Exploring access to scientific literature using content-based image retrieval

Thomas M. Deserno[a,b,1], Sameer Antani[b], and Rodney Long[b]

[a] Department of Medical Informatics,
Aachen University of Technology (RWTH), 52057 Aachen, Germany
[b] U. S. National Library of Medicine, U. S. National Institutes of Health,
8600 Rockville Pike, Bethesda, MD 20894, USA

ABSTRACT

The number of articles published in the scientific medical literature is continuously increasing, and Web access to the journals is becoming common. Databases such as SPIE Digital Library, IEEE Xplore, indices such as PubMed, and search engines such as Google provide the user with sophisticated full-text search capabilities. However, information in images and graphs within these articles is entirely disregarded. In this paper, we quantify the potential impact of using content-based image retrieval (CBIR) to access this non-text data. Based on the Journal Citations Report (JCR), the journal *Radiology* was selected for this study. In 2005, 734 articles were published electronically in this journal. This included 2,587 figures, which yields a rate of 3.52 figures per article. Furthermore, 56.4% of these figures are composed of several individual panels, i.e. the figure combines different images and/or graphs.

According to the Image Cross-Language Evaluation Forum (ImageCLEF), the error rate of automatic identification of medical images is about 15%. Therefore, it is expected that, by applying ImageCLEF-like techniques, already 95.5% of articles could be retrieved by means of CBIR. The challenge for CBIR in scientific literature, however, is the use of local texture properties to analyze individual image panels in composite illustrations. Using local features for content-based image representation, 8.81 images per article are available, and the predicted correctness rate may increase to 98.3%. From this study, we conclude that CBIR may have a high impact in medical literature research and suggest that additional research in this area is warranted.

Keywords: Content-Based Image Retrieval (CBIR), Scientific Literature, Information System Integration, Radiology, Data Mining, Information Retrieval

1. INTRODUCTION

Content-based image retrieval (CBIR) has long been identified as a technology with the potential for significant impact for management of and retrieval from a large collection of images [1, 2]. As a basic principle of CBIR, images are internally represented by numerical features, which are extracted directly from the image pixels. These features are stored in the database, as a signature, along with the images, and are indexed for rapid access. At retrieval time, the query-by-example (QBE) paradigm is usually applied [3]. Here, the user presents a sample image or pattern, and the system computes the numerical features, compares them to those stored in the database, and returns all images with similar features. It is obvious that the quality of response depends on (1) the features representing the image and (2) the distance or similarity measure that is used to compare features from different images. Regarding the features, different approaches are used [4]:

- *Global* image features are defined as those that are computed on the entire image, e.g., histogram representation of the image. As such only one signature is related to each image. Using global features, the

---

[1] Corresponding author: Priv.-Doz. Dr. Thomas M. Deserno (nob: Lehmann), Assoc. Prof., Department of Medical Informatics, Aachen University of Technology (RWTH), Pauwelsstr. 30, D - 52057 Aachen, Germany, email: deserno@ieee.org; web: http://irma-project.org/deserno, phone: +49 241 80 88793, fax: +49 241 80 33 88793.

semantic gap between the low-level feature extraction by machine and the high-level scene interpretation by humans tends to be wide. However, global features have been successfully applied for automatic image categorization according to the imaging modality, body region, viewing direction, and the biomedical system imaged [5, 6, 7].

- *Local* features are defined as those that are computed at prominent image regions, e.g. texture or shape features localized at a particular image region. This results in a number of signatures are related to each image increasing capability of CBIR techniques to focus on particular aspects of the image content. A similar assessment is made by Tagare et al. stating that the information contained in medical images is local [8], and hence, local features may narrow the semantic gap.

- *Relational* features have not yet been applied routinely, but the concept has been discussed in the literature [9, 10]. The idea is to capture the spatial and/or temporal relationships between the image regions of interest (ROI), such as distance, direction, and size relationships. Clearly, relational features are most similar to the scene interpretation by humans, and consequently, the semantic gap will be smallest.

The distance or similarity measure is usually specific to a particular feature. For instance, the Jensen-Shannon divergence [11] is used for histogram-based features, while the Mahalanobis or Euclidian distances are applied for vector-based signatures. However, the most important characteristic of a distance is whether it is a metric. According to Traina et al., a distance function $d(A,B)$, of features $A \neq B \neq C$, which is a metric, must satisfy [12]:

1. *reflexivity*, i.e., $d(A,A) = 0$,

2. *non-negativity*, i.e., $d(A,B) > 0$,

3. *symmetry*, i.e., $d(A,B) = d(B,A)$, and

4. *triangle inequality*, i.e., $d(A,B) + d(B,C) \geq d(A,C)$

An overview of features and distances is given by [13], and can be taken from the results of the ImageCLEF campaign [5, 6, 7].

