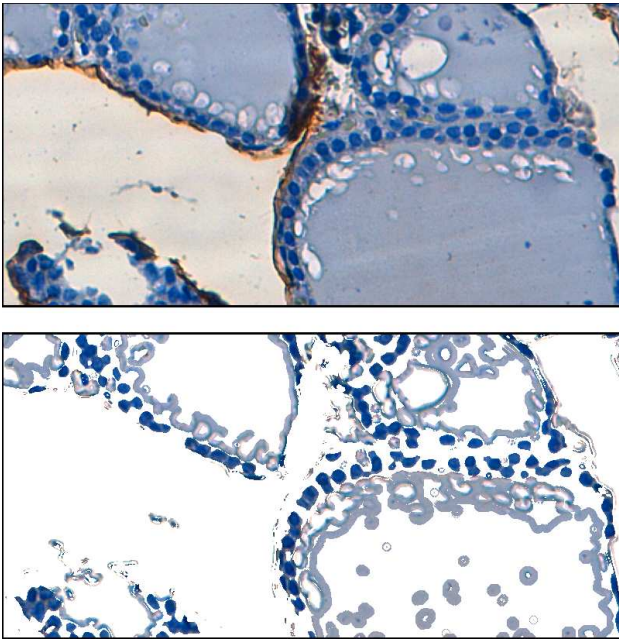


Fig. 2. Example of structures distinguishing normal tissue of thyroid gland and ovary: original image (top) and key morphological structures highlighted (bottom).

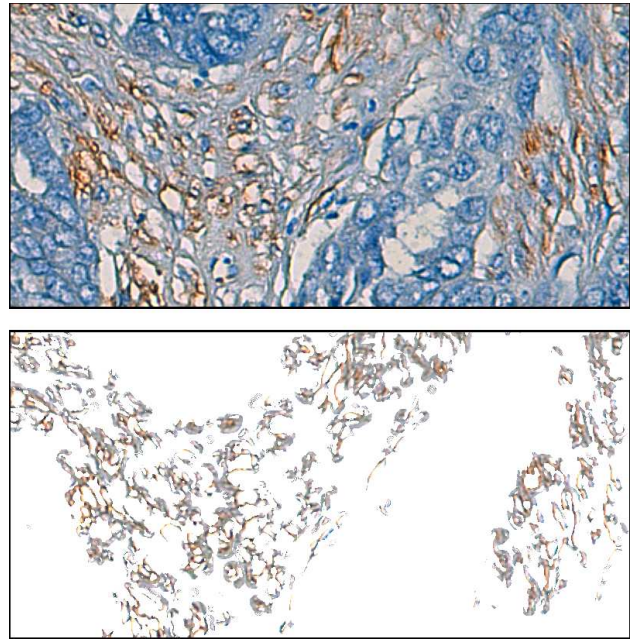


Fig. 3. Example of structures distinguishing tumor tissue of thyroid gland and ovary: original image (top) and key morphological structures highlighted (bottom).

predictable) regularities as well as morphological relationships that potentially new for biomedical experts.

Fig. 2 shows an example of image regions that distinguish non-tumor areas of thyroid gland and ovarian tissue (which are demonstrated on a thyroid image). In particular, it was found here that the total area (ie., the number) of cell nuclei (see dark blue blobs surrounding large circular-shaped follicles) is about 3.38 times greater in thyroid gland compared to ovary (significance is as high as $t = 26.6$). Similar interpretations can be supplied for other morphological structures highlighted in Fig. 2.

Fig. 3 demonstrates another regularity which is related to the tumor tissue. The automatically detected region of given ovarian tumor image clearly related to the lymphatic vessels and the associated numbers say that their area is 1.65 times greater in the ovary compared to the thyroid (just $t = 6.48$, no such vessels in total).

IV. CONCLUSIONS

Results obtained with this study allow to draw the following conclusions:

1. The unsupervised pixel-level image mining procedure suggested in this study allows to detect and to visualize image regions distinguishing large image classes.
2. It was found that normal (more precisely non-tumor) tissue regions of the ovary and thyroid gland are more dissimilar than respective tumor regions. This is in line with more general biomedical regularity suggesting that being affected by a strong pathology, tissues and organs tend to become more similar to each other compared to their natural age- and gender-related variability in healthy individuals.
3. The triplet-based grayscale and color co-occurrence descriptors suggested in this work does not diminish each

other and should be considered as complimentary tools for describing spatial structure of histological images.

The formal textural regularities found with this study needs to be examined by biomedical experts in order to evaluate their semantic novelty and to draw relevant biomedical conclusions.

ACKNOWLEDGMENT

This work was funded by the ISTC project B-1682 and CRDF grant under contract BEM1-4028-MK-08.

REFERENCES

- [1] B. M. Erovic, C. Neuchrist, U. Berger, K. El-Rabadi, and M. Burian, "Quantitation of microvessel density in squamous cell carcinoma of the head and neck by computer-aided image analysis," *Wiener Klinische Wochenschrift*, vol. 117, no. 1, pp. 53-57, 2005.
- [2] P. Carmeliet, "Angiogenesis in life, disease and medicine," *Nature*, vol. 438, pp. 932-936, 2005.
- [3] W. Hsu, M. Lee, and J. Zhang, "Image mining: Trends and developments," *Journal of Intelligent Information Systems*, vol. 19, no. 1, pp. 7-23, 2002.
- [4] H. H. Malik, "Efficient algorithms for clustering and classifying high dimensional text and discretized data using interesting patterns," PhD Thesis, School of Arts and Sciences, Columbia University, 2008, 176 p.
- [5] V. Kovalev and M. Petrou, "Multidimensional co-occurrence matrices for object recognition and matching," *Graphical Models and Image Processing*, vol. 58, no. 3, pp. 187-197, 1996.
- [6] V. A. Kovalev, M. Petrou, and Y. S. Bondar, "Using orientation tokens for object recognition," *Pattern Recognition Letters*, vol. 19, no. 12, pp. 1125-1132, 1998.
- [7] V. A. Kovalev, F. Kruggel, H.-J. Gertz, and D. Y. von Cramon, "Three-dimensional texture analysis of MRI brain datasets," *IEEE TMI*, vol. 20, no. 5, pp. 424-433, 2001.
- [8] V. A. Kovalev, M. Petrou, and J. Suckling, "Detection of structural differences between the brains of schizophrenic patients and controls," *Psych Res*, vol. 124, pp. 177-189, 2003.
- [9] V. Kovalev and M. Petrou, *Handbook of Medical Image Processing and Analysis*, 2nd ed. San Diego, CA, USA: Academic Press, 2009, ch. High order statistics for tissue segmentation, pp. 245-257.