

Content-based Image Retrieval of Histology Images

Alexander Dmitruk, Vassili Kovalev, Ihar Safonau and Aliaksandr Prus

United Institute of Informatics Problems, Belarus National Academy of Sciences, Gomel, Belarus

Telephone(Fax): +375-232-70-09-39

Email: dmitruk@newman.bas-net.by, vassili.kovalev@gmail.com, ihar.safonau@gmail.com, alex_prus@tut.by

Abstract—Our goals included developing and comparative assessing of descriptors for content based retrieval on the large medical image database. Test database consisted of 4000 color histological images which were divided into four equal classes: ovary non-tumor/tumor, thyroid non-tumor/tumor. Both color and grayscale version of extended multidimensional co-occurrence matrices defined on pixel triplets were used as image descriptors. Experimental evaluation of the retrieval performance of the system illustrates its feasibility to serve as a possible research tool to study medical images.

I. INTRODUCTION

Content-based image retrieval (CBIR) is the process of retrieving images by directly using image visual characteristics. In comparison to text-based image retrieval which uses textual language to describe image content, it has significant limitations, especially in medical domain, since image data cannot be fully described textually. Traditionally, CBIR uses visual characteristics, such as color, texture or shape to represent image content and to retrieve images from databases which are visually similar to a query image. One of the CBIR challenges is to bridge the semantic gap between low-level features extracted automatically from images by computers and the high-level interpretation by human. Therefore CBIR for medical databases has the potential to assist clinical decision-making, research, training and easy-to-use retrieval of medical images [1].

II. MATERIALS AND METHODS

Original cases were selected from the database of the Postmortem Laboratory at the Minsk City Hospital for Oncology, Belarus. All tumor tissue samples were immunohistochemically processed using D2-40 mesothelial marker for highlighting tumor angiogenesis. The images were acquired using a Leica DMD108 light microscope at $200\times$ magnification from histological slides containing tumor probes obtained from patients with confirmed diagnosis of ovarian cancer (23 patients, 2000 images) and thyroid gland cancer (the same number of patients and images). A final test database consisted of 4000 histological 24-bit color images of 2048×1536 pixels. All the images were divided into 4 equal classes: ovarian tumor/non-tumor tissues, and thyroid tumor/non-tumor tissues; 1000 images in each class. Representative images of all 4 classes are shown in Fig. 1.

Image object recognition is always based on a model of the object at some level of abstraction. At the lowest level of abstraction (no abstraction at all), an object can be modeled as a whole image and compared, pixel by pixel, against a raw input image. Selecting proper abstraction level is often not trivial task. In pure histograms when calculating cumulative sum of given intensities, geometrical information is completely lost. Classical Haralick's matrices, first introduced in 1973 [2], add geometric information to the ordinary histogram. Co-occurrence matrix for a gray-level image can be represented as a matrix of relative frequencies $P(i, k, d)$ with which the pixels with gray levels (i, k) occur in the image by distance d apart.

Generalized co-occurrence matrices extend Haralick's matrices with additional image characteristics: intensity difference, gradients or angles between them [3]. Even 2D silhouette recognition was implemented using co-occurrence matrices, where relative angles between tangent vectors to the object boundaries were utilized as one of its dimensions [4].

In our experiments image pixel triplets (equilateral triangles with certain gray levels/colors located at their vertices) were considered as elementary image structures. Geometrically, these structures are covering the entire image and being summarized in form of corresponding co-occurrence matrix (multi-dimensional histogram) describe in statistical manner the spatial pixel relationships, i.e. the image texture. Two matrix types are referred to by abbreviations of their corresponding dimensions with the letter I denoting the binned pixel intensity, C —paletted pixel color, and D —the inter-pixel distance.

Invariance for both rotation and reflection was achieved by matrix reduction to a triangular form. In other words sequence order of intensities at sides of triangles becomes unimportant. To decrease the number of elements in matrix and simplify data, feature vector was extracted from the co-occurrence matrix.

The descriptors compared were matrices defined on triplets of pixels describing spatial occurrence of both, color and gray scale intensities, namely: $CCCD$ and $IIID$ matrices with 16 intensity levels and inter-pixel distances ranged from 1 to 5. Since image descriptors employ intensity pixel characteristics, original color images were converted to gray-level images using well-known RGB to Gray transform as well as color space reduction using median-cut method.

Color/intensity binning was performed to reduce color

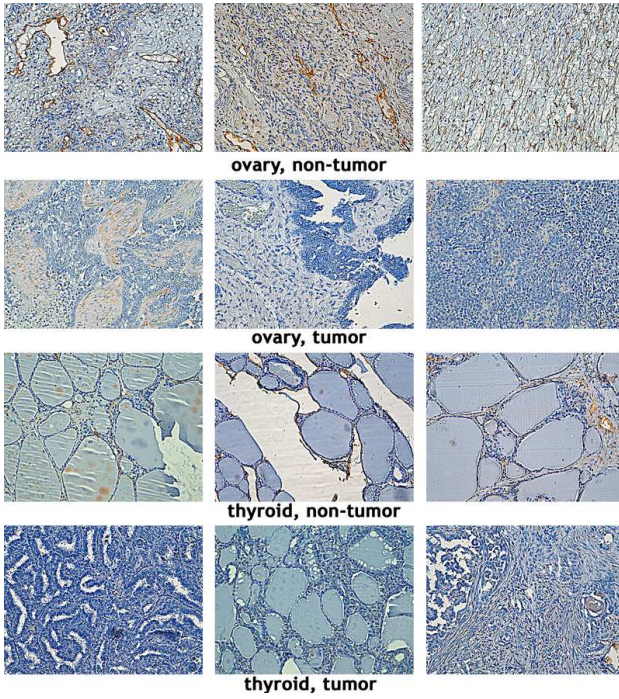


Fig. 1. Original histology images of ovarian and thyroid non-tumor/tumor tissues.

