Bulletin, February/March 2007

Special Section

Image Retrieval in Medicine: The ImageCLEF Medical Image Retrieval Evaluation

by William Hersh and Henning Müller

William Hersh is affiliated with the Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon. He can be reached by email at hersh<at>ohsu.edu. Henning Müller is with the Section of Medical Informatics, University & Hospitals of Geneva, Geneva, Switzerland. She can be reached by email at henning.Mueller<at>sim.hcuge.ch.

Many areas of medicine are highly visually oriented, such as radiology and dermatology, yet there is surprisingly little research that has been done investigating how medical personnel use and find images, especially in digital format. In particular, image retrieval is far less developed in medicine than other areas of information retrieval (IR), such as literature searching and consumer health information access. Whereas many Internet users, from laypersons to biomedical professionals, perform text searching routinely, few search for images on a regular basis. This paper describes the ImageCLEF medical image retrieval challenge evaluation, which has developed test collections for system-oriented evaluation of image retrieval systems and algorithms. ImageCLEF is part of the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), a challenge evaluation that operates on an annual cycle using test collections. The first running of the medical image retrieval track was in 2004; since then, there have been regular offerings of the track in 2005 and 2006 with more planned for 2007 and beyond.

Image retrieval systems generally take two approaches to indexing and retrieving data. One is to index and retrieve textual annotations associated with images. Several commercial systems, such as Google Images (images.google.com) and Flickr (www.flickr.com), employ this approach. A second approach, called visual or content-based, employs image processing techniques to features in the images, such as color, texture, shape and segmentation.

Each approach to indexing and retrieval of images has its limitations. Researchers such as Joergensen and Le Bozec have described the limitations of purely textual indexing of images for retrieval, such as the inability to capture synonymy, conceptual relationships or larger themes underlying their content. One effort to improve the discipline of image indexing in medicine has been the Health Education Assets Library (HEAL) project, which aims to standardize the metadata associated with all medical digital objects, but its adoption remains modest at this time. Visual indexing and retrieval also have their limitations. In a recent review article of content-based image retrieval applied in biomedicine, Henning Müller and her colleagues noted that image processing algorithms to automatically identify the conceptual content of images have not been able to achieve the performance of IR and extraction systems applied to text.

Another problem with all image retrieval research has been the lack of robust test collections and realistic query tasks that allow comparison of system performance. A few initiatives exist for certain types of visual information retrieval (for example, TRECVid for retrieval of video news broadcasts), but none have focused on the biomedical domain. [For a discussion of TRECVid please see the article by Alan Smeaton elsewhere in this special section of the *Bulletin*.] Test collections have been used extensively to evaluate IR systems in biomedicine. A number of test collections have been developed for document retrieval in the clinical domain. More recently, focus has shifted to the biomedical research domain in the Text Retrieval Conference (TREC) Genomics Track.

ImageCLEF Medical Image Retrieval Test Collection

In standard system-oriented IR research, test collections consist of three components: content items that actual users are interested in retrieving; topics that represent examples of their real information needs; and relevance judgments that denote which content is relevant (i.e., should be retrieved) to which topic. For the content of our collection, we set out to develop one of realistic size and scope. We aimed to use collections that already existed and did not intend to modify them (e.g., improve them with better metadata) other than organizing them into a common structure for the experiments. As such, we used the original annotations, which were not necessarily created for image retrieval. We obtained four collections of images that varied in both subject matter and existing annotation. Consistent with the nature of CLEF, they were annotated in different languages.

Tables 1 and 2 describe the collections used in the 2005 and 2006 ImageCLEF medical image retrieval task. The Casimage collection consists of clinical case descriptions with multiple association images of a variety of types, including radiographs, gross images and microscopic images. While most of the case descriptions are in French, some are in English and a small number contain both languages. The Mallinckrodt Institute of Radiology (MIR) collection consists of nuclear medicine images, annotated around cases in English. The Pathology Education Instructional Resource (PEIR) is a large collection of pathology images (gross and microscopic) that are tagged using the HEAL format in English. PathoPIC is another pathology collection that has all images annotated in longer German and shorter English versions.

Table 1 - Collection origin and types for ImageCLEFmed 2005 library.

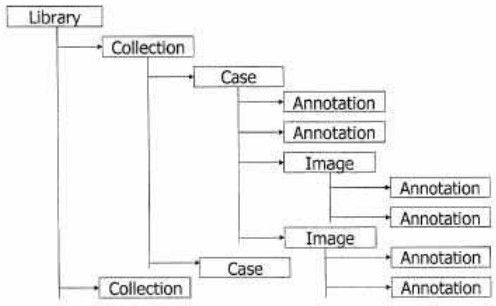
Collection Name	Image Type(s)	Annotation Type(s)	Original URL
Casimage	Radiology and pathology	Clinical case descriptions	www.casimage.com/
Mallinckrodt Institute of Radiology (MIR)	Nuclear medicine	Clinical case descriptions	http://gamma.wustl.edu/ home.html
Pathology Education Instructional Resource (PEIR)	Pathology and radiology	Metadata records from HEAL database	http://peir.path.uab.edu/; www.healcentral.org/
PathoPIC	Pathology	Image description – long in German, short in English	http://alf3.urz.unibas.ch/ pathopic/e/intro.htm

Table 2 - Items and sizes of collections in ImageCLEFmed 2005 library.

