

UDC 519.7

S.N. Potapov¹, N.Ye. Kulishova²¹ KhNMU, Kharkiv, Ukraine, s.n.potapov@rambler.ru;² KhNURE, Kharkiv, Ukraine, kunonna@mail.ru;

K-MEANS APPROACH IN TUMORS CELL COLOR SEGMENTATION IN LAB COLOR SPACE

Digital microscope images are becoming increasingly important in the diagnosis of serious diseases such as cancer. The observations are carried out on the immunohistochemical preparations that change color under the influence of specific markers. Automatic selection of the marked areas and their analysis allows identifying the disease in its early stages. For the detection of marked areas a two-step segmentation method using k-means is proposed: on the first stage a choice of all possible marked image areas is carried out, and on the second - their partition according to the marker expression level. Segmentation is performed in the color space Lab, which allows compensating colors differences caused by marking chemical reactions variations. To evaluate the expression level using - lightness. The marker expression level is evaluated by one of the color space coordinates - lightness L.

SEGMENTATION, K-MEANS, LAB COLOR SPACE, IMMUNOHISTOCHEMICAL MARKERS

Introduction

Currently, immunohistochemistry is one of the basic techniques that are used in modern pathological studies and widely used both in scientific research and in clinical practice. However, in the scientific literature still there are studies that use qualitative and semi-quantitative methods of evaluation only. It certainly reduces the quality and, as a consequence, the information content of the study, making it quite subjective.

During quite a long time, various methods for semiautomatic and automatic computer image processing are constantly being invented, patented and perfected to give the quantitative data. Methods of computer analysis of digital color images can reduce, if not eliminate the subjectivity of the study, and obtain reliable quantitative data. Thereby the immunohistochemistry became objective in, for example, diagnostic, prognostic and research tasks, making studies more accurate. However, there are some problems in digital images quantifying that relate to immunohistochemical preparations manufacturing techniques imperfections. They are: photographic equipment and a microscope quality, different thickness of tissue sections, reaction visualization time, which is always selected empirically, the lack of common indicators and parameters for quantifying immunohistochemical reactions. And most importantly - just as no two people are alike, just as no two identical tissues that are subjected to immunohistochemical studies. That is the problem with many variables.

Thus, for the realization of high-quality and objective analysis of digital images we need to continue to develop methods that will most effectively carry out image processing and getting more accurate and informative quantitative data by the introduction of such methodological approaches that will undoubtedly lead to an improvement in the data interpretation quality.

The purpose of this work: to develop a computer quantitative evaluation method for the calculation of the

tissue structures area of interest by segmenting color digital images of histological micropreparations.

Under the image segmentation we understand the process of its decomposition into component parts, having meaningful sense: objects, their boundaries or other informative fragments that have characteristic geometric features, and so on.

In general, segmentation is step partition of a plane, where the original image function $f(x, y)$ is defined, on the k non-empty associated subsets s_i ($i = [1, k]$) in accordance with some predicate P . This predicate P is defined on the set $S = \{s_1, s_2, \dots, s_k\}$ and receive true value when any pair of points in each subset s_i satisfies the homogeneity of certain criteria (for example, the criterion of homogeneity, based on an assessment of the maximum difference between the single pixel brightness or average brightness value calculated on the relevant area).

In this context, image segmentation is reducing to pixel feature vectors clustering. In [1] it is two basic statistical clustering techniques: hierarchical and iterative (partitions).

Hierarchical methods are procedures that create a series of nested partitions, based on the proximity of the data matrix. By initial clusters division techniques for hierarchical technology it can be distinguish two basic strategies: agglomerate (initializing each object as separate cluster) and divide (initially all objects belong to the same cluster).

As a result of hierarchical agglomerative methods, all classified objects will belong to the same cluster. Therefore, to obtain the relevant partitions, it is necessary to consider different «slices» constructed hierarchy.

The basic idea of iterative methods (splitting method) is finding a single division patterns (objects) in clusters by assigning each object to the cluster, the distance to which center is minimum [1].

