Isolation of Tumor Areas of Histological Images for Assessment of Quantitative Parameters

Vassili Kovalev[a], **Valery Malyshev**[a], **Artem Piddubnyi**[b], **Alona Moskalenko**[c], **Anatolii Romaniuk**[b*]

[a] *United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NAS of Belarus), Surganov str, 6, Minsk, 220012, Belarus*
[b] *Sumy State University, Department of Pathology, Rymskiy-Korsakov str. 2, Sumy, 40007, Ukraine*
[c] *Sumy State University, Department of Computer Science, Rymskiy-Korsakov str. 2, Sumy, 40007, Ukraine*

Abstract

The research object are biomedical whole slide histological images of breast cancer.

The aim of the work is to develop methods, algorithms and basic elements of a software for automatic search of tumor sites, adaptive assessment of the immunohistochemical markers expression and quantitative assessment of the analysis results. At this stage, we have analyzed difficulties of whole slide histological scans analysis. We have developed algorithms for background separation and color normalization. A search algorithm has been implemented for semi-automatic selection of tumor areas on whole slide histological images.

Keywords

whole slide images, segmentation, clustering, image processing, breast cancer

1. Introduction

The analysis of whole slide images is an extremely laborious process, so the implementation of automated diagnostic algorithms in this area is relevant. However, histological whole slide images have a number of features that complicate the development of such algorithms. These features are: a high level of tissue diversity both in one image and between different images, hierarchy and a large amount of graphic information [1]. The development of an algorithm and its software implementation for the automatic selection of tumor areas in histological whole slide images is a difficult task [2]. Based on that, pre-processing of whole slide images is required. It should include normalization of the color distribution in whole slide histological images and the selection of the image area with specific order of tissue localization to reduce the operating time and to prevent the analysis of the background. A search algorithm for semi-automatic detection of tumor areas by detection of various image descriptors has been developed and implemented.

Features of the whole slide histological images analysis

We used whole slide images of scanned tissue samples with background illumination. Hematoxylin and eosin stained slides, as well as immunohistochemistry samples were used as markers. There are many manufacturers of histological scanners capable to make whole slide images of various modalities. Each scanner saves and compresses the image in different formats. It slows down the development of algorithms. For example, DICOM format was developed in radiology to solve this problem. The standardization of data format in the whole slide histological imaging industry is also very active now.

[*] IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine, November 19–21, 2020, Växjö, Sweden

EMAIL: vassili.kovalev@gmail.com (A. 1); malyshevalery@gmail.com (A. 2); a.piddubny@med.sumdu.edu.ua (A. 3); a.moskalenko@cs.sumdu.edu.ua (A, 4); pathomorph@gmail.com (A, 5).
ORCID: 0000-0002-8154-5875 (A. 1); 0000-0002-8737-8879 (A. 2); 0000-0002-6508-0131 (A. 3); 0000-0003-3443-3990 (A. 4); 0000-0003-2560-1382 (A, 5).

Most formats have hierarchy, so the image is stored as 2D blocks to significantly speed up access to small square areas. So image areas are stored in a lower resolution so there is no need to load all small blocks to show the entire image [3] (Figure 1).