Collection Name	Cases	Images	Annotations	Annotations by Language	File Size (tar archive)
Casimage	2076	8725	2076	French - 1899 English - 177	1.28 GB
MIR	407	1177	407	English - 407	63.2 MB
PEIR	32319	32319	32319	English - 32319	2.50 GB
PathoPIC	7805	7805	15610	German - 7805 English 7805	879 MB

The images and annotations are organized into a single library, which is structured as shown in Figure 1. The entire library consists of multiple *collections*. Each collection is organized into *cases* that contain one or more related *images* and may include one or more *annotations*, which can consist of metadata and/or a textual annotation. Each individual image may also have annotations. (In these collections, annotations occur either at the case or image level, but not both.)

Figure 1. Structure of test collection library.



For the 2005 track, we developed 25 topics for the test collection consisting of a textual information needs statement and one to three index images. For the 2006 track, another 30 similar topics with index images were developed. In both years, the topics were classified based on *topic categories* reflecting whether they were more amenable to retrieval by visual, textual or mixed algorithms. Eleven topics were

visually oriented (1-11), 11 topics were mixed (12-22) and three topics

American Society for Information Science and Technology
1320 Fenwick Lane, Suite 510, Silver Spring, Maryland 20910, USA
Tel. 301-495-0900 / Fax: 301-495-0810 / E-mail: asis@asis.org

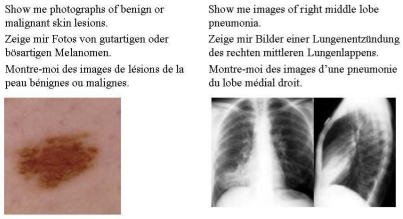


Figure 2. Example of visually (left) and semantically (right) oriented topics from the test collection.

The experimental process for each year was conducted by providing each participating research group with the collection and topics. They then carried out *runs*, consisting of the same retrieval approach applied to all the topics. Groups were allowed to submit as many runs as they desired, but were required to classify them based on whether the run used manual modification of topics (automatic vs. manual vs. interactive) and whether the system used visual retrieval, text retrieval or both (visual vs. textual vs. mixed).

The final component of the test collection was the relevance judgments. As with most challenge evaluations, the collection was too large to judge every image for each topic. So as is commonly done in IR research, we developed pools of images for each topic consisting of the top-ranking images in the runs submitted by participants. The relevance assessments were performed by physicians who were also graduate students in OHSU biomedical informatics program. The number of topics assessed by each judge varied depending on how much time they had available, but varied from four to eight topics. Some judges also performed duplicate assessment of other topics.

Once the relevance judgments were done, we could then calculate the results of the experimental runs submitted by ImageCLEF participants. We used the trec_eval evaluation package (available from trec.nist.gov), which takes the output from runs (a ranked list of retrieved items for each topic) and a list of relevance judgments for each run (called *qrels*) to calculate a variety of relevance-based measures on a per-topic basis that are then averaged over all the topics in a run. The trec_eval package includes MAP (mean average precision, binary preference (B-Pref) [19], precision at the number of relevant images (R-Prec) and precision at various levels of output from 5 to 1000 images (e.g., precision at 5 images, 10 images, etc., up to 1000 images). We also released the judgments so participants could perform additional runs and determine their results.

Results

A complete description of the results of the 2005 and 2006 tracks is beyond the scope of this paper. Overall summaries of the results can be obtained in the track overview papers from 2005 and 2006. Similar to many challenge evaluations, the primary evaluation measure has been MAP. A number of interesting findings can be briefly summarized:

- **A variety of different approaches produce comparable MAP.**
- **Image retrieval by textual methods (for example, searching over the text of annotations) appears to be more robust than visual approaches, as the textual results degrade less poorly for topics amenable to visual retrieval than visual results do for topics amenable to textual retrieval. In other words, visual retrieval alone fares poorly for topics that are not explicitly amenable to visual techniques.**
- **Common to most IR test collections, results vary widely by topic.**
- **Also common to most challenge evaluations, there are substantial variations in relevance judgments.**
- **MAP may not be the ideal retrieval measure for the image retrieval task, since it tends to be influenced by the full spectrum of all retrieval images. As most users may only desire a few quality images from their searches, a more precision-oriented measure may better reflect users' situations.**

Discussion

The ImageCLEF medical image retrieval tasks have developed a large test collection and attracted research groups who have brought a diverse set of approaches to a common goal of effective image retrieval. Not only did these groups learn from their own experiments,

but other researchers will subsequently be able to improve image retrieval by using the test collection that will now be available.

This work has some limitations. First, like all test collections, the topics were artificial and may not be realistic or representative of how real users would employ an image retrieval system. Likewise, the annotation of the images may not be representative of how image annotation is done generally or represent best practice. And as with all test collections, the pools generated for relevance assessment only represent images retrieved by the techniques of the participating research groups. As such, there could be other retrieval techniques that would retrieve other images that may be relevant.

