Journal of Digital Imaging

The ImageCLEFmed Medical Image Retrieval Task Test Collection

William Hersh, MD,[1] Henning Müller,[2,3] and Jayashree Kalpathy-Cramer[1]

A growing number of clinicians, educators, researchers, and others use digital images in their work and search for them via image retrieval systems. Yet, this area of information retrieval is much less understood and developed than searching for text-based content, such as biomedical literature and its derivations. The goal of the ImageCLEF medical image retrieval task (ImageCLEFmed) is to improve understanding and system capability in search for medical images. In this paper, we describe the development and use of a medical image test collection designed to facilitate research with image retrieval systems and their users. We also provide baseline results with the new collection and describe them in the context of past research with portions of the collection.

KEY WORDS: Medical image retrieval, imageCLEF, test collection, recall, precision

INTRODUCTION

Images have a variety of uses in health care and biomedical research. Despite their widespread use, however, little is known about how those who use them search for and/or manage them. Two small analyses have found that the image use tends to be related to the "role" of the user, such as clinician, educator, and researcher[1,2]. As there are growing numbers of image collections and search interfaces proliferating on the World Wide Web as well as closed networks, we believe it is important to understand user needs as well as provide systems that meet those needs.

The goal of the ImageCLEF (www.imageclef.org) medical image retrieval task (ImageCLEFmed) is to improve understanding and system capability in search for medical images[3]. ImageCLEF is a part of the Cross-Language Evaluation Forum (CLEF, www.clef-campaign.org), an evaluation forum for information retrieval (IR) from diverse languages[4]. CLEF itself is an outgrowth of the Text Retrieval Conference (TREC, trec.nist.gov), an evaluation forum for general text retrieval systems[5].

CLEF and TREC build on the tradition of *challenge evaluations* that have been used historically to assess the performance of IR systems, where realistic *test collections* are developed that simulate real-world retrieval tasks and enable researchers to assess and compare system performance[6]. 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 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

[1]*From the Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd., BICC, Portland, OR 97239, USA.*

[2]*From the University Hospitals and University of Geneva, Geneva, Switzerland.*

[3]*From the Business Information Systems, University of Applied Sciences Western Switzerland, Sierre, Switzerland.*

Correspondence to: William Hersh, MD, Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd., BICC, Portland, OR 97239, USA; tel: +1-503-4944563; fax: +1-503-4944551; e-mail: hersh@ohsu.edu

Online publication 3 September 2008

doi: 10.1007/s10278-008-9154-8

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 researcher or developer in order to evaluate the effectiveness of new tools.

Challenge evaluations in IR usually employ test collections to measure how well systems or algorithms retrieve relevant items. The most commonly used evaluation measures are *recall* and *precision*. Often, there is a desire to combine recall and precision into a single aggregate measure. Although many approaches have been used for aggregate measures, the most frequently used one in TREC and CLEF has been the *mean average precision* (MAP)[7].

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[8,9]. More recently, focus has shifted to the biomedical research domain in the TREC Genomics Track[10]. Test collections are also used increasingly for image retrieval outside of medicine[11].

ImageCLEFmed has run for 3 years through the corresponding yearly cycles of CLEF (2005–2007). As with most text collections, we aimed to make the content and search topics as realistic as possible. From 2005 to 2007, ImageCLEF featured a medical retrieval task based around ad-hoc retrieval. The collection of images came from four sources initially, with two additional ones added in the third year. Each collection was used "as is"; that is, its annotations are used from the original source. This paper describes the recent effort by the project to consolidate the 3 years of test collections into a single collection that aims to provide a test bed for evaluating systems and algorithms that perform medical image retrieval. In the following sections, we describe the content, topics, relevance judgments, evaluation methods, baseline results, lessons learned, and future plans for the merged collection and our work.

CONTENT

The conceptual structure of the content of the ImageCLEFmed test collection is as follows. 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 (e.g., Health Education Assets Library [HEAL] tagging) and/or a textual annotation. All of the images and annotations are stored in separate files. An Extensible Markup Language (XML) file contains the connections between the collections, cases, images, and annotations. Figure 1 shows a graphical depiction of the library, while Figure 2 shows the XML metadata format.

