BIG MEDICAL DATA: IMAGE MINING, RETRIEVAL, AND ANALYTICS

V.A. Kovalev

Заведующий лабораторией анализа биомедицинских изображений ОИПИ НАН Беларуси, кандидат технических наук, Республика Беларусь

A.A. Kalinovsky

Научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси, Республика Беларусь

Biomedical Image Analysis Department, United Institute of Informatics Problems National Academy of Sciences of Belarus, vassili.kovalev@gmail.com

This paper is devoted to the key issues associated with handling of the content of very large databases of natively digital medical images. A particular attention is drawn to the problem of examining image content in order to generate new knowledge, which is lately referred to as the image mining problem. Other important questions discussed in the paper are the content-based medical image retrieval and image analytics that are aimed at automation of recent patient diagnosis and treatment technologies. It is also argued that in many occasions such tasks of big medical image data analysis can be solved using co-occurrence image descriptors of different sorts.

Introduction: Medical Imaging coming into the Big Data camp

Image data are substantially different from the conventional data and documents. The main difference is that images consist of a large number of data elements (pixels, voxels) which typically play no role being considered separately. Instead, they collectively shape up certain spatial configurations in 2D or 3D what in turn represent some meaningful patterns or image objects. Consequently, images are typically stored as solid, unstructured arrays. Suitable quantitative features describing the image content should be derived from these data arrays prior to any comparison, statistical analysis, categorization, and other kinds of high-level manipulations [1].

In the last decade, medical imaging domain has demonstrated very strong and outrunning growth. It is observed in both the huge amount of medical images accumulated in the leading healthcare centers worldwide and advances on their acquisition, store, content-based retrieval, and utilization of machine learning algorithms [2]. All these achievements have promoted modern medical imaging research and practice onto the front line of Big Data problem.

Medical image databases. The image data involved in present studies were sub sampled from 3 different databases. Their content is briefly described below and illustrated in Fig. 1.

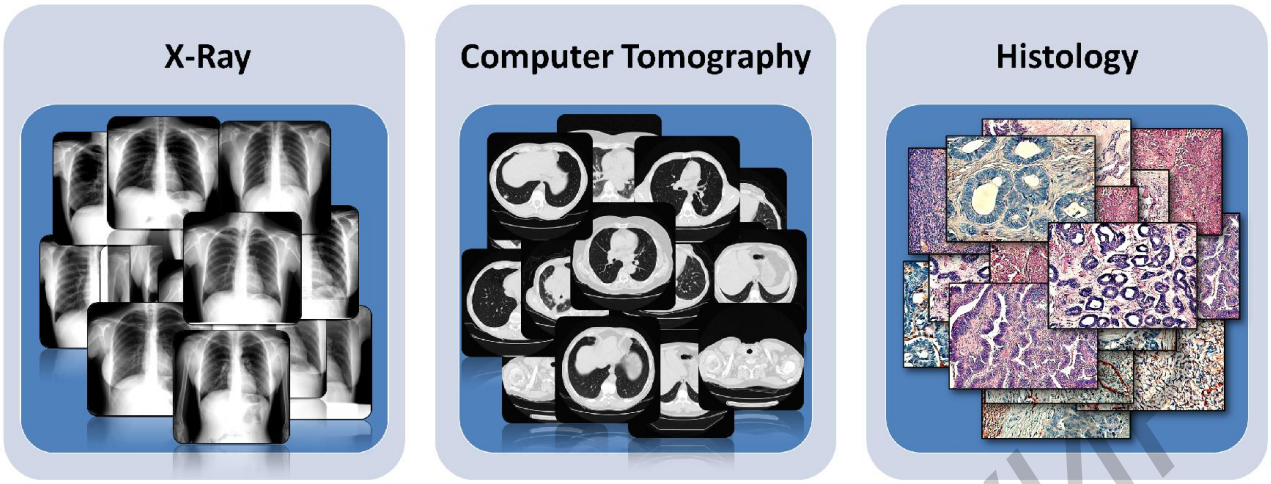


Fig. 1. Medical image databases used in presented studies

DB-1. An image archive containing about 1.5 million of natively-digital x-ray chest images taken from 700 000 people resulted from screening of lung diseases. Tens of thousands of these images are accompanied by descriptions given by professional radiologists. Preliminary diagnosis is often available as well. By our best knowledge, this is a world-largest image collection of its sort.

DB-2. A sample of 9000 3D computer tomography (CT) images consisting of about 1 400 000 axial image slices, i.e. 2D images of 512x512 pixels in size acquired from 5 000 of patients suffering from lung tuberculosis.

DB-3. A collection of about 100 000 color histology images obtained with the help of recent digital optical microscopes and optical microscopy scanners of Leica family using different antibodies and colorizing techniques. These images represent tissue samples of 120 cancer patients.

Lung shape mining

The purpose of this study was to introduce and examine an approach for mining 2D projective shape of human lungs from very large x-ray image archives and to discover some new and interesting gender- and age-related regularities in the lung shape and size.

Lung image segmentation. In order to extract lungs' shape, we have developed and implemented a novel multi-step lung segmentation procedure. The procedure has been mostly capitalized on the following three basic steps: detecting a bounding box covering the lungs area, fitting scalable mask of lungs and employing two bunches of rays drawn from the lung centers for detecting lung borders. In addition, a registration-based lung segmentation procedure was also developed which capitalized on selection of most similar image pre-segmented by an expert and applying non-rigid body deformation to the study image for fitting it to the target one in order to get the

lung region segmented. Results of lung segmentation are illustrated in Fig. 2 using a small sample of male and female subjects.

