cytological screening, image databases,
early detection of bladder cancer, CBIR

Bogusław Dariusz PIĘTKA[1], Annamonika DULEWICZ[1], Paweł KUPIS[1]

IMPROVING QUALITY OF CYTOLOGICAL SCREENING IN EARLY DETECTION OF MALIGNANCY ASSOCIATED CHANGES

The paper deals with an image database organization and utilization in computer-aided cytology. To illustrate the idea we take as an example the problem of bladder cancer early detection based on urine cytology. In spite of its diagnostic potential for discovering malignancy associated changes (MAC) at the cell level it seems to be underestimated. There is common view that sensitivity of the method, especially for early cancer stages, is relatively low. We depict here just one but significant direction of our works that aims to support pathologists making the diagnosis more accurate and reliable. The key idea relies on automatic searching for MAC by comparing nuclear chromatin structure of objects in a smear with a collection of sample patterns contained in a pathomorphological image database.

1. BIOMEDICAL IMAGING

Many tasks that traditionally required human expertise and human visual skills are still carried out manually. Simply, in spite of recent impressive advancements in image processing methods and computer algorithms these objectives are still too difficult to be fully automated. Many good examples can be found in medical diagnostics. One of such field is cytology and the need of inspecting visually thousands of microscopic images. On the other hand, automation of the process is really an important issue. The need to apply computer-based solutions arises for a number of reasons. (1) The need to quantify findings for comparative investigations. (2) Highly trained thus expensive experts are utilized for tedious, visual tasks. (3) It is natural for human experts to exhibit inter- and intra-observer variability. (4) The amount of data to be processed is too high for the effective application of humans.

Medical imaging as a discipline, in its widest sense, is part of biological imaging and incorporates radiology, nuclear medicine, endoscopy, medical thermography, medical photography and microscopy [1]. Worth noticing may be that in 2010 some 5 billion medical imaging studies were done worldwide.

2. CONTENT-BASED IMAGE RETRIEVAL SYSTEMS

Actually, early pre-CBIR solutions were not generally based on visual features but on the textual annotations of images. In other words, images were first annotated with text and then searched using a text-based approach from traditional database management systems. Comprehensive survey of the text-based image retrieval methods can be found in [2]. Text-based image retrieval uses traditional database techniques to manage images. However, since automatically generating descriptive texts for a wide spectrum of images is not feasible, most text-based image retrieval systems require manual annotation of images. Obviously, annotating images manually is a cumbersome and expensive task for large image databases and is often subjective, context-sensitive and incomplete. As a result, it is difficult for the traditional text-based methods to support a variety of task-dependent queries.

Real content-based image retrieval (CBIR), a technique that uses visual contents to search for pictures from large-scale image databases according to users' interests, has been an active and fast advancing research area since the 1990s. During the past decade, remarkable progress has been made in both theory and system development. One of the first and well known was commercially available IBM's QUBIC (QUery By Image Contents). Other general-purpose solutions, to name a few, are Virage, Candid,

[1] Laboratory of Fundamentals of Computer-Aided Image Diagnostics, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Ks. Trojdena 4, 02-109 Warsaw, Poland.

Photobook, Netra, Blobworld, PicHunter, GIFT, Viper, WIPE and Compass. Most of them may be found and some even tried, as there are demonstration versions available on the web.

Most of CBIR systems have very similar architecture for browsing, archiving, indexing, comparing images, extracting visual features, storing and efficiently retrieving these features. They use common tools for distance measurements or similarity calculation and exhibit similar type of Graphical User Interface (Fig.1). Moreover, what seems even more interesting, there is one more, very important common trait. All of them utilize different visual features which are still low-level compared to high-level concepts contained in images. They do not necessarily correspond to objects in images or the semantic concepts or structures that a user is interested in. Although many authors speak of semantic or cognitive image retrieval, in the end this has not yet been realized with visual features alone. It often comes down to connecting visual low-level features with textual high-level ones. Visual features may be classified into primitive ones such as color or shape, logical features such as identity of objects shown and abstract features such as significance of scenes depicted. Again, all currently available systems use only primitive features unless manual annotation is coupled with the visual features. Even systems using segments and local features, such as listed earlier Blobworld, are still far away from identifying objects reliably. Actually, no system offers interpretation of images or even medium level concepts. This loss of information from an image to a representation by features is called the semantic gap. The situation is surely not satisfactory and the semantic gap definitely accounts for part of the rejections to use image retrieval applications. However, the technology can still be valuable when users understand advantages and problems. The more a retrieval application is specialized for a certain, limited domain, the smaller the gap can be made by using domain knowledge. Thus, at the current state of the art there is a good reason to work on and develop specialized rather than general-purpose CBIR systems and applications. One of such an approach, dedicated for urine cytology, will be shown in a moment.