1. K-means clustering

The most used method is the hierarchical clustering by k-means [2, 3].

Basically k-means algorithm is based on minimizing an objective function, which is the sum of squares of distances from all cluster points to the center. The objective function, based on the least squares sum criteria, is defined as follows:

$$J = \sum_{i=1}^K \sum_{C \in S_i} w(C) \|C - \bar{C}_{S_i}\|^2, \quad (1)$$

where \bar{C}_{S_i} denotes the cluster center;

K - the number of clusters;

$w(C)$ - the weight of the point C;

$\|\bullet\|^2$ - an expression of the quadratic norm to calculate the distance between two points.

In the classical segmentation methods [4] local characteristics are smoothed and treated as vectors in a metric space, thus describing each image area by average characteristics vector (center). As a difference measure in this approach is used most often square (weighted) Euclidean distance.

When we have a data set X, K-means algorithm minimizes the objective function iteratively. This process consists of several steps:

Step 1. Select K initial centers $\bar{C}_{S_1}, \bar{C}_{S_2}, \dots, \bar{C}_{S_K}$.

Step 2. On the t-th iteration step to distribute the elements of X between K clusters according to the ratio:

$$C \in S_i^t \quad \text{if} \quad \|C - \bar{C}_{S_i}^t\|^2 < \|C - \bar{C}_{S_j}^t\|^2 \quad (2)$$

for all $j = 1, 2, \dots, K, j \neq i$,

where S_i^t is the set of points, for which $\bar{C}_{S_i}^t$ is the center of the cluster.

In other words, the cluster S_i^t is filled with dots, for which performs the condition

$$S_i^t = \left\{ C \left\| \|C - \bar{C}_{S_i}^t\|^2 < \|C - \bar{C}_{S_j}^t\|^2 \right\} \right\} \quad (3)$$

for all $j = 1, 2, \dots, K, j \neq i$.

Step 3. The new cluster centers $\bar{C}_{S_i}^{t+1}$ are calculated on the results of step 2, to reduce the objective function. The new centers are formed by the ratio:

$$\bar{C}_{S_i}^{t+1} = \frac{\sum_{C \in S_i^t} w(C) \times C}{\sum_{C \in S_i^t} w(C)}. \quad (4)$$

Step 4. If all the cluster centers did not change with increasing iteration step, the procedure is stopped. Go to step 2.

The behavior of this algorithm strongly depends on the value of K, the choice of cluster centers and the geometrical properties of the original data.

However, the simplicity of the method it has provided a wide application in pattern recognition problems, image processing and computer vision. Review of the method features and algorithms of its implementation can be found, for example, in [5].

2. LAB color space

One of the basic uniformity criteria for grouping pixels into clusters for image segmentation is the color. Historically, in the first works the RGB color space was used to describe the data [6, 7]. However, this color space is not a good description of human color vision characteristics, so its application to the segmentation tasks is not always effective. So far there are many examples of other color spaces use to describe color for image segmentation [8, 9, 10, 11].

Among the used color spaces there are HSV, Y'I'Q', XYZ, L*u*v* and Lab. Due to their different characteristics they are used for problems of segmentation and further analysis of medical images. It should be noted that the color coordinates of the HSV and Y'I'Q' space, calculated using the RGB values of pixels brightness [10]. RGB is a device-dependent space, i.e. the value of RGB color coordinate depends on the type of device, reproducing color (camera, monitor). Therefore, the use of spaces HSV, Y'I'Q' reduces the accuracy of the segmentation methods. Device-independent color spaces XYZ, L*u*v* and Lab are deprived this lack - they are based on the descriptions of the standard observer properties as color matching functions and connected only with human visual system action specialties.