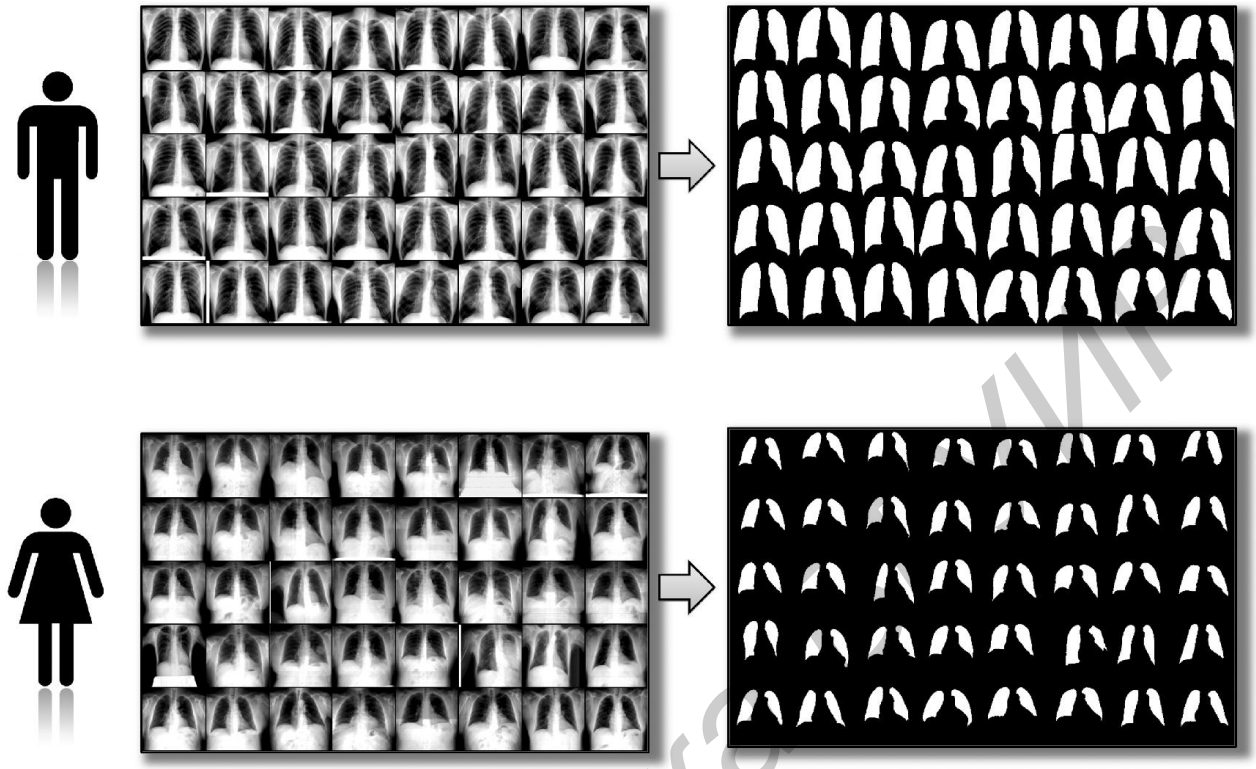


Fig. 2. Examples of original chest x-ray images (left panels) and their lung components extracted by the segmentation procedure (binary lung masks on the right) for male and female subjects

Creating study groups. The first study group of images called G1 was formed out for mining lung shape distinctions associated with age in different age categories or age "classes". It consisted of three sub-groups conditionally named as young (20-30 years), mid-aged (40-50 years) and aged (60-70 years) healthy subjects. Each sub-group included images of 6930 subjects (3465 pairs of male-female subjects with same age, 315 males and 315 females per age year), total 20790 images in the group G1. The second sub-group G2 has been created for mining both age- and gender-related lung shape regularities. It was covering the wide life span between the 20 and 80 years for both genders. This age range corresponds to 60 age intervals from 20 to 79 complete years each. A total of 9000 male-female pairs were collected from the image repository, 150 pairs for each age year. Thus, the group G2 consisted of 18000 x-ray images of the chest of 18000 different subjects aged 20 to 79 years, 300 images per age year (150 males plus 150 females), 9000 males and 9000 females in total. Finally, an auxiliary group G3 was created explicitly from female subjects aged 20 to 57 years, 1016 persons per age year, 38608 females in total.

Integral shape features. A number of commonly recognized shape features were calculated for the left and right lungs of every subject. They include lung area, di-

