computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

BIOMEDICAL IMAGE SEARCH AND RETRIEVAL ALGORITHMS

O. Berezsky [1], G. Melnyk [2], Yu. Batko [3]

[1] Ternopil National Economic University, 11, Lviska st., Ternopil, 46004, Ukraine, ob@tneu.edu.ua
[2] Ternopil National Economic University, 11, Lviska st., Ternopil, 46004, Ukraine, mgm@tneu.edu.ua
[3] Ternopil National Economic University, 11, Lviska st., Ternopil, 46004, Ukraine, programer_tang@ukr.net

Abstract: *In this paper algorithm for search the tumour cells images in a database is developed. This algorithm based on shape and colour image features.*

Keywords: *colour, shape, contour, query by form, image databases, image segmentation.*

1. INTRODUCTION

Today every biomedical research includes the processing of a big amount of textual and image data. For storing and retrieval of medical data researchers uses various databases. For textual information searching different databases are used, for example MEDLINE, PubMed. They include metadata on medical articles being classified into categories. Searching engines like NCBI-Entrez [1], Medic8 [2] indexes trusted medical websites on the internet, paper titles, keywords and authors names.

There are two main approaches for image searching, either in local database or internet: classification defined by expert and the content based image retrieval. The first approach means that images are classified on some categories: disease, organ, tissue type, magnification etc. The second approach means that images are segmented into homogeneous regions, based on some rules, and their different features are calculated.

IBM’s Query By Image Content (QBIC) [3] search technology allows user to find images by selecting colours from a palette or by sketching shapes on a canvas. Or, refine existing search results by requesting all images with comparable visual attributes.

Increasing size of images database requires effective search mechanism. For it’s implementation it is necessary to solve automatic image description task and indexation of the extracted information. Image segmentation considerably improves speed of image search in the case when it contains the group of objects.

Images are segmented into homogeneous regions. During segmentation the features of each region are calculated and stored in database.

As shown in work [8] image analysis contains such stages: the determination of edges, the segmentation, the determination of texture features and representation, the organization of their features. The goal of this paper is the determination of edge and colour features of tumour cells images and developing the algorithm for search of these images in a database.

2. OBJECTS SELECTION ALGORITHM AND DETERMINATION OF SHAPE FEATURE

Processing of image means realization of certain actions for achievement of certain result over it. The shape of objects is important image feature for content-based image retrieval. An important class of shape analysis algorithms is based on the representation of the contour of objects. Today a lot of algorithms of determination of the contour of object are based on the analysis of the function of brightness.

Let’s consider some area as an object on entrance image, where brightness changes in certain ranges.

For determination of points of image we will use threshold segmentation. A threshold for segmentation is determined on 10% higher a from a minimum brightness’ image. The threshold is determined on the basis of analysis of histograms of previous and apriority information: brightness of the background is less than brightness of the objects. For determination of points, that do not belong to the background such equation is used (1):

$$g(x,y)=\begin{cases}(x,y)\in R, f(x,y)>T\\(x,y)\notin R, f(x,y)\le T\end{cases} \tag{1}$$

where (x,y) – coordinates of a current point, that belongs to the region R, $f(x,y)$ – luminance function, T – threshold.

The advantage of the algorithm is the high fast-acting and the exactness of determination of informing points. The disadvantage of the approach is the sensuality on the noise influence, that can be negatively reflected on the results of the segmentation process.

Let us name an indissoluble curve which links maximum pixels the contour of a 2D object.

The selection of informative pixels on an entrance image happens by means of comparison of brightness. As a result of processing, binary image, where informing pixels are encoded as 1 and background as 0. Algorithm "Radial Sweep" with "Jacob's stopping criterion" is used in the toolbox for contour tracing:

1) choice of go-off point;

2) moving clockwise, a choice of next (neighboring) pixel, that does not belong to the background, but borders with it;

3) verification in the presence of next (neighboring) point, that meets the conditions of the point belonging to the contour is carried out. If not, then transition to the point 5;

4) a recoil is conducted at the positive result of the point 3 (the co-ordinates of active point did not test the changes). Background status is appropriated to the point, and a previous point is elected as active and transition to the point 2 is carried out;

5) verification upon completion of determination (shorting) of the contour is executed, the active point got back in the initial position with the use of the Jacobs' criterion. If a condition is not executed, – transition to the point 2;

6) after the receiving of reserved contour (the program successfully selected some integral field of image, which is different from the background), appropriation of cell identifier to every point of image of limited by contour is conducted, to avoid the repeating of working up the given pixels, and also for facilitation of subsequent prosecution of image and selected field. In case of appropriation completion – transition to the point 1.