We have a variety of future plans, starting with ImageCLEF 2007. Our main plan is to enlarge the image collection and develop a new set of topics. Additional future plans include carrying out user experiments on two fronts: one to see how users interact and perform with real systems using this collection and also to better elicit user information needs to develop even more realistic topics. With these experiments, we will also aim to assess performance measures to determine which are more representative for real tasks. This will be done by assessing which measures are best associated with the information needs of real users in specific searching situations.

We have created a large image retrieval test collection that will enable future research in this area of growing importance to biomedicine. We have also identified some observations that warrant further study to optimize the performance of such systems. The growing prevalence of images used for a variety of biomedical tasks makes imperative the development of better image retrieval systems and an analysis of how they are used by real users. The ImageCLEF test collections, with both system-oriented and user-oriented research around them, will contribute to further advances in this active research area.

Acknowledgements
This work was supported by a supplement to National Science Foundation (NSF) grant ITR-0325160. We also acknowledge the European Commission IST projects program in facilitating this work (through the Semantic Mining Network of Excellence, grant 507505) and the Swiss National Funds (grant 205321-109304/1). Instructions for obtaining the data described in this paper can be obtained from the ImageCLEFmed website (http://ir.ohsu.edu/image/).

For Further Reading

TREC - General
Voorhees, E.M. & Harman, D.K. (Eds). (2005, September). Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

TREC – Selected Topics
Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., et al. (2006). The CLEF 2005 Cross-Language Image Retrieval Track. In C. Peters, F. Gey, J. Gonzalo, H. Mueller, G. Jones, M. Kluck, B. Magnini, & M. Rilke (Eds.). Accessing multilingual information repositories. 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers [pp, 535-557]. [Lecture Notes in Computer Science, v. 4022]. Berlin, Springer-Verlag.

Hersh, W.R. (2001). Interactivity at the Text Retrieval Conference (TREC). Information Processing and Management, 37, 365-366.

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D. F. (2006, March 13). Enhancing access to the bibliome: The TREC 2004 Genomics Track. Journal of Biomedical Discovery and Collaboration, 1, no.3. Retrieved December 31, 2006, from www.j-biomed-discovery.com/content/1/1/3.

Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E.. & Hersh, W. (In press). Overview of ImageCLEFmed 2006 medical retrieval and information tasks. In Evaluation of multi-lingual and multi-modal information retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, 20-22 September, 2006, Berlin: Springer-Verlag.

Image Retrieval Technology
Candler, C. S., Uijtdehaage, S. H., & Dennis, S. E. (2003). Introducing HEAL: The Health Education Assets Library. Academic Medicine, 78, 249-253.

Jorgensen, C. (1999). Retrieving the unretrievable in electronic imaging systems: Emotions, themes, and stories. In B.E. Rogowitz & Thrasyvoulos N. Pappas (Eds.) Human Vision and Electronic Imaging IV (pp. 348-355). [Proceedings of the SPIE—International Society for

Optical Engineering, 3644]. Bellingham, WA: SPIE.

LeBozec, C., Zapletal, E., Jaulent, M. C., Heudes, D., & Degoulet, P. (2000). Toward content-based image retrieval in a HIS-integrated PACS. In Proceedings of the AMIA 2000 Annual Symposium, Los Angeles, CA, November 4-8, 2000 (pp. 477-481). Philadelphia: Hanley & Belfus.

Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based information retrieval systems in medical applications – Clinical benefits and future directions. International Journal of Medical Informatics, 73, 1-23.

Rui, Y, Huang, T. S., & Chang, S. F. (1999). Image retrieval: Past, present and future. Journal of Visual Communication and Image Representation, 10, 39-62.

Image Retrieval Test Collections
Horsch, A., Prinz, M. Schhneider, S., Sipilia, O., Spinnier, K., Vallee, J. P., et al. (2004). Establishing an international reference image database for research and development in medical image processing. Methods of Information in Medicine, 43, 409-412.

Hersh, W. R., Buckley, C., Leone, T., & Hickam, D. H. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In W. B. Croft & C. J. van Rijsbergen, Eds. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 3-6 July, 1994. (pp. 192-201). London: Springer-Verlag.

Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallee, J. P., & Ratib, O. (2004). Casimage project: a digital teaching files authoring environment. Journal of Thoracic Imaging. 19, 105-108.

Wallis, J. W., Miller, M. MI, Miller, T. R., & Vreeland, T.H. (1995). An Internet-based nuclear medicine teaching file. Journal of Nuclear Medicine, 36, 1520-1527.

Jones, K. N., Kreisle, R., Geiss, R., Holliman, J., Lill P., & Anderson, P. G. (2002). Group for Research in Pathology Education “online” resources to facilitate pathology instruction. Archives of Pathology and Laboratory Medicine, 126, 346-350.

Glatz-Krieger, K., Glatz, D., Gyusel, M., Dittler, M., & Mihatsch, M. J. (2003). Web-based learning tools in pathology [German]. Pathologe, 24, 394-399.