ImageCLEFmed: Developing a Test Collection for Biomedical Image Retrieval

William Hersh, MD[1], Jeffery Jensen, BS[1], Henning Müller, PhD[2],
Paul Gorman, MD[1], Patrick Ruch, PhD[2]
[1]Oregon Health & Science University, Portland, OR, USA
[2]University & Hospitals of Geneva, Geneva, Switzerland

The goal of our EU-NSF project was to improve image retrieval in the biomedical domain through the development of a robust collection. Our work succeeded in developing a high-quality test collection that enabled the image retrieval research of ourselves and others. It has also cemented a collaboration that will likely continue beyond the end of the specific funding.

Introduction

Image retrieval is a poor stepchild to other forms of information retrieval (IR). Whereas a broad spectrum of Internet users, from laypeople to biomedical professionals, perform text searching routinely [1], fewer (though a growing number) search for images on a regular basis. While development of image retrieval approaches and systems began as a research field 20 years ago, progress has been stalled for multiple reasons. One problem is the inability of image processing algorithms to automatically identify the content of images in the manner that information retrieval and extraction systems have been able to do so with text [2]. A second problem is the lack of robust test collections and realistic query tasks that allow comparison of system performance [2, 3].

The lack of useful test collections is one of the motivations for the ImageCLEF initiative, which aims to build test collections for image retrieval research. ImageCLEF is a part of the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), a challenge evaluation for information retrieval from diverse languages [4]. CLEF itself is an outgrowth of the Text Retrieval Conference (TREC, trec.nist.gov), a forum for evaluation of text retrieval systems. TREC and CLEF operate on an annual cycle of test collection development and distribution, followed by a conference where results are presented and analyzed.

The goals of TREC and CLEF are to build realistic test collections that simulate retrieval tasks and enable researchers to assess the performance of their systems and compare their results with others [5]. The goal of test collection construction is to assemble a large collection of *content* (documents, images, etc.) that resemble collections used in the real world. Builders of test collections also seek a sample of realistic *tasks* to serve as *topics* that can be submitted to systems as *queries* to retrieve content. The final component of test collections is relevance judgments that determine which content is relevant to each topic. A major challenge for test collections is to develop a set of realistic topics that can be judged for relevance to the retrieved items. Such benchmarks are needed by any development team (for research or business purposes) in order to evaluate the effectiveness of new tools.

The first author of this paper has worked for many years in the development of test collections, particularly in the biomedical domain [6, 7]. He is the chair of the TREC Genomics Track, which is devoted to evaluation document retrieval in the genomics domain [8]. His main National Science Foundation (NSF) Information Technology Research (ITR) grant funds the TREC Genomics Track and research surrounding it. Likewise, the European Union (EU) partner SemanticMining Network of Excellence is experienced in developing test

collection tools as well [9]. The supplement for EU-US collaboration funds the development of the ImageCLEF medical image test collection within the CLEF 2005 framework.

The ImageCLEF 2005 medical image retrieval test collection

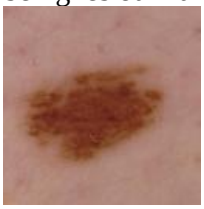
Our goal in developing the image test collection for ImageCLEF was to represent various "axes" of image retrieval searches [10]. One axis was whether the search was visual (aiming to find images based on features of one or more index images), semantic (aiming to find images about a subject, such as a disease or anatomical location), or had aspects of both. Other axes focused on content-related areas:

- Imaging modality: One or more of radiograph, CT, MRI, nuclear medicine, ultrasound, photograph, illustration, gross pathology, micro pathology
- Anatomical location: One or more of lungs, bone, heart, brain, abdomen (liver, stomach, intestines, kidney), blood, skin
- Diagnoses: tuberculosis, fracture, infarction, leukemia, genetic disorder
- Findings: enlargement, tumor, infiltrate, lesion

We developed 25 topics based on these axes. Eleven topics were visually oriented, three topics were semantically oriented, and eleven topics were mixed. Each topic had one or more associated index images. Because the images were variously annotated in English, German, or French, the topics were translated into all three languages. (See Figure 1 for examples.)

The images and annotations were organized into a library, which was structured as shown in Figure 2. The entire library consists of multiple *collections*. Each collection is organized into *cases* that represent a group of related *images* and *annotations*. Each case consists of a group of images and an optional annotation. Each image is part of a case and has optional associated annotations, which consist of metadata and/or a textual annotation. Tables 1 and 2 describe the collections used in the 2005 task.

Show me photographs of benign or malignant skin lesions.
Zeige mir Fotos von gutartigen oder bösartigen Melanomen.
Montre-moi des images de lésions de la peau bénignes ou malignes.