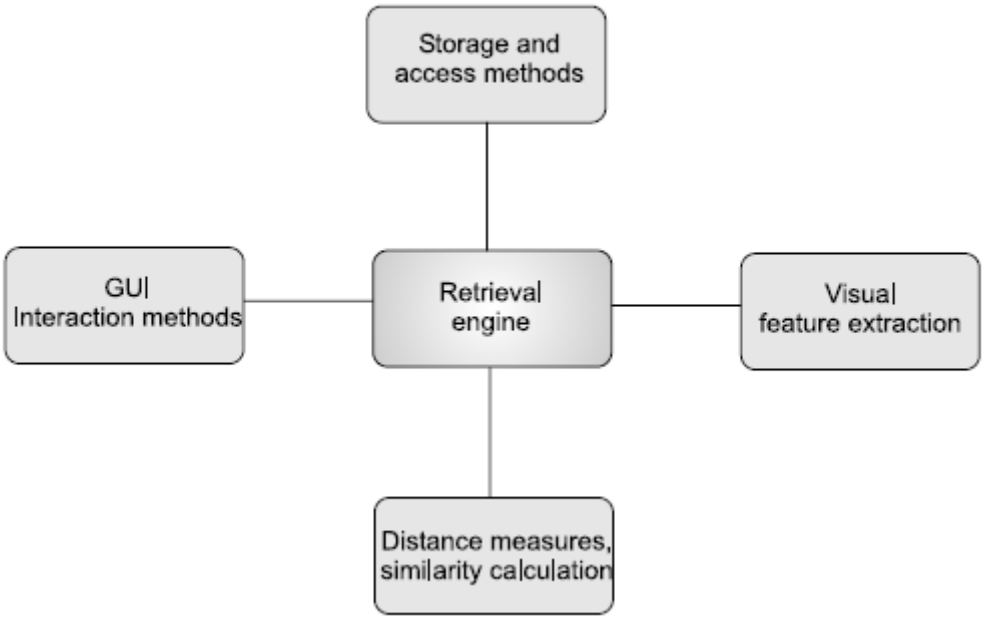


Fig. 1. Common components of a content-based image retrieval system.

Generally, an information processing system consists of two major types of entities: documents and queries. A document can be of any format with any kind of internal structure organizing the information it carries. An information retrieval system treats all documents in collection as unstructured files. Documents are represented with their surrogates. A surrogate is usually a concise representation of the original document being processed for particular purposes and contains only a small portion of the information that the complete document carries. Two examples of document surrogates are article abstracts and image icons. They are more suitable for efficient storage, fast access and processing.

A query is an indication of user's needs. It can take many different forms ranging from Boolean query that is very concise to a detailed specification of the kind of documents the user needs. The query may also be one or more sample documents (such as image queries) that are set as examples for the kind of desired documents. Researchers tend to think of queries and document surrogates as the special kinds of documents that bear many of the same characteristics as regular documents. The retrieval process is to find documents that match the query document based on the specified criteria and processing procedure that is applied to both query and documents in the archive. This vision helps to bring the image retrieval problem and many derived techniques into the big picture of information retrieval.

With image retrieval engine all images in the archive are usually preprocessed by the machine to generate signature files to get retrieval operations faster. Document surrogates should be seen as constituent part of the image database that remains relatively static. On the query side, query-by-example has gained wide acceptance as a major image database-querying scheme. In this approach, the query is not used for direct comparison. Instead, the query image is processed first, following the same procedure, to generate the "query surrogate" which is then compared with the document surrogates from the database (Fig. 2).

