Content-based Image Retrieval for Scientific Literature Access

T. M. Deserno[1, 2]; **S. Antani**[2]; **L. Rodney Long**[2]
[1]Department of Medical Informatics, Aachen University of Technology (RWTH), Aachen, Germany;
[2]U. S. National Library of Medicine, U. S. National Institutes of Health, Bethesda, Maryland, USA

Keywords

Content-based image retrieval (CBIR), scientific literature, information system integration, radiology, data mining, information retrieval

Summary

Objectives: An increasing number of articles are published electronically in the scientific literature, but access is limited to alphanumerical search on title, author, or abstract, and may disregard numerous figures. In this paper, we estimate the benefits of using content-based image retrieval (CBIR) on article figures to augment traditional access to articles.
Methods: We selected four high-impact journals from the Journal Citations Report (JCR) 2005. Figures were automatically extracted from the PDF article files, and manually classified on their content and number of sub-figure panels. We make a quantitative estimate by projecting from data from the Cross-Language Evaluation Forum (ImageCLEF) campaigns, and qualitatively validate it through experiments using the Image Retrieval in Medical Applications (IRMA) project.
Results: Based on 2077 articles with 11,753 pages, 4493 figures, and 11,238 individual images, the predicted accuracy for article retrieval may reach 97.08%.
Conclusions: Therefore, CBIR potentially has a high impact in medical literature search and retrieval.

Correspondence to:
Prof. Dr. Thomas M. Deserno, né Lehmann
Professor of Medical Informatics
Department of Medical Informatics
Aachen University of Technology (RWTH)
Pauwelsstr. 30
52057 Aachen
Germany
E-mail: deserno@ieee.org
Web: http://irma-project.org/deserno

Methods Inf Med 2009; 48: 371–380
doi: 10.3414/ME0561
received: April 11, 2008
accepted: January 12, 2009
prepublished: July 20, 2009

1. Introduction

Content-based image retrieval (CBIR) has long been identified as a key technology with the potential for significant impact for the management of and the retrieval from large collections of images [1, 2]. With respect to medical and health information, Haux has postulated a paradigm shift from mainly alpha-numeric data in hospital information systems (HIS) to images [3]. Typical image collections studied in biomedical CBIR are collections in picture archiving and communication systems (PACS) and research studies [4, 5]. Applications of medical CBIR systems appear in the fields of computer-aided diagnosis, evidence-based medicine, case-based reasoning, and medical training [1, 4–8]. In recent research, the diagnostic fields range from mammography [9], high-resolution computed tomography [10], or dynamic PET images [11] to application-unspecific annotation and classification tasks [12, 13]. Grid computing has been suggested to perform the remarkable computational load that is associated with CBIR applications [14, 15].

However, there are other fields in medicine that can benefit from such techniques. In particular, a huge amount of medical images, figures, drawings, 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 bitmapped figure images 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. Background

As a basic principle of CBIR, images are internally represented by numerical features, which are extracted directly from the image pixels (bitmap). 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 [16]. 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 the 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.

The distance or similarity measure is usually specific to a particular feature. For instance, the Jensen-Shannon divergence [17] is used for histogram-based features, while the Mahalanobis or Euclidean distances are applied for vector-based signatures. Several approaches are used to compute signatures [18]:

- *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 *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 biological system imaged [19–21].

- *Local* features are defined as those that are computed in prominent image regions, e.g. texture or shape features localized at a particular region of interest (ROI). This results in a number of signatures that are related to each image, and the capability of CBIR techniques to focus on particular aspects of the image content is increased. A similar assessment is made by Tagare et al. stating that the information contained in medical images is local [6], and hence, local features may further narrow the semantic gap.
- *Relational* features have not yet been applied routinely, but the concept has been discussed in the literature [7, 22]. The idea is to capture the spatial and/or temporal relationships between the image regions of interest (ROIs), such as distance, direction, and size relationships. Clearly, relational features are most similar to the scene interpretation by humans.
- *Hybrid* features include a combination of text and image features. Such a combination can benefit image retrieval, especially when supporting text information is available.

An overview of features and distances is given by [4], and can be taken from the results of the ImageCLEF[a] campaign [19–21].

3. Methods

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

3.1 Selection of Journals

We first selected a representative set of scientific journals as a data source for studying the effect of CBIR on article retrieval. Using the impact factor that is published in the Institute for Scientific Information (ISI) Journal Citation Reports (JCR) [23] as an indicator of journal importance, we selected within the three best-ranked journals with the most articles published electronically from 2005. Specifically, we applied the following four JCR journal categories:

1. *All*, to capture the most important journal;
2. *Radiology*, since radiology is the medical discipline that arguably produces the highest number of diagnostic images;
3. *Dentistry*, as one of the medical disciplines that is also closely related to medical imaging; and
4. *Medical informatics*, since this field is of high interest to our own research.

The selected journals were *New England Journal of Medicine* (ISSN 0028-4793), *Radiology* (ISSN 0033-8419), the *Journal of Dental Research* (ISSN 0022–0345), and the *Journal of the American Medical Informatics Association* (*JAMIA*) (ISSN 1067-5027). The *New England Journal of Medicine* was ranked third, but had 309 cited articles. In contrast, the *CA: A Cancer Journal for Clinicians* and the *Annual Review of Immunology* have slightly higher impact factors, but contain a significantly lower number of "cited articles" (20 and 29, respectively[b]), which is a contra-indication for our purposes.

3.2 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 "Advanced – Extract all images" and "Advanced – Batch processing" of the Adobe Acrobat Professional 7.0 software, all PDF-embedded bitmaps were automatically extracted as individual image files and stored in the lossless-compressed Portable Network Graphics (PNG) format.

