УДК 611.018:004.38

CONTENT-BASED IMAGE RETRIEVAL FROM WHOLE SLIDE HISTOLGY SCANS

V.A. KOVALEV, PhD
Head of the laboratory of Biomedical Images Analysis, United Institute of Informatics Problems (UIIP) of the National Academy of Sciences of Belarus

D. Voynov
Student Department of Biomedical Informatics BSU

R. Roskach
Student Department of Biomedical Informatics BSU

A. Yarashevich
Student Department of Biomedical Informatics BSU

G. Jarin
Student, Department of Biomedical Informatics BSU

Belarus State University, Department of Biomedical Informatics, Republic of Belarus
E-mail: vassili.kovalev@gmail.com

Abstract. This paper presents results that were obtained during solving problem of searching stained regions of tissue. The object of research is whole slide histology scans of tissue sections, that were damaged by cancer. The main idea is transformation of colored image to a gray-scale one, then splitting it to tiles for which different descriptors (intensity histogram, co-occurence matrix) are calculated. The similarity map is obtained by applying comparing of precalculated descriptors of tiles with descriptor of tile that represents damaged region of tissue using metrics (L1, L2, Chi-metric) and visualization using existing color maps. The conclusion was that this methodology can be effectively used to perform query searching fast even on non-specialized computers.

Keywords: content-based image retrieval, histology, whole slide images, image descriptors, similarity maps.

1. Introduction and Problem Statement. According to large work [1], the full digitalization of the microscopic evaluation of stained tissue sections in histopathology has become feasible in recent years because of advances in slide scanning technology and cost reduction in digital storage. Advantages of digital pathology include remote diagnostics, immediate availability of archival cases, and easier consultations with expert pathologists. Along with this, the possibility for computer-aided diagnosis could provide certain advantages being used as one of the ways to provide so-called "second opinion". However, recently it is commonly realized that one of the biggest challenges for pathologists is managing the huge amounts of data which are generated by modern microscope scanners daily [2, 3] As stated by Dr. Wenyi Luo and Prof. Lewis Hassell from Department of Pathology, University of Oklahoma Health Center [2], although glass slides provide highly efficient ways to

convey information needed to make the initial diagnoses, they are often inefficient, expensive, and time-consuming when it comes to physical management, consultation, education, and research. Modern Digital Pathology provides more efficient and cost-effective means of presenting, transmitting, archiving and transporting pathology image data. The whole slide imaging is now the primary means of pathology image capture, store, and evaluation which used for both diagnostic [2-4] and education (e.g [5]) purposes. A number of use cases presented in [2] demonstrate a variety of ways that Digital Pathology can be used to facilitate pathology practice and education. The whole-slide imaging of tissue probes produces very large color histological images the size of which typically ranges between 50000x50000 and 200000x200000 pixels. As a consequence, such large image size induces various problems of different kinds which can be conditionally subdivided into technical and methodological ones.

The purpose of this study is to examine possibility of the use of Content-Based Image Retrieval (CBIR) technology [6] for fast searching of images most similar to certain regions of whole-slide images in framework of query-by-example paradigm.

2. Image descriptors

2.1. Idea of descriptor. First of all the problem of searching means comparison. Digital images itself are very hard to compare: every trivial method is inconsistent. For example pixel-by-pixel comparison obviously incorrect as it is completely unstable to rotations, parallel moves and other natural transformations. So, here comes the idea of image features. Basic considered image features are: shape, color and texture. Now the task is to extract features from image and compare. But both extraction and comparison of features are also complicated tasks: at first there's no strict definition of the word "feature", at second features (in human understanding) also can be difficult to compare. One of the best solutions is constructing so-called "vector of features" (not the features, that are presented above). The key ideas are:

–We calculate different values of image, that describe the basic features and are stable to natural transformations

–These values are easy to compare and comparison gives meaningful results

These values we call descriptors. So we can compare images by comparing calculated descriptors. If these descriptors are chosen right the quality of image retrieval will be high.

