Colon Cancer Diagnosis on Digital Tissue Images

Zoltán Kerekes, Zoltán Tóth, Sándor Szénási, Zoltán Vámossy, Szabolcs Sergyán
Óbuda University/Institute of Software Engineering, Budapest, Hungary
{wheelernik, ctsp0vf}@gmail.com,
{szenasi.sandor, vamossy.zoltan, sergyan.szabolcs}@nik.uni-obuda.hu

***Abstract*—The purpose of this project is to develop a software which can be an aid for difficult colon cancer diagnosis and using this system the patients can be helped with an early diagnosis. The aim can be achieved with processing and analysing microscopic tissue images. This paper contains the basic knowledges related to the project and the description of the developed system.**

The implemented algorithms determines the locations and features of glands and save these information for the subsequent diagnosis. One of the most important algorithm in this project is the Color Structure Code, which performs a color based segmentation and the output is the starting point of the further process.

***Index Terms*—medical image processing, colon cancer diagnosis, gland detection, nucleus detection**

I. Introduction

First of all the structure of the colon had to be known. The color structure has multiple layers, the most important of them is the mucosal layer, because the lesions like infections and tumors are evolve from there. In therms of process the important histological components can be seen on Fig. 1. The difference between the healthy and infected mucosal structure can be seen by the lesions of the histological components. The components are e.g. the glands, the goblet cells inside the glands, the goblets cells on the epidermis and the nuclei.

Some similar projects were analysed, which helped us to set up a starting point and these projects drew our attentions for a lot of difficulties that we have to face through the tissue analysis. In [1] the algorithms checked the size of the input image, and if it was too large then it was split into multiple sub-pictures. These sub-pictures were processed on multiple threads with GPGPU, and this method could accelerated the processing time, which was really long in default. After the testing phase, it was found that the algorithms resulted more matches than the real number of the nuclei, thus it was not possible to make an accurate diagnosis with the project.

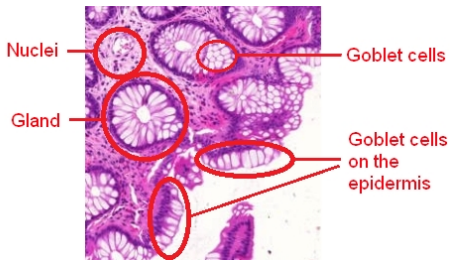


Fig. 1. Structure of the tissue image. The difference among nuclei, glands, goblet cells and goblet cells of the epidemis can be found easily.

In [2] the processing is separated into multiple parts. The preprocessing contains thresholding and noise filtering. For the nuclei segmentation the watershed [3] algorithm was used which had an overflow problem so they had to use some correction to adjust it. For the diagnosis they analysed many features which has a really long processing time. The disadvantage of the program is that several steps required manual configuration, which could slow down the processing time.

In paper [4] the circle-fit is the primary and fundamental procedure and the primitive objects defined by this algorithm will be the analysed components. With the usage of LAB color space the classification and separation of the histological structures was much easier. Thereby we used the LAB color space in out projects for the analysis of the histological structures.

In paper [5], [6] and [7] new approaches are mentioned which can be applied as a more efficient way of feature descriptors as the previous ones.

II. Our approach

We developed two necessary module in the need of diagnosis, one for the gland detection and the other for the nuclei detection. For the gland detection we used color segmentation with HSV and LAB color space, tresholding, then using the connected components method we could identify the glands as independent objects. For the nuclei detection and separation two procedures were implemented. In the implementation phase the statements of [8], [9], [10] were taken into account.

The structure of our algorith can be seen on Fig. 2. In following subsections the different parts of the algorithm will be shown.

A. Gland detection

The digital tissue image contains a lot of noise and homogeneous area that hinders the successful gland detection and segmentation. In the preprocessing phase we use an algorithm that blurs the similar color shades in HSV color space, so the significantly different colors can be separated into independent objects. Thus the white interiors of the glands can be clearly identifiable.

With the Color Structure Code (furthermore CSC) [11] we can blur the pixels with similar colors, so in most cases the boundaries of glands can be separated well from the other parts of the mucosal as it can be seen in Fig. 3.

