Retrieving Similar Cases from the Medical Literature –The ImageCLEF experience

Jayashree Kalpathy-Cramer[a], Steven Bedrick[a], Saïd Radhouani[a], William Hersh[a], Ivan Eggel[b], Charles E. Kahn Jr.[c], Henning Müller[b]

[a]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Sciences University, Portland, OR, USA;*
[b]*University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland;*
[c]*Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA*

Abstract

An increasing number of clinicians, researchers, educators and patients routinely search for relevant medical images using search engines on the internet as well as in image archives and PACS systems. However, image retrieval is far less understood and developed compared to text-based searching. The ImageCLEF medical image retrieval task is an international challenge evaluation that enables researchers to assess and compare techniques for medical image retrieval using test collections.

In this paper, we describe the development of the ImageCLEF medical image test collection, consisting of a database of images and their associated annotations, as well as a set of realistic search topics and relevance judgments obtained using a set of experts. 2009 was the sixth year for the ImageCLEF medical retrieval task and had strong participation from research groups across the globe. We will provide results from this year's evaluation and discuss the successes that we have had as well as challenges going forward.

Keywords:

Image retrieval, Information Storage and Retrieval, Content Analysis and Indexing, Systems and Software

Introduction

Image retrieval is a burgeoning area of research in medical informatics [1, 2]. With the increasing utilization of digital imaging in all aspects of health care and medical research, there has been a substantial growth in the number of images being created every day in healthcare settings. Consequently, there is a critical need to manage the storage and retrieval of these image collections, whether they are stored in Picture Archival and Communication Systems (PACS), in patient health records, or on the web. Effective image annotation and retrieval can be useful in the clinical care of patients, education and research [2, 3]. Image retrieval can be used by clinicians to generate differential diagnoses, monitor patient response to therapy, and for quality control. Medical students and residents have also indicated that effective image retrieval can be useful for self-education [4], and other practitioners report using image retrieval systems for patient education, as well. Data-mining of large image collections can provide useful information for researchers. Examples include prevalence of certain findings including polyps during routine screening [5], visual characteristics associated with malignancy in mammography [6, 7], and prediction of response to radiation therapy based on FDG-PET [8].

Many areas of medicine, such as radiology, dermatology, and pathology are visually oriented, yet surprisingly little research has been done investigating how clinicians use and find images. In particular, medical image retrieval techniques and systems are under-developed in medicine when compared with their textual cousins. In particular, the field has suffered from a lack of evaluation opportunities and avenues for researchers to use to compare and measure their systems' performance. The lack of standardized test collections is an especially large problem facing medical image retrieval researchers.

Text retrieval, on the other hand, has a long history of evaluation campaigns, in which different groups use a common test collection to compare the performance of their different methods. The best-known such campaign is the Text REtrevial Conference (TREC), which has been running continuously since 1992. There have been several offshoots from TREC, including the Cross-Language Evaluation Forum (CLEF). CLEF operates on an annual cycle, and has produced numerous test collections since its inception in 2000. While CLEF's focus was originally on cross-language text retrieval, it has grown to include multimedia retrieval tracks of several varieties. The largest of these, ImageCLEF, first began in 2003 as a response to the aforementioned need for standardized test collections and evaluation forums and has grown to become today's pre-eminent venue for image retrieval evaluation.

ImageCLEF itself also includes several sub-tracks concerned with various aspects of image retrieval; one of these tracks is the subject of the present paper: the medical retrieval task. This medical retrieval task was first run in 2004, and has been repeated each year since.

The medical image retrieval track's test collection began with a teaching database of 8,000 images. Since then, it has grown to a collection of over 66,000 images from several teaching collections, as well as a set of topics that are known to be well-suited for textual, visual or mixed retrieval methods. In 2008, images from the medical literature were used for the first time, moving the task one step closer towards applications that can be of interest in clinical scenarios. Several user studies have been performed to study the image searching behavior of clinicians. These studies have been used to inform the development of the task over the years, particularly to help identify realistic search topics. In 2009 we introduced a case-based retrieval task as we continue to strive for scenarios that more closely resemble actual clinical work-flows.

A major goal of ImageCLEF has been to foster development and growth of multimodal retrieval techniques: i.e., retrieval techniques that combine visual, textual, and other methods to improve retrieval performance. Traditionally, image retrieval systems have been text-based, relying on the textual annotations or captions associated with images [9]. Several commercial systems, such as Google Images (images.google.com) and Yahoo! images (http://images.yahoo.com), employ this approach.