Typical image collections studied in biomedical CBIR are collections in Picture Archiving and Communication Systems (PACS) and research studies [13, 14]. Applications of medical CBIR systems have been made in the fields of computer-aided diagnosis, evidence-based medicine, case-base reasoning, and medical training [1, 8, 9, 13, 14].

However, there are other fields in medicine which typically have large medical image archives. In particular, a huge amount of figures, graphs, images, and case examples is published in scientific literature, and the number of scientific journals that are published electronically is increasing explosively. The aim of this work is to evaluate and estimate the impact of state-of-the-art medical CBIR integrated with text-based searches for retrieval of scientific literature. That is, we investigate the use of bitmaps within the journal articles as additional information for retrieval. Results from this study will support development of techniques for CBIR of figures and image types specific to scientific literature.

2. METHODS

In this section, we describe the study design, journal selection, procedure of extraction and classification of the illustrations, the database and programs used for evaluation, and the methodology used.

2.1. Selection of Journal

We first selected a representative journal as a data source for studying the effect of CBIR on article retrieval. Using the impact factor that is published in the ISI Journal Citation Reports (http://scientific.thomson.com/products/jcr/) as an indicator of journal importance, we selected the best-ranked journal from 2005 in the category *Radiology,* since radiology is the medical discipline that produces a very high number of diagnostic images.

2.2. Extraction of Illustrations

Usually, electronic publication of journal articles makes use of the Portable Document Format (PDF), which is an open file format created and controlled by Adobe Systems Inc. (San Jose, CA, USA) for representing two-dimensional documents in a device and resolution-independent, fixed layout. Using Adobe Acrobat Professional 7.0, all PDF-

embedded bitmaps were automatically extracted as individual image files and stored in the lossless Portable Network Graphics (PNG) format.

2.3. Classification of Illustrations

The variety of illustrations in scientific literature is very large. Frequently, diagnostic images are combined in panels, annotated with text and drawings, and composed together with schematic graphs, diagrams or other types of illustrations. With respect to content-based image analysis, the number and kind of images, graphs, drawings, and photographs, the frequency of annotations, as well as the presence of color are important. We defined the following major classes of illustrations (Fig. 1):

- *diagnostic image*, i.e., an original image as obtained from any medical imaging modality (e.g., radiography, microscopy, endoscopy, sonography), that may be color or black and white and annotated.

- *diagnostic visualization*, i.e., a color or grayscale computed visualization of medical image data, such as a three-dimensional (3D) direct volume rendering of computed tomography (CT) or magnetic resonance imaging (MRI) data.

- *photograph*, i.e., any type of an optical static image, which, again, may be in color or grayscale and show devices, medical objects or situations, persons, or portraits.

- *screen shot*, i.e., any illustration showing a computer screen, window, or a part thereof.

- *graph*, i.e., any visualization of numerical data such as plots, curves, as well as block or pie charts.

- *diagram*, i.e. any kind of functional or block diagram, scheme, or mind map.

- *drawing*, i.e., any type of manual drawings.

- *multi-panel figure*, i.e., a composition of different parts, which may be composed of strictly medical, non-medical, or mixed panels, and may be presented in color or grayscale. If one of the panels is color, the entire illustration is labeled as color. Also, the number of panels is recorded.

2.4. Database and Reference Categorization

All illustrations were analyzed manually for the number and composition of diagnostic images included as figures. This evaluation was carried out using the Image Retrieval in Medical Applications (IRMA) framework (http://irma-project.org). In particular, the IRMA Web-based interfaces for reference categorization were used for computer-assisted coding of illustrations [9, 15].

The hierarchical multi-axial IRMA code [16] was extended to capture the characteristics of illustrations in scientific papers. All images within the IRMA system are related to an ABCD code which is composed of 4 labels, viz. the body region (A-natomy) and biomedical system (B-iosystem) imaged, the imaging modality (C-reation), and the view (D-irection). In order to classify published illustrations, the C-axis of the IRMA code was used. Figure 2 shows the resulting part of the IRMA code. Note that the third digit of the code always distinguishes color vs. grayscale, which allows easy summation over all categories. Also, the last digit for the multi-panel images denotes the number of panels.