Since the *New England Journal of Medicine* provides direct access to the article illustrations (http://content.nejm.org/search_figures.dtl), we omitted the procedure of PNG extraction and downloaded all illustrations directly from the Web in the lossy-

Table 1 Journals selected for the study

	New England Journal of Medicine	Radiology	Journal of Dental Research	JAMIA	Sum
Impact factor 2005	**44.016**	**5.377**	**3.192**	**4.339**	
Total pages	5,582	4,308	1,197	666	11,753
Total articles	1,152	738	195	87	2,172
Available PDF articles	1,061	734	195	87	2,077
Extracted PNG figures	1,221	2,587	465	220	4,493
Resulting figure panels	2,630	6,469	1,826	313	11,238

[a] The Cross Language Evaluation Forum (CLEF, http://www.clef-campaign.org/) promotes research and development in multilingual information access by i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and ii) creating test-suites of reusable data, which can be employed by system developers for benchmarking purposes. In the CLEF campaigns, image and video data is regarded as certain language, too.

[b] 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 items.

compressed format of the Joint Photographic Experts Group (JPEG). All extracted image files were stored on the Image Retrieval in Medical Applications (IRMA) system (http://irma-project.org).

▶Table 1 summarizes the number of pages, articles, and extracted figures that are used in this study. In total, 2077 articles with more than 10,000 pages and 4493 figures were included in the analysis.

3.3 Classification of Illustrations

The variety of figures in scientific literature is very large. They vary in figure layout, image type (e.g., line illustration, x-ray, histology), and imaged content. Frequently, diagnostic images are combined as sub-figure panels, annotated with text and drawings, and composed together with schematic graphs, diagrams or other types of illustrations. For content-based image analysis it is important to understand the number and kinds of images, graphs, drawings and photographs, the frequency of annotations, and the presence of image color. In order to assess these parameters systematically, we defined the following major classes of figure images (▶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 grayscale 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

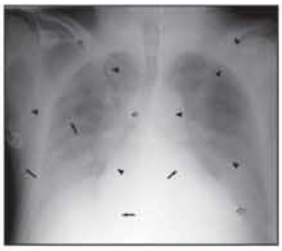
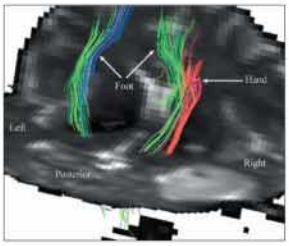
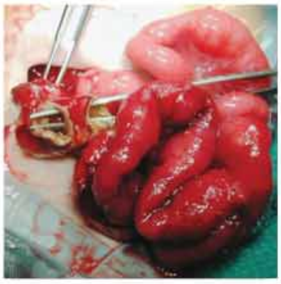
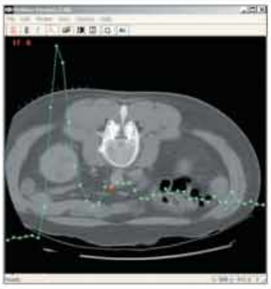
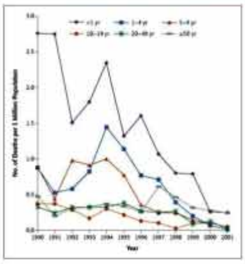
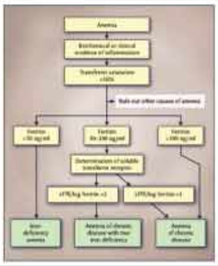
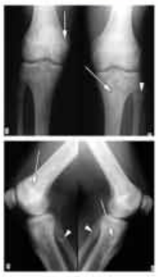
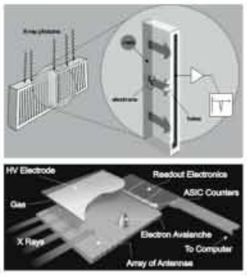
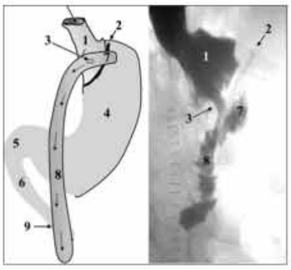
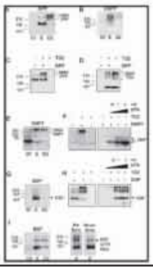
Example image	Image code	Example image	Image code
	9112 figure diagn. image grayscale annotation		9220 figure diagn. vis. color unspecified
	9323 figure photograph color object		9420 figure screen shot color unspecified
	9520 figure graph color unspecified		9620 figure diagram color unspecified
	9710 figure illustration color unspecified		9814 figure multi-p. med. grayscale 4 parts
	9912 figure multi-p. non-med. grayscale 2 parts		9a12 figure multi-p. mix. grayscale 2 parts
	9b12 figure protein spot grayscale annotated		9c13 figure other grayscale scan. equat.