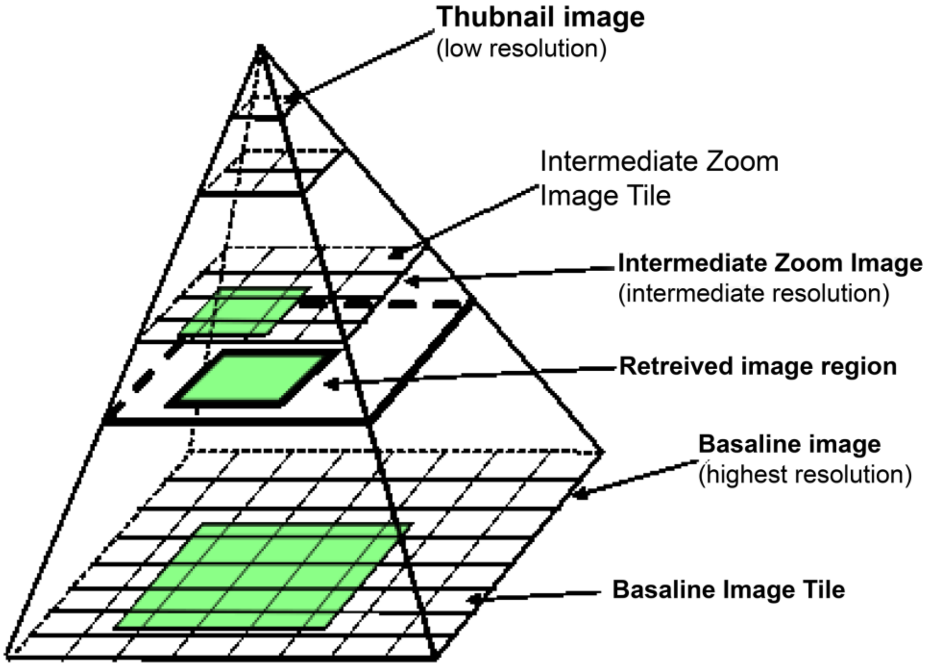

Figure 1. The structure of whole slide image in the computer memory [4]

Machine learning methods have to work with images with various artifacts and color variability. Artifacts can appear both during the tissue samples processing and scanning. There could be a variability of chemical markers, tissue defects, scanner artifacts. Some artifacts are related to the scanner by itself. The scanner takes a picture at maximum magnification and then collects a whole slide image, so artifacts such as lighting and focus differences, stitching and calibration problems may appear [3] (Figure 2). In a couple with the high variability of tissue images, this complicates the application of deep learning methods images. In addition, the high resolution of whole slide images makes it impossible to apply deep learning methods "directly" for images. Thus, most existing solutions use the division of whole slide images into small areas that are used in the dataset. A short list of problems to be solved is presented below:

- high images resolution;
- lack of context (if the image is viewed in sections);
- image artifacts;
- variability of color distribution because of lighting and staining differences;
- variability of tissue samples.

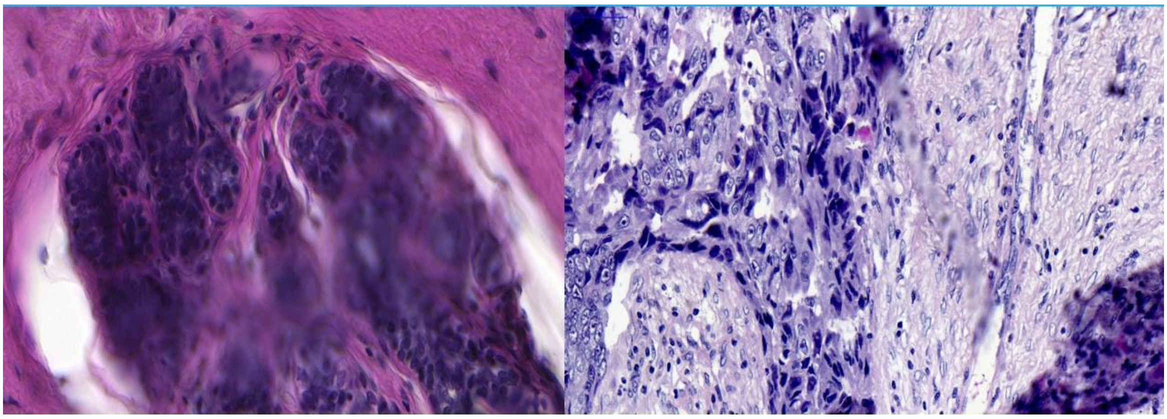

Figure 2. Artifacts in whole slide images

Normalization of whole slide images

Color normalization of histological whole slide images is necessary because of different scanning conditions, such as different scanners characteristics and illumination, different amounts of chemical markers for tissue staining.

The following algorithm is the main method for color normalization of whole slide images:

- exclusion of areas that are not suitable for color scheme detection (for example, a background with no tissue);
- change of RGB color model to another;
- detection of basic vectors (their linear combination defines the color model);
- the reverse transformation of these vectors into RGB color model to show the primary color combinations of image (Figure 3);
- transformation of the image into this colors (Figure 4);
- replacement of the primary colors with the reference ones before the start of the algorithm;
- inverse conversion from the concentration values of the primary colors and RGB reference color values to RGB color model.