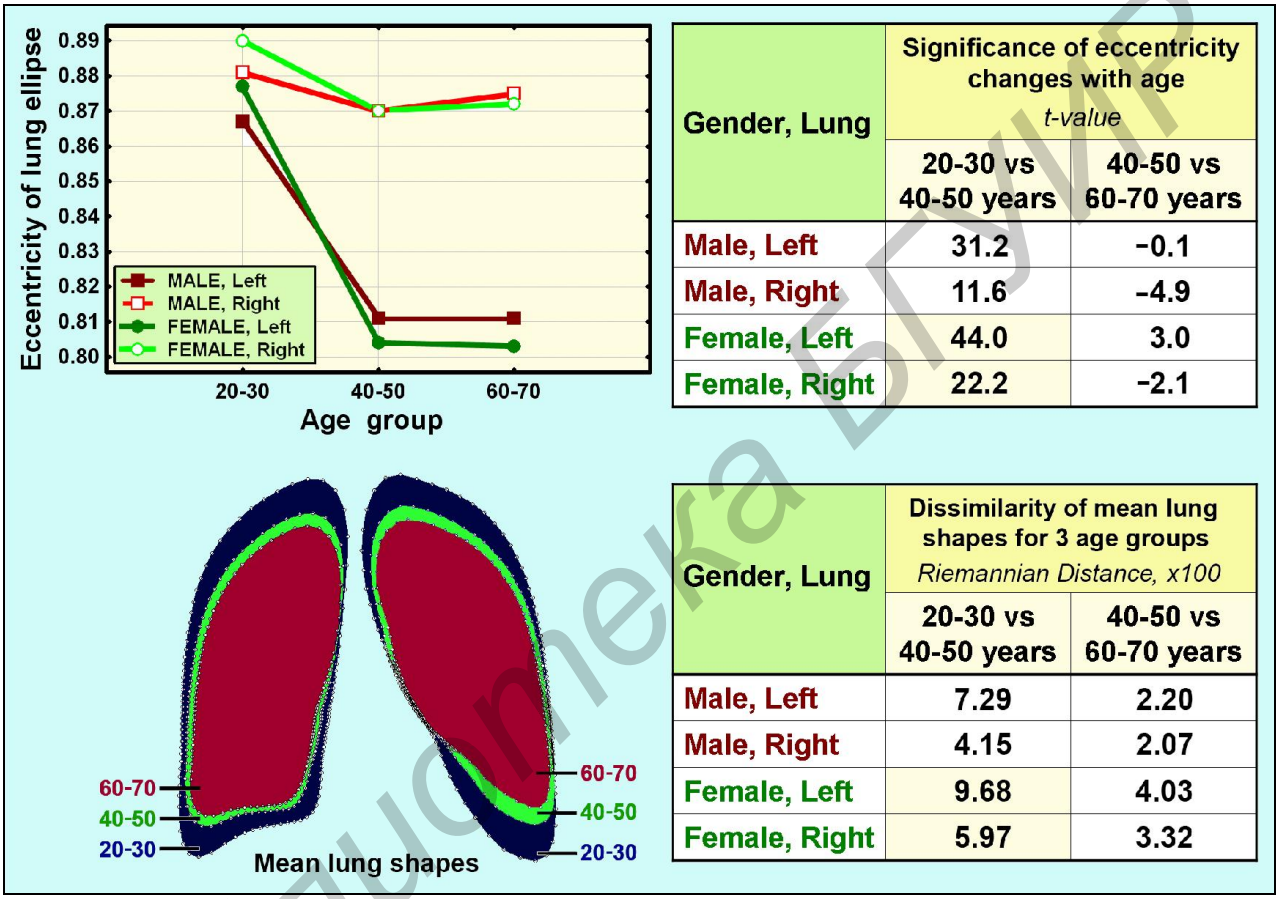
mensions of boxing rectangle, boundary length, compactness defined in its usual way i.e., as a squared boundary length divided by the area as well as the size of major axis, minor axis and the eccentricity of the ellipse with equivalent area. The ellipses were fitted to the lung contours using general linear model and the statistical confidence ellipse tools. The ellipse eccentricity feature was computed based on major and minor half-axes. In addition, the lung contour itself being represented by the vector of lengths of 500 rays ordered counter-clockwise naturally served as a polar signature of lung shape allowing computing such standard features as statistical moments.

Shape analysis methods. Statistical shape analysis is a geometrical analysis from a set of shapes in which statistics are measured to describe geometrical properties from similar shapes or different groups, for instance, the difference between normal and pathological bone shapes, etc. The statistical shape analysis involves methods for the geometrical study of objects where location, rotation and scale information can be removed. The key tasks of shape analysis are to obtain a measure of distance between two shapes, to estimate average shapes from a sample and to estimate shape variability in a sample [3]. In this work we have used 2D version of procrustes shape analysis and implemented in form of the `shapes` software package in framework of R, a language and environment for statistical computing.

The procrustes analysis consider objects made up from a finite number of points in N dimensions which are called landmark points. The shape of an object is considered as a member of an equivalence class formed by removing the translational, rotational and scaling components. Specifically, the following basic algorithms of procrustes analysis were used: calculating Riemannian distance between two shapes, Bookstein's baseline shape registration, testing for mean shape differences of groups of lung shapes with the help of Hotelling's T^2 and Goodall's F tests. These tests were developed for examining differences in mean shape between the two independent populations and involve complex eigenanalysis and iterative generalized procrustes analysis for two dimensions. In addition, when studying the age-related lung shape changes, a regression model with broken-line relationships suggested by Muggeo was used. The method is aimed to estimate linear and generalized linear models having one or more segmented relationships in the linear predictor.

Results: Shape of lung ellipses in different periods of life. Statistical assessment of differences between ellipses fitted to the lungs of subjects belonging to different age groups has revealed a bit more complicated pattern of age-related changes compared to the lung areas. Although the size of major and minor ellipse axes generally behave in a way similar to the lung area, i.e., decreases with age, the reduction rate varied significantly reflecting corresponding variations in global shape of lungs.

Since the eccentricity feature captures mutual relationships between the two axes and describes the global elongated shape of lungs in relative units, it is worth to consider here the eccentricity instead of raw axes values. As it can be easily seen from Fig. 3, the eccentricity exhibits the non-linear character of age-related changes even more sharply than the lung area. It is especially true for the left lung, the eccentricity of which drops down dramatically from young (20-30) to mid-aged (40-50) periods of life and remains nearly unchanged over the second gap from 40-50 to 60-70 years.