2.2. Calculating descriptors of whole-slide images. Also we need to consider another aspect: the object of research is whole-slide histology scans. This means that the content of the whole image is presented when deeply zoomed:

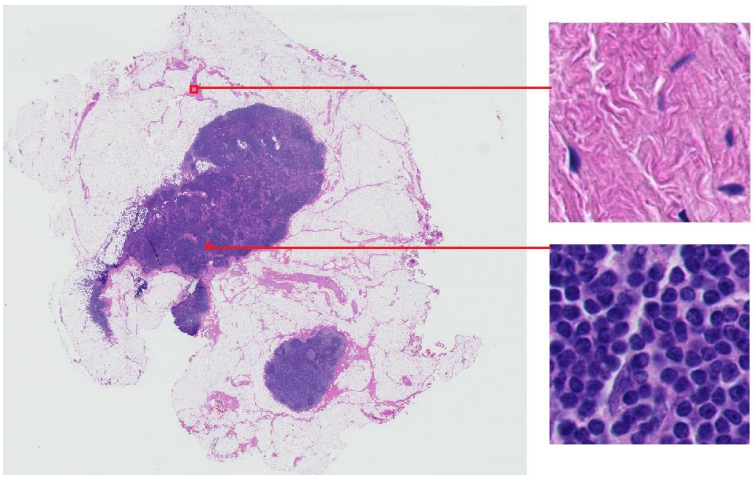

Figure 1: Difference of content on whole-slide image

So, for better retrieval image is cut on tiles – 256x256 pieces of initial image. They are considered as independent images and descriptors are calculated for them separately. As a result the process

of searching occurs on one whole-slide image: for chosen tile we search for a similar ones and then display the results. Now lets consider descriptors used in this project: intensity histogram and co-occurence matrix. These descriptors are constructed for gray-scale images, but we have colored ones. That means that the first step of calculation is proper transformation of images.

2.3. Transformation to gray-scale images. Each RGB pixel is reduced: 3 higher bits are taken from R channel, 2 higher bits from G, 3 from B and merged. As a result pixel transforms to a byte of information and the whole colored image transforms to a gray-scale one. Such a transformation spawns a loss of information, but its the small one. Numbers (3, 2, 3) are chosen because of precalculated histograms of color channels: they've shown that differences of red and blue shades are bigger than the green.

2.4. Descriptor: intensity histogram. Calculation: each pixel is integer from 0 to 255. The descriptor is histogram of that values – vector, which entry v_i is number of pixels equal to "i".

Intensity histogram is classic descriptor. It's very simple and fast to calculate, can be easily reduced by binning values, but its unnecessary because it doesn't take much memory. The result shows the colors variety on a tile, so its a color feature descriptor.

2.5. Descriptor: co-occurrence matrix. Calculation: we count number of co-occurrences of each pixel with each other. The descriptor is matrix, where the element in i-th row and j-th column is number of occurrences of pixels eqaul to "i" with pixels equal to "j" on distance 1. For each pixel we have 8 neighbor pixels – this is 8 co-occurrences.

This descriptor is also known as GLCM (Gray-Level Co-Occurence Matrix). The result is 256x256 matrix which takes much memory, but most entries are zero (or insignificantly small). So we transform this matrix in a vector and remove such entries. Now it takes acceptable memory and can be easily and effectively compared with other such descriptors.

3. Searching the most similar images

3.1. Similarity function. So, now we have calculated feature vectors of user selected tile and the other tiles. And the question is, how do we properly compare two integer vectors? In order to do that we have introduced a so-called similarity function that, for two vectors yields a degree of their similarity – number in [0; 1] interval. The function is defined as follows:

$$r(x, y) = 1 - \frac{d(x, y)}{D} \tag{1}$$

where: d(x, y) is some metric and D is maximal value that this metric can reach. It's hard to predict which metric will be better in our case, so we have used different ones.

3.2. Metrics: L1 or the taxicab distance. We calculate it as follows:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

It may seem odd, but this metric gave us the best results in comparison with others.