In this work, two approaches for normalization of the color distribution were tested. Both of them used the transformation of the RGB value of the image pixels into the optical density values according to the formula (1) [4]:

$$OD = -\log_{10}(I), \tag{1}$$

Pixels with too low optical density were not analyzed, as their signals were considered as the background with a white color. The first algorithm used SVD decomposition. After that vectors normalization according to the length of the vectors, angles between vectors and directions of the SVD expansion were analyzed [5]. In the second algorithm, instead of SVD decomposition, the covariance matrix of RGB image channels were used and vectors were obtained from the eigenvectors of this matrix [6]. Angles were measured as angles between defined by coordinates in the new space vectors and the resulting vectors. The 1st and 99th percentiles of the obtained vectors angles distribution in comparison to defined basis vectors were used as the main vectors. Further transformations are described in the generalized algorithm above.

Thus, the color scheme of all images corresponds to the same base colors of chemical markers. Further, algorithms for image normalization were applied, in particular equalization of the image brightness histogram and deletion of the 1st and 99th percentiles of image pixel intensities to partially avoid the various image artifacts influence. These two normalization algorithms were applied to the Y channel in the YCbCr image representation, since only the image contrast is normalized, and the color scheme should remain the same.

Figure 3. Colors of the whole slide images region component (hematoxylin, eosin, residual)

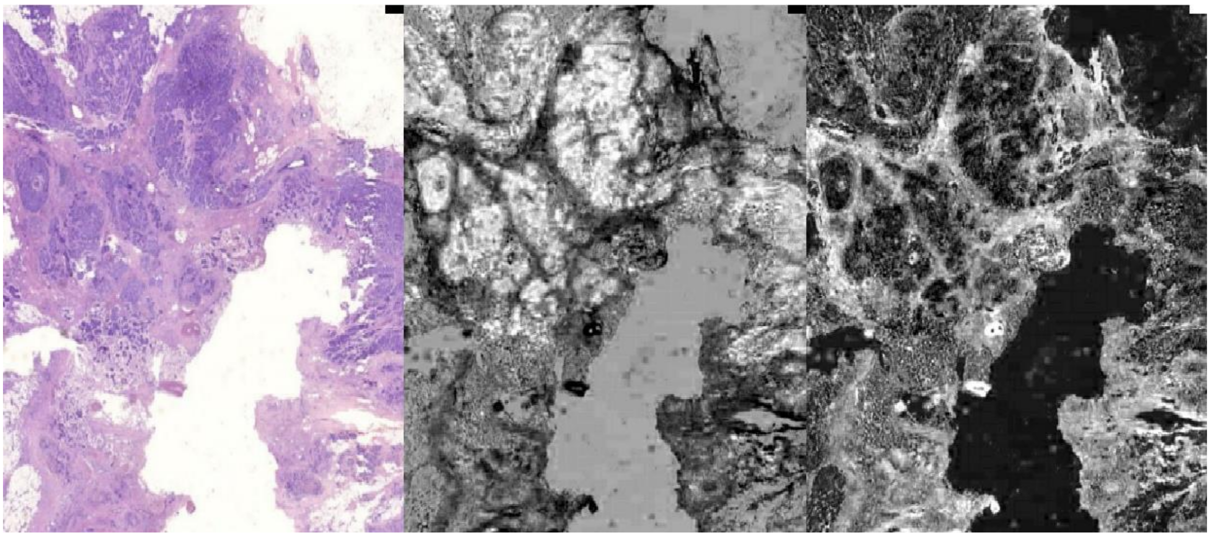

Figure 4. Decomposition of a whole slide image area into components corresponding to chemical markers. Left to right: original image, hematoxylin component, eosin component

Segmentation of a tissue sample in an image

The second essential component is a separation of the tissue image area from the main background.

In this work, a number of algorithms have been used to achieve this result. Considering large dimensions of image, for segmentation we used the whole-slide image layer with smallest magnificatoin (16 times lower than maximum available for the image). First, we blurred an image to increase the smoothness and stability of the final regions. To separate background we utilized S channel of the HSV image representation cause it shows the color saturation in each pixel, which is optimal for detection of mainly white or black background. The algorithm sequentially applies flood fill algorithm to saturation map of the image (S channel in HSV representation of image) starting from empty pixels of the images, considering that each pixel can belong to only one region. After that step we acquire a large number (more than several hundreds of regions). In order to merge the regions into several large ones the algorithm calculates various descriptors for the founded regions. At the moment, descriptors include image channels histograms in RGB, HSV color spaces as well as histograms of histology markers intensities obtained by applying decomposition algorithm. Instead of the descriptor itself, we used the result of its transformation by the principal component method to optimize the clustering algorithm time. This allows to reduce the number of descriptor elements to 16 and significantly speed up the calculations. Afterwards we applied the K-means algorithm to cluster regions into 3 groups. We supposed that background and tissue regions will be in separate groups due to highly different saturation values. The third groups was added to prevent mistakes due to artifacts. So the third group have to be merged with background or tissue regions group. This decision is taken based on relation between groups average saturation values. This method has a lot of parameters, which allowed us to adjust the number of areas and algorithm quality. These parameters are the size of the minimum area, connectivity, the maximum allowable difference between pixel intensities within the same region, etc.