In our application, two types of contour representation are used: the matrix of boundary coordinates and centroid distance functions.

The object contour representation with the help of the matrix of boundary coordinates is a simple way to deliver information. The matrix of boundary coordinate is given by:

X	Y
x_1	y_1
x_2	y_2
…	
x_N	Y_N

where N – total number of pixels on the boundary.

The centroid distance function is the contour representation which will be used in extracting the shape features. It is defined to be the distance of boundary pixels from the centroid (x_c; y_c) (2):

$$R(S)=\sqrt{(x_s-x_c)^2+(y_s-y_c)^2} \tag{2}$$

where N – amount of the points which belong edge of object, (x_c,y_c), – centroid of the object, (x_c,y_c), coordinate points which belong edge of object.

Analysis and classification of the object's shape is one of the important tasks at artificial perception. Determination of the object's shape is based on the analysis of the object's features. Basic metrical feature for authentication object's shape are area and perimeter (length of the edge) of the object. The features of the area and perimeter are invariant to displacement and rotation, but not invariant to scaling. However the proper correlations of the area and the perimeter is invariant to all types of affine transformations.

Perimeter is the length of the object's edge, which is based on the calculation of distances between neighbour edge points:

$$P=\sum_{i=1}^{N-1}\sqrt{(x_i-x_{i+1})^2+(y_i-y_{i+1})^2} \tag{3}$$

where N – number of points which form the edge of object, (x_i,y_i) and $(x_i+1,\ y_i+1)$ – coordinates of two neighbour points which belong to edge of object.

Area of object is the sum of all points which belong to the object and determined as:

$$S=\sum_{i=1..N,j=1..M}(x_i,y_j) \tag{4}$$

Let's consider the point of the crossing of two diagonals of rectangle (vertex of rectangle are minimum and maximal coordinates of object) as the centre of object and calculated on the following formula [7]:

$$x_c\approx\frac{(x_{\min}+x_{\max})}{2}; \tag{5}$$

$$y_c\approx\frac{(y_{\min}+y_{\max})}{2}, \tag{6}$$

where $x_{\min}, x_{\max}, y_{\min}, y_{\max}$ – minimum and maximal values of abscissas and ordinates which limit area the object, respectively.

The longest chord (large axe) $D_{\max}$ is a line which connects two points of the object's edge where distance between them is maximal (x_i, y_i) and (x_j, y_j):

$$d\left[(x_i, y_i), (x_j, y_j)\right] \to \max \tag{7}$$

Smallest axe $D_{\min}$ is a line that perpendicular to the longest axe. The lines, which are conducted through the ends of longest and smallest axes create a rectangle (base rectangle) with the maximal area which includes for itself an object's edge.

Eccentricity is a relation of longest and smallest axes of the object (invariant for affine transformations):

$$D_{ek} = \frac{D_{\max}}{D_{\min}} \tag{8}$$

The circularity (invariant under affine transformations) of an object is defined as [8]:

$$T = 4\pi\left(\frac{S}{P^2}\right) \tag{9}$$

This attribute is also called the thinness ratio. A circle-shaped object has a circularity 1; oblong-shaped objects possess a circularity of less than 1.

Then the coordinates of center of masses of the object is calculated as

$$x_c = \frac{1}{N}\sum_{f(x,y)\in R} x\,, \tag{10}$$

$$y_c = \frac{1}{N}\sum_{f(x,y)\in R} y \tag{11}$$

where N – amount points which belong to the object's edge, (x_c,y_c) – coordinates of the center of masses of object, $f(x,y)$ – coordinates points which belong to the object; R – region of interest.

The result of the analysis of the object's features is the objects' classification.

3. COLOUR FEATURE REPRESENTATION

Colour histogram is a popular colour representation scheme that has been used in many image retrieval applications. It works quite well in quantifying global colour content in images.

For similar objects search quantized histogram is used. Every cell on a cytology image is represented by the limited colour palette. For the vector quantization of colours in the image of each individual cell, k-means algorithm is used in RGB colour space. The number of clusters is defined experimentally and equal to 8.

The colour feature of cell image is formed as a set of colour values of each cluster and it's relative area [8]. Length of the coloured feature depends on the defined number of clusters. Let's denote: A is the first image, K^A – number of clusters for an image A, I_i denotes the colour value of i cluster, S_i is a relative area of i cluster. The colour feature f_A for image A is defined as:

$$\mathrm{f}_A = \left\{(I_i, S_i) \mid I_i \in \{R,G,B\},\ R,G,B \in \{0,1,...,255\},\right.$$
$$\left. 0 \le S_i \le 1, \sum_{i\in K} S_i = 1, 1 \le i \le K^A \right\}$$

This representation schema is compact enough and extracts the most meaningful and distinct colours in the image of cell.