Gender, Lung	Significance of eccentricity changes with age *t-value*	
	20-30 vs 40-50 years	40-50 vs 60-70 years
Male, Left	31.2	-0.1
Male, Right	11.6	-4.9
Female, Left	44.0	3.0
Female, Right	22.2	-2.1

Gender, Lung	Dissimilarity of mean lung shapes for 3 age groups *Riemannian Distance, x100*	
	20-30 vs 40-50 years	40-50 vs 60-70 years
Male, Left	7.29	2.20
Male, Right	4.15	2.07
Female, Left	9.68	4.03
Female, Right	5.97	3.32

Fig. 3. The significance of general lung shape differences with respect to age groups as measured by features of fitted ellipses (top two panels) and the magnitude of shape changes measured using Riemannian distance between shapes (bottom two panels). It can be seen that most changes happened between 20-30 and 40-50 years and these changes are more significant in female subjects comparing to males

Similar trend can be observed for the right lung too but with a considerably lower confidence. In fact, the mean eccentricity values even slightly growing up after 40-50 years but the growth rate is close to the border of statistical significance (at this point it is good to remember that degree of freedom here is as high as df=6928 and the commonly-accepted minimal threshold for statistical significance is p<0.05 what approximately corresponds to $abs(t) > 2.0$. The significance rates supplied with the table depicted on the top-right quarter of Fig. 3 as well as pictures of mean group shapes accompanied by their dissimilarity values (see bottom two panels of Fig. 3)

provide further quantitative evidences for the discovered regularity. In everyday words, all these numbers testify for a conclusion that during the ageing the lung shape tends to "round up" and this process is mostly accomplished by the age of about 50 years. Such a behavior is more prominent in the shape of left lung.

It should be noted that due to the simplicity of the analysis of changes in lung size associated with both normal ageing and gender-related differences, corresponding part of the research is omitted. However, the results of such analysis are included in the concluding section given below.

Conclusions. As a result of this study, the following conclusion regarding the age-related lung changes and differences associated with gender can be drawn.

(a) The suggested image mining approach allows managing large collections of x-ray data, reliably extracting projective lung shape, and running 2D shape mining procedures for discovering new regularities from large image databases.

(b) It was found that the lung projective area declines with age in a non-linear way. The significance scores of lung reduction from moderate 40-50 to elderly 60-70 years were nearly twice as low as from young 20-30 to mature 40-50 periods of life. The temporal pattern of lung size reduction in females can be roughly described as "plateau--slope--plateau". The accelerated decline starts around 33-35 years and lasted till 48-50 where the process begins to slow down.

(c) The procrustes analysis suggest that similar to the size, the largest portion of lung shape changes occurs from young (20-30) to mid-aged (40-50) period and the magnitude of these changes in female subjects is always greater than in males. During the ageing, the lung shape tends to "round up". This process is mostly accomplished by the age of 50 years. Such a behavior is more prominent in shape of the left lung.

Content-based retrieval of CT images

Content-based image retrieval that is searching images similar to given query example is known as a promising technology for retrieving images from large non-annotated image archives for about fifteen years. However, despite certain achievements in general domains such as retrieving holiday photographs, the problem of computer-assisted searching for similar images and cases in medical image archives remains largely unsolved. The purpose of this section is to present current research results on application of content-based image retrieval techniques for assisting in managing large collections of CT images of tuberculosis patients. Fig. 4 explains the image hierarchy of the CT archive employed in the content-based similarity retrieval.

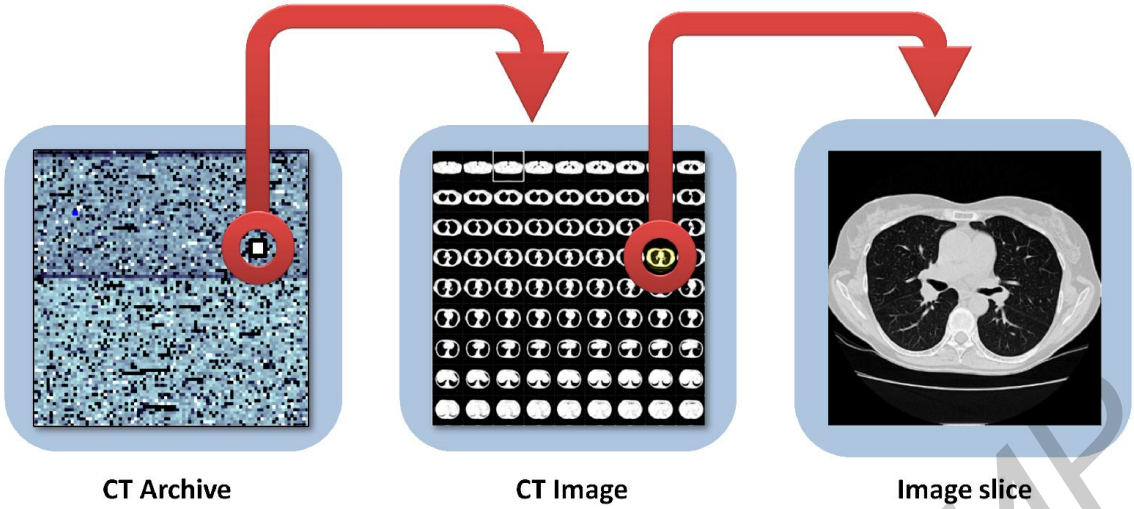


Fig. 4. Computed tomography (CT) image hierarchy of the image archive containing 1 400 000 2D images which is used in content-based image retrieval task. See the web site [6] for examining the quality of similarity retrieval

Regardless of a great diversity of diagnosis assistance tasks and image modalities used for medical diagnosis and treatment purposes, we have employed a uniform approach for representing the image content. The approach is based on the idea of extended multi-dimensional multi-sort co-occurrence matrices suggested in [4] for 2D images and for describing volumetric structure of 3D (CT, MRI, SPECT, PET and other kinds of computer tomography) volumetric images, which are introduced in [5] later on. Both image pixel pairs and pixel triplets, i.e., equilateral triangles with certain gray levels/colors located at their vertices are 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) they describe in statistical manner the spatial pixel relationships.