The image library for ImageCLEFmed 2005 and 2006 consisted of the first four collections listed in Tables 1 and 2 (Casimage, MIR, PEIR, and PathoPIC). In 2007, we added the latter two collections listed in those tables (myPACS and CORI). Table 1 describes the image collections, their image and annotation types, and their origins, while Table 2 lists the numbers of images and annotations (including amounts in each language) as well as the archived file size. Figure 3 shows an example case from the Casimage collection, demonstrating how multiple different images and image types can be part of a case. However, note that the largest collection, PEIR, is not organized into cases per se (or, using our framework, has one image per case). The image library for the consolidated test collection will be the entire library, which is the same as that used for ImageCLEFmed 2007.

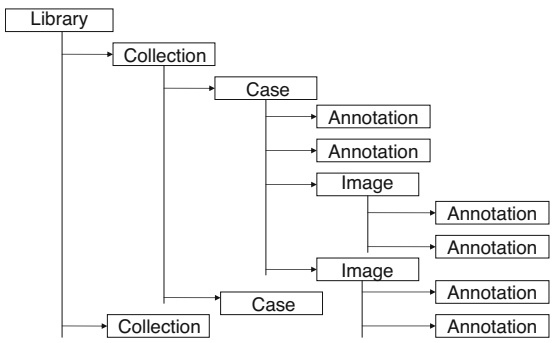


Fig. 1. Structure of the ImageCLEFmed test collection content.

```
<library>
  <collection>
    <name>name-text</name>
    <cases>
      <case>
        <id>identifier-text</id>
        <images>
          <image>
            <id>identifier-text</id>
            <imagefile>file-name-text</imagefile>
            <annotation lang=" ">file-name-text</annotation
            <annotation lang=" ">file-name-text</annotation>
          </image>
        <images>
        <annotation lang=" ">file-name-text</annotation>
        <annotation lang=" ">file-name-text</annotation>
      </case>
    </cases>
  </collection>
</library>
```

Fig. 2. Structure of ImageCLEFmed XML metadata format for the content.

TOPICS

A total of 85 topics were developed over 2005–2007 for ImageCLEFmed. Topics were generated from a variety of real-world Internet medical search engine logs. Each topic was provided with an information needs statement in English, French, and German, as well as one or more relevant *index images* for use by visual retrieval systems. We classified each topic as visual, textual, or mixed because we discovered early on that the results on different tasks varied by whether the topic was amenable to visual or textual retrieval.

There were 25 topics in 2005 and 30 each in 2006 and 2007. In each year, each topic was numbered from 1, i.e., 1–25 in 2005 and 1–30 in 2006 and 2007. In the consolidated test collection, the topics from 2005 were numbered 1–25, those from 2006 numbered 26–55, and those from 2007 numbered 56–85. A sample topic from the consolidated collection is shown in Figure 4.

RELEVANCE JUDGMENTS

Relevance judgments in ImageCLEFmed have been performed by physicians who are also students in the Oregon Health and Science University biomedical informatics graduate program. They were paid an hourly rate for their work. The pools for relevance judging were created by selecting the top ranking images from all submitted runs. The actual number selected from each run varied by year but was usually about 30–40, with the goal of having pools of about 800–1,200 images in size for judging. Judges were instructed to rate images in the pools as definitely relevant, partially relevant, or not relevant.

For the consolidated test collection, we needed to perform relevance judgments for the new 2007 images applied to the 2005 and 2006 topics since the new images had not been used in those years. We also rejudged several topics in totality after two people reviewed the original judgments and found them to be of poor quality. For all judging in 2007, judges were asked to adhere to the following instructions:

1. Note that a topic can refer to one or more of the following: (a) an imaging modality, (b) an anatomical location, (c) a view, and/or (d) a disease or finding. An image should only be considered relevant if it meets all the terms mentioned explicitly in the topic (i.e., should be an AND, not an OR). For instance, in the topic

Table 1. ImageCLEF Medical Image Retrieval Task (ImageCLEFmed) Image Collections, Image and Annotation Types, and their Origins