For two images A and B, colour features are defined as $\mathrm{f}_A = \{(I_i^A, S_i^A) \mid 1 \le i \le K^A\}$ and $\mathrm{f}_B = \{(I_j^B, S_j^B) \mid 1 \le j \le K^B\}$. Let's define the distance between any two colours as Euclidean distance:

$$W(I_i^A, S_j^B) = \left\|I_i^A - I_j^B\right\| = \sqrt{(R_{I_i^A} - R_{I_j^B})^2 + (G_{I_i^A} - G_{I_j^B})^2 + (B_{I_i^A} - B_{I_j^B})^2} \tag{12}$$

Let's find a colour I_k^B, from the image B which has the minimum distance to the colour I_i^A:

$$k = \arg\min_{j\in K^B} W(I_i^A, I_j^B) \tag{13}$$

The found k is used to calculate the distance measure:

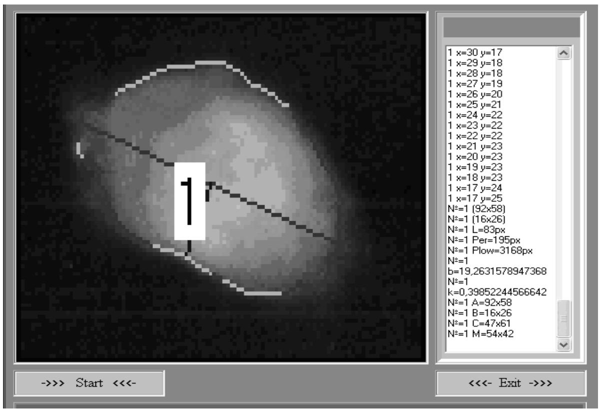
$$D[(I_i^A, S_i^A), \mathrm{f}_B] = \left|S_i^A - S_k^B\right| \cdot W(I_i^A, I_k^B) \tag{14}$$

between the i feature element (I_i^A, S_i^A) and feature f_B. The relative areas S are used as weighting coefficients. Thus, for each colour in the image A the is closest colour in the image B is found. Distance $D[(I_j^B, S_j^B), \mathrm{f}_A]$ can be calculated in the same way. The distance between images A and B is defined as follows:

$$d(A,B) = \sum_{i\in K^A} D[(I_i^A, S_i^A), \mathrm{f}_B] + \sum_{i\in K^B} D[(I_j^B, S_j^B), \mathrm{f}_A] \tag{15}$$

Now we describe the search algorithm to find the best match between colour feature of query image A and colour feature of image B in the database based on distance *d* calculation .

Input: two RGB images *A* and *B* and number of clusters *K*.



```
1 x=30 y=17
1 x=29 y=18
1 x=28 y=18
1 x=27 y=19
1 x=26 y=20
1 x=25 y=21
1 x=24 y=22
1 x=23 y=22
1 x=22 y=22
1 x=21 y=23
1 x=20 y=23
1 x=19 y=23
1 x=18 y=23
1 x=17 y=24
1 x=17 y=25
№=1 (92x58)
№=1 (16x26)
№=1 L=83px
№=1 Per=195px
№=1 Plow=3168px
№=1
b=19,2631578947368
№=1
k=0,39852244566642
№=1 A=92x58
№=1 B=16x26
№=1 C=47x61
№=1 M=54x42
```

Fig. 1 – Example of the cell contour extraction.

Algorithm:

1) On the first step provide the clusterization of both images *A* and *B* using the *k*-means method with *K* clusters. Get the label matrices L^A, L^B, and colours of every cluster I^A, I^B.

2) On the basis of labelling L^A, L^B compute segment relative size S_i^A i S_i^B for both images.

$$S_i = \frac{L_i}{N * M},$$

where *N* – width and *M* – height of image. Save f_{ar}, f_B.

3) Inside the loop for $1 \le i \le K$ for each I_i^A use (12) and (13) to find the segment label *k* which has the closest colour. Use (14) to calculate the distance measure between the current feature element (I_i^A, S_i^A) and feature f_B . Calculate the sum:

$$\sum_{i \in K^A} D[(I_i^A, S_i^A), f_B]$$

4) Execute the third step for an input image *B*.

5) Calculate the distance *d* between images *A* and *B* using (15).

Now we use *d* for dissimilarity measure during the search.