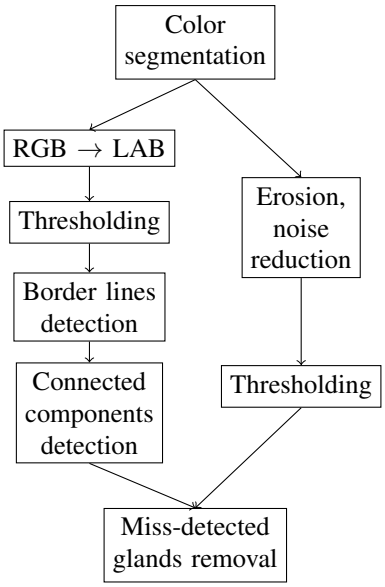
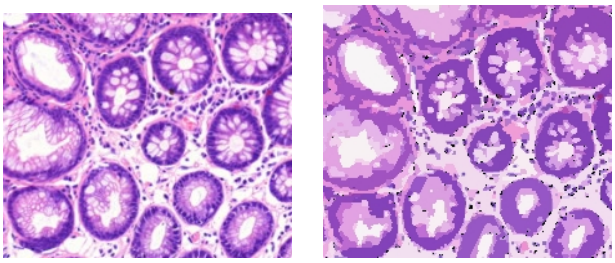


Fig. 2. The flowchart of our algorithm

The CSC is based on the Hartmann-like hierarchical region growing method, and the so-called islands compose the hierarchical structure in different levels. (See Fig. 4.) The island on the very first level contains seven pixel (one selected in the middle and 6 neighbours). On the next level this hexagonal component does not consist of seven pixel, but the former islands build-up a higher level new island. The building of this structure continues to that point, where the whole image is involved in one structure.

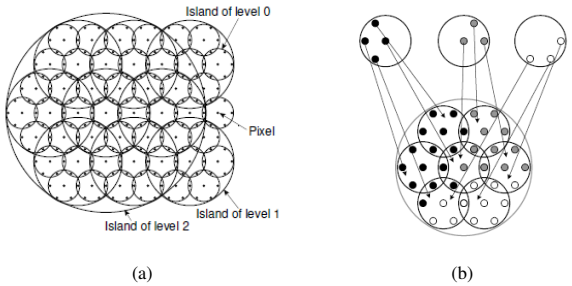
After segmentation the processed image was converted into LAB color space. In this converted image the homogeneous background areas and the contours of the glands can be separated well, and it does not depend the type of the used painting method.

In the LAB color space the A and B components determine the color which is independent from the luminousness. The A parameter means the red-green and the B component the blue-yellow transition. The L component is the luminance. A threshold value is determined considering the ratio of the average intensity of homogeneous regions and the inhomogeneous regions. The result is an image which contains the contour



(a) (b)

Fig. 3. The effect of the CSC segmentation algorithm. (a) The original image, (b) the same image after using CSC algorithm.



(a) (b)

Fig. 4. The structure of CSC [11]. (a) The splitting phase of the algorithm using hexagonal pixel structure. (b) Composing segments from the homogeneous pixels.

curves of glands. In our tests nearly fifty tissue images were used to determine the appropriate threshold value.

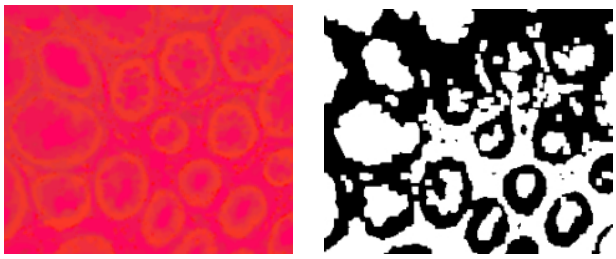
We have found that in threshold value determination only the A component has to be taken into account, because this parameter differs with the same value among the background and gland boundary regions. The result is binarised towards of the further process. The result of thresholding can be seen in Fig. 5.

Because the intensity of some pixels in the contour line is close to the intensity of background pixels, so the boundaries do not constitute closed figures. To generate the connected components the opening morphological operation was used.