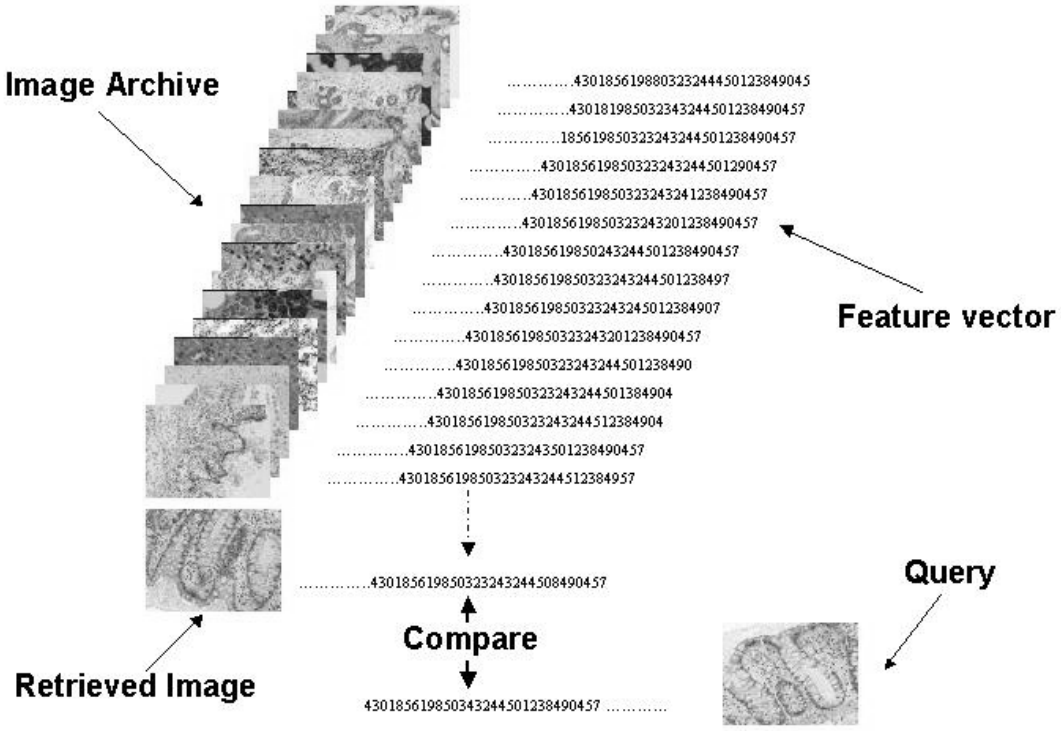


Fig. 2. The idea of retrieving biomedical images from an archive.

The image search and classification tasks are conducted as follows: 1) each image is broken down into a feature vector of (usually) numerical parameters; 2) image comparison is based on vector arithmetic; 3) database search is to find the feature vectors that have the minimal distance (difference) to the query image feature vector. A graphical illustration of the process is depicted below in Fig .3.

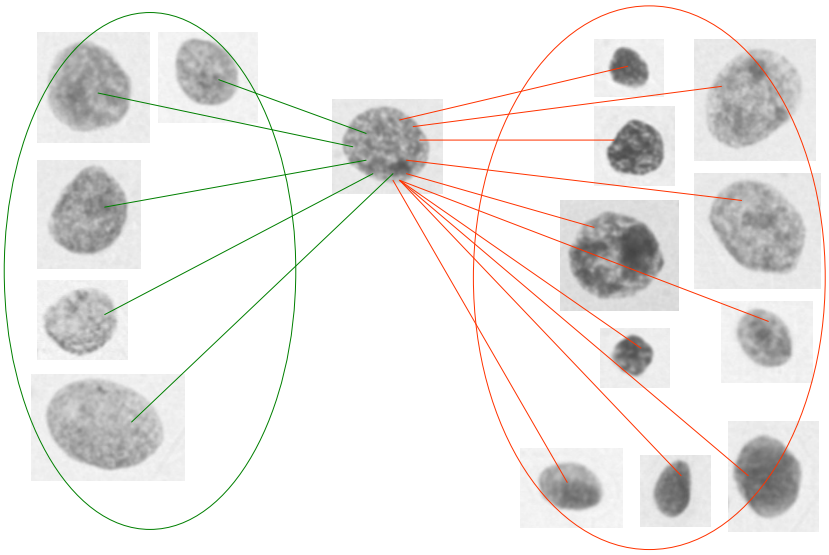


Fig. 3. Computing distances to classify an object to the nearest database type (normal vs. suspicious).

3. DESIGNING CBIR FOR COMPUTER-AIDED URINE CYTOLOGY

As the paper is devoted to image databases rather than biomedical investigations the medical descriptions are very limited here. However, there must be a short introduction into the bladder cancer basics as some diagnostic issues influence the shape of the system.