In total, a sufficiently reliable algorithm for analysis of immynohistochemistry and hematoxylin and eosin stained tissue samples was developed. The steps of the algorithm are shown in Figure 5.

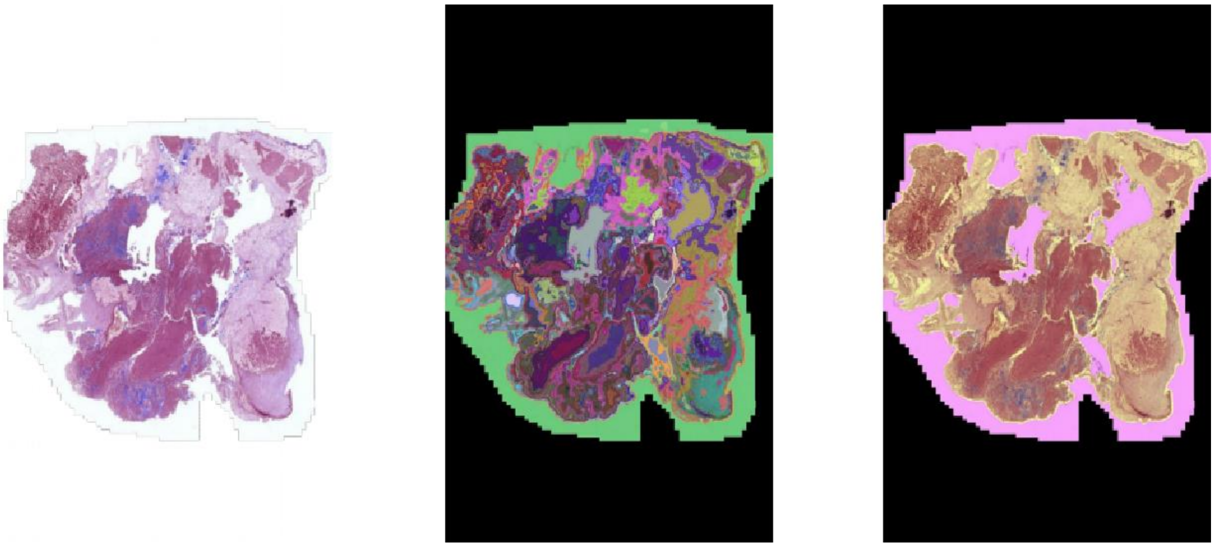

Figure 5. An algorithm of whole slide image: a) original image, b) found regions, c) combined regions of highlighted area with tissue sample

Semi-automatic segmentation algorithm

We have developed a semi-automatic algorithm for the selection of the tumor area at whole slide histological images. A search of similar areas by various descriptors and metrics was the main idea of the algorithm. We implemented the algorithm as web-service. Hence we include details of client-server communication in the algorithm description.

After the image upload to the server, the image is divided into small square regions of equal dimension, which are called tiles. Every region represents a small area of the whole-slide image, and the tile is the smallest unit for the algorithm to process. Thus we create the database of whole-slide image descriptors by calculating different descriptors. Next paragraph will cover the descriptors, which were used in the algorithm because their selection has a high impact on the results of the algorithm. The process of descriptor calculation takes 5-20 minutes depending on the size of the image and the descriptor. Such duration times are large in comparison with the expected search query response time. Thus, after descriptor calculation user can select one region of rectangular shape and any size to use it as a base for the search algorithm. After receiving the selected region coordinates and required descriptor from the user, the server retrieves selected region from the whole-slide image and calculates its specified descriptor. Using that descriptor the server can find distance metric from the selected region descriptor to descriptors of each tile. Distance metric can be L1 metric, L2 metric or correlation metric. Afterwards, the server ranks tiles based on the minimum distance between descriptors and top 10 tiles are sent to the user with a distance map for the whole image (Figure 6). As a result, the user can assess the obtained results.