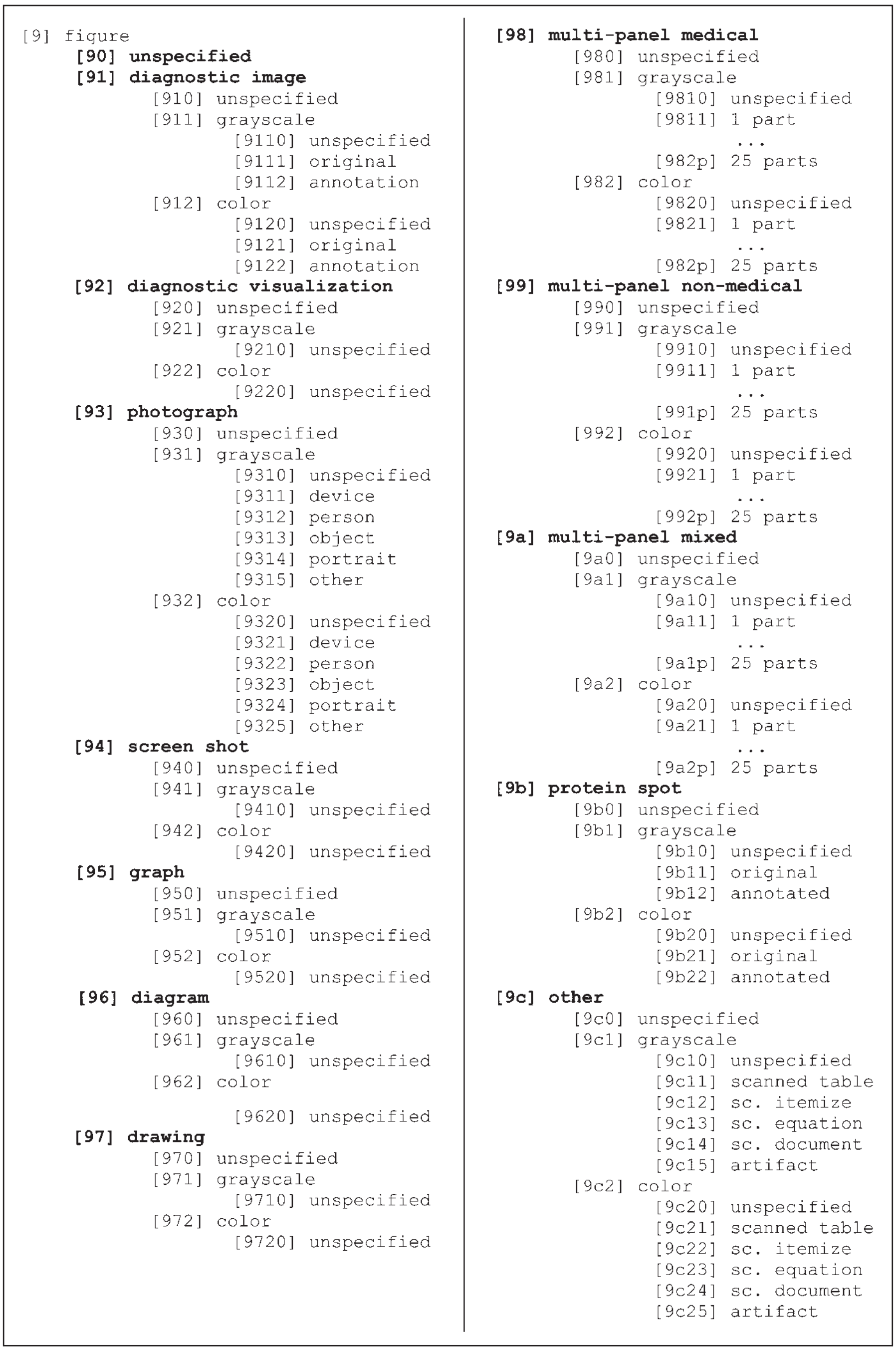
Fig. 1 Example illustrations from all major categories. The codes refer to Figure 2.

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;

- *protein spot*, i.e., a special type of multi-panel figures, where the high number of spots (panels) frequently is ambiguous, and therefore, not countable. Therefore, figures with protein spots are counted entirely as one individual image.

3.4 Database and Reference Categorization

All figure images were analyzed manually for the number and composition of diagnostic images included as figures. This labeling 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 [7, 24].



```
[9] figure
    [90] unspecified
    [91] diagnostic image
         [910] unspecified
         [911] grayscale
               [9110] unspecified
               [9111] original
               [9112] annotation
         [912] color
               [9120] unspecified
               [9121] original
               [9122] annotation
    [92] diagnostic visualization
         [920] unspecified
         [921] grayscale
               [9210] unspecified
         [922] color
               [9220] unspecified
    [93] photograph
         [930] unspecified
         [931] grayscale
               [9310] unspecified
               [9311] device
               [9312] person
               [9313] object
               [9314] portrait
               [9315] other
         [932] color
               [9320] unspecified
               [9321] device
               [9322] person
               [9323] object
               [9324] portrait
               [9325] other
    [94] screen shot
         [940] unspecified
         [941] grayscale
               [9410] unspecified
         [942] color
               [9420] unspecified
    [95] graph
         [950] unspecified
         [951] grayscale
               [9510] unspecified
         [952] color
               [9520] unspecified
    [96] diagram
         [960] unspecified
         [961] grayscale
               [9610] unspecified
         [962] color

               [9620] unspecified
    [97] drawing
         [970] unspecified
         [971] grayscale
               [9710] unspecified
         [972] color
               [9720] unspecified
```

```
[98] multi-panel medical
     [980] unspecified
     [981] grayscale
           [9810] unspecified
           [9811] 1 part
                  ...
           [982p] 25 parts
     [982] color
           [9820] unspecified
           [9821] 1 part
                  ...
           [982p] 25 parts
[99] multi-panel non-medical
     [990] unspecified
     [991] grayscale
           [9910] unspecified
           [9911] 1 part
                  ...
           [991p] 25 parts
     [992] color
           [9920] unspecified
           [9921] 1 part
                  ...
           [992p] 25 parts
[9a] multi-panel mixed
     [9a0] unspecified
     [9a1] grayscale
           [9a10] unspecified
           [9a11] 1 part
                  ...
           [9a1p] 25 parts
     [9a2] color
           [9a20] unspecified
           [9a21] 1 part
                  ...
           [9a2p] 25 parts
[9b] protein spot
     [9b0] unspecified
     [9b1] grayscale
           [9b10] unspecified
           [9b11] original
           [9b12] annotated
     [9b2] color
           [9b20] unspecified
           [9b21] original
           [9b22] annotated
[9c] other
     [9c0] unspecified
     [9c1] grayscale
           [9c10] unspecified
           [9c11] scanned table
           [9c12] sc. itemize
           [9c13] sc. equation
           [9c14] sc. document
           [9c15] artifact
     [9c2] color
           [9c20] unspecified
           [9c21] scanned table
           [9c22] sc. itemize
           [9c23] sc. equation
           [9c24] sc. document
           [9c25] artifact
```

Fig. 2 IRMA code extension for classification of illustrations. The major categories are displayed in bold face.

The hierarchical, multi-axial IRMA code [25] was extended to capture the characteristics of illustrations in scientific papers. All images within the IRMA system are related to an A-B-C-D code which is composed of four labels, viz. the body region (A-natomy) and biological system (B-iosystem) imaged, the imaging modality (C-reation), and the view (D-irection). In order to classify the nature of published illustrations, we used the C-axis of the IRMA code. Figure 2 shows the resulting part of the IRMA code. Note that the third digit of the code always distinguishes color from grayscale, which allows easy summation over all categories. Also, the last digit for the multi-panel images denotes the number of panels.