Show me images of right middle lobe pneumonia.
Zeige mir Bilder einer Lungenentzündung des rechten mittleren Lungenlappens.
Montre-moi des images d'une pneumonie du lobe médial droit.

Figure 1 - Example of visually (left) and semantically (right) oriented topics from the test collection

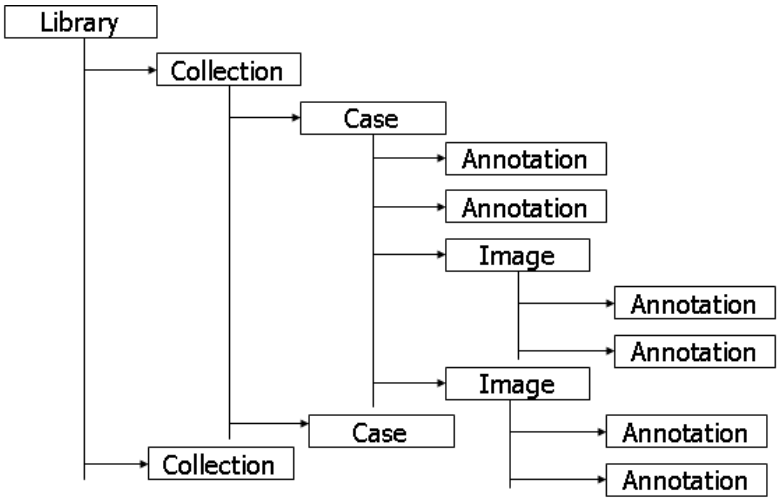


Figure 2 - Structure of test collection library.

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

Collection Name	Image Type(s)	Annotation Type(s)	Original URL
Casimage [11]	Radiology and pathology	Clinical case descriptions	http://www.casmiage.com/
Mallinckrodt Institute of Radiology (MIR) [12]	Nuclear medicine	Clinical case descriptions	http://gamma.wustl.edu/home.html
Pathology Education Instructional Resource (PEIR) [13]	Pathology and radiology	Metadata records from HEAL database	http://peir.path.uab.edu
PathoPIC [14]	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 final component of the test collection was the relevance judgments. As with most challenge evaluations, the collection was too large to judge every image relative to each topic. So as is commonly done in IR research, we developed "pools" of images ranked highest in the runs submitted by all the participants in the experiments. There were 13 research groups who took part in the task and submitted a total of 134 official runs. From

these runs, we developed pools of about 800 images each to be judged for every topic. The judgments were done by eight physicians recruited from the student body of the graduate program in biomedical informatics at Oregon Health & Science University. Each image was judged at least once, and several thousand were judged more than once to assess interjudge reliability (results forthcoming).

Once the relevance judgments were done, we could then calculate the results of the experimental runs submitted by ImageCLEF participants. We were also able to release the judgments so they could perform additional runs and determine their results. As is done in most IR evaluations, the primary evaluation measure was mean average precision (MAP). This is an aggregate measure that balances the values of recall and precision. It is calculated by taking the average precision at each point of recall (relevant item retrieved) for a given topic, and then taking the mean over all of the topics.

Results

To determine the relative efficacy of different techniques, we classified the different experimental runs into categories. The first axis of the category was whether the run had any human intervention in construction of the query, i.e., automatic (A) vs. manual (M). The second axis was whether the run used visual retrieval features (V), textual retrieval features (T), or a mixture of the two (M). Table 3 shows the best performing runs for each of the categories based on these two axes. Figure 3 shows the results graphically for each of the categories. One obvious conclusion is that visual retrieval techniques performed poorly with semantic queries, bringing down their overall performance.