2.5. Evaluation

Obviously, the number and kind of figures can be counted directly. However, IRMA classification of multi-panel illustrations is limited, since each panel can be from a different major category, but within the IRMA framework, each image is uniquely assigned to its appropriate IRMA code. Therefore, the percentages resulting for IRMA codes 91xx to 97xx were used to predict the corresponding numbers for the multi-panes (IRMA code 98xx to 9axx).

For ground truth to estimate the impact of CBIR-based literature research, we refer to the Cross-Language Evaluation Forum (CLEF) image campaign. In recent years, ImageCLEF (http://ir.shef.ac.uk/imageclef/) has served as a forum for determining the state of the art in annotating images. Since 2005, a competitive medical image retrieval task has been defined for CBIR researchers; this task is based on the IRMA reference image dataset. In a first approximation that is based on the count of illustrations, the error rates from CLEF are used to predict expected results of images extracted from the published articles.
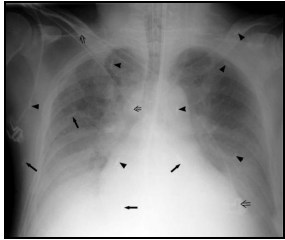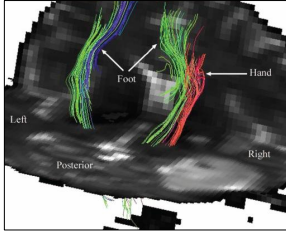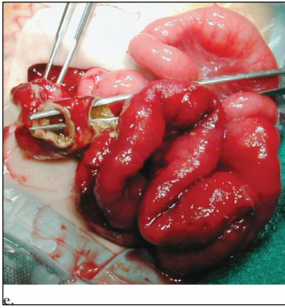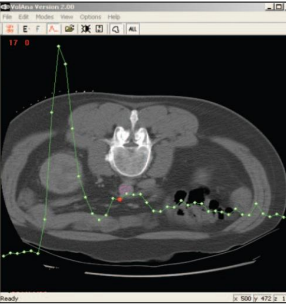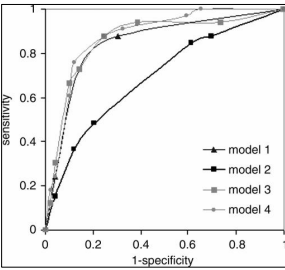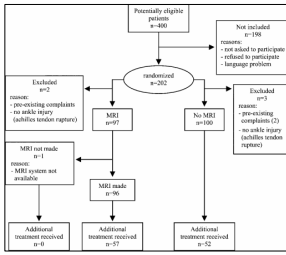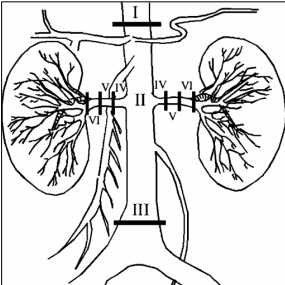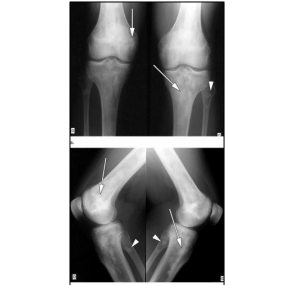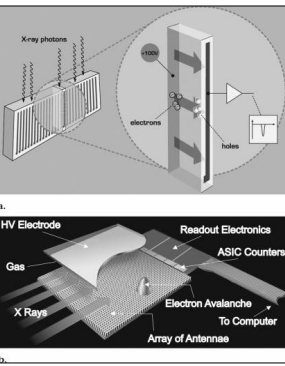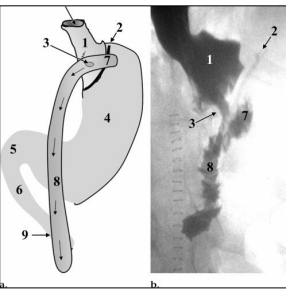
Example image	Image code	Example image	Image code
	9112 figure diagnostic image grayscale annotation		**9220** figure diagnostic visualization color unspecified
	9323 figure photograph color object		**9420** figure screen shot color unspecified
	9510 figure graph grayscale unspecified		**9610** figure diagram grayscale unspecified
	9710 figure illustration grayscale unspecified		**9814** figure multi-panel medical grayscale 4 parts
	9912 figure multi-panel non-medical grayscale 2 parts		**9a12** figure multi-panel mixed grayscale 2 parts