The generated boundaries of glands give the most glands in the image. The further task is the separation and identification. The sequential connected component analysis method assigns unique color coordinates for all connected pixel sett, and these coordinates are the bases of further identifications. For all components the area was calculated, and too big (backgrounds) and too small (noises) components do not consider in the following process.

B. Boundary generation and matching

The glands on the boundary of the tissue were not detected after the earlier presented algorithm. For detection of these glands further process was necessary. In the first step noise reduction, then edge detection and connected component analysis were applied. After skipping the small regions the boundary lines of glands were detected. As the last step dilation was used to enhance the boundary lines.



(a) (b)

Fig. 5. Result of thresholding in LAB color space. (a) The image in LAB color space, (b) the same image after thresholding

TABLE I
EXPERIMENTAL RESULT VALUES WITH 10 TISSUE SAMPLES.

	True positive	False positive	True negative	False negative	Precision	Sensitivity	Specificity	Accuracy
1. tissue	74%	14%	11%	0%	0.84	1.00	0.86	0.92
2. tissue	77%	9%	11%	4%	0.90	0.95	0.90	0.93
3. tissue	79%	15%	0%	10%	0.84	0.88	0.88	0.87
4. tissue	73%	17%	3%	7%	0.81	0.91	0.83	0.87
5. tissue	61%	17%	20%	3%	0.78	0.95	0.82	0.87
6. tissue	73%	13%	8%	5%	0.85	0.93	0.86	0.89
7. tissue	70%	14%	11%	5%	0.83	0.94	0.85	0.89
8. tissue	72%	15%	7%	6%	0.83	0.93	0.85	0.88
9. tissue	85%	8%	5%	2%	0.91	0.98	0.92	0.95
10. tissue	68%	17%	8%	6%	0.80	0.92	0.83	0.87

Using this method the glands on the border of the tissue was detected as it can be seen in Fig. 6.

C. Removal of miss-detected glands

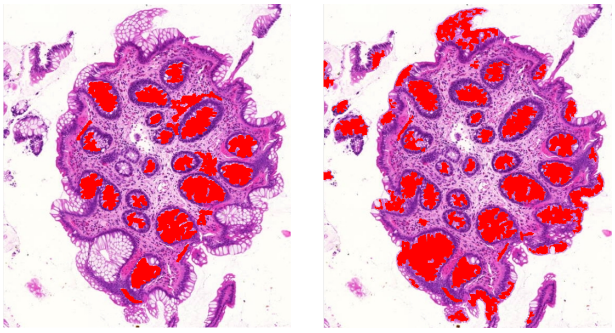
Using the color based segmentation the real glands can be found very well, but unfortunately many non-glands are detected as well. For avoid miss-detection only those regions are considered as glands where few nuclei can be found inside the region. Due to this the image of nucleus regions is generated. Then the number of nuclei in a gland region is counted. If the fraction of the nuclei in a gland candidate region is too high, then this region is not be considered as a gland region in the following process. The result of this method can be seen in Fig. 7.

D. Nucleus detection using the HSV color space

In the first approach the nuclei were detected considering the size and the specific dark color of them. Since the color based calculations are not precise enough the images were converted into the HSV color space.

In this color space the nuclei have significantly high saturation value, so using a well-determined threshold value these can be retrieved. An adaptive algorithm was implemented, which search connected components with small size.

The measure of badly detected nuclei was very low, but adjacent nucleus generates a bigger size component and these are not retrieved in the size based detection process.



(a) (b)

Fig. 6. Glands detection at the tissue border. (a) The originally detected glands, (b) the glands using the boundary line detection

E. Nucleus detection applying color segmentation

At object determination the segmentation algorithms are very useful. In two other papers of the authors [12], [13] the region growing algorithm was applied implemented in parallel environment.

In another approach the Color Structure Code method was applied. Using this method around the dark nucleus black color segments were appeared. These segments can be removed with erosion. The obtained pixels can be used for the description of nucleus positions, but do not give information about the size and shape of them.

The advantage of this approach is that the nucleus region can be determined with bigger precision, but some dark region was detected in the glands as well.