Of 4493 bitmap files that were available, only 4418 bitmap files are considered with the following breakdown into five categories (for each category, count and IRMA codes are shown in parentheses, see Table 2):

1. an individual medical image, visualization or protein expression (547; 91** + 92** + 9b**),
2. a combination of medical images (1451; 98**),
3. an individual graph, diagram, drawing, or photograph (1483; 93** + … + 97**),
4. a combination of several graphs (703; 99**), and
5. a combination of medical images and graphs within a single figure file (234; 9a**).

Of the 75 images that were excluded: 57 (PNG or JPEG) files contained scanned documents, lists, equations or tables (9b*1 + … + 9b*4), and another 18 contain artifacts (9b*5), e.g., a single line that is used to separate text blocks but do not represent an illustration.

It can be further deduced from Table 2 that, i) with a frequency of more than 55% (98**+ … + 9b**; 2435 of 4418 = 55.12%), medical images and/or graphs found in the literature were composed of figures with multiple image panels, and ii) the majority of illustrations are still published in grayscale (**1*; 4493 – 18 ar-

Table 2 Figure items extracted from the PDF articles

		New England Journal of Medicine				Radiology				Journal of Dental Research				Journal of the American Medical Informatics Association JAMIA				Sum			
IRMA code	Name of figure category	Items #	Items %	Panels #	Panels %	Items #	Items %	Panels #	Panels %	Items #	Items %	Panels #	Panels %	Items #	Items %	Panels #	Panels %	Items #	Items %	Panels #	Panels %
91**	diagnostic image	65	5.32	65	2.47	408	15.77	408	6.13	15	3.23	15	0.82	0	0.00	0	0.00	488	10.86	488	4.34
92**	diagnostic visualization	3	0.25	3	0.11	9	0.35	9	0.14	0	0.00	0	0.00	0	0.00	0	0.00	12	0.27	12	0.11
93**	photograph	87	7.13	87	3.31	80	3.09	80	1.24	4	0.86	4	0.22	19	8.64	19	6.07	190	4.23	190	1.69
94**	screen shot	0	0.00	0	0.00	2	0.08	2	0.03	1	0.22	1	0.05	42	19.09	42	13.42	45	1.00	45	0.40
95**	graph	234	19.16	234	8.90	470	18.17	470	7.72	58	12.47	58	3.18	39	17.73	39	12.46	801	17.83	801	7.13
96**	diagram	189	15.48	189	7.19	55	2.13	55	0.85	26	5.59	26	1.42	69	31.36	69	22.04	339	7.55	339	3.02
97**	drawing	53	4.34	53	2.02	40	1.55	40	0.62	11	2.37	11	0.60	4	1.82	4	1.28	108	2.40	108	0.96
98**	multi-panel medical	237	19.41	759	28.86	1,095	42.33	4,015	62.07	117	25.16	717	39.27	2	0.91	8	2.56	1,451	32.29	5,499	48.93
99**	multi-panel non-medical	276	22.60	936	35.59	272	10.51	787	12.17	122	26.24	452	24.75	33	15.00	119	38.02	703	15.65	2,294	20.41
9a**	multi-panel mixed	58	4.75	285	10.84	93	3.59	540	8.35	82	17.63	513	28.09	1	0.45	2	0.64	234	5.21	1,340	11.92
9b**	protein spot	17	1.39	17	0.65	1	0.04	1	0.02	28	6.02	28	1.53	1	0.45	1	0.32	47	1.05	47	0.42
9c**	others	2	0.16	2	0.08	62	2.40	62	0.96	1	0.22	1	0.05	10	4.55	10	3.19	75	1.67	75	0.67
Sum		1,221	100.00	2,630	100.00	2,587	100.00	6,469	100.00	465	100.00	1,826	100.00	220	100.00	313	100.00	4,493	100.00	11,238	100.00
9c*5	artifact	0	0.00	0	0.00	18	0.70	18	0.28	0	0.00	0	0.00	0	0.00	0	0.00	18	0.40	18	0.16
else	regular image	1,221	100.00	2,630	100.00	2,569	99.30	6,451	99.72	465	100.00	1,826	100.00	220	100.00	313	100.00	4,475	99.60	11,220	99.84
Sum		1,221	100.00	2,630	100.00	2,587	100.00	6,469	100.00	465	100.00	1,826	100.00	220	100.00	313	100.00	4,493	100.00	11,238	100.00
**1*	grayscale	415	33.99	757	28.78	2,255	87.78	5,269	81.68	359	77.20	1,142	62.54	167	75.91	214	68.37	3,196	71.42	7,382	65.79
**2*	color	806	66.01	1,873	71.22	314	12.22	1,182	18.32	106	22.80	684	37.46	53	24.09	99	31.63	1,279	28.58	3,838	34.21
Sum		1,221	100.00	2,630	100.00	2,569	100.00	6,451	100.00	465	100.00	1,826	100.00	220	100.00	313	100.00	4,475	100.00	11,220	100.00
91**–97**	individual	631	51.76	631	24.01	1,064	42.14	1,064	16.61	115	24.78	115	6.30	173	82.38	173	57.10	1,983	44.88	1,983	17.76
98**–9b**	multi-panel	588	48.24	1,991	75.99	1,461	57.86	5,343	83.39	349	75.22	1,710	93.70	37	17.62	130	42.90	2,435	55.12	9,180	83.34
Sum		1,219	100.00	2,628	100.00	2,525	100.00	6,407	100.00	464	100.00	1,825	100.00	210	100.00	303	100.00	4,418	100.00	11,163	100.00
91*1+ 9b*1	original	31	37.80	31	37.80	69	16.87	69	16.87	1	2.33	1	2.33	0	0.00	0	0.00	101	18.98	101	18.98
91*2+ 9b*2	annotated	51	62.20	51	62.20	340	83.13	340	83.13	42	97.67	42	97.67	1	100.00	1	100.00	434	81.12	434	81.12
Sum		82	100.00	82	100.00	409	100.00	409	100.00	43	100.00	43	100.00	1	100.00	1	100.00	535	100.00	535	100.00