Bladder cancer occurs more often than one might think. In fact, it is the fourth most common cancer among men and the ninth most common among women in USA and most European countries. Fortunately, the majority of bladder tumors do not grow rapidly. Average period of the disease development is evaluated from 10 to 30 years [3, 4, 5] and early growths can be treated without major surgery. Thus, most patients with bladder tumors are not at risk of developing the cancer that will spread and become life threatening if one but crucial condition is met – it must be recognized in early stage of development. Precise cytological screening is capable of detecting cancerous cells in voided urine or bladder wash before they tend to form a tumor. On the other hand, the number of qualified pathologists will never be sufficient to manage the mass screening tasks manually. This is the reason of our efforts to support the work by means of computerized tools. The aim of support in our system is not to suggest a diagnosis but, after automatic analysis of microscopic smear, to present objects and smear regions that potentially contain malignancy associated changes. A simplified illustration of how the idea is realized can be seen in Fig. 4.

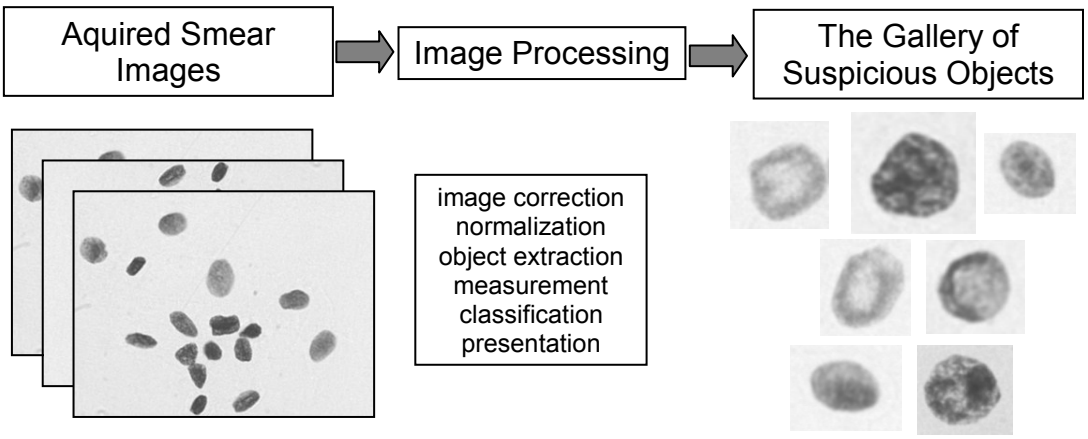


Fig. 4. Image acquisition, digital processing and presentation of suspicious objects.

And now we are approaching the essence of things. There are many different sources of diagnostic information that should be taken into account by urologists making diagnosis. A computer, which knowledge is limited to cytological images, should not decide what is normal state and what is not. The system is only to supply additional data to that already in hand to allow for better diagnosis. The reference database, created and modified by experts and used as a basis for classifying cytological objects, should not define NORM vs. PATHOLOGY a-priori. It should be divided into different classes containing distinctive types of objects. Let it be pathologist, who decides what phenomena and types of morphological alterations are really important in case of particular patient. We just give her/him support supplying visual and quantitative information about the rate of different object classes encountered in a cytological sample under study. The idea of type-based classification is shown in Fig. 5.

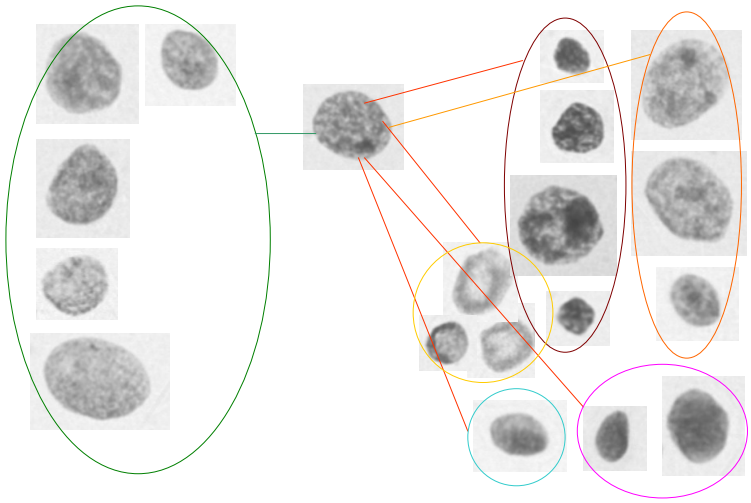


Fig. 5. Classifying to different object types without norm-pathology discrimination.