III. TESTS AND RESULTS

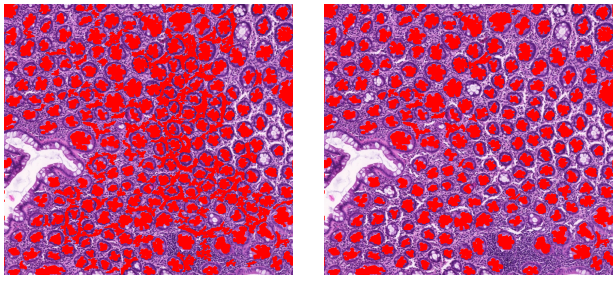
The main function of our system is the gland detection. Our approach was tested on ten representative tissue images.

For analysing the efficiency of our system the frequently used measures of information retrieval were used [14]. These are the true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The inherited measures were used as weel:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$



(a) (b)

Fig. 7. The influence of miss-detected glands removal. (a) The result without using the removal of miss-detected glands. (b) The result using the removal of miss-detected glands.

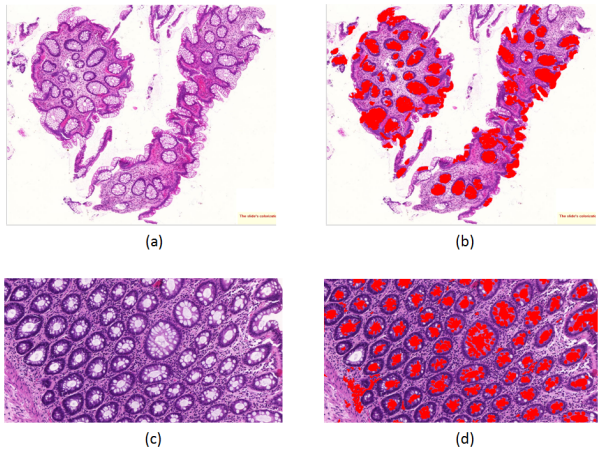


Fig. 8. (a) and (c) presents the original tissue images; (b) and (d) the respective images after applying our method

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

In Table I the mentioned result values can be seen. The advantage of our system is that there are very few number of those glands which are not detected. But the number of false positive cases is rather high, it means there are relatively big numbers of miss-detected objects in the image. The accuracy feature has to enhanced from the inherited property values, it is the most important feature in our test. The average value of this feature is 0.89, so a rather good gland detection can be implemented with our software.

IV. Conclusions

In Fig. 8-9 original tissue images and the obtained results can be seen.

On the ground of the tests our algorithm acts accurately in the most part of tissue images. The fraction of non-detected glands is averagely 6.2%, even the ratio of miss-detected glands is 18.8%. The separated components outside the gland are considered as miss-detected glands, but the most of the can be easily excluded.

Our project is based on a color-based approach. This has that advantage the detection is invariant for the shape, position and size of glands, whereas our algorithms in order to separate different structures consider only the color differences relative to the different histological structure. The disadvantage of the approach is that it is not suitable to for the separation of different merged glands or glands merged by the background. Another disadvantage is that complex correction algorithms needed to achieve precise results.

Overall, declared that our implemented system has a great advantage compared to other implementations in higher detection rate of true glands.

As an additional development possibility we wish to take into account the application of nuclei detection algorithm in [15], which provides a more accurate detection rate of nuclei. In addition higher hit ratio of nuclei our approach may serve more effective results as well.