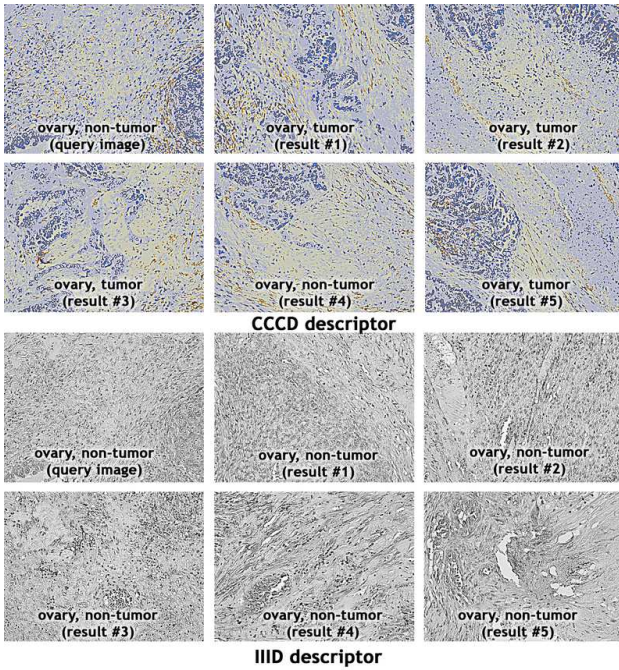


Fig. 2. Retrieval results using two types of descriptors but the same query image. Images are shown in a manner they were used in *CCCD* and *IIID* descriptors (paletted and gray-level images).

noise influence and simplify descriptor calculations.

One of the colors/intensities was labeled as background and skipped during matrix calculation.

Image retrieval was performed by comparing the descriptor (i.e., the co-occurrence matrix) of given query example with descriptors of every image stored in the database using $L1$ distance metric and selecting the top- N most similar cases.

TABLE I
RETRIEVAL PERFORMANCE FOR *CCCD* AND *IIID* DESCRIPTORS ON 4000 IMAGES DATABASE, %.

Top-1 results		
	<i>CCCD</i>	<i>IIID</i>
Ovary non-tumor	93.7	92.3
Ovary tumor	95	91.1
Thyroid non-tumor	94.8	95.5
Thyroid tumor	95.6	91.7
Top-10 results		
	<i>CCCD</i>	<i>IIID</i>
Ovary non-tumor	91.5	89.97
Ovary tumor	93.1	84.73
Thyroid non-tumor	94.5	93.22
Thyroid tumor	93.63	86.17
Top-20 results		
	<i>CCCD</i>	<i>IIID</i>
Ovary non-tumor	90.52	88.66
Ovary tumor	91.8	82.01
Thyroid non-tumor	93.54	92.26
Thyroid tumor	92.76	83.35

III. RESULTS

Both types of descriptors revealed good performance when searching for histological images on query-by-example paradigm (Table I, Fig. 2). Since the four-dimensional matrices are very large, the principal component analysis (PCA) method was applied. As a result, it was found that selecting 22 components for *CCCD* and 12 for *IIID* descriptors covering 95 percent of variance instead of 500 and 575 original variables (matrix elements) provides approximately the same retrieval accuracy with less than 5 percent variation.

IV. CONCLUSION

Generalized co-occurrence matrices showed very promising results as similarity measures of histological images. Considering color over intensity characteristics didn't give significant gain in retrieval performance. The proposed method may be employed for medical image retrieval for assisting the cancer diagnosis processes.

ACKNOWLEDGMENT

This work was funded by the ISTC project B-1682 and CRDF grant under contract BEM1-4028-MK-08.

REFERENCES

- [1] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical application-sclinical benefits and future directions," *International journal of medical informatics*, vol. 73, pp. 1-23, 2004.
- [2] Haralick, R.M., Shanmugan, K.S., and Dinstein, I., "Textural features for image classification," *Transactions on Systems. Man and Cybernetics*, vol. 6, pp. 610-621, 1973.
- [3] V. Kovalev and M. Petrou, "Multidimensional co-occurrence matrices for object recognition and matching," *Graphical Models and Image Processing*, vol. 58, pp. 187-197, 1996.
- [4] Kovalev, Vassili A. , Petrou, Maria, and Bondar, Yaroslav S., "Using orientation tokens for object recognition," *Pattern Recognition Letters*, vol. 19, no. 12, pp. 1125-1132, 1998.