Although text-based information retrieval methods are mature and well-researched, they are limited by the quality of the annotations applied to the images. There are other important limitations facing traditional text retrieval techniques when applied to image annotations: 1) image annotations are subjective and context sensitive, and can be quite limited in scope or even completely absent; 2) manually annotating images is labor and time intensive, and can be very error prone; 3) image annotations are very "noisy" if they are automatically extracted from the surrounding text; and 4) there is far more information in an image than can be abstracted using a limited number of words.

Advances in techniques in computer vision have led to a second family of methods for image retrieval: content-based image retrieval (CBIR). In a CBIR system, the visual contents of the image itself are mathematically abstracted and compared to similar abstractions of all images in the database. These features could include the color, shape or texture of images. Typically, such systems present the user with an ordered list of images that are visually most similar to the sample (or "query") image.

Materials and Methods

The traditional system-oriented IR evaluation process depends on a test collection made up of three parts: a "collection" of content items (articles, images, videos, etc.) that are to be retrieved; a set of "topics" representing potential queries or information needs that are to be answered by searching over the collection's content items; and a set of "gold standard" relevance judgments describing an expert's (or several experts') opinion as to which content items are relevant for each of the search topics.

ImageCLEF Medical Image Retrieval Test Collection

For the first several years, the ImageCLEF medical retrieval test collection was an amalgamation of several teaching case files in English, French, and German [9, 10]. In both 2008 and 2009, the Radiological Society of North America (RSNA) made a subset of its journals' image collections available for use by participants in the ImageCLEF campaign. The 2009 database contained a total of 74,902 images, the largest collection yet.

All images were taken from the journals *Radiology* and *RadioGraphics*, both published by the RSNA. The ImageCLEF collection is similar in composition to that powering the "ARRS GoldMiner"[1] search system [11]. This collection constitutes an important body of medical knowledge from the peer-reviewed scientific literature, and includes high quality images with annotations. Images are associated with specific published journal articles, and as such may represent either an entire figure or a component of a larger figure. In either event, the image's annotations in the collection will contain the appropriate caption text. These high-quality annotations enable textual searching in addition to content-based retrieval using the image's visual features. Furthermore, as the PubMed IDs of each image's article are also part of the collection, participants may access bibliographic metadata such as the MeSH (Medical Subject Headings) terms created by the National Library of Medicine for PubMed.

Creation of Realistic Search Topics

Our goal in creating search topics for the ImageCLEF medical retrieval task has been to identify typical information needs for a variety of users. In the past, we have used search logs from a different medical websites to identify topics. This year again search topics were identified by surveying actual user needs. The starting point for this year's topics was a user study conducted at Oregon Health & Science University (OHSU) during early 2009. This study was conducted with 37 medical practitioners in order to understand their needs, both met and unmet, regarding medical image retrieval. During the study, participants were given the opportunity to use a variety of medical and general-purpose image retrieval systems, and were asked to report their search queries.

In total, the 37 participants used the demonstrated systems to perform a total of 95 searches using textual queries in English. We randomly selected 25 candidate queries from the 95 searches to create the topics for ImageCLEFmed 2009. We added to each candidate query 2 to 4 sample images from the previous collections of ImageCLEFmed, which represented visual "queries" for content-based retrieval. Additionally, we provided French and German translations of the original textual description for each topic. Finally, the resulting set of topics was categorized into three groups: 10 visual topics, 10 mixed topics, and 5 semantic topics. This classification was performed by the organizers based on their knowledge of the capabilities of visual and textual search techniques, prior experience with the performance of textual and visual systems at ImageCLEF medical retrieval task, and their familiarity with the test collection. The entire set of topics was finally approved by a physician. An example of a "visual" topic can be seen in Figure 1 while that of a "textual" topic is shown in Figure 2.

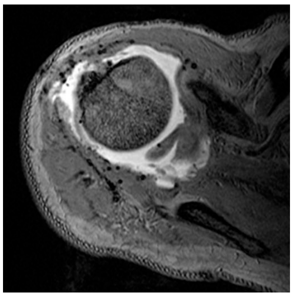


Figure 1: A "visual" topic: "MR Images of rotator cuff"

[1] http://goldminer.arrs.org/

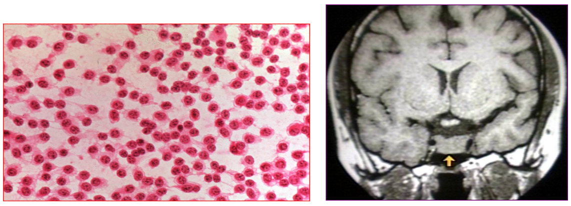


Figure 2: A "semantic" topic: "Pituitary adenoma"