Thus, the image retrieval task was accomplished by way of comparing image descriptor that is a vector of selected elements of conventional histograms or co-occurrence matrices of given query example with descriptors of every image stored in the database. Such a comparison is typically done by calculation of Manhattan (L1) or Euclidean (L2) distance metric and selecting from database the top-*N* most similar cases, which are treated as image searching results.

The retrieval performance of various image descriptors was examined on the large database containing 1 400 000 images described above. The current image retrieval version is available for experimentation and testing on-line from an experimental web-site [6]. Introductory explanations on how to use the CT image retrieval engine are provided in Fig. 5.

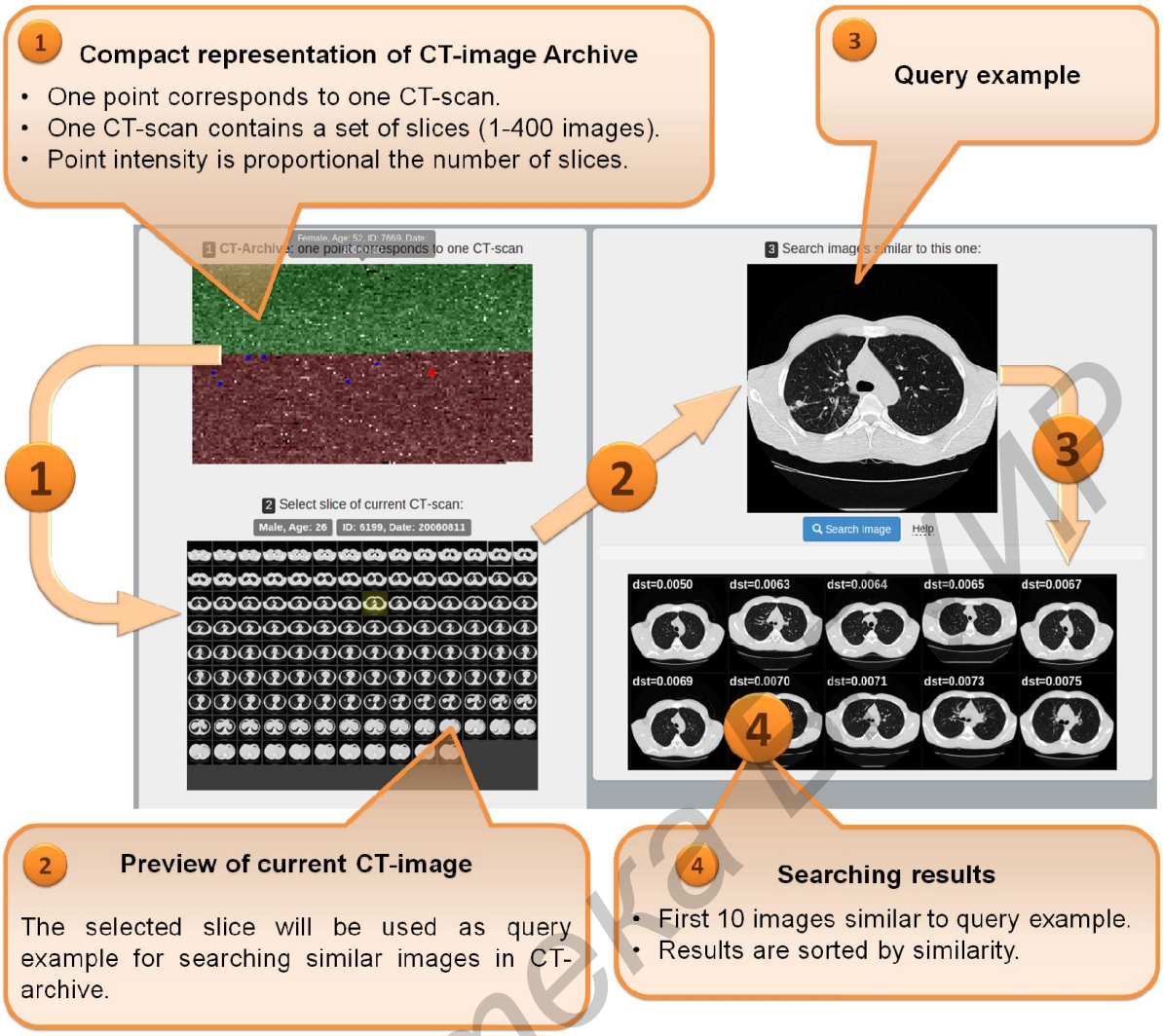


Fig. 5. A schematic representation of help-screen of the on-line CT image retrieval engine

The readers are kindly asked to test the image searching engine and to provide corresponding feedback both on the visual quality of retrieving similar CT slices and the searching time which is necessary to compare descriptor of selected query example with the rest 1 400 000 items using *L1* metric.

Image analytics in ovarian cancer research

The problem of angiogenesis in malignant tumors. Over the last several years numerous factors involved in the development and progression of solid cancers have been identified. Among these, tumor angiogenesis is certainly one the most important owing to the fact that, in order to grow and to develop, a tumor needs to be supplied with a vascular system following its growth. It is anticipated that angiogenesis research will probably change the face of medicine in the next decades [7], with more than 500 million people worldwide predicted to benefit from pro- or anti-angiogenesis treatments. Towards this end, a method is suggested for identification and visualization of histology image structures relevant to the key characteristics of the state of cancer patients. The method is based on a multi-step procedure which in-

cludes calculating image descriptors, extracting their principal components, correlating them to known object properties and mapping disclosed regularities all the way back up to the corresponding image structures they found to be linked with.