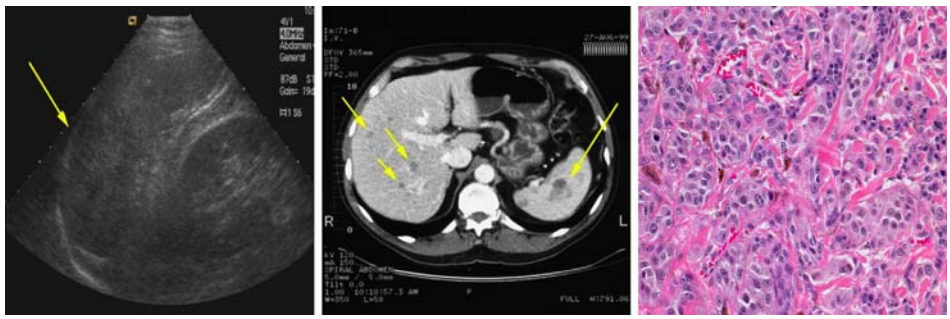
Collection name	Image type(s)	Annotation type(s)	Original URL
Casimage	Radiology and pathology	Clinical case descriptions	http://www.casimage.com/
Mallinckrodt Institute of Radiology (MIR)	Nuclear medicine	Clinical case descriptions	http://gamma.wustl.edu/home.html
Pathology Education Instructional Resource (PEIR)	Pathology and radiology	Metadata records from HEAL database	http://peir.path.uab.edu/
PathoPIC	Pathology	Image description—long in German, short in English	http://alf3.urz.unibas.ch/pathopic/e/intro.htm
MyPACS	Radiology	Clinical case descriptions	http://www.mypacs.net/
Clinical Outcomes Research Initiative (CORI) Endoscopic Images	Endoscopy	Clinical case descriptions	http://www.cori.org/

Table 2. ImageCLEF Medical Image Retrieval Task (ImageCLEFmed) Numbers of Images and Annotations (Including Amounts in Each Language) as Well as the Archived File Size

Collection Name	Cases	Images	Annotations	Annotations by Language	File Size (tar archive)
Casimage	2,076	8,725	2,076	French—1,899 English—177	1.28 GB
MIR	407	1,177	407	English—407	63.2 MB
PEIR	32,319	32,319	32,319	English—32,319	2.50 GB
PathoPIC	7,805	7,805	15,610	German—7,805 English—7,805	879 MB
myPACS	3,577	15,140	3,577	English—3,577	390 MB
CORI	1,496	1,496	1,496	English—1,496	34 MB
Total	47,680	66,662	55,485	French—1,899 English—45,781 German—7,805	5.15 GB

"CT liver abscess," only computed tomography (CT) scans showing a liver abscess should be considered relevant. Pathology or magnetic resonance imaging images of liver abscesses should not be considered relevant. Images of other abscesses should not be considered relevant. An X-ray image associated with an annotation that refers to a need for a CT scan in the future should not be considered relevant.

2. When a photograph is the desired imaging modality, i.e., it says "image of" or picture of," only photographic images should be considered relevant. Although, technically, microscopic images of histology/pathology may be considered to be photographs, in this context, they should not be considered relevant.
3. Pathology in the query refers to pathological images (microscopic/gross pathology), not the state of being abnormal.
4. Refer to the sample images provided with each topic for a better understanding of desired imaging modalities.



Images

Case annotation

ID : 4272
Description: A large hypoechoic mass is seen in the spleen. CDFI reveals it to be hypovascular and distorts the intrasplenic blood vessels. This lesion is consistent with a metastatic lesion. Urinary obstruction is present on the right with pelvocaliceal and uretreal dilatation secondary to a soft tissue lesion at the junction of the ureter and baldder. This is another secondary lesion of the malignant melanoma. Surprisingly, these lesions are not hypervascular on doppler nor on CT. Metastasis are also visible in the liver.
Diagnosis: Metastasis of spleen and ureter, malignant melanoma
Clinical Presentation: Workup in a patient with malignant melanoma. Intravenous pyelography showed no excretion of contrast on the right.

Fig. 3. An example ImageCLEFmed case from the Casimage collection, including images and annotation.

```
<topic>
 <number>55</number>
 <EN-description>Show me images of findings with Alzheimer's Disease.
</EN-description>
 <DE-description>Zeige mir Bilder von Fällen mit einer Alzheimer Diagnose.
</DE-description>
 <FR-description>Montre-moi des images d'observations avec la maladie
            d'Alzeimer.</FR-description>
 <year>2006</year>
 <query-images>
  <image>images2006/3-10a.jpg</image>
  <image>images2006/3-10b.jpg</image>
 </query-images>
 <query-type>semantic</query-type>
</topic>
```

Fig. 4. Topic 55 from the consolidated ImageCLEFmed test collection.