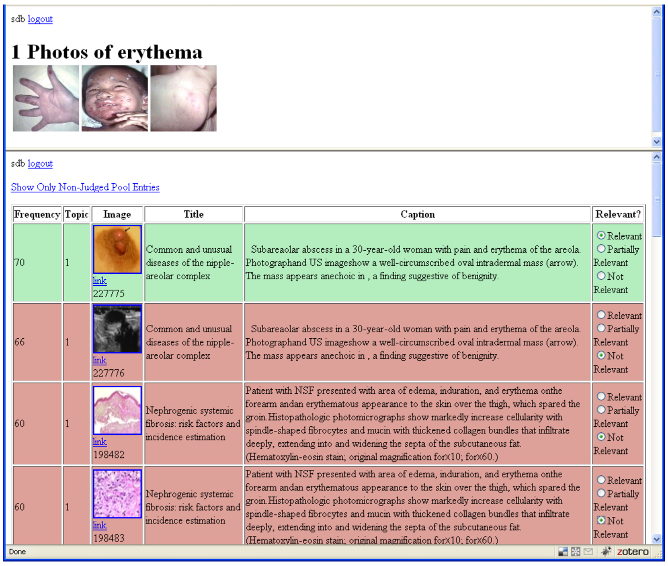
In 2009, we also introduced "case-based" topics as part of an exploratory task whose goal was to create search topics that are potentially more aligned with the information needs of an actual clinician in practice. These topics were meant to simulate the use case of a clinician who is diagnosing a difficult case, and has information about the patient's demographics, list of presenting symptoms, and imaging studies, but not the patient's final diagnosis. Providing this clinician with articles from the literature that deal with cases similar to the case (s)he is working on ("similar" based on images and other clinical data on the patient) could be a valuable aide to creating differential diagnosis or identifying treatment options.

These case-based search topics were created based on cases from the French teaching file Casimage, which contains cases (including images) from radiological practice. Ten cases were pre-selected, and a search with the final diagnosis was performed against the 2009 ImageCLEF data set to make sure that there were at least a few matching articles. Five topics were finally chosen. The diagnoses and all information about the chosen treatment were removed from the cases to simulate the aforementioned situation of a clinician dealing with a difficult diagnosis. However, in order to make the judging more consistent, the relevance judges were provided with the original diagnosis for each case.

Relevance Judgments

During 2008 and 2009, relevance judgments were made by a panel of clinicians using a web-based interface. Due to the infeasibility of manually reviewing 74,900 images for 30 topics, the organizers used a TREC-style "pooling" system to reduce the number of candidate images for each topic to approximately 1,000 by combining the top 40 images from each of the participants' runs. Each judge was responsible for between three to five topics, and sixteen of the thirty topics were judged multiple times (in order to allow evaluation of inter-rater agreement).

For the image-based topics, each judge was presented with the topic as well as several sample images as shown in Figure 3. For the case-based topics, the judge was shown the original case description and several images appearing in the original article's text. Besides a short description for the judgments, a full document was prepared to describe the judging process, including what should be regarded as relevant versus non--relevant. A ternary judgment scheme was used, wherein each image in each pool was judged to be ``relevant'', ``partly relevant'', or ``non-relevant''. Images clearly corresponding to all criteria were judged as ``relevant'', images whose relevance could not be safely confirmed but could still be possible were marked as ``partly relevant'', and images for which one or more criteria of the topic were not met were marked as ``non-relevant''. Judges were instructed in these criteria and results were manually verified during the judgment process.