4. WORK ALGORITHM AND EXPERIMENTAL RESULTS

The current implementation of Image Retrieval Application uses a colour and form for a search and retrieving cells on cytological images [9]. The system is developed in Borland Delphi v7.0, to provide client platform independence. The database contains up to 500 colour images from the gallery of tumour cells' pictures.

Image segmentation and each tumour cells' features calculation is provided during database creation.

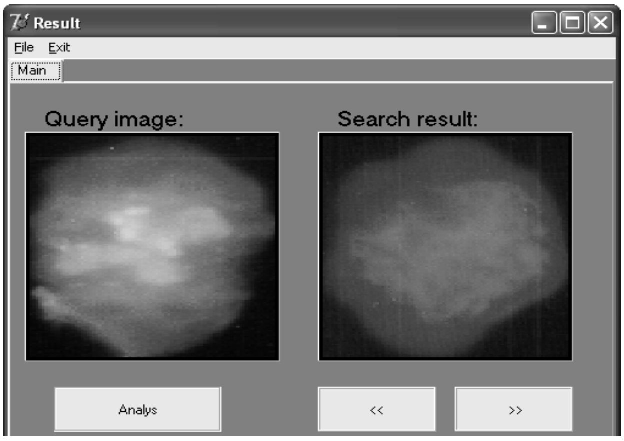
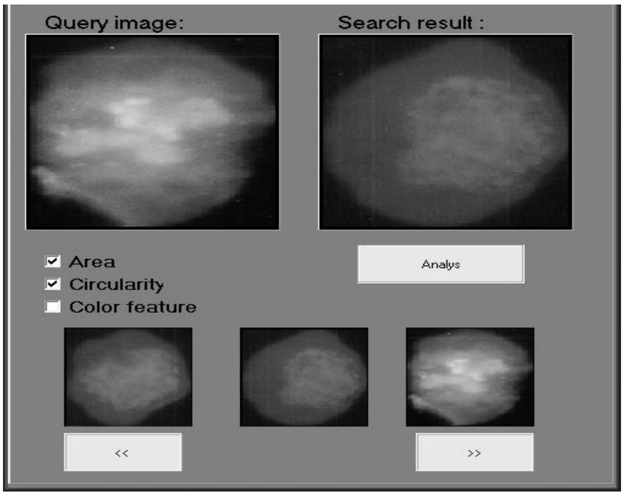


Fig. 2 – Window of to display the search results.

The shape features are stored in the program as a vector. Such representation simplifies a subsequent computational process. A colour feature is stored as a table where for each cluster, in colour space, a value of *R,G,B* components and cluster area *S* is stored. The preliminary evaluation of cells similarity is calculated based on distance between them in colour space. After it, the array of candidates is formed. A final decision about similarity is made comparing input image cell shape feature to every cell from array of candidates. Those candidates, which similarity to input image is greater than some predefined threshold are accepted as similar.

The selection of objects takes place as on the basis of common similarity between objects and after separate signs (by the group of signs). Example of area and circularity features based objects search is shown on Fig. 3:



☑ Area
☑ Circularity
☐ Color feature

Fig. 3 – Cell image retrieval using only area and circularity features.

The database search of similar objects takes place after the following algorithm:

1) Determination of criteria of search and thresholds of similarity;

2) Verification of similarity between *i* feature query image and and feature *n* database image.

3) If the *n* image of data-base dissatisfies the criterion of similarity, it is eliminated from the subsequent analysis.

4) If in a data-base there were no objects, that satisfy to the criteria of search, transition of p 6. is executed

5) If verification is executed after all signs, transition of p 6, differently an *i+1* sign gets out and is executed transition to p 3.

6) Output of the search results.

Complication of algorithm of selection is calculated as:

$$SKL = N + \sum_{i=1}^{Q-1} n_i$$

where N – is an amount of the objects in a data-base, Q – is an amount of features of chosen for comparison, n_i – is an amount of objects, that are remained in a data-base after verification on *i* feature.

The features database (Fig. 4) is created at the time of images gallery analysis and is stored in a plain text form. For every image file (file) there is stored it's name (name), number of cells (numcells), and features of each cell (cell). And for each cell the features of shape (shape) and colour (colour_feature) are stored.

As a query for the search in the database, the image of one cell is used. The search is provided by comparing features. Some typical search results are shown on Fig. 5, which illustrates, that using described features considerably improves the results of search and retrieving on the medical images galleries.