The first descriptor was the color distribution histogram. For this one, the image RGB channels were combined into one by merging the values. The R channel value used the first 3 bits of the number, the G channel value used the next 2 bits, and the B channel value used the last 3 bits. After that, a histogram with 256 bins was built. To increase the efficiency, the size of the histogram was restricted by 256 elements.

Further, an improved algorithm for descriptor calculation by the histogram was developed. It was an adaptive color histogram, which is a histogram of the colors distribution from a 256-color palette and consists of the number of elements equal to the size of the palette. Palette was prepared by color quantization through clustering of meaningful pixel colors from different whole-slide images.

The last descriptor was the color co-occurrence matrix. This matrix was built also with a palette of colors. So, i and j) element of the matrix has the number of color pixels in the i-th place in the palette, which border the pixel of the color in the palette at j) matrix element in the m place.

All descriptors were normalized to the resolution of the area. A search for similarities and descriptors makes it possible to build a similarity map (Figure 6) for a selected area, allows to find similar tumor areas within the same image.

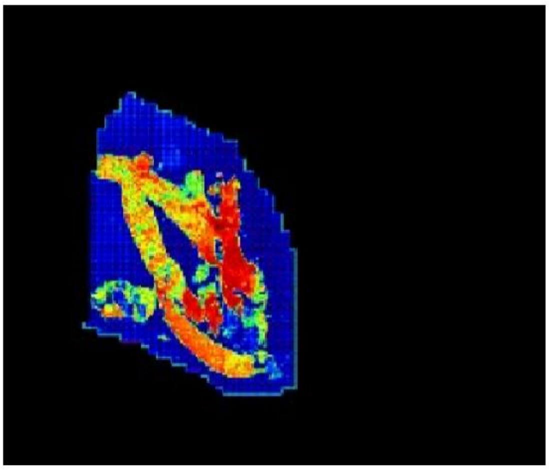

Figure 6. Example of a similarity map

Conclusions

We developed two algorithms for subtasks of the main problem. The first segmentation algorithm reliably selects tissue. The algorithm can be used to reduce the area of the image, which will be processed what results in faster work of all pipeline. In contrast to the first algorithm of segmentation, the semi-automatic algorithm allows user control of search parameters and the region of interest. Selection of the region of interest allows to select not only tissue but different types of tissue and highlight them on the images. For example, selection of a small region of interest with normal tissue will result in more intense highlight for normal tissue than malignant one, what can be noticed easily on the heatmap.

We analyzed the features and problems of whole slide image processing and analysis. Two algorithms have been proposed to solve some problematic aspects. The search algorithm for identification of the chemical markers color allows the normalization of the color space of a whole slide histological image. The whole slide tissue segmentation algorithm reduces the area for processing and reduces the computation time. An algorithm for semi-automatic tumor areas search by the build of a similarity heat map of the selected region to the rest of the image regions was developed.

Acknowledgements

This research has been performed with the financial support of joint UKRAINE-BELARUS R&D project «Development of an automated program for differential diagnosis of breast tumors with a morphometric evaluation of the receptor status of cancer cells» in 2019-2020.

The authors declare no conflict of interest.

References

1. V. Kovalev, Y. Diachenko, V. Malyshev et al. Comparative features of open source software products for the development of an automated breast cancer diagnostic program. EUMJ (2019) 377-385. - DOI: https://doi.org/10.21272/eumj.2019;7(4):377-385.

2. V. Gargin, R. Radutny, G. Titova, et al. “Application of the computer vision system for evaluation of pathomorphological images”, in: 2020 IEEE 40th International Conference on Electronics and Nanotechnology, ELNANO 2020 - Proceedings, pp 469-473. doi:10.1109/ELNANO50318.2020.9088898.

3. N. Dimitriou, O. Arandjelovic, P.D. Caie. Deep Learning for Whole Slide Image Analysis: An Overview. Frontiers of Medicine (2019) 6:264. doi: 10.3389/fmed.2019.00264

4. DICOM Whole Slide Imaging. – URL: http://dicom.nema.org/Dicom/DICOMWSI/.

5. M. Macenko et al. “A method for normalizing histology slides for quantitative analysis”, in IEEE International Symposium on Biomedical Imaging: From Nano to Macro 2009 - Proceedings, pp 1107-1110.

6. P. Bankhead et al. QuPath: Open source software for digital pathology image analysis. Scientific Reports (2017) 7(1) 16878.