5. Synonyms of terms should be considered relevant in the topic. For instance, any MeSH synonyms of the search terms should be considered relevant. As an example, cholangiocarcinoma is a synonym of bile duct cancer. But, on the other hand, the liver/biliary system/pancreas should not be considered synonymous with the entire gastrointestinal system.

EVALUATION

Most IR challenge evaluations are based on the fundamental measures of recall and precision. *Recall* is the proportion of the total number of relevant images retrieved from a collection for a topic:

$$\text{Recall} = \frac{\text{number of relevant images retrieved}}{\text{total number of relevant images in collection}} \quad (1)$$

As the total number of relevant images in real world-sized collections is impossible to know, the measure of *relative recall* is usually employed, where the total number of relevant images in a collection is estimated by all those identified by multiple different searches on the topic.

Precision is the proportion of relevant searches retrieved:

$$\text{Precision} = \frac{\text{number of relevant images retrieved}}{\text{total number of images retrieved}} \quad (2)$$

The usual approach in test collections for using recall, precision, or the aggregated measures to be described next is to calculate the mean across all topics in a test collection. Two or more different systems can then be compared for statistical significance using repeat measures analysis of variance.

As recall and precision tend to vary inversely (i.e., high recall searches tend to have lower precision and vice versa, although there is not an

Table 3. Name and Description of Baseline Runs

Run name	Description
cons_as_is_no_parse_or	Topics file using the default Ferret search without any modifications
cons_as_is_custom	Removal of stop words from query, including standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task, such as *including*, *show*, *me*, *images*, *containing*, and *showing*
topics_parse_OR	Removal of stop words from query, including standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task. Query terms then combined using Boolean OR. Parsing performed to recognize parts of speech and aggregate noun phrase components
topics_mod_parse_OR	Removal of stop words from query, including standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task. Additional stop words removed, including imaging modalities such as *CT* or *MR*. Query terms then combined using Boolean OR and parsing performed
topics_mod_parse_man2	Manual modification of textual queries to add synonyms and regularize language
cons_mixed2	Removal of stop words from query, including standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task. Additional stop words removed. Query terms then combined with Boolean OR. Modality detection on the images in the database used to resort the results
cons_umls_auto	Removal of stop words from query, including standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task. Query expansion performed using UMLS concepts
topics_parse_AND	Removal of stop words from query, including the standard English stop words provided by Ferret as well as common words specific to the ImageCLEF task. Additional stop words removed. Query terms then combined using Boolean AND
GE_GIFT8_b	A purely visual run using the MedGIFT system

Table 4. Results of Baseline Runs for MAP, Precision at Ten Images, and Precision at 30 Images

Run Name	MAP	Precision at 10	Precision at 30
cons_as_is_no_parse_or	0.2073	0.3506	0.2859
cons_as_is_custom	0.2413	0.4235	0.3322
topics_parse_OR	0.2443	0.4200	0.3416
topics_mod_parse_OR	0.2228	0.3776	0.3424
topics_mod_parse_man2	0.1668	0.3541	0.3055
cons_mixed2	0.2672	0.5047	0.4071
cons_umls_auto	0.1968	0.3329	0.2776
topics_parse_AND	0.1013	0.2753	0.1988
GE_GIFT8_b	0.0414	0.1906	0.1318

explicit mathematical tradeoff), there have been a variety of measures developed to aggregate the two measures, such as the *F* measure and recall-precision table. The most commonly used measure now, however, is MAP, which despite its name is more of a measure of recall across the entire retrieval output[7]. MAP is calculated by taking the mean of *average precision* (AP) across all topics. AP is calculated for a single topic by calculating the average of precision at each point a relevant document is retrieved or, for relevant documents not retrieved, a value of 0. The resulting MAP value varies between 0 and 1 and provides an aggregate and comparable measure of system performance.

BASELINE RESULTS

With the new consolidated collection, we performed baseline runs using some common basic image retrieval techniques. These results will allow comparison with new and different experimental approaches. Our image retrieval system is based on Ferret, a Ruby port of the widely used open-source Lucene search engine (lucene.apache.org)[12]. The queries were sent to the search engine as given in the topics file or after a variety of modifications, performed automatically as well as manually, as described in Table 3. This table also lists one purely visual run that used the open-source MedGIFT system[13]. Table 4 shows the results of these runs with three different measures. In addition to MAP, we show precision calculated at 10 and 30 images retrieved. Table 5 MAP results broken down by the year of ImageCLEF and the topic types for all the years.