Figure 1: Example illustrations from all major categories. The codes refer to Figure 2.

```
[9] figure                                    [97] drawing
        [90] unspecified                               [970] unspecified
        [91] diagnostic image                          [971] grayscale
                [910] unspecified                              [9710] unspecified
                [911] grayscale                        [972] color
                        [9110] unspecified                     [9720] unspecified
                        [9111] original        [98] multi-panel medical
                        [9112] annotation              [980] unspecified
                [912] color                            [981] grayscale
                        [9120] unspecified                     [9810] unspecified
                        [9121] original                        [9811] 1 part
                        [9122] annotation                              ...
        [92] diagnostic visualization                          [982p] 25 parts
                [920] unspecified                      [982] color
                [921] grayscale                                [9820] unspecified
                        [9210] unspecified                     [9821] 1 part
                [922] color                                            ...
                        [9220] unspecified                     [982p] 25 parts
        [93] photograph                        [99] multi-panel non-medical
                [930] unspecified                      [990] unspecified
                [931] grayscale                        [991] grayscale
                        [9310] unspecified                     [9910] unspecified
                        [9311] device                          [9911] 1 part
                        [9312] person                                  ...
                        [9313] object                          [991p] 25 parts
                        [9314] portrait                [992] color
                        [9315] other                           [9920] unspecified
                [932] color                                    [9921] 1 part
                        [9320] unspecified                             ...
                        [9321] device                          [992p] 25 parts
                        [9322] person          [9a] multi-panel mixed
                        [9323] object                  [9a0] unspecified
                        [9324] portrait                [9a1] grayscale
                        [9325] other                           [9a10] unspecified
        [94] screen shot                               [9a11] 1 part
                [940] unspecified                              ...
                [941] grayscale                                [9a1p] 25 parts
                        [9410] unspecified             [9a2] color
                [942] color                                    [9a20] unspecified
                        [9420] unspecified                     [9a21] 1 part
        [95] graph                                             ...
                [950] unspecified                              [9a2p] 25 parts
                [951] grayscale                [9b] other
                        [9510] unspecified             [9b0] unspecified
                [952] color                            [9b1] grayscale
                        [9520] unspecified                     [9b11] scanned table
        [96] diagram                                           [9b12] sc. itemize
                [960] unspecified                              [9b13] sc. equation
                [961] grayscale                                [9b14] sc. document
                        [9610] unspecified                     [9b15] artefact
                [962] color                            [9b2] color
                        [9620] unspecified                     [9b21] scanned table
                                                               [9b22] sc. itemize
                                                               [9b23] sc. equation
                                                               [9b24] sc. document
                                                               [9b25] artefact
```

Figure 2: IRMA code extension for classification of illustrations. The major categories are displayed in red.

3. RESULTS

According to the JCR 2005 Science Edition [17] the journal *Radiology* (ISSN 0033-8419) has the highest impact of 5.377 in the category "Radiology". In total, 738 articles are listed on the Web (http://radiology.rsnajnls.org/), and 734

| IRMA code	Name	# Items	% Items	# Panels	% Panels
91**	diagnostic image	409	15.80	409	6.33
92**	diagnostic visualization	9	0.35	9	0.14
93**	photograph	80	3.09	80	1.24
94**	screen shot	2	0.08	2	0.03
94**	graph	470	18.17	470	7.72
96**	diagram	55	2.13	55	0.85
97**	drawing	40	1.55	40	0.62
98**	multi-panel medical	1,095	42.33	4,015	62.07
99**	multi-panel non-medical	272	10.51	787	12.17
9a**	multi-panel mixed	93	3.59	540	8.35
9b**	others	62	2.40	62	0.96
Sum		**2,587**	**100.00**	**6,469**	**100.00**
**1*	grayscale	2,273	87.86	5,287	81.73
**2*	color	314	12.14	1,182	18.27
Sum		**2,587**	**100.00**	**6,469**	**100.00**
91*1	original	69	16.87	1,091	16.87
91*2	annotated	340	83.13	5,378	83.13
Sum		**409**	**100.00**	**6,469**	**100.00**

Table 1: Results. Estimates are displayed in red.

