Using Medline Queries to Generate Image Retrieval Tasks for Benchmarking

Henning MÜLLER[ac], Jayashree KALPATHY-CRAMER[b], William HERSH[b], Antoine GEISSBUHLER[a]
[a]*Medical Informatics Service, University & Hospitals of Geneva, Geneva, Switzerland*
[b]*Oregon Health and Science University, Portland, OR, USA*
[c]*Business Information Systems, University Of Applied Sciences, Sierre, Switzerland*

Abstract. Medical visual information retrieval has been a very active research area over the past ten years as an increasing amount of images is produced digitally and made available in the electronic patient record. Tools are required to give access to the images and exploit the information inherently stored in medical cases including images. To compare image retrieval techniques of research prototypes based on the same data and tasks, ImageCLEF was started in 2003 and a medical task was added in 2004. Since then, every year a database was distributed, tasks developed, and systems compared based on realistic search tasks and large databases. For the year 2007 a set of almost 68'000 images was distributed among 38 research groups registered for the medical retrieval task. Realistic query topics were developed based on a log file of Medline. This log file contains the queries performed on Pubmed during 24 hours. Most queries could not be used as search topics directly as they do not contain image-related themes, but a few thousand do. Other types of queries had to be filtered out as well, as many stated information needs are very vague; for evaluation on the other hand clear and focused topics are necessary to obtain a limited number of relevant documents and limit ambiguity in the evaluation process. In the end, 30 queries were developed and 13 research groups submitted a total of 149 runs using a large variety of techniques, from textual to purely visual retrieval and multi-modal approaches.

Keywords. Content-based image retrieval, classification, medical image retrieval, benchmarking, topic development

Introduction

The availability of large amounts of medical images in the electronic patient record and their use by clinicians who are often no specialists in radiology creates a need to develop new tools. Content-based image retrieval (CBIR) is a technique that aims at retrieving images based on their visual content instead of textual metadata [1]. In the non-medical domains this technique is frequently used to access and browse in large visual information repositories. It is used to complement text-based search particularly when little metadata is available. In the medical domain image retrieval has been discussed as important for a fairly long time [2,4]. It was used for image classification [3] as well as for retrieval with a large variety of visual features such as shapes, colors, and textures [4,5,6]. First clinical tests also showed that diagnosis performance can be increased, particularly for non-specialists by making available additional data [7].

As medical image retrieval is in the process to transform from research prototypes towards real applications it is important to separate good techniques from techniques performing poorly. For this goal of technology assessment, benchmarks have existed in domains such as information retrieval for a very long time (i.e. TREC[1], Text Retrieval Conference). In image retrieval, benchmarks have started to develop from 2002, only. First benchmarks include TRECVid that started in TREC and has the goal to evaluated mainly video retrieval and ImageCLEF[2] that is part of the Cross Language Evaluation Forum (CLEF[3]). Since 2004, ImageCLEF has a medical retrieval task with the goal to evaluate medical visual information retrieval. An overview of the ImageCLEF medical task can be found in [12]. Once databases are available for participants, one of the hardest parts in organizing such a retrieval benchmark is the development of realistic search topics based on which the systems can be compared. For other domains the visual information retrieval behavior of user aws described in several articles, for example journalists searching for visual information to illustrate articles [9].

For ImageCLEF, several approaches of query topic development have been applied: In 2004, a domain expert defined search tasks that were typical for him; in 2005, topics were developed based on several surveys performed among image users [11], and in 2006, the topics were based on the log files of a medical image search engine on the Internet [10]. Topic development has several constraints. The goal is to have specific and non-ambiguous information needs and facilitate the job of relevance judges to judge for relevance or non-relevance. If the number of relevant images is too high, a comparison of several techniques becomes more difficult, because participating systems only submit a limited number of images and many images can lead to a strong variation by chance. On the other hand an information need has to correspond at least to a few relevant images in the database to allow for an evaluation at all.

These constraints make the topic development task often hard and time-consuming. Having an image retrieval system in clinical use and use such a system's log file would be the easiest option but currently no such system is available for us.

1. Methods

1.1. Base data used

The base data used includes a full day of queries performed at the Pubmed[4] web side during 24 consecutive hours. The log file contained 2,689,166 queries. A more detailed analysis of this log file can be found in [8]. Most queries are in no connection with the visual field, and thus we discard them. Only information needs regarding visual content are of interest for us. Taking into account the large number of queries in the log file a manual analysis is not possible for the entire set but only for a subset of the data.

[1] http://trec.nist.gov/
[2] http://www.imageclef.org/
[3] http://www.clef-campaign.org/
[4] http://www.pubmed.gov/

1.2. Rules for filtering

Obtaining only image-related topics is difficult. The easiest way is filtering for queries in connection a modality. We filtered out queries containing the following key words: (image, video, media, xray, x-ray, CT, MRI, PET, tomography, ultrasound, ultrasound, endoscopy). Before applying these filters all plural forms were normalized to singular.