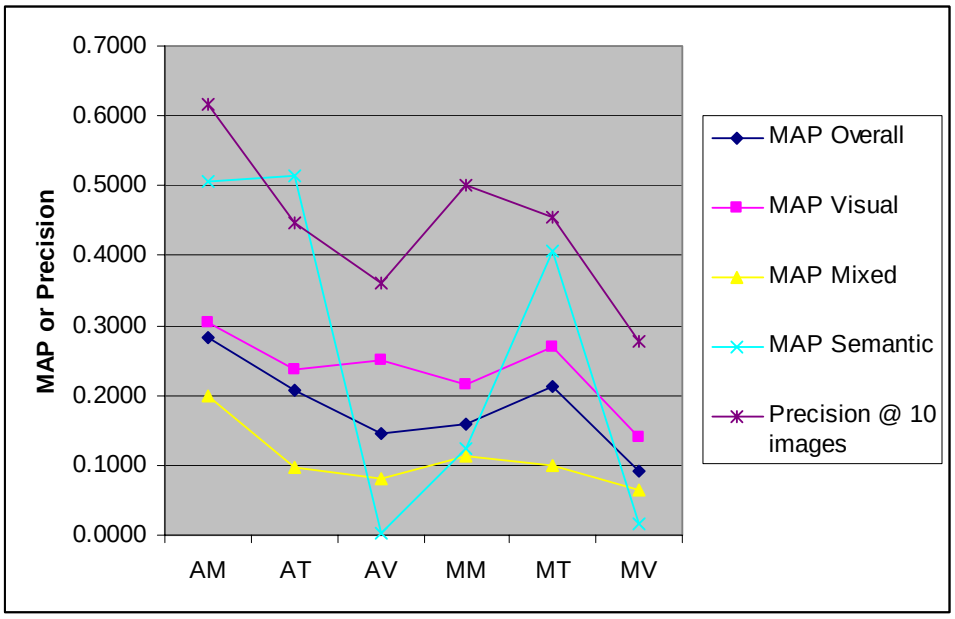


Figure 3 - Results Best results (based on MAP for all topics) in each category, with run identifier; MAP for all topics, visual topics only, mixed topics only, and semantic topics only; and precision at 10 images.

Logistical Issues of the Collaboration

This collaboration was undertaken with the SemanticMining Network of Excellence and involved only academic partners. We were able to carry out this project with few logistical problems. This was probably because our project took place in the context of the larger CLEF initiative. The work went extremely well and we reached all the goals we set for ourselves. The only reason why the work might not have otherwise succeeded was that US and EU agencies typically do not fund trans-Atlantic collaborations. But because they did this type of funding, we were able to complete our work successfully. Without such trans-national funding, we would not have been able to carry out this work. As such, we would recommend NSF and EU authorities consider funding collaborative efforts like this that would not otherwise be possible due to usual policies of research funding not crossing national borders.

Fortunately there are no intellectual property issues involved in this collaboration, because the test collection we have produced will be available to anyone who signs a data usage form, which basically says they will not commercialize the collection or put it on a public Web site where others can get it. Providing unfettered access to a test collection is a key factor in ensuring its widespread use for additional research.

Future Plans

Because the results of ImageCLEF just became available at the deadline for this paper, only a modest amount of data analysis could be done. Future analysis will be undertaken soon, however, aiming to determine what approaches lead to best retrieval for various types of topics. The development of this test collection will also likely lead to future research and improved image retrieval systems in the future.

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 (NoE 507505).

References

1. Madden M and Rainie L, *America's Online Pursuits: The changing picture of who's online and what they do*. 2003, Pew Internet & American Life Project: Washington, DC, http://www.pewinternet.org/pdfs/PIP_Online_Pursuits_Final.PDF.
2. Müller H, et al., *A review of content-based image retrieval systems in medical applications-clinical benefits and future directions.* International Journal of Medical Informatics, 2004. 73: 1-23.
3. Horsch A, et al., *Establishing an international reference image database for research and development in medical image processing.* Methods of Information in Medicine, 2004. 43: 409-412.
4. Braschler M and Peters C, *Cross-language evaluation forum: objectives, results, achievements.* Information Retrieval, 2004. 7: 7-31.
5. Sparck-Jones K, *Reflections on TREC.* Information Processing and Management, 1995. 31: 291-314.

6. Hersh WR, et al. *OHSUMED: an interactive retrieval evaluation and new large test collection for research*. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1994. Dublin, Ireland: Springer-Verlag. 192-201.
7. Hersh WR, *Interactivity at the Text Retrieval Conference (TREC).* Information Processing and Management, 2001. 37: 365-366.
8. Hersh W, et al. *TREC 2004 genomics track overview*. *The Thirteenth Text Retrieval Conference (TREC 2004)*. 2004. Gaithersburg, MD: NIST. in press. http://trec.nist.gov/pubs/trec13/papers/GEO.OVERVIEW.pdf.
9. Clough P, Müller H, and Sanderson M, *The CLEF cross language image retrieval track (ImageCLEF) 2004*, in *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, Peters C, et al., Editors. 2004, Springer-Verlag: Heidelberg, Germany. 597-613.
10. Müller H, et al. *Evaluation axes for medical image retrieval systems - the ImageCLEF experience*. *ACM Multimedia*. 2005. Singapore. in press.
11. Rosset A, et al., *Casimage project: a digital teaching files authoring environment.* Journal of Thoracic Imaging, 2004. 19: 103-108.
12. Wallis JW, et al., *An Internet-based nuclear medicine teaching file.* Journal of Nuclear Medicine, 1995. 36: 1520-1527.
13. Jones KN, et al., *Group For Research In Pathology Education "online" resources to facilitate pathology instruction.* Archives of Pathology and Laboratory Medicine, 2002. 126: 346-350.
14. Glatz-Krieger K, et al., *Web-based learning tools in pathology [German]*. Pathologe, 2003. 24: 394-399.