PDF files are available[2]. Images, graphs, drawings, and figures are extracted from these. It was found that, frequently, medical images and/or graphs were composed of multiple images or image panels. In particular, 2587 bitmaps are available, and 418, 1095, 647, 272, and 93 display an individual medical image or visualization (# 91** + # 92**), a combination of medical images (# 98**), an individual graph, diagram, drawing, or photograph (# 93** + … + # 97**), a combination of several graphs (# 99**), and a combination of medical images and graphs within a single figure file (# 9a**), respectively (Tab. 1). In addition, 44 PNG files contain scanned documents, itemizes, equations or tables (# 9b*1 + … + # 9b*4), and another 18 (0.7%) contain artifacts (# 9b*5), i.e., a single line that is used to separate text blocks but do not represent an illustration.

As can be further deduced from Table 1, the majority of illustrations are still published in grayscale (87.86%). If multi-panel illustrations that contain at least one colored component were counted as if all components are colored, the number of grayscale panels is still above 80%. Similarly, the majority of illustrations are annotated with text, arrows, or other symbols which may cover image information and affect the textural feature extraction.

The results were correlated with the results of the ImageCLEF competition to predict the relevance of CBIR for literature access. In ImageCLEFmed 2005, leave-one-out experiments based on 10,000 radiographs of 51 categories were conducted. Applying state-of-the-art CBIR techniques, i.e., *global* texture features are used to represent the image content, error rates of about 12% were reported [5], while ImageCLEFmed 2006 with 116 categories for 11,000 radiographs yielded about 14% for an optimal classifier combination [7]. Using 15% as an estimate, and since, in average, 3.52 figures are contained in an article; the expected error rate for CBIR-assisted literature retrieval may decrease to approximately 4.5%. Using *local* features for content-based image representation, 8.81 images per article are available, and the predicted error rate may further decrease to 1.7%.

---

[2] Note that these numbers differ from JCR, where the number of "articles" is defined as the number of published items in the shown year that comprise the scholarly contribution of the journal. This number is also called "citable items" to indicate that these items in the journal are the ones most likely to be incorporated into the further research literature through citation. This number includes all research reports, reviews or mini-reviews, and scholarly and extensively referenced commentary. News, editorials, letters to the editor, and other materials, while they fulfill a vital function in the journal itself, are not considered "citable," and are in fact rarely cited. Therefore, from the 12 issues of *Radiology* in 2005, a total of 667 articles are included in the Web of Science, and 501 are considered as citable item.
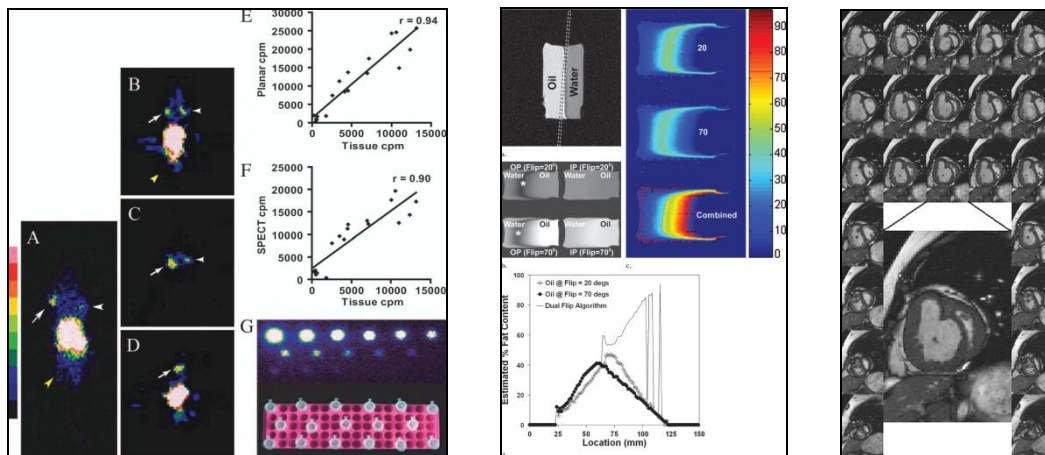
Figure 3: Examples for complex multi-panel illustrations.

4. DISCUSSION

Content-based image retrieval has not yet been suggested to support retrieval of scientific literature. This idea is novel and to the best of our knowledge, an application does not exist yet. Based on data collected from the 2005 volumes of *Radiology,* we estimated the impact of medical CBIR quantitatively, using ImageCLEFmed campaign as ground truth.