XYZ values are calculated based on the results of visual stimuli spectral measurements and, in turn, are the basis for finding the color coordinates L*u*v* and Lab. For color specifications most widely used is Lab space, which coordinates are calculated by the following formula [12]:

$$\begin{aligned} L &= 25 \left(\frac{100Y}{Y_0} \right)^{1/3} - 16, \\ a &= 500 \left[\left(\frac{X}{X_0} \right)^{1/3} - \left(\frac{Y}{Y_0} \right)^{1/3} \right], \\ b &= 200 \left[\left(\frac{Y}{Y_0} \right)^{1/3} - \left(\frac{Z}{Z_0} \right)^{1/3} \right], \end{aligned} \quad (5)$$

where X, Y, Z - the coordinates of a given color; X_0, Y_0, Z_0 - coordinates of the nominal standard light source white color stimulus. Coordinate L displays lightness visual stimulus, and a and b coordinates - its chromaticity.

In this color space the difference between stimuli is computed using Euclidean distance [12]:

$$\Delta E = \left[\Delta L^2 + \Delta a^2 + \Delta b^2 \right]^{1/2} \quad (6)$$

This formula after its introduction into use in 1976, has been repeatedly improved and modified, but in this form it is the basis for modern colorimetric technique and is commonly used to assess the accuracy of color reproduction.

Therefore, the use of this formula as a color difference metrics for the segmentation algorithm for color is natural and justified.

3. Proposed segmentation algorithm

Distinction of digital color images with microscopic images of histological preparations is essential instability of the color content. Due to the fact that for different substances under the action of the same marker the chemical reactions take place in different ways, the optical properties of the preparations allow judging the nature of the reaction.

To automate the image processing the k-means segmentation method with reference cluster centers coordinates could be used, since objects of the same type - the cell nucleus, membranes, cell fusion products, etc. are present on all pictures. However, this approach proved to be unacceptable, because the properties of different tissue structures preparations greatly differ from one another, and these differences lead to different results of tissue reactions for markers. Therefore, color coordinates of the same types of tissue are substantially differs on digital images. Examples of Lab coordinate values for the different image areas are shown in Table 1.

Table 1.

Lab coordinate values for the different image areas

Sample number	Color area type	Lab coordinates
1.	Strong marker expression	36; 22; 14
2.	Strong marker expression	48; 21; 26
3.	Strong marker expression	39; 26; 22
4.	Strong marker expression	49; 19; 25
5.	Medium marker expression	50; 13; 16
6.	Medium marker expression	61; 10; 16
7.	Medium marker expression	58; 9; 21
8.	Light marker expression	61; 10; 16
9.	Light marker expression	58; 11; 11
10.	Background	70; 0; 10
11.	Background	91; 2; 1

Furthermore, in marked regions it is necessary to select different expression levels - from light to dark, and determine their relative area.

All these factors lead to the necessity of the observer participation in the segmentation procedure to select the areas for further analysis, and to extract areas, not related to tumor tissue image. Because of variations in the staining reactions the shade of light level marker, for example, can be distinguish from the interstitial background shades visually only. So the proposed segmentation is performed in two stages: at the first stage the investigated image is divided by k-means to a marked region, background, nuclei and membranes. The observer visually evaluates these areas and determines which ones present the desired marker color. At the second stage the selected area is divided by k-means into three levels of marker expression: level of strong, medium and light expression. General procedure diagram is shown in Fig. 1.

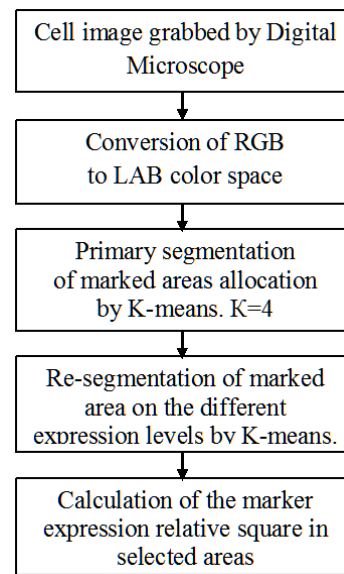


Fig. 1. Flow chart for obtaining clustered image

4. Experimental results

To test the effectiveness of the proposed approach sets of images (digital photos) of various testicular germ cell tumors were used. These images display the results of histological sections staining by various immunohistochemical markers. The photos were made on the microscope Olympus BX-41 with Olympus DP-Soft program (Version 3: 1). Some images contain marked tissues, and the relative area of these regions is varied on different photos. Examples of images are shown in Fig. 2. The main differences between images are as follows: the expressed markers can be determined in a variety of cell and tissue structures (nuclei, cells cytoplasm, intercellular tissue, membrane structures); expression of the marker area, as well as its intensity, in different tumors may vary considerably.