Among the frequent results of this filtering we made sure that at least two, or better three of the four following axes are fulfilled to result in topics that are specific enough:

- anatomic region;
- modality;
- pathology;
- visual observation (such as an enlarged heart).

Candidate topics that were obtained after these steps were subsequently used to perform test queries with several systems to make sure that at least a few results images are available in the data. Finally, the queries are ranked with respect to their visualness, meaning how much they seem to correspond to a visual search system or a text search system. Final goal is to develop 30 topics, ten of each category visual, textual, mixed.

1.3. Finding images corresponding to the query topics

In ImageCLEF, all query topics are created from a task description in three languages (English, French, German) with at least two accompanying images. After 30 query topics were identified, we needed to find at least 2 images per topic that correspond to the information needs and that can be used as input for visual search systems. We used Google image search to find such images and made sure that the copyright allows at least for a use in a non-commercial research environment. For a few of the topics it turned out to be hard to find query images and thus these queries were discarded.

2. Results

2.1. Most frequent queries

Table 1. The most frequent queries overall in the log file.

Query	Frequency
Finasteride	3601
Ibuprofen and toxicity not gastrointestinal	3421
One and a half syndrome	1751
#1 and #2	1242
Hypertension	801
Osteoclast tab12	767
Influenzae	765
Diabetes	640
Cancer	552
Heart	481

Table 1 shows the ten most frequent queries in the log file. It can easily be seen that none of the queries corresponds to a visual information need. This underlines the need for filtering. It also becomes clear that there is an extremely large number of different queries performed with Pubmed. Even the most frequent queries are rare compared to the overall number of queries performed. Most queries are very specific.

2.2. Image-related topics

Table 2. The most frequent queries containing topics related to visual information needs

Query	Frequency
MRI	58
Ultrasound	42
Otitis media	37
fMRI	33
Cardiac MRI	20
Endoscopy	20
Walsh CT	18
Lung ultrasound	15
Capsule endocopy	15
Ultrasound for thyroid disorders	15

Table 2 shows the most frequently found query entries that contain the keywords that were chosen for filtering of visual topics. Again, it can be seen that only a few of them could actually be used for a retrieval benchmark making further processing necessary. This further processing is manual, going through the long list of visual queries and filtering out those containing at least two of the axes mentioned above. Frequencies of image-related terms are shown in Table 3.

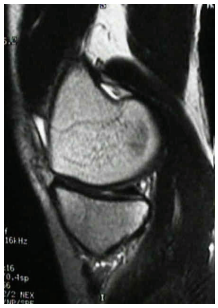
Table 3. Overall number of queries with image related terms.

Term	Frequency
Image	1275
Video	298
Media	4774
Xray/x-ray	822
CT	230752
MRI	5578
Ultrasound	2140
Endoscopy	571

2.3. Final topics

Finally, 30 topics were distributed among the participants of ImageCLEF, divided into the categories visual, semantic, and mixed topics depending on whether they seem adapted for visual or textual retrieval systems. The classification was performed by a domain expert familiar with both techniques.

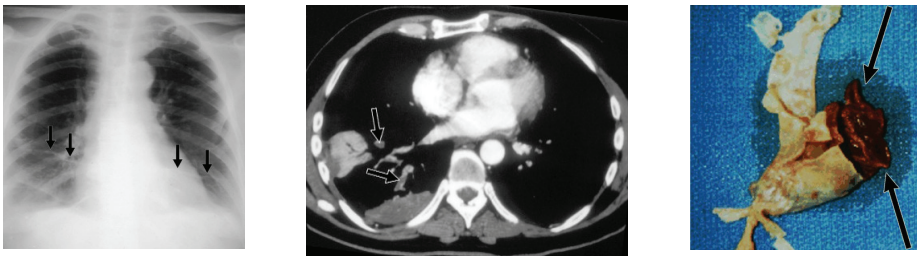




</TOPIC>
<ID>**S9**</ID>
<EN-DESCRIPTION>**medial meniscus MRI**</EN-DESCRIPTION>
<DE-DESCRIPTION>**MR des medialen Meniskusses**</DE-DESCRIPTION>
<FR-DESCRIPTION>**IRM du ménisque interne**</FR-DESCRIPTION>
</TOPIC>

Figure 1 shows an example for a visual topic and Figure 2 and example for a semantic topic more aimed at textual search engines. It can be seen that in the first case results images are expected to be rather homogeneous whereas in the second car the visual variety among the results can be extremely high.