The caveat in the estimate we present is that, further work is required to take into account the composite image panels, which need automatic decomposition into individual, but related, images. Figure 3 shows that multi-panel illustrations may have complex composition. In particular, heuristic assumptions such as "panels are separated by horizontal or vertical lines", or that "all panels are of the same size and aspect ratio", are demonstrably false. Therefore, one challenge for CBIR in scientific literature is use of *local* image characteristics to separate and analyze individual image panels in composite images. Another challenge is seen in the annotations which are frequently overlaid on the images. In order to cope with these annotations, *robust* texture-based indexing methods must be developed.

Future steps in this research will be the evaluation of state-of-the-art techniques in a realistic use case for these articles. Furthermore, additional journals must be evaluated. Nonetheless, we believe that such technology will have significant impact in the retrieval approaches for structured scientific literature.

5. CONCLUSION

Content-based image retrieval is an active field of research in medical informatics. However, the current view of research is limited to images that are obtained directly from the imaging modality. In this paper, we suggest to extend the idea of medical CBIR to scientific literature and to medical informatics in general. By analyzing 2,587 images from the 2005 volumes of the journal *Radiology* we have shown that CBIR may improve the quality of literature retrieval through use of robust local image features. Therefore, indexing of ROIs should be addressed more seriously in future research. If effective CBIR techniques can be developed for images in scientific articles, retrieval of these articles may be significantly enhanced. CBIR could be used as an additional component along with familiar text-based retrieval, such as that currently used in scientific databases such as SPIE Digital Library, IEEE Xplore, and PubMed.

ACKNOWLEDGEMENT

REFERENCES

1. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22(12): 1349-80
2. Vailaya A, Figueiredo MAT, Jain AK, Zhang HJ. Image classification for content-based indexing. IEEE Transactions of Image Processing 2001; 10(1): 117-30
3. Niblack W, Barber R, Equitz W, Flickner M, Glasman E, Petkovic D, Yanker P, Faloutsos C, Taubin G. The QBIC project: Querying images by content using color, texture, and shape. Proceedings SPIE 1993; 1908: 173-87
4. Lehmann TM, Antani S, Long LR. Gaps in content-based image retrieval. Proceedings SPIE 2007; 6516: in press in this issue.
5. Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W: The CLEF 2005 cross language image retrieval track. Lecture Notes in Computer Science 2006; 4022: 535-558.
6. Deselaers T, Müller H, Clough P, Ney H, Lehmann TM: The CLEF 2005 automatic medical image annotation task. International Journal of Computer Vision 2007; in press.
7. Müller H, Deselaers T, Lehmann T, Clough P, Hersch W: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. Lecture Notes in Computer Science 2007; in press.
8. Tagare HD, Jaffe CC, Duncan J. Medical image databases: A content-based retrieval approach. Journal of the American Medical Informatics Association - JAMIA 1997; 4(3): 184-98
9. Lehmann TM, Güld MO, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein BB: Content-based image retrieval in medical applications. Methods of Information in Medicine 2004; 43(4): 354-361.
10. Fischer B, Winkler B, Thies C, Güld MO, Lehmann TM: Strukturprototypen zur Modellierung medizinischer Bildinhalte. In: Handels H, Erhardt J, Horsch A, Meinzer HP, Tolxdorff T (eds.) Bildverarbeitung für die Medizin 2006, Springer-Verlag, Berlin 2006; 71-75 [in German]
11. Puzicha J, Rubner Y, Tomasi C, Buhmann J: Empirical evaluation of dissimilarity measures for color and texture. Proceeding ICCV 1999; 2: 1165–1173
12. Traina C Jr, Traina AJ, Araujo MR, Bueno JM, Chino FJ, Razente H, Azevedo-Marques PM: Using an image-extended relational database to support content-based image retrieval in a PACS. Computer Methods and Programs in Biomedicine 2005; 80(1): 71-83
13. Müller H, Michoux N, Bandon D, Geissbuhler A: A review of content-based image retrieval systems in medical applications. Clinical benefits and future directions. International Journal of Medical Informatics 2004; 73(1): 1-23
14. Lehmann TM, Güld MO, Deselaers T, Keysers D, Schubert H, Spitzer K, Ney H, Wein BB: Automatic categorization of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics 2005; 29(2): 143-155.
15. Lehmann TM, Plodowski B, Spitzer K, Wein BB, Ney H, Seidl T: Extended query refinement for content-based access to large medical image databases. Proceedings SPIE 2004; 5371: 90-98.
16. Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. Proceedings SPIE 2003; 5033: 440-451.
17. The Thomson Corporation (ed). ISI Journal Citation Reports 2005, Science Edition 2006.