It is evident that the objects in the images have different sizes, marked area is characterized by different structural properties and color, and the only constant feature of these areas is «brown color» in the subjective view. Some techniques were applied to these images: a proposed two-step segmentation algorithm using the k-means in the Lab space and one-step segmentation by k-means with the presentation of images in grayscale and RGB values (Fig. 3).

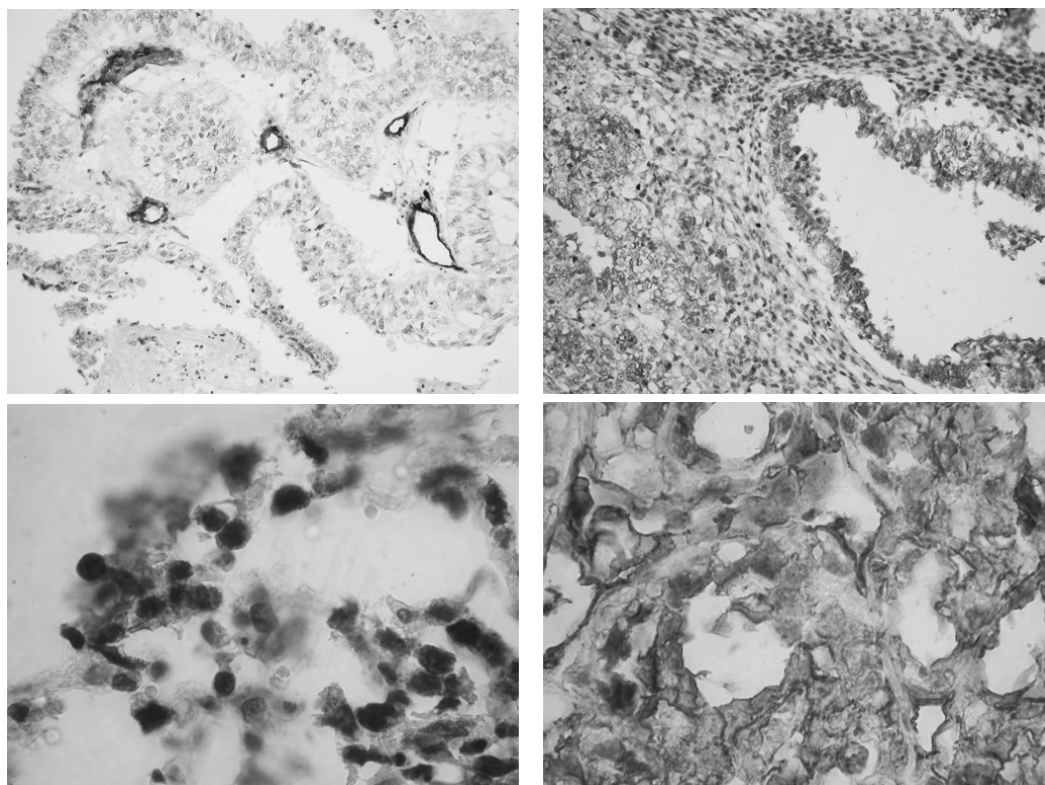


Fig. 2. Image samples

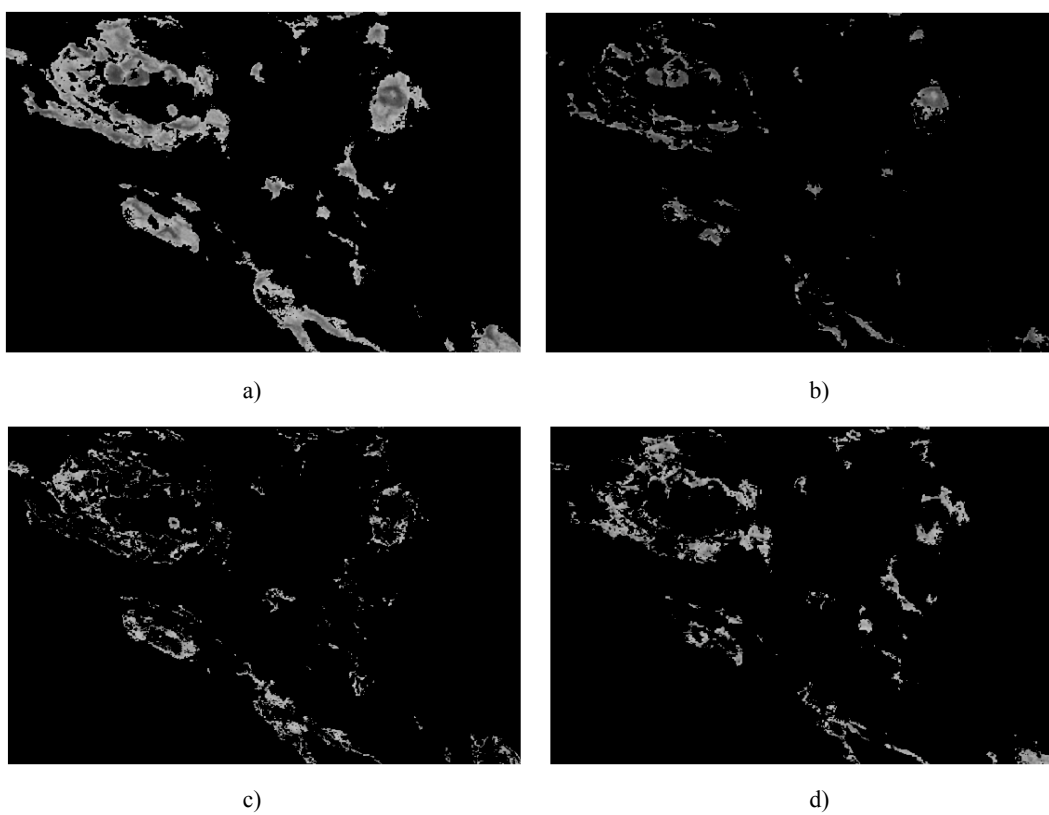


Fig. 3. Allocation of the marker expression levels areas in the re-segmentation:
a) original image for the re-segmentation, the total marker area is 0.113;
b) the marker expression level 1, marked relative area is 0,031;
c) the marker expression level 2, marked relative area is 0.036;
d) the marker expression level 3, marked relative area is 0,046

Numerous examples show, that the use of grayscale presentation don't allow to separate the labeled tissue fields from the unmarked areas. RGB representation leads to the fact that selected components of the image, though different in brightness (and respectively in lightness L), but they are also mixed marked and unmarked tissues. In contrast to these approaches, the proposed algorithm in the first phase of segmentation allows you to split the image onto marked and unmarked tissue areas (Fig. 6), and when re-segmentation – to divide the marked area by a marker expression levels (Fig. 7). On the resulting image for distinct regions the relative marker area S_R is determined by:

$$S_R = \frac{N_{NNZ}}{N \times M}, \quad (7)$$

where N_{NNZ} - number of non-zero pixels in the marker expression level image;

$N \times M$ - image size (pix.).

Found in this way the expression marker relative area allows you to objectively evaluate the basic biological properties of the tumor.

Conclusions

The paper considers the problem of allocation of marked tissue regions in testicular germ tumors cell digital color images and finding of their relative areas. To solve this problem the two-step segmentation algorithm using k-means in the Lab color space was proposed. Experimental studies confirm the high efficiency of the algorithm in conditions of the image different regions color characteristics instability.