These results show that using the text of the topic as the query and use of Boolean AND for query terms gave the poorest results. Parsing using part-of-speech tagging and aggregation of noun phrase components provided modest gains, with synonym expansion from the Unified Medical Language System (UMLS) Metathesaurus achieving lesser improvement. The best improvement to results came from modality detection, a technique we used to achieve the best results in ImageCLEF 2007[14,15]. The purely visual run fared poorly, similar to results obtained in the individual years of ImageCLEF.

LESSONS LEARNED

A number of lessons have been learned from the first three years of the ImageCLEF medical retrieval task. In each year, about 12–15 research groups have participated in the task, providing a

Table 5. Results of Baseline Runs for MAP Broken Down by ImageCLEF year (2005—Topics 1–25, 2006—Topics 26–55, 2007—Topics 56–85) and Topic Type (Textual, Visual, and Mixed)

Run name	Topics 1–25	Topics 26–55	Topics 56–85	Textual topics	Visual topics	Mixed topics
cons_as_is_no_parse_or	0.1216	0.1443	0.3416	0.3338	0.1331	0.1876
cons_as_is_custom	0.1673	0.2025	0.3417	0.3754	0.1668	0.2162
topics_parse_OR	0.1665	0.2012	0.3523	0.3857	0.1684	0.2154
topics_mod_parse_OR	0.1660	0.1959	0.2970	0.3760	0.1656	0.1663
topics_mod_parse_man2	0.1627	0.1685	0.2620	0.3309	0.1419	0.1333
cons_mixed2	0.1907	0.2298	0.3744	0.4010	0.1949	0.2461
cons_umls_auto	0.1570	0.1772	0.2712	0.3516	0.1720	0.1335
topics_parse_AND	0.1059	0.0912	0.2892	0.3018	0.0802	0.1643

diverse array of techniques assessed and allowing a diverse retrieval pool for relevance judging. The lessons learned can be viewed at the system or user (or application) level.

At the system level, we have seen that a wide variety of approaches can be used for image retrieval. The major approaches center around textual (also called semantic or concept-based) and visual (also called content-based) techniques. There is no question that textual retrieval techniques alone are more robust than visual techniques alone. That is, textual techniques provide more consistent performance across a wide variety of topics. Although purely visual techniques fare poorly on many topics, especially those judged as amenable to textual retrieval, they have shown to perform better on topics amenable to visual retrieval, especially in combination with textual techniques.

One visually oriented technique performing particularly well has been the automated determination of image modality, which then allows effective filtering of retrieval output[14,15]. Detection of other types of concepts in topics and image text has also been shown to improve retrieval performance[16–18]. One approach to visual retrieval has performed relatively well by making use of a great deal of machine learning[19].

On the user level, our knowledge is less complete. One hint that further research is necessary emanates from the observation that recall-oriented measures like MAP can sometimes give different results than more precision-oriented measures that focus on the output that a typical user is likely to focus on, which is the top 10–30 images. Some experiments have shown that certain techniques, such as fusion of textual and visual search results, give a lower overall MAP but achieve a higher precision starting at five images and continuing to 30 images[20,21]. This finding is significant by virtue of the fact that many real-world users of image retrieval may explore output solely in this range, i.e., do not need to retrieve the full set of relevant images. For example, a clinician or educator seeking a "few good cases" is likely to be satisfied by a relatively small number of relevant images, even if tens or hundreds exist in the database, similar to millions who search information on the Internet and who do not require a complete list of the available information.

FUTURE DIRECTIONS

The ImageCLEFmed challenge evaluation has produced a medical image retrieval collection with 66,662 images and their annotations, 85 topics categorized by amenability to visual, textual, or mixed retrieval, and about 800–1200 relevance judgments per topic. In this paper, we have described the construction, components, and baseline evaluation statistics with the collection. Our hope is that additional researchers will use the collection to evaluate new approaches to medical image retrieval in the future.

Our work in evaluating medical image retrieval is not ending with the production of this test collection. ImageCLEFmed has continued in 2008, with a new collection of images based on a large subset of the Goldminer™ collection (goldminer.arrs.org)[22]. We are also undertaking user studies to further elaborate what users do with image retrieval systems and to determine what evaluation metrics are most effective in measuring their priorities. This will be operationalized by providing our test collections in several different image retrieval systems to users covering the many different user roles and asking them to pose queries with the aim of understanding how they define success and what measures can be used to capture it.