The technical problem of searching links between pathological image structure and variables, characterizing the state of patient. In a typical setup there is a patient database available which contains both image data of different modalities as well as non-visual patient characteristics such as general social data, clinical observations, results of laboratory tests, history of personal and family diseases, etc. Then technically the problem is posed as finding statistically significant associations between the morphological image structures presented in form of suitable quantitative features and database variables containing the patient records. Such correlations can be found in a straightforward manner using, for example, conventional approach of feature extraction followed by a multivariate statistical analysis for identifying significant links between these two. However, this is only possible with *a priori* research hypothesis in hands which presumes certain connections between the specific, pre-defined image structures and some patient characteristics. Being developed, implemented, and successfully applied to the input data, this approach leaves researcher with only particular results and image structures that have been extracted and examined. For instance, our preliminary study exploiting this approach was attempting to attribute tumor vessel development visualized with the help of D2-40 marker to patients' conditions. To this end, the vessel network was segmented, characterized by five quantitative features, and correlated to the patient state. However, despite certain time and other resources were spent, it gave very particular and rather modest results.

Thus, in this context it is worth to consider an alternative, exploratory approach which is aimed at detecting the whole bunch of objectively existing correlations between the histology image structures and patient state first and separate investigating their novelty together with the underlying biomedical substrate afterwards. Such an approach may conditionally be categorized into the image mining research area. In much the same way as data mining, the image mining can be understood as the process of extracting hidden patterns from images. More specifically, image mining deals with extraction of implicit knowledge, image data relationship or other patterns not explicitly stored in the image. Given that the histological image analysis is the task that difficult to automate due to its structural sophistication, it appears promising to examine the wide-cut image mining techniques for discovering the links we are interested in.

Materials and Methods. The image descriptors employed are extended 4D color co-occurrence matrices counting the occurrence of all possible pixel triplets located at the vertices of equilateral triangles of different size. The method is demonstrated on a

sub-sample of 952 color histology images of 2048x1536 pixels in size which were acquired using digital Leica DMD108 microscope. They included 272 routine hematoxylin-eosin stained diagnostic images (4 images for each patient) and 680 images of tissue probes (10 per patient) immuno-histochemically processed with D2-40 marker highlighting lymphogenesis. In addition, an auxiliary test database containing 4000 color images (35 Gigabytes) was created which equally represents the following 4 classes: ovary tumor, ovary non-tumor, thyroid tumor, thyroid non-tumor (Fig. 6).

Once the co-occurrence matrices are calculated, the very common strategy is to calculate Haralick's features next and to use them for image characterization, clustering, etc. However, this traditional procedure may not be followed here at least because Haralick's features cannot be mapped back to the original images as the second introductory condition requires. On the contrary, the matrix elements themselves may be mapped back [8] but there are too many of them to satisfy the well known statistical condition of limited number of variables which avoids pseudo-correlations. The solution is to apply the Principal Component Analysis (PCA) method for extracting a limited number of uncorrelated features from matrices. Thus, the method supposes calculating 4D co-occurrence matrices, extracting principal components, correlating them to patients' state, selecting significant ones, projecting selected components back to co-occurrence matrix elements, and finally using them for visualizing the image structures we are looking for.

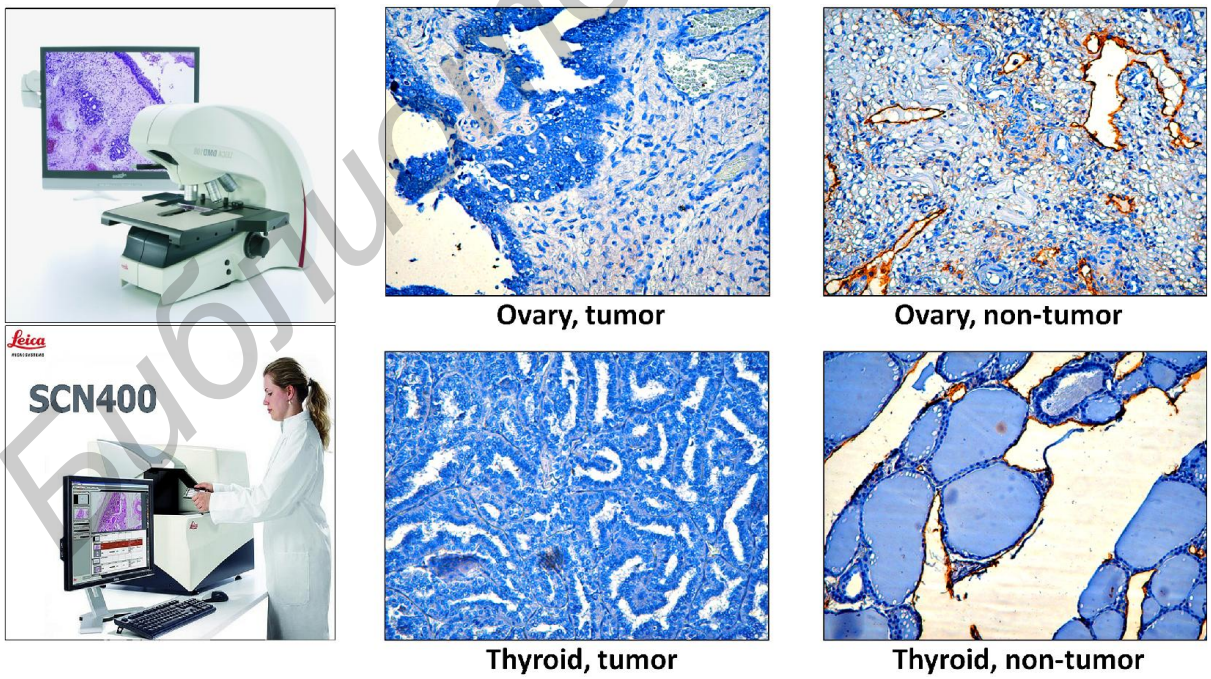


Fig. 6. Examples of histological images representing the four classes used for analytic algorithms of discovering links between the pathological image morphology and clinical data describing the state of cancer patients.