References: 1. The Handbook of Pattern Recognition and Computer Vision (4th Edition). Ed. by Chen C.H. – World Scientific Publishing Co., 2010. – 1004 p. 2. Lloyd, S.P. Least squares quantization in PCM / S.P. Lloyd // Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical statistics Meeting Atlantic City, NJ, 1957. 3. MacQueen, J. Some methods for classification and analysis of multivariate observations / J. McQueen // Proc. 5th Symp. Math Statist Prob., 1967. – Pp. 281–297. 4. Shapiro L.G. Computer Vision / L.G. Shapiro, G.C. Stockman – New Jersey, PrenticeHall, 2001: – 617 p. 5. Burney, S.M.A. K-Means Cluster Analysis for Image Segmentation / S.M.A. Burney, H. Tariq // International Journal of Computer Applications. – Vol. 96. – No.4. – 2014. 6. Lucchese, L. Color Image Segmentation: A State of the Art Survey (invited paper) / L. Lucchese, S.K. Mitra. // Image Processing, Vision and Pattern Recognition, Proc. Of Indian National Science Academy. – vol. 67, A. – No.2. – 2001. – P. 207-221. 7. Cheng, H.D. Color image segmentation: Advances and prospects / H.D. Cheng, X.H. Jiang, Y. Sun, J.L. Wang. // Pattern Recognition. – 34. – 2001. – Pp. 2259 – 2281. 8. Paldevbhai, P.J. Color Image Segmentation for Medical Images using L*a*b* Color Space / P.J. Baldevbhai, R.S. Anand // IOSR Journal of Electronics and Communication Engineering (IOSRJECE), Vol. 1, Issue 2, 2012. – Pp. 24-45.

9. Mohapatra, S. Unsupervised Blood Microscopic Image Segmentation and Leukemia Detection using Color based Clustering / S. Mohapatra, D. Patra, S. Satpathy // International Journal of Computer Information Systems and Industrial Management App. Vol. 4, 2012. – Pp. 477-485. 10. Praveena, S.M. Optimization Fusion Approach for Image Segmentation using K-Means Algorithm / S.M. Praveena, Dr.I. Vennila // International Journal of Computer App. – Vol. 2. – No.7. – 2010. 11. Azam, B. Color Based Segmentation of White Blood Cells in Blood Photomicrographs Using Image Quantization / B. Azam, R. J. Qureshi, Z. Jan, T.A. Khattak // Research Journal of Recent Sciences. Vol. 3(4), 2014. – Pp. 34-39. 12. Fairchild, M.D. Color Appearance Models. 3-d Edition / M.D. Fairchild. – John Wiley & Sons, Ltd. – 2013. - 439 p.

Поступила в редколлегию 15.04.2016

УДК 519.7

Метод К-средних в сегментации по цвету изображений опухолевых клеток в цветовом пространстве Lab / С.Н. Потапов, Н.Е. Кулишова // Бионика интеллекта: науч.-техн. журнал. – 2016. – № 1(86). – С. 85-89.

В статье рассматривается проблема определения относительной площади маркированных опухолевых тканей на цифровых микрофотографиях. Предлагается выделять отмеченные области с помощью двухэтапной сегментации на основе метода К-средних в цветовом пространстве Lab.

Ил. 3. Библиогр.: 12 назв.

УДК 519.7

Метод К-средних в сегментации за кольором зображень опухолевих клітин в кольоровому просторі Lab / С.М. Потапов, Н.С. Кулішова // Біоніка інтелекту: наук.-техн. журнал. – 2016. – № 1(86). – С. 85-89.

У статті розглядається проблема визначення відносної площі маркованих пухлинних тканин на цифрових мікрофотографіях. Пропонується виділяти відзначені області за допомогою двоетапної сегментації на основі методу К-средних у кольорному просторі Lab.

Іл. 3. Бібліогр.: 12 найм.