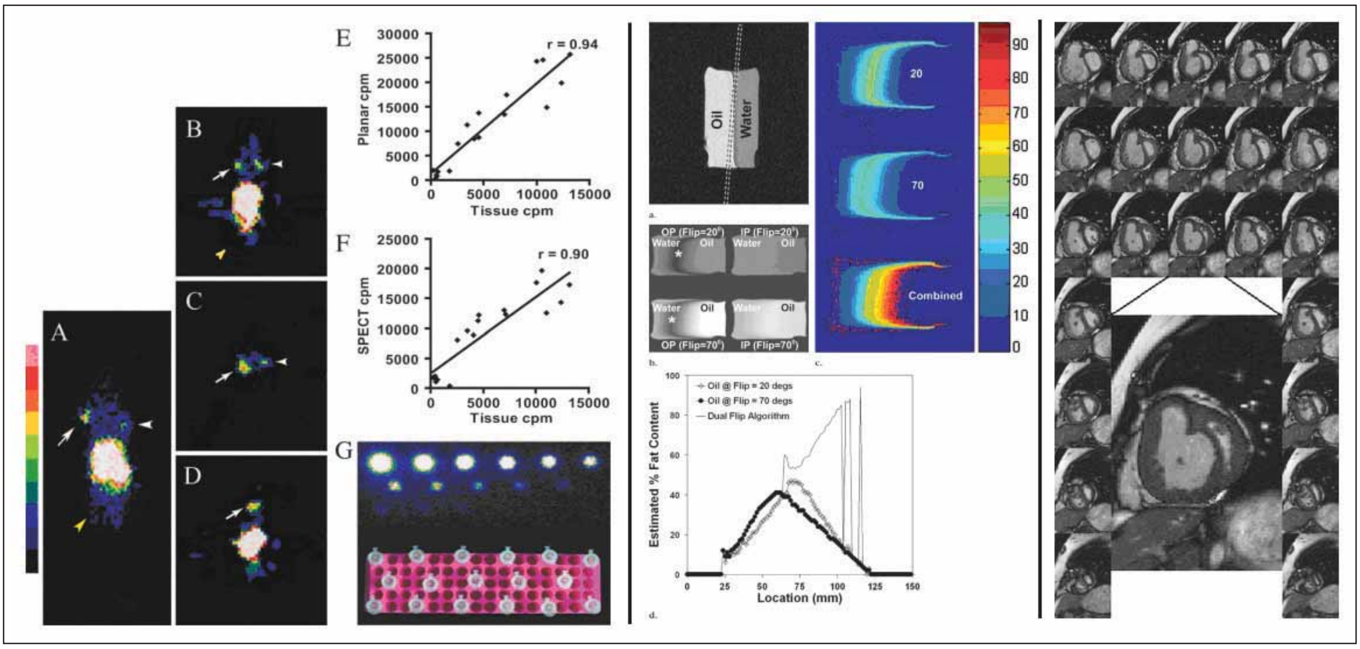


Fig. 3 Examples for complex multi-panel illustrations

tifact images with IRMA code 9c*5 = 4475; 3196 of 4475 = 71.42%;). Even 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 65%. Similarly, a majority of diagnostic images and protein spots are annotated with text, arrows, or other symbols (91*2 + 9b*2; 434/535 = 81.12%) that may cover image information and affect the image texture feature extraction.

In total, 11,163 useful figure panels were extracted from 11,753 pages giving us one (0.95) figure panel per article page, and two (4418/2077 = 2.13) figures and over five (11,163/2077 = 5.37) individual figure panels per article.

3.5 Evaluation

To quantitatively estimate the impact of CBIR-based literature research, we useddata from the Cross-Language Evaluation Forum (CLEF) image campaign as a ground truth. In recent years, ImageCLEF (http://ir.shef.ac.uk/imageclef/) has served as an international 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 [19–21].

In a first approximation that is based on the count of illustrations, the error rates from CLEF are used to compute an expected error rate for *article* retrieval based on global-feature CBIR. We note two points: 1) using global-feature CBIR we would expect to be able to retrieve figures, but not individual figure panels (the entire figure – possibly multi-panel – is treated as a single image, for this retrieval); and 2) we expect that, the greater the number of figures per article, the greater our chances of successfully retrieving the article by CBIR. However, the relationship among the global signatures of the various figures in a single article is complex, and we know of no published research that has created an empirically- or theoretically-based model to explain this relationship, or how this relationship may be factored into CBIR error rates. In this paper, we will calculate the expected error rate for article retrieval under the assumption that the error rate will decrease *by a linear factor* as the number of figures per article increases. Future research is required to amend or refine this assumption.

To qualitatively demonstrate the impact of CBIR-based literature research, a *global* signature was calculated for all journal figures, following the methodology used by the IRMA research group in its submission to the ImageCLEF 2005 competition [26]. As a typical example, we selected the *coronal chest radiograph* (IRMA code: A-B-C-D = 500-000-1123-1**, where * denotes a wildcard) as an image class likely to appear frequently in the literature based on the fact that this is the most frequent imaging procedure at Aachen University Hospital. Then, for each individual category in the ImageCLEF 2006 database that matched this IRMA code, we randomly selected an image. We then used these images as QBE examples for retrieval from the article database.

4. Results

In this concept paper we argue from previously-published work, and from one retrieval experiment, that our proposed method has a reasonable expectation of success in enhancing searching of published scientific literature by incorporating image searching. We rely on image search data from the ImageCLEF competition. For the "quantitative evaluation" below, to get a first-order estimate of how image search might enhance literature search, we treat article search results as being linearly related to the number of fig-