sdb logout

1 Photos of erythema

sdb logout

Show Only Non-Judged Pool Entries

Frequency	Topic	Image	Title	Caption	Relevant?
70	1	link 227775	Common and unusual diseases of the nipple-areolar complex	Subareaolar abscess in a 30-year-old woman with pain and erythema of the areola. Photographand US imageshow a well-circumscribed oval intradermal mass (arrow). The mass appears anechoic in , a finding suggestive of benignity.	● Relevant ○ Partially Relevant ○ Not Relevant
66	1	link 227776	Common and unusual diseases of the nipple-areolar complex	Subareaolar abscess in a 30-year-old woman with pain and erythema of the areola. Photographand US imageshow a well-circumscribed oval intradermal mass (arrow). The mass appears anechoic in , a finding suggestive of benignity.	○ Relevant ○ Partially Relevant ● Not Relevant
60	1	link 198482	Nephrogenic systemic fibrosis: risk factors and incidence estimation	Patient with NSF presented with area of edema, induration, and erythema onthe forearm andan erythematous appearance to the skin over the thigh, which spared the groin.Histopathologic photomicrographs show markedly increase cellularity with spindle-shaped fibrocytes and mucin with thickened collagen bundles that infiltrate deeply, extending into and widening the septa of the subcutaneous fat. (Hematoxylin-eosin stain; original magnification forX10; forX60.)	○ Relevant ○ Partially Relevant ● Not Relevant
60	1	link 198483	Nephrogenic systemic fibrosis: risk factors and incidence estimation	Patient with NSF presented with area of edema, induration, and erythema onthe forearm andan erythematous appearance to the skin over the thigh, which spared the groin.Histopathologic photomicrographs show markedly increase cellularity with spindle-shaped fibrocytes and mucin with thickened collagen bundles that infiltrate deeply, extending into and widening the septa of the subcutaneous fat. (Hematoxylin-eosin stain; original magnification forX10; forX60.)	○ Relevant ○ Partially Relevant ● Not Relevant

Done

Figure 3: Web interface used for relevance judgments

As mentioned, we had sufficient judges to perform multiple judgements on many topics, both image-based and case-based. Inter-rater agreement was assessed using the kappa metric, given as:

$$\kappa = \frac{P(O)-P(E)}{1-P(E)} \qquad (1)$$

where $P(O)$ is the observed agreement between judges, and $P(E)$ is the expected (random) agreement. It is generally accepted that a $\kappa<0.7$ is good and sufficient for an evaluation. The score is calculated using a 2x2 table for the relevances of images or articles. These were calculated using both "lenient" and "strict" judgment rules. Under the lenient rules, "partly relevant" judgment was counted as "relevant"; under strict rules, "partly relevant" judgments were considered to be "not-relevant".

In general the agreement between the judges was fairly high (with a few exceptions), and our 2009 overall average κ is similar to that found during other evaluation campaigns.

Participation

For the medical retrieval task, the participation remained similar to the previous year with 37 registrations. 17 of the participants submitted results to the tasks. We had six first-time participants in 2009, which we consider to be a very positive development.

A total of 124 valid runs were submitted, 106 of which were submitted for the image-based topics, while 18 were submitted for the case-based topics. The number of runs per group was limited to ten per subtask and case-based and image-based topics were seen as separate subtasks in this view. Participants were requested to provide information about each run that they submitted. Runs could be classified as "textual", "visual" or "mixed" depending on the type of search engine used. They could also be classified as automatic, manual or feedback depending on the level of user interaction. The number of runs by run type can be seen in Table 1 below.

Table 1 – Number of runs by run type

Number of Runs		Run Type		
		automatic	feedback	manual
Retrieval Type	Textual	52	7	
	Visual	15	1	
	Mixed	25	3	2
	N/A	1		

Results

The metrics used to evaluate the runs include the mean averge precision (MAP), early precision (e.g. P@5, p@10) and bpref), measures that have historically been used for TREC and other challenge evaluations [12]. As was the case in the recent past, the focus of many participants in this year's ImageCLEF was primarily on text-based retrieval methods (as opposed to visual techniques). Almost half the runs submitted were automatic and textual. The increasingly semantic topics, combined with a database containing high-quality annotations, produced an evaluation environment better-suited for text-based image retrieval, and this fact was not lost on the participants. Only a few participants submitted visual runs, and those runs that were submitted were small in number and generally performed poorly, as can be seen from the average of the MAPs of the runs in table 2.

Table 2 – Mean average precision by run type

Average MAP		Run Type		
		automatic	feedback	manual
Retrieval Type	Textual	0.27	0.26	
	Visual	0.01	0.01	
	Mixed	0.20	0.26	0.19

Mixed-media runs performed similarly to textual runs in terms of mean average precision. That said, mixed runs that effectively combined visual and textual retrieval approaches typically outperformed the corresponding purely textual runs when considering metrics such as early precision, as can be seen in Table 3 where some mixed automatics runs demonstrated high early precision.

Case-based topics were introduced for the first time, and only a few groups participated. Runs submitted for case-based topics performed slightly worse than those submitted for image-based topics.