Note that since principal components are uncorrelated, there is no need to apply complicated and somewhat risky multivariate statistical analysis methods. Searching for significant links can be done by straightforward univariate correlations or with the help of Student's *t*-test according to the feature type.

Along with the statistical data analysis methods described above, the classification accuracy of recognition of 4 classes of textural pathological images was assessed analytically using the Support Vector Machines (SVM) and Random Forests classifiers.

Results. Original RGB images were converted into the *Lab* space with Euclidean color dissimilarity metrics and the number of colors was reduced down to 24 bins using the median cut algorithm preserving most important colors. Thus, the 3D color co-occurrence sub-matrices CCC with a fixed inter-pixel distance *d* contain 24^3=13824 cells. Given that elements above leading diagonal are zeros, the number of effective matrix elements was N_E=2600. Equilateral triangles with side lengths D={1,3,5} were considered so that the total number of elements of completed CCCD matrices was 7800. Cumulative CCCD matrices computed over all the images of each patient were vectorized constituting an input PCA data table with 68 rows and 7800 columns. PCA resulted in extracting 27 principal components (PCs) in case of matrices of routine images and 38 PCs in case of D2-40 images under condition of covering 95.0% of variance. The first components cover 55.7% and 26.5% of variances respectively. These results suggest that structural variability of D2-40 images is substantially higher compared to routine ones. Being correlated with patients' data, 27 PCs of routine images have produced a total of 43 events of correlation significant at $p<0.01$. Same procedure being applied to 38 PCs derived from descriptors of D2-40 images with highlighted lymphatic vessels resulted in detecting 47 significant links between these features and patient state records.

Detailed investigation of significant correlation has revealed that some of them were easily deductible from existing knowledge whereas other are suggestive for novelty and certainly interesting from both scientific and practical points of view. For instance, in case of routine images the significant links between PCs and the following patient data appears to be very promising: development of distant metastases ($p<0.001$), the degree of cancer tissue differentiation ($p<0.007$), the number of miscarriages ($p<0.0001$), and the number of chemotherapy trials ($p<0.000002$, r=-0.543). The negative correlation of the length of borders highlighted in the figure with the number of trials may be explained by the fact that more spacious tumor structure is typical for relatively "young" tumors which are chemically treated first compared to "old" ones which removed immediately. Images of tissue processed by D2-40 endo-

thelial marker have demonstrated similar behavior disclosing a number of interesting links.

Besides the results obtained with the mutual analysis of image features and variables of clinical database characterizing the patients, there was developed an electronic atlas of histological images of normal and cancerous tissue (Fig. 7).