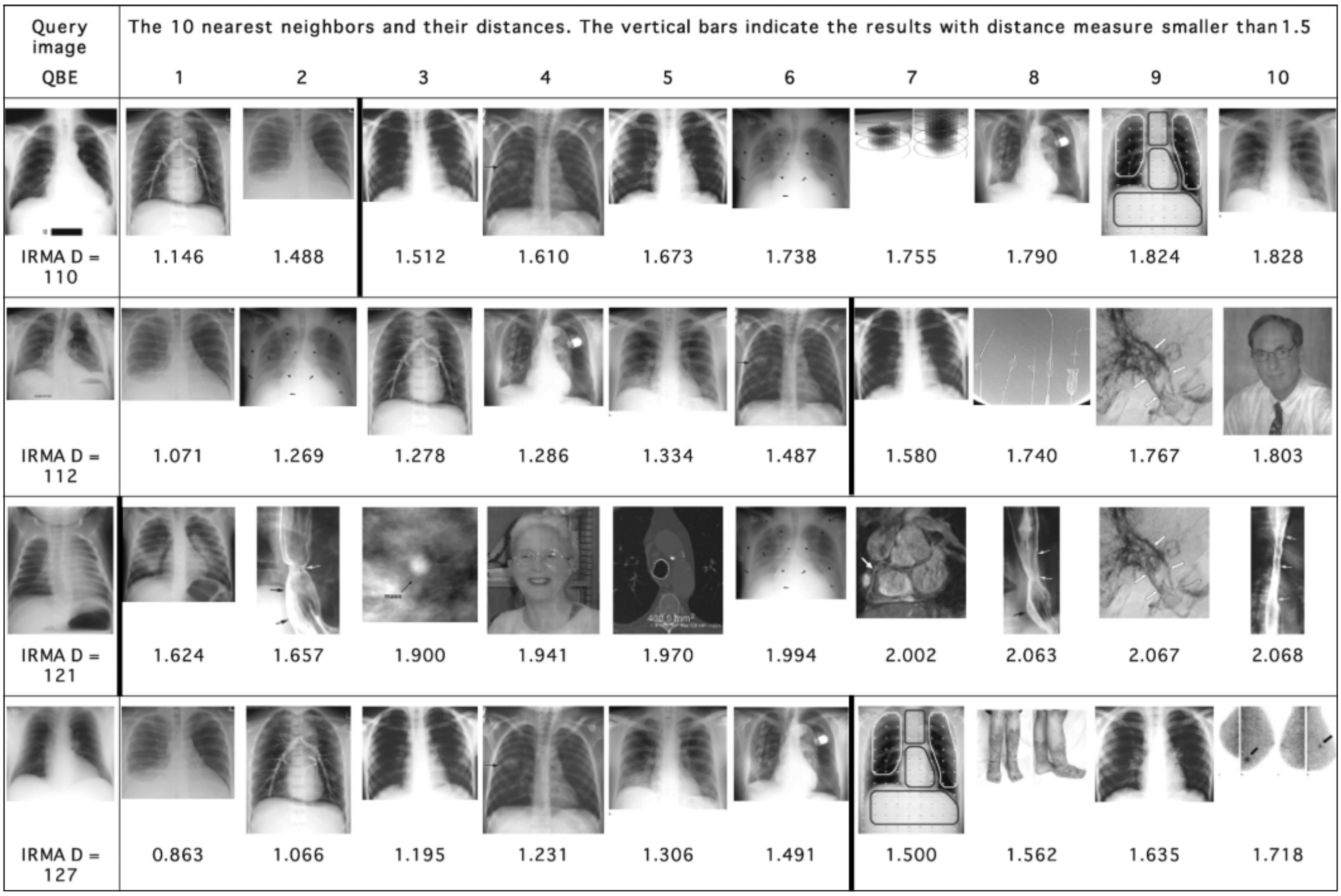


Fig. 4 Qualitative evaluation based on the IRMA framework

ures per article. For the “qualitative evaluation” we again use ImageCLEF data, and show that the results of one experiment suggest that image-enhanced retrieval may tend to exhibit a good precision rate, and may also return valid results that would not be found by text-only searches.

4.1 Quantitative Evaluation

Our 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 in 51 categories were conducted. Applying state-of-the-art CBIR techniques that use global texture-based signatures to represent the image content, error rates of about 12% were reported [19], while ImageCLEFmed 2006 with 116 categories for 11,000 radiographs yielded a 14% error rate for an optimal classifier combination [21]. Based on these results we chose the number 15% as a reasonable expected error rate when using global-feature CBIR for image retrieval.

Under the assumption of linearity, then, since we have on average, two figures in an article; the expected error rate for CBIR-assisted literature retrieval is expected to decrease to 7.5%. If we used local features for content-based image representation, we would presumably have retrieval capability to the individual panels in the figures; in this scenario, five sub-figure images per article are available, and the predicted error rate may further decrease to 3%, or a predicted accuracy of 97%.

For *Radiology* only, the predicted error rates are 15/(2525/734) = 4.36% and 15/(6407/734) = 1.72% for global and local signatures, respectively. Using CBIR-supported access to *JAMIA*, error rates of 15/(210/87) = 6.22% and 15/(303/87) = 4.31% can be predicted, respectively.

It is of interest to note that, if we apply this same methodology to compare error rates for all four of the journals we examined, then, for global signatures, where the number of items/article is the determinative factor, the journals rank, from smallest to greatest error rate: *Radiology*, *JAMIA*, *Journal of Dental Research*, and *New England Journal of Medicine*, with respective average numbers of items/article of (3.44, 2.41, 2.38, and 1.15) and resulting error rates of (4.36, 6.22, 6.30, and 13.04). However, journals with the highest occurrence of figures do not necessarily have the highest occurrence of sub-figures (panels). For local signatures, where the number of panels/article is the factor of interest, the ranking is as follows: *Journal of Dental Research*, *Radiology*, *JAMIA*, and *New England Journal of Medicine*, with respective average panels/article of (9.36, 8.73, 3.48, and 2.48), and resulting error rates of (1.60, 1.72, 4.31, and 6.05).

There are several aspects and limitations of the above analysis which require investi-

Table 3 Article titles corresponding to the first line of Figure 4

Response no.	Title of article
1	**Chest Radiograph**y with a Digital Flat-Panel Detector: Experimental Receiver Operating Characteristic Analysis
2	Case 92 from the Department of Radiology, University of Wisconsin Hospital and Clinics
3	Case 90 from the Departments of Pediatrics and Radiology, Hospital Universitario Central de Asturias
4	Interpretation of **Chest Radiograph**s in Infants with Cough and Fever
5	Case 90 From the Departments of Pediatrics and Radiology, Hospital Universitario Central de Asturias
6	Detectability of Catheters on Bedside **Chest Radiographs**: Comparison between LiquidCrystal Display and High-Resolution Cathode-Ray Tube Monitors
7	Multi–Detector Row CT Systems and Image- Reconstruction Techniques
8	Medical Mystery The Answer
9	Comparative Scatter and Dose Performance of Slot-Scan and Full-Field Digital **Chest Radiograph**y Systems
10	Radiographic-Clinical Correlation in Severe Acute Respiratory Syndrome: Study of 1373 Patients in Hong Kong

gation to move from the conceptual understanding of our work to a refined assessment of what is achievable in practice. These include the observations that 1) the image retrieval rates obtainable in ImageCLEF may be different from those obtainable in the more general medical literature, and 2) the assumption of linear improvement in article search results as a function of number of figures per article has been made to create a first-order assessment of what may be achievable, and will doubtless be modified as experimental data accumulates from actual implementations.