Table 3 – Maximum precision@5 by run type

Maximum P@5		Run Type		
		automatic	feedback	manual
Retrieval Type	Textual	0.73	0.61	
	Visual	0.09	0.06	
	Mixed	0.71	0.74	0.62

A kappa analysis between several relevance judgments for the same topics showed that there were differences between judges but that agreement was generally high. There were, however, a few judges that had significant disagreements with other judges. Additionally, feedback that we received from the judges indicated that the level of expertise of the judge in the specific area being searched affects their leniency with the relevance judgment process. The relevance judgments from judges with markedly different opinions were not used for calculating the final results. Interestingly, as has been found in the text retrieval domain, the overall rankings of the systems remain relatively stable even with using relevance judgments from different judges. However, the topic of relevance judging and the role of the judge (student, resident, general practitioner, expert radiologist, etc.) while evaluating the relevance of an image is of significant interest to us and one that we are investigating further.

Very few participants submitted interactive and manual runs as most participants seem to prefer batch processing with automatic text-based approaches, leading to primarily system-based evaluation. However, the role of the user in the retrieval process is important and we continue to encourage participants to introduce interactivity into their search systems and runs.

Conclusions

The ImageCLEF medical image retrieval campaign have been quite successful in attracting international researchers by providing a test collection that can be used to evaluate the performance of both text-based and content-based image retrieval systems. The test collection has grown from 8,000 images in 2004 to over 74,900 images in 2009. The participants in ImageCLEF have had interesting but diverse approaches to the addressing the problem of effective image retrieval in the medical domain. However, the collaborative nature of the forum and the annual workshops have fostered a community of participants that have willingly shared their techniques and even resources to the common goal of improving access to clinical images.

This work has some limitations. First, like all test collections, the topics were artificial. However, since they grew out of the results of a user study, we feel that they are reasonable and valid examples of clinician information needs and language use. Another limitation is the pools uses for relevance judgments reflect the runs submitted by the participants. Images that may have been retrieved by other techniques or were not the top hits would not be evaluated.

Going forward, we plan to expand the case-based topics as we believe that they more closely simulate the experiences of real user. We will continue to encourage participants to improve their multimodal techniques by making available the best visual and textual runs from past years in an effort to identify optimal ways of combining them. We are also continuing our user studies to better understand the needs of real users and to assess the validity of the performance measures used in these evaluation campaigns. The role of the user in assessing relevance continues to be of interest to us.

Acknowledgements

This work was supported in part by a supplement to National Science Foundation (NSF) grant ITR-0325160, National Library of Medicine Training grant (2T15 LM07088), the Swiss National Funds (grant 200020-118638/1) and the BeMeVIS project of the HES-SO. Instructions for obtaining the data described in this paper can be obtained from the ImageCLEFmed website (http://www.imageclef.org). The authors thank the Radiological Society of North America for contributing the use of images published in the journals *Radiology* and *RadioGraphics*.

References

[1] Tagare HD, Jaffe CC, Duncan J. Medical image databases: A content-based retrieval approach. J Am Med Inform Assoc 1997;4(3):184-98.

[2] 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.

[3] Muller H, Despont-Gros C, Hersh W, Jensen J, Lovis C, Geissbuhler A. Health care professionals' image use and search behaviour. Proceedings of Medical Informatics Europe (MIE 2006), Maastricht, Netherlands 2006:24-32.

[4] A qualitative task analysis of biomedical image use and retrieval; Imageclef/MUSCLE workshop on image retrieval evaluation, vienna, austria. 2005d.

[5] Lieberman DA, Holub J, Eisen G, Kraemer D, Morris CD. Prevalence of polyps greater than 9 mm in a consortium of diverse clinical practice settings in the united states. Clinical Gastroenterology and Hepatology 2005;3(8):798-805.

[6] Orel SG, Kay N, Reynolds C, Sullivan DC. Radiology 1999;211(3):845-50.

[7] Freer TW, Ulissey MJ. Radiology 2001;220(3):781-6.

[8] Kapoor V, McCook BM, Torok FS. Radiographics 2004;24(2):523-43.

[9] Hersh WR, Müller H, Jensen JR, Yang J, Gorman PN, Ruch P. Advancing biomedical image retrieval: Development and analysis of a test collection. J Am Med Inform Assoc 2006;13(5):488-96.

[10]Müller H, Kalpathy-Cramer J, Kahn CE Jr, Hatt W, Bedrick S, Hersh W, Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task, Springer Lecture Notes in Computer Science 5706, pages 500-510, 2009.

[11]Kahn CE Jr, Thao C. GoldMiner: a radiology image search engine, AJR Am J Roentgenol, 188(6):1475–1478, June 2007.

[12]Voorhees EM, Harman DK, eds. TREC: Experiment and evaluation in information retrieval, Cambridge, MA: MIT, 2005

Address for correspondence

kalpathy@ohsu.edu