Figure 1. A visual query topic



```
<TOPIC>
<ID>s9</ID>
      <EN-DESCRIPTION>pulmonary embolism all modalities</EN-DESCRIPTION>
      <DE-DESCRIPTION>Lungenembolie alle Modalitäten</DE-DESCRIPTION>
      <FR-DESCRIPTION>Embolie pulmonaire, toutes les formes</FR-DESCRIPTION>
</TOPIC>
```

Figure 2. A semantic query topic

2.4. Outcome of the benchmark

38 research groups from all continents and over 20 countries registered for the medical retrieval task and obtained access to the data and query topics. 13 groups finally submitted results that were compared. In total, 149 runs were submitted in the benchmark and evaluated by the organizers.

Several performance measures were calculated and compared to show various aspects of the systems. Lead performance measure is as in many other benchmarks Mean Average Precision (MAP). For the first time in 2007, best results were obtained by a fully textual retrieval system. On the other hand, for the first time, a fully visual system had extremely good results. This shows that mixed media retrieval has still a very large potential and that much can still be learned in this domain (combining the best visual and textual techniques did lead to a significantly higher result). Mixed media approaches also had by far the best results in early precision, which is most often the performance measure really important for a system user. More on the outcome of the medical benchmark in 2007 can be found in [13].

3. Conclusions

Benchmarks have become an extremely important part of many research domains, particularly in fields where basic research is transforming into applied research and where several prototypes are available. In these fields, a benchmark can focus research work into certain directions and objectively compare several techniques on the same bases, something that is otherwise most often impossible.

ImageCLEFmed has shown its importance through a large participation from worldwide research groups that is increasing every year. Only through realistic search

topics that correspond well to the available databases it is possible to motivate a large number of participants as they need to see an important advantage to participate in such an event and spend their time on the proposed tasks. The evaluation of a log file of the Pubmed search engine allowed us to develop 30 search topics that were subsequently used to compare the results of the participating groups. It is clear that most searches in such a log file are not for visual content, and even those information needs that contain visual requirements are often too vague to be used directly. Much manual intervention is necessary to obtain realistic and usable information needs. Tests with the data base are necessary to make sure that images corresponding to these needs are present.

In the end, such realistic query topics can help to create acceptance for visual retrieval techniques in medical institutions by showing their performance. Good techniques can be identified and compared with other techniques based on realistic tasks. Once first systems are in routine use in clinical institutions, their log files can even lead to topics better-adapted to real visual information needs.

References

[1] AWM. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans on Pattern Anal Mach Intell* **22** (200), 1349-1380.

[2] H Müller, N Michoux, D Bandon, A Geissbuhler, A review of content-based image retrieval systems in medicine – clinical benefits and future directions, *Int J Med Inform* **73** (2004), 1-23.

[3] TM Lehmann, MO Güld, C Thies, B Fischer, K Spitzer, D Keysers, H Ney, M Kohnen, H Schubert, BB Wein, Content-based image retrieval in medical applications, *Methods Inf Med* **43** (2004), 354-361.

[4] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, Z. Protopapas, Fast and effective retrieval of medical tumor shapes, *IEEE Trans Knowl Data Eng,* **10** (1998), 889-904.

[5] LHY Tang, R Hanka, HHS Ip, A review of intelligent content-based indexing and browsing of medical images, *Health Inform J* **5** (1998), 40-49.

[6] C-R Shyu, CE Brodley, AC Kak, A Kosaka, AM Aisen, LS Broderick, ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases, *Comput Vis Image Underst* **75** (1999), 111-132.

[7] AM Aisen, LS Broderick, H Winer-Muram, CE Brodley, AC Kak, C Pavlopoulou, J Dy, CR Shyu, A Marchiori, Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment, *Radiology* **228** (2003), 265-270.

[8] JR Herskovic, LY Tanada, W Hersh, EV Bernstam, A day in the life of Pubmed: Analysis of a typical day's query log, *J Am Med Inform Assoc*, 14 (2007), 212-220.

[9] M Markkula, E Sormunen, Searching for photos - journalists' practices in pictorial IR, *The challenge of image retrieval*, Newcastle upon Tyne, UK, 1998.

[9] H Müller, C Boyer, A Gaudinat, W Hersh, A Geissbuhler, Analyzing web log files of the Health On the Net HONmedia Search Engine to define typical image search tasks for image retrieval evaluation, *Medinfo* **12** (2007), 1319-1323.

[10] W Hersh, J Jensen, H Müller, P Gorman, P Ruch, *A qualitative task analysis for developing an image retrieval test collection,* International Workshop on Image and Video Retrieval Evaluation, Vienna, 2005, p. 11-16.

[11] W Hersh, H Müller, J Jensen, J Yang, P Gorman, P Ruch, Advancing biomedical image retrieval: development and analysis of a test collection, *J Am Med Inform Assoc* **13** (2006), 488-496.

[12] H Müller, T Deselaers, E Kim, J Kalpathy-Cramer, TM. Deserno, P Clough, W Hersh, Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks, *Lect Notes Comput Sci*, 2008 - in press.