4.2 Qualitative Evaluation

With 3270 out of 11,000 images, frontal chest radiographs occur most frequently in the ImageCLEFmed 2006 database. In particular, four classes of images matching the IRMA code mask A-B-C-D = 500-000-1123-1** were present: 1011 radiographs in posterioanterior (PA) projection (IRMA code: D = 110), 51 PA-images in expiration (D = 112), 86 x-rays in anterioposterior (AP) projection in inspiration (D = 121), and 2122 AP-images captured supine (D = 127). For each IRMA category, one QBE image was selected randomly from the ImageCLEFmed 2006 database. In separate queries, we then used the IRMA system to search this ImageCLEFmed database for images similar to these four input images.

Figure 4 shows the respective ten best image matches that were retrieved from the article figures using a global-feature signature. As can be observed, the majority of responses are coronal chest radiographs. Taking into account the similarity between the QBE image and the response images, and disregarding all responses with a dissimilarity larger than 1.5, only chest radiographs were retrieved for all four QBE images. This suggests that image-based article searching may be expected to have a practical level of precision, i.e. the query tends to return relevant results.

The query based on the unspecified lateral radiograph (IRMA code: D = 110) was analyzed in more detail. Table 3 shows the titles of the ten corresponding articles. As it can be observed, only four of the titles contain the keyword *chest radiograph*, and hence, only these four would be retrieved using a text-based search strategy. In other words, the correspondence between the title of the article and its figure content is 40%. However, based on the image-based query, five additional relevant papers were found, with only one irrelevant response. This observation suggests that image-based article searching may be expected to have an enhanced level of recall as compared to text-based searching of images in articles.

5. Discussion

Content-based image retrieval has not yet been suggested as a technique to enhance traditional (text-based) approaches for retrieval of scientific literature. This idea is novel and to the best of our knowledge an application does not exist yet. As a pilot experiment and a feasibility test we previously analyzed the figures published in the 2005 volume of *Radiology* [27]. In this paper, we formulate a model to compute a quantitative estimate of the impact of medical CBIR using ImageCLEFmed campaign as ground truth, and a qualitative estimate using experimental results from IRMA framework as a test-bed. We apply it to an extended data collection from the 2005 volumes of the *New England Journal of Medicine*, *Radiology*, the *Journal of Dental Research* and the *Journal of the American Medical Informatics Association* (*JAMIA*).

Park et al. have shown that the size of the reference database as well as the composition of references significantly affects the results in content-based medical image retrieval schemes [9]. This particularly holds for the ImageCLEF campaign that we used as a baseline to estimate the impact of CBIR literature access. Keeping this in mind, we can conclude from the quantitative analysis that CBIR may significantly improve scientific document retrieval from electronically published journals as a complement to traditional text indexing methods. Furthermore, the predicted error rates that were reported in this paper are determined for the best match only. Since CBIR-aided literature search systems may not only respond an individual article, but a list of matching articles, the recall might further increase when seeking for specific articles. For instance, error rates are reported to decrease from 15% to 7% and 5% if the best n matches are considered, n = 1, 5, and 10, respectively [5].

From the qualitative analyses – although using the IRMA code for annotation and modeling of ground truth may put a bias on the results – we can conclude that figure information in digital multimedia documents provides additional information to the traditionally indexed entities such as title and abstract of the article. This finding is consistent with the report by Christiansen et al [28]. Based on more than 1900 PDF files, which were downloaded from more than 20 different journals, the authors found correspondence of figure caption and title and abstract content in only about 700 documents (37%). As such, caption analysis and figure image analysis can add valuable information for enabling relevant retrieval. This idea is supported by experiments in classifying document figure images by modality (radiograph, chart, photograph, etc.) and utility (diagnostic, procedural, outcome, etc.) together with figure captions, descriptions in full text of the article [29, 30].

However, several assumptions have been made in stating the benefit of CBIR in scientific document retrieval. First, there is a need for indexing the image information (features, respective signatures). Currently, the ISI impact factor is based on 6500 scientific journals. PubMed contains about 16 million articles, which are daily increased by 2000 to 4000, each of it providing more than one figure for CBIR. For image retrieval from a literature database of substantial size to be practical, the image indexing effort should require minimal human effort and, ideally, should be fully automatic. The image retrieval results cited in this paper were done by fully-automatic methods used in the ImageCLEF [19–21] campaign.

Second, the figures that appear in the articles are often composed of several panels or sub-figures. These sub-figures are labeled using letters or roman numerals, for example. These composite figures need to be decomposed into individual, but related, sub-panels while maintaining references to the original, multi-panel figure. Initial efforts have been taken in this direction as reported in [31] and [32], where a related problem is raised from certain imaging modalities in dentomaxillofacial radiology. With this step, *global* and *local* image features can be extracted and indexed.

Third, these features need to be indexed with figure caption information along with the traditionally indexed items, such as title and abstract in a manner so as to improve overall article relevance. We note that a study in indexing dermatological images using terms extracted from the UMLS® metathesaurus was conducted at the U.S. National Library of Medicine. It was found that controlled vocabularies such as UMLS, SNMI (SNOMED), and RCD (Read Thesaurus) may be useful in indexing medical images [33]. These indices may help serve as a bridge between extracted image features and traditional text indices used for scientific articles. Other useful characteristics include annotations on images in articles and their correspondence with the caption or article text. It has been found that while detection of these annotations is feasible, their recognition through use of OCR techniques is challenging [31]. This is primarily due to image resolution and contextual dependence of robust OCR techniques.

Finally, we want to point out that pathology must also be captured in the signatures to fully explore the benefit from CBIR methods, when this technology is routinely applied to literature research. Therefore, future research must address the content-related gaps [18].