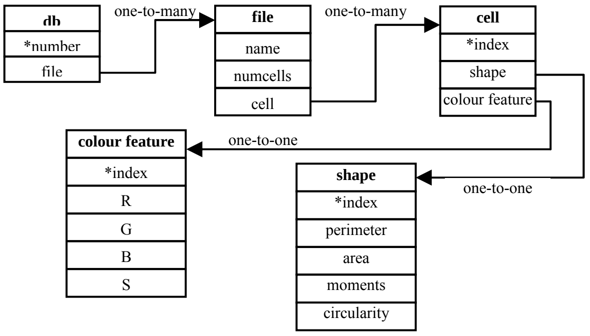


Fig. 4 – Features database structure.

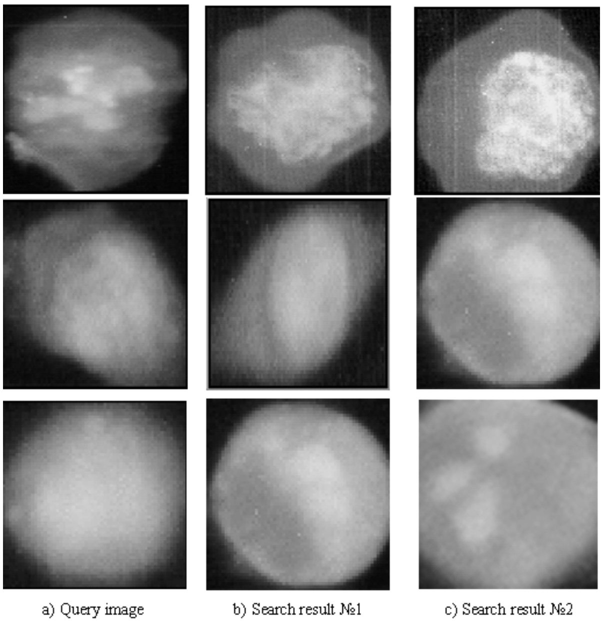


Fig. 5 – Results of cell image retrieval using the features of the colour and shape.

5. CONCLUSION

In this work the algorithm of the cytological images search is developed, which is based on the contour and texture feature extraction, and a designed application for search and retrieving the tumour cells' images is described.

6. ACKNOWLEDGEMENTS

This work was supported by project IOSU-10-2008-"B" "Information-analytical system for research and diagnosis of tumour (cancer) cells based on their images analysis".

7. REFERENCES

[1] http://www.ncbi.nlm.nih.gov/Entrez

[2] www.medic8.com

[3] http://wwwqbic.almaden.ibm.com

[4] W. K. Pratt Digital Image Processing: PIKS Inside, Third Edition, New York:John Wiley & Sons, Inc., 2001. – 736 p.

[5] O.M. Berezsky, Yu.M. Batko Analysis of algorithms image processing //Visnyk Lviv Polytechnic National University. Computer science and information technologies – №565. – 2006. – pp. 212-216.

[6] V.A. Soufer Image processing computer methods – M.: FIZMATLIT, 2003. (In Russian) – 784p.

[7] Ya.A.Fyrman, A.V Krevetskiu, A.K. Peredreev, A.A. Rozhencov, R.G. Khavizov, I.L. Egoshuna, A.N. Leykhin Intriduction in contour analysis: image and signal processing application, 2nd ed. – M.:Fizmatlit, 2003. (In Russian) – 592p.

[8] Wei-Ying Ma, B. S. Manjunath. NeTra: A Toolbox for of Navigating Large Image Databases //Multimedia Systems – vol.7 – №3. – 1999 – pp. 184-198.

[9] O. Berezsky, G.Melnyk, Yu. Batko. Image search and retrieval application. *Proceedings of the International Conference on Computer Science and Information Technologies (CSIT'2007)*, Lviv, Ukraine, 27-29 September 2007, pp. 121-123.

__Oleh Berezsky,__ graduated from the automatics faculty of Lviv polytechnic institute in 1985. Received his Ph.D degree in 1996. Associate professor of Information and Calculating Systems and Control Department of Faculty of Computer and Informational Technologies of Ternopil National Economic University since 2001.

Areas of scientific interests: image analysis and synthesis.

__Grygoriy Melnyk,__ received the B.S. degree from the Ternopil state economic university in 2004, and the M.S. of computer systems and network degree from the Ternopil state economic university in 2005.

His research interests include, image processing, computer vision, texture image analysis and synthesis, gpu processing, artificial intelligence.

__Yuriy Batko,__ received the B.S. degree from the Ternopil state economic university in 2004, and the M.S. computer systems and network degree from the Ternopil state economic university in 2005.

His research interests include image processing, computer vision, contour analysis.