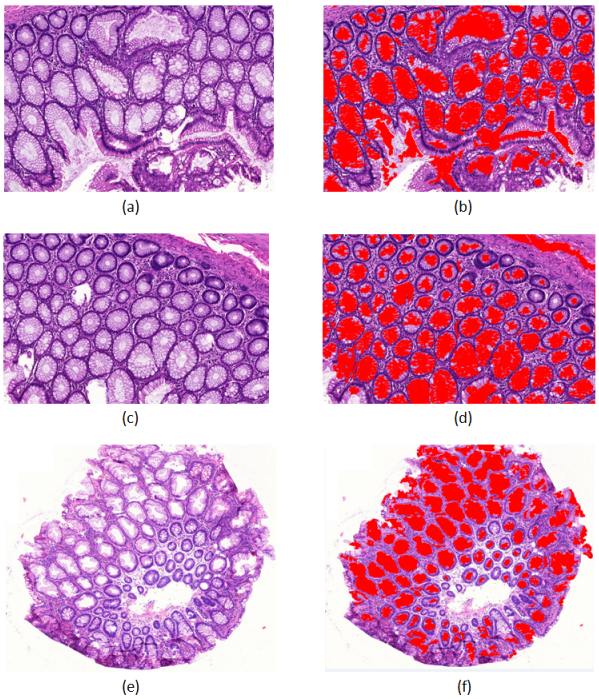


Fig. 9. (a), (c) and (e) presents the original tissue images; (b), (d) and (f) the respective images after applying our method

Acknowledgement

The project was realized through the support of the European Union, with co-financing of the European Social Fund TÁMOP-4.2.1.B-11/2/KMR-2011-0001 and TÁMOP-4.2.2/B-10/1-2010-0020.

References

[1] A. Reményi, S. Szénási, I. Bándi, Z. Vámossy, G. Valcz, P. Bogdanov, S. Sergyán, and M. Kozlovszky, "Parallel biomedical image processing with GPGPU-s in cancer research," in *3rd IEEE International Symposium on Logistics and Industrial Informatics*, August 25-27. 2011, pp. 245–248.

[2] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey," Rens Selaer Polytechnic Institute, Department of Computer Science, Tech. Rep., September 2005.

[3] V. T. DeVita Jr, T. S. Lawrence, and S. A. Rosenberg, *Cancer: Principles and Practice of Oncology*. Lippincott Williams and Wilkins, 2010, vol. 1.

[4] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Medical Image Analysis*, vol. 14, pp. 1–12, 2010.

[5] A. Rövid and T. Hashimoto, "Improved high dynamic range image reproduction method," *Acta Polytechnica Hungarica*, vol. 4, no. 3, pp. 49–59, 2007.

[6] A. Rövid, L. Szeidl, and P. Várlaki, "The HOSVD based domain and the related image processing techniques," *International Journal of Applied Mathematics and Informatics*, vol. 5, no. 3, pp. 157–164, 2011.

[7] G. Györök, M. Makó, and J. Lakner, "Combinatorics at electronic circuit realization," *Acta Polytechnica Hungarica*, vol. 6, no. 1, pp. 151–160, 2009.

[8] J. Tick, "Business process-based initial modeling at software development," in *11th International Symposium on Applied Machine Intelligence and Informatics*, Herlany, Slovakia, 2013, pp. 141–144.

[9] J. Tick and Z. Kovács, "P-graph based workflow synthesis," in *12th International Conference on Intelligent Engineering Systems*, Miami, Florida, USA, February 25-29 2008, pp. 249–253.

[10] J. Tick, "Potential application of p-graph-based workflow in logistics," *Aspects of Computational Intelligence: Theory and Applications: Revised and Selected Papers from the 6th IEEE International Symposium on Applied Computational Intelligence and Informatics*, vol. 2, pp. 71–82, 2012.

[11] V. Rehrmann and L. Priese, "Fast and robust segmentation of natural color scenes," in *Computer Vision - ACCV'98*, ser. Lecture Notes in Computer Science, R. Chin and T.-C. Pong, Eds. Springer Berlin Heidelberg, 1997, vol. 1351, pp. 598–606.

[12] S. Szénási and Z. Vámossy, "Implementation of distributed genetic algorithm for parameter optimization in cell nuclei detection project," *Acta Polytechnica Hungarica*, in press.

[13] ——, "Evolutionary algorithm for optimizing parameters of GPGPU based image segmentation," *Acta Polytechnica Hungarica*, in press.

[14] G. Chowdhury, *Introduction to Modern Information Retrieval (3rd Edition)*. Facet Publishing, 2010.

[15] S. Szénási, Z. Vámossy, and M. Kozlovszky, "GPGPU-based data parallel region growing algorithm for cell nuclei detection," in *12th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, Hungary, November 21-22. 2011, pp. 493–499.