6. Conclusion

Content-based image retrieval for biomedical images is an active field of research in medical informatics. However, the current view of research is limited to diagnostic or medical research purposes operating on image databases that are developed for that specific purpose. Images occur frequently in biomedical scientific literature and article retrieval may benefit from application of this technology. In this concept paper, we have sought to justify extending the idea of medical CBIR to access the scientific literature and to medical informatics in general. In addition, coupling this with traditional text retrieval methods may significantly enhance the search experience as well as quality of retrieved results. We note that our work is currently at the concept level and requires further critical analysis and investigation, but we argue that through our work in analyzing nearly 4500 images from more than 11,000 article pages, and computing estimates from retrieval experiments on similarly sized, comparable databases, we have shown that CBIR may improve the quality of literature retrieval, in particular, through use of robust local image features. We believe that our work, though still at the conceptual stage, indicates that experimental investigation of image-enhanced medical literature retrieval is strongly justified. 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.

Acknowledgment

We would like to acknowledge the editorial and critical review inputs we have received for this article and to thank the editor and each of the reviewers. This research was supported (in part) by the Intramural Research Program of the U.S. National Institutes of Health (NIH), U.S. National Library of Medicine (NLM), and the Lister Hill National Center for Biomedical Communications (LHNCBC).

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): 1349–1380.
2. Vailaya A, Figueiredo MAT, Jain AK, Zhang HJ. Image classification for content-based indexing. IEEE Transactions of Image Processing 2001; 10 (1): 117–130.
3. Haux R. Health information systems. Past, present, future. International Journal of Medical Informatics 2006; 75 (3–4): 268–281.
4. 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.
5. 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.
6. 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): 184–198.
7. 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 Inf Med 2004; 43 (4): 354–361.

8. Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective automated indexing of findings and diagnoses in radiology reports. J Biomed Inform 2001; 34 (4): 262–273.
9. Park SC, Sukthankar R, Mummert L, Satyanarayanan M, Zheng B. Optimization of reference library used in content-based medical image retrieval scheme. Medical Physics 2007; 34 (11): 4331–4339.
10. Scott G, Shyu CR. Knowledge-driven multidimensional indexing structure for biomedical media database retrieval. IEEE Transactions on Information Technology in Biomedicine 2007; 11 (3): 320–331.
11. Kim J, Cai W, Feng D, Wu H. A new way for multidimensional medical data management: Volume of interest (VOI)-based retrieval of medical images with visual and functional features. IEEE Transactions on Information Technology in Biomedicine 2006; 10 (3): 598–607.
12. Rahman MM, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Transactions on Information Technology in Biomedicine 2007; 11 (1): 58–69.
13. Greenspan H, Pinhas AT. Medical image categorization and retrieval for PACS using the GMM-KL framework. IEEE Transactions on Information Technology in Biomedicine 2007; 11 (2): 190–202.
14. Hassan K, Tweed T, Miguet S. A multi-resolution approach for content-based image retrieval on the Grid-application to breast cancer detection. Methods Inf Med 2005; 44 (2): 211–214.
15. Montagnat J, Breton V, E Magnin I. Partitioning medical image databases for content-based queries on a Grid. Methods Inf Med 2005; 44 (2): 154–160.
16. 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: 173–187.
17. Puzicha J, Rubner Y, Tomasi C, Buhmann J. Empirical evaluation of dissimilarity measures for color and texture. Proceeding ICCV 1999; 2: 1165–1173.
18. Deserno TM, Antani S, Long R. Ontology of gaps in content-based image retrieval. J Digit Imaging 2008; online-first, DOI 10.1007/s10278–007–9092-x.
19. 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.
20. Deselaers T, Müller H, Clough P, Ney H, Lehmann TM. The CLEF 2005 automatic medical image annotation task. International Journal of Computer Vision 2007; 74 (1): 51–58.
21. Müller H, Deselaers T, Deserno TM, Clough P, Kim E, Hersch W. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. Lect Notes Comput Sci 2007; 4730: 595–608.
22. 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. Berlin: Springer-Verlag; 2006. pp 71–75 (in German).
23. The Thomson Corporation (ed). ISI Journal Citation Reports 2005, Science Edition 2006; (http://scientific.thomson.com/products/jcr/).
24. Lehmann TM, Plodowski B, Spitzer K, Wein BB, Ney H, Seidl T. Extended query refinement for content-based access to large medical image databases. Procs SPIE 2004; 5371: 90–98.
25. 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.
26. Güld MO, Thies C, Fischer B, Lehmann TM. Content-based retrieval of medical images by combining global features. Lecture Notes in Computer Science 2006; 4022: 702–711.
27. Deserno TM, Antani S, Long LR. Exploring access to scientific literature using content-based image retrieval. Procs SPIE 2007; 6516: OL1-OL8.
28. Christiansen A, Lee DJ, Chang Y. Finding relevant PDF medical journal articles by the content of their figures. Procs SPIE 2007; 6516: OK1-OK12.
29. Demner-Fushman D, Antani S, Thoma GR. Automatically finding images for clinical decision support. Proc. IEEE International Conference on Data Mining, Workshop on Data Mining in Medicine 2007. pp 139–144.
30. Névéol A, Deserno TM, Darmonic SJ, Oliver M, Güld, Aronson AR. Natural language processing vs. content-based image analysis for medical document retrieval. Journal of the American Society for Information Science and Technology 2009; 60 (1): 123–134.
31. Antani S, Demner-Fushman D, Li J, Srinivasan BV, Thoma GR. Exploring use of images in clinical articles for decision support in evidence-based medicine. To appear in Proc. IS&T/SPIE Electronic Imaging: Document Recognition and Retrieval, 2008.
32. Lehmann TM, Molander B, Güld MO, Thies C, Gröndahl HG. Content-based access to oral and maxillofacial radiographs. Dentomaxillofacial Radiology 2007; 36 (6): 328–335.
33. Woods JW, Sneiderman CA, Hameed K, Ackermaan MJ, Hatton C. Using UMLS metathesaurus concepts to describe medical images: dermatology vocabulary. Computers in Biology and Medicine 2006; 36 (1): 89–100.