4. IMPLEMENTATION

Although the paper is not intended for detailed description of image analysis and classification it seems reasonable to mention here that the image processing is carried out by means of a method known as SGF (*Statistical Geometric Features*). Shortly speaking, it relays on input image decomposition by thresholding into a stack of binary images (Fig.6). From every elementary image several geometric parameters of connected regions are computed. The parameters describe size and spatial distribution of different chromatin fractions. This way we get statistical distributions as a function of the threshold. Several statistics derived from those distributions become descriptors of the chromatin texture under study.

Fig. 6. The idea of nuclear chromatin texture analysis by means of *Statistical Geometric Feature.*

Pathologists selected several types of cytological objects and they were included into our database (Fig.7). Each class in the database consists of several representatives

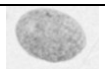
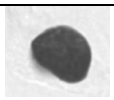
Class code	Typical object	Number of representatives
C0		11
C1		7
C2		5
C3		9
C4		5
C5		6
C6		3

Fig. 7. General view of the pathomorphological image database contents.

The algorithms and procedures described above were implemented in a model of interactive system for computer-aided diagnosing of bladder cancer *(NeoSniffer)*. The look of a final screen containing gallery of possible cancer symptoms is shown in Fig. 8.

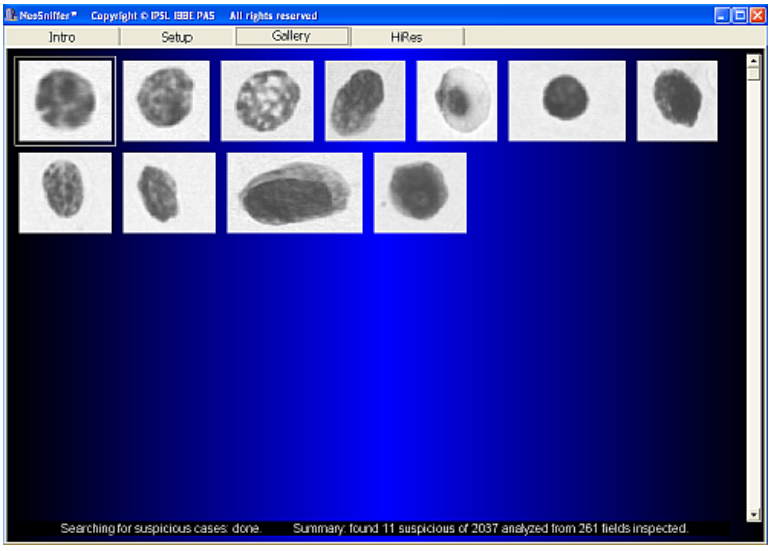


Fig. 8. Objects with malignancy associated symptoms in *NeoSniffer* output gallery.

Although the work on pathomorphological image database is rather preliminary it seems the way we choose is the right one. There are no fixed rules of classifying cytological objects. Everything depends on actual patient and therapeutic context. Pathologist takes into consideration many prerequisites and different sources of information. Physician points interesting objects in our database and the system finds them in a sample and finally they are presented on a computer screen.

BIBLIOGRAPHY

[1] BAERT A.L., Encyclopedia of Diagnostic Imaging, Berlin: Springer-Verlag, 2008.
[2] CHANG SK, HSU A., Image information systems: where do we go from here?, IEEE Trans. On Knowledge and Data Engineering, 1992, Vol.5, No.5, pp. 43-442.
[3] DULEWICZ A., PIĘTKA D., JASZCZAK P., NECHAY A., SAWICKI W., PYKAŁO R., KOŹMIŃSKA E. BORKOWSKI A., Computer Identification of Neoplastic Urothelial Nuclei from the Bladder, Analiytical and Quantitative Cytology And Histology, 2001, Vol. 23, No. 5.
[4] DULEWICZ A., PIĘTKA B.D., JASZCZAK P., A Trial of Practical Computer Analysis of Urothelial Nuclei for Cancer Detection, Horizons in Cancer Research, Vol.6: Progress in Bladder Cancer Research, MALLOROY A.M. (Editor), Nova Biomedical Books, New York, 2005.
[5] JASZCZAK P., DULEWICZ A., NECHAY A., PIĘTKA D., Selected algorithms for computer analysis of microscope scans in the system for early detection of urinary bladder cancer, 6th European Conference on Engineering and Medicine – ESEM-2001, Belfast, 2001, pp. 214-217.