3.3. Metrics: L2 or the Euclidean distance. We calculate it as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

3.4. Metrics: Chi metric or the chi-squared distance. We calculate it as follows:

$$d(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i} \quad (4)$$

3.5. Creating similarity maps. So, now we have a number representing the similarity to the desired tile for each tile of our image. We want to visualize those similarities. That is we want to map each similarity number to some color thereby obtaining a so-called similarity map. We have done that by constructing a vector of all similarities, reshaping it according to the image, normalizing and binning the resulting data to one of the existing color maps.

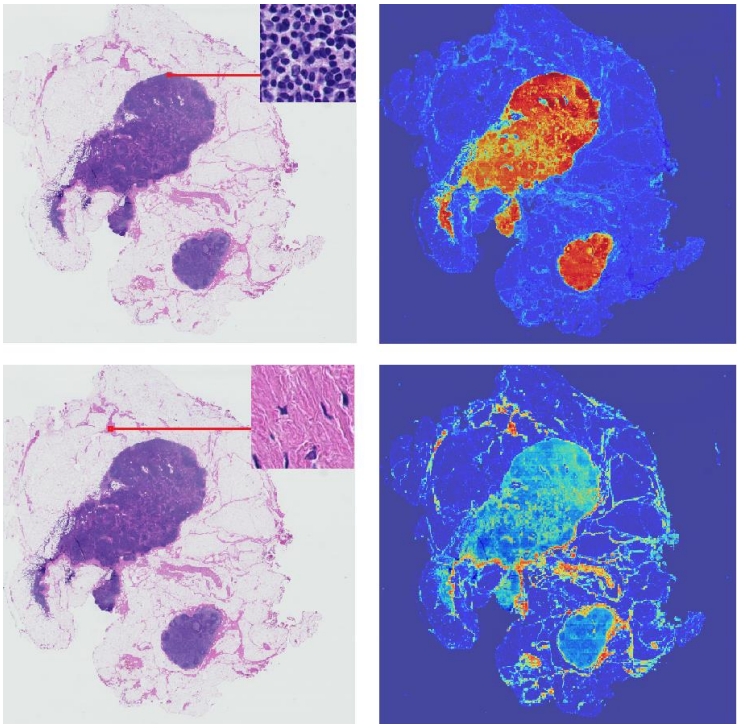

Figure 2. Similarity maps with different queries

References

[1]. Bejnordi B.E., Veta M., van Diest P.J., van Ginneken B., Karssemeijer N., Litjens G., van der Laak J.A.W. M., and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases inWomen With Breast Cancer, Journal of the American Medical Association (JAMA), vol. 318, No 22, pp. 2199-2210. doi: 10.1001/jama.2017.14585.

[2]. Luo W., Hassell L.A. Use Cases for Digital Pathology. In: Digital Pathology: Historical Perspectives, Current Concepts and Future Applications, J.K. Kaplan, K.F.L. Rao (Eds), Springer International Publishing, ISBN 978-3-319-20378-2, 2016, pp. 5-15.

[3]. Beckwith B.A. Standards for Digital Pathology and Whole Slide Imaging. In: Digital Pathology: Historical Perspectives, Current Concepts and Future Applications, J.K. Kaplan, K.F.L. Rao (Eds), Springer International Publishing, ISBN 978-3-319-20378-2, 2016, pp. 87-97.

[4]. Eccher A., Neil D., Ciangherotti A., Cima L., et al. Digital reporting of whole-slide images is safe and suitable for assessing organ quality in preimplantation renal biopsies, Human Pathology, vol. 47, Issue 1, 2016, pp. 115-120.

[5]. Saco A., Bombi J.A., Garcia A., Ramírez J., Ordi J. Current Status of Whole-Slide Imaging in Education. Pathobiology, vol. 83, 2016, pp.79–88.

[6]. Kovalev V., Volmer S. Color co-occurrence descriptors for querying-by-example, Int. Conf. on Multimedia Modeling, Oct. 12-15, Lausanne, Switzerland, IEEE Comp. Society Press, pp. 32-38, 1998 (best paper award).