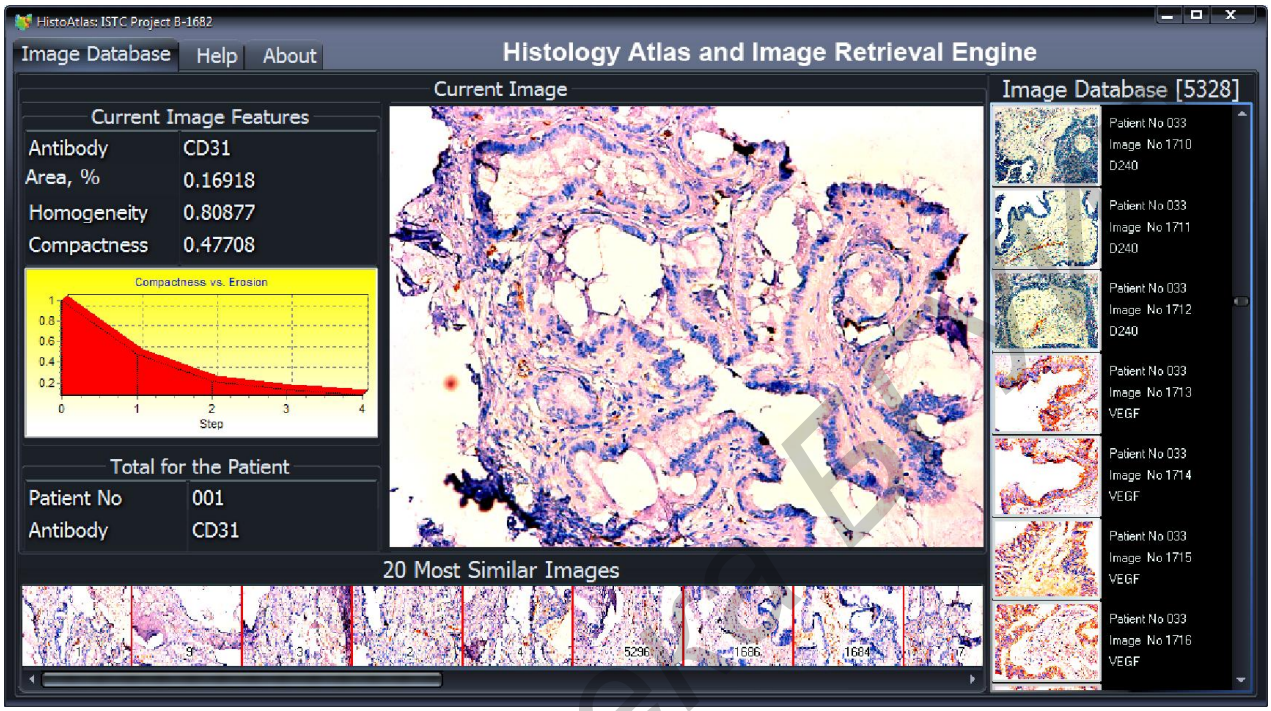


Fig. 7. The main screen of the histology atlas with imbedded content-based image retrieval functionality

The atlas includes tissue images stained using different techniques which present conventional pathological features as well as processes of angiogenesis and lymphogenesis in tumors. On the top, the atlas also incorporates certain content-based image retrieval functionality which allows for a quick image database search for the most similar cases. This option is implemented using color co-occurrence image descriptors that characterize color textures quantitatively.

Finally, the abilities of IIID co-occurrence matrices computed using grayscale version of the images were also assessed. Despite some promising correlations were found, an ambiguity was revealed. In particular, when certain IIID matrix element was mapped back to the grayscale images, it highlights structures of biologically different sorts. This is because two or more substantially different image colors were converted down to one single gray level.

Assessing the general inter–class differences. A preliminary statistical assessment of inter-class differences was performed by way of applying unpaired Student's $t-$test separately to every variable (matrix element). In case of color descriptors, N = 456 out of 500 variables were found to be significantly different at $p<0.05$ when

comparing OvaryNorm–ThyroidNorm classes and N=447 variables for the pair of tumor classes OvaryTumor–ThyroidTumor. It was revealed that normal tissue of the two organs is more dissimilar (mean t=13.4, STD =6.8) than their respective cancerous regions (mean t=9.94,STD=7.3). The use of grayscale descriptors leads to the essentially same but notably more distinct results: N=559, mean t=30.5, STD = 19.2 for the pair of normal tissue classes OvaryNorm–ThyroidNorm and again much lower mean t = 11.3, STD = 7.7 for tumor regions. Some more sophisticated multivariate statistical methods were also tried but they provide results well compared with the above straightforward $t-$tests.

Classification quality. A number of experiments on pair-wise classification of normal and pathological tissue images were performed using different kinds of input features and different classifiers (SVM with 10-fold cross-validation and Random Forests) in order to get a clear picture of general inter-class differences. The image recognition results are summarized in the Tab. 1 below.

Tab. 1. Recognition accuracy of histology image classes, %

Input image features	**Image classes**	**Classifier**	
		SVM	**Random Forests**
Original co-occurrence features (~500 variables)	Ovary Norm vs. Thyroid Norm	97.6	96.4
	Ovary Tumor vs. Thyroid Tumor	93.9	93.3
Principal Components (26 variables)	Ovary Norm vs. Thyroid Norm	96.2	94.0
	Ovary Tumor vs. Thyroid Tumor	94.2	94.2

As it can be easily seen from the table, all the classification results are reasonably uniform with non-tumor tissue classes (conditionally called here as “Norm”) separated better than pathological ones. This is in agreement with the above statistical tests.

Conclusions. As a result of computational experiments, a number of associations between the patients’ conditions and morphological image structures were discovered including both easily explainable and the ones whose biological substrate remains obscured. It was also found that non-tumor tissue regions of the ovary and thyroid gland are more dissimilar than respective tumor regions. This is in line with a more general biomedical regularity suggesting that being affected by a strong pathology, tissues and organs tend to become more similar to each other.

Thus, the suggested method may be considered as a promising tool capable of an automatic identification and visualization of histological image structures relevant to the cancer patient conditions. However, since there is no intrinsic mechanism for semantic assessing the resultant links detected by the method, an expert-based evaluation of the novelty and biological substrate of the result is necessary.

Acknowledgements. This work was partly funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project OISE-14-60497-1.

References

1. Nixon M.S., Aguado A.S. Feature Extraction and Image Processing for Computer Vision, Third Edition, Academic Press, ISBN-13 978-0123965493, 2012. – 632 p.

2. Gao X., Müller H., Deserno T.M. Integration of Medical Images into the Digital Hospital, *The Open Medical Informatics J,* vol. 2011, No 5, pp. 17–18.

3. Dryden I.L. and Mardia K.V. Statistical Shape Analysis. John Wiley & Sons, New York, Sep 1998, ISBN 0-471-95816-6, 376 p.

4. Kovalev V. and Petrou M. Multidimensional Co-occurrence Matrices for Object Recognition and Matching, *Graphical Models and Image Processing,* vol. 58, No. 3, May, pp. 187-197, 1996.

5. Kovalev V.A., Kruggel F., Gertz H.-J., and von Cramon D.Y. Three-dimensional Texture Analysis of MRI Brain Datasets, *IEEE Transactions on Medical Imaging,* vol. 20, No. 5, May, pp. 424-433, 2001.

6. http://imlab.grid.by/ last visited 20 April 2015.

7. Carmeliet P. Angiogenesis in life, disease and medicine, *Nature,* vol. 438, 2005, pp. 932-936.

8. Kovalev V.A., Petrou M., and Suckling J. Detection of structural differences between the brains of schizophrenic patients and controls, *Psychiatry Research: Neuroimaging,* vol. 124, pp. 177-189, 2003.