ACKNOWLEDGEMENTS

This work was supported by a supplement to National Science Foundation (NSF) grant ITR-0325160. We also acknowledge the Swiss National Funds (grant 200020-118638/1). Instructions for obtaining the data described in this paper can be obtained from the ImageCLEFmed Website (http://ir.ohsu.edu/image/).

REFERENCES

1. Hersh WR et al: A qualitative task analysis of biomedical image use and retrieval. In: MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation. Vienna, Austria, 2005. http://muscle.prip.tuwien.ac.at/workshop2005_proceedings/hersh.pdf

2. Müller H et al: Health care professionals' image use and search behaviour. In: Proceedings of Medical Informatics Europe 2006, Maastricht, The Netherlands, 2006, pp 24–32. http://www.sim.hcuge.ch/medgift/publications/MIE2006_Mueller.pdf

3. Hersh WR, et al: Advancing biomedical image retrieval: Development and analysis of a test collection. J Am Med Inform Assoc 13:488–496, 2006

4. Braschler M, Peters C: Cross-language evaluation forum: Objectives, results, achievements. Inf Retr 7:7–31, 2004

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

6. Sparck-Jones K: Reflections on TREC. Inf Process Manag 31:291–314, 1995

7. Buckley C, Voorhees EM: Retrieval system evaluation. In: Voorhees EM, Harman DK Eds. TREC: Experiment and Evaluation in Information Retrieval. Cambridge, MA: MIT, 2005, pp. 53–75

8. Hersh WR et al: OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Springer, Dublin, Ireland, 1994, pp 192–201

9. Hersh WR: Interactivity at the Text Retrieval Conference (TREC). Inf Process Manag 37:365–366, 2001

10. Hersh WR, et al: Enhancing access to the bibliome: The TREC 2004 Genomics Track. J Biomed Discov Collab 1:3, 2006. http://www.j-biomed-discovery.com/content/1/1/3

11. Clough P et al: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval—Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Springer Lecture Notes in Computer Science, Alicante, Spain, 2006. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/cloughOCLEF2006.pdf

12. Gospodnetic O, Hatcher E: Lucene in Action, Greenwich, CT: Manning, 2005

13. Müller H et al: The use of MedGIFT and EasyIR for ImageCLEF 2005. In: Accessing Multilingual Information Repositories—6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Springer Lecture Notes in Computer Science, Vienna, Austria, 2005, pp 724–732

14. Kalpathy-Cramer J, Hersh W. Medical image retrieval and automatic annotation: OHSU at ImageCLEF 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 2007. http://www.billhersh.info/imageclef-07-ohsu.pdf

15. Kalpathy-Cramer J, Hersh W. Automatic image modality based classification and annotation to improve medical image retrieval. In: MEDINFO 2007—Proceedings of the Twelfth World Congress on Health (Medical) Informatics. IOS, Brisbane, Australia, 2007, pp 1334–1338

16. Chevallet JP, Lim JH, Radhouani S. A structured visual learning approach mixed with ontology dimensions for medical queries. In: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Springer Lecture Notes in Computer Science, Springer, Vienna, Austria, 2005

17. Lacoste C, et al. IPAL knowledge-based medical Image retrieval in ImageCLEFmed 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval—Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Springer Lecture Notes in Computer Science, Alicante, Spain, 2006. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/lacosteCLEF2006.pdf

18. Lacoste C, et al: Medical image retrieval based on knowledge-assisted text and image indexing. IEEE Trans Circ Syst Video Technol 17:889–900, 2007

19. Deselaers T et al. FIRE in ImageCLEF 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 2007. http://www.clef-campaign.org/2007/working_notes/deselaersCLEF2007.pdf

20. Jensen J, Hersh W: Manual query modification and data fusion for medical image retrieval. In: Accessing Multilingual Information Repositories—6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Springer Lecture Notes in Computer Science, Vienna, Austria, 2005, pp 673–679. http://medir.ohsu.edu/~hersh/imageclef-OHSU-05.pdf.

21. Hersh W, Kalpathy-Cramer J, Jensen J. Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval—Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Springer Lecture Notes in Computer Science, Alicante, Spain, 2006, pp 660–669. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/hershCLEF2006.pdf

22. Kahn CE, Thao C: GoldMiner: A radiology image search engine. Am J Roentgenol 188:1475–1478, 2007