Overview of the ImageCLEFmed 2008 medical image retrieval task

Henning Müller[1,2], Jayashree Kalpathy–Cramer[3], Charles E. Kahn Jr.[4], William Hatt[3], Steven Bedrick[3], William Hersh[3]

[1] Medical Informatics, University Hospitals and University of Geneva, Switzerland
[2] University of Applied Sciences Western Switzerland, Sierre, Switzerland
[3] Oregon Health and Science University (OHSU), Portland, OR, USA
[4] Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA
henning.mueller@sim.hcuge.ch

Abstract

2008 was the fifth year for the medical image retrieval task of ImageCLEF, one of the most popular tracks within CLEF. Participation continued to increase in 2008. A total of 15 groups submitted 111 valid runs. Several requests for data access were also received after the registration deadline.

The most significant change in 2008 was the use of a new database containing images from the medical literature. These images, part of the Goldminer collection, were from the RSNA journals Radiology and Radiographics. Besides the images, the figure captions and the part of the caption referring to a particular sub figure were supplied to the participants. Access to the full text articles in HTML was also provided, as was each article's Medline PMID (PubMed Identifier). An article's PMID could be used to obtain the officially assigned MeSH (Medical Subject Headings) terms. Unlike previous years, this year's collection was entirely in English, as it was obtained from English-language medical literature. However, the topics were, as in previous years, supplied in German, French, and English. The topics used in 2008 were a subset of the 85 topics used in 2005-2007. Thirty topics were made available, ten in each of three categories: visual, mixed, and semantic.

As in previous years, most groups concentrated on fully automatic retrieval. However, three groups submitted a total of seven manual or interactive runs; these runs did not show a substantial increase in performance over the automatic approaches. In previous years, multi–modal combinations were the most frequent submissions. However, in 2008 only half as many mixed runs as purely textual runs were submitted. Very few fully visual runs were submitted, and the ones submitted performed poorly. This may be explained in part by the heavily semantic nature of the 2008 topics.

The best MAP scores were very similar for textual and multi–modal approaches, whereas early precision performance was clearly better for the multi-modal approaches.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image Retrieval, Performance Evaluation, Image Classification, Medical Imaging

1 Introduction

ImageCLEF[1] [1, 2, 5] started within CLEF[2] (Cross Language Evaluation Forum, [6]) in 2003. A medical image retrieval task was added in 2004 to explore domain–specific multilingual visual information retrieval and also multi–modal retrieval by combining visual and textual features for retrieval. A medical retrieval task and a medical image annotation task have been part of ImageCLEFmed since 2005 [5].

This paper reports on the medical retrieval task whereas additional papers describe the four other tasks of ImageCLEF. More detailed information can also be found on the task web pages for ImageCLEFmed. A detailed analysis of a previous medical image retrieval task is available in [3].

2 The medical retrieval task in 2008

The main change in the medical retrieval task in 2008 was the use of a new database. The search tasks remained essentially the same as in the previous years. The collection distributed to the participants included the images and the captions, as published in the medical journals. URLs to access the full text of the journal article were also made available to the participants.

2.1 Registration and participation

As in previous years, registration for the medical retrieval task increased in 2008, albeit slowly. Several of the groups registered solely to obtain the test collection in order to use it as training data for their algorithms, rather than actually participating in the competition. In the end, 15 research groups submitted a total of 130 runs. Groups were asked to not submit more than ten runs in 2008 (different from previous years) so as not to bias the pools too much towards any single group.

There were significant problems with many of the 130 initial runs: some were submitted in incorrect formats; several runs were duplicated; and there were runs that provided search results for only a subset of the thirty topics. These problems were corrected in collaboration with the authors as much as was possible, resulting in 111 valid runs that were used to generate the pools that were finally judged for relevance. The following groups submitted valid runs:

- Hungarian Acadamy of Sciences, Budapest, Hungary;
- National Library of Medicine (NLM), National Institutes of Health NIH, Bethesda, MD, USA;
- Bania Luka University, Bosnia-Hercegovina;
- MedGIFT group, University of Geneva, Switzerland;
- Natural Language Processing group, University Hospitals of Geneva, Switzerland;
- GPLSI group, University of Alicante, Spain;

[1] http://www.imageclef.org/
[2] http://www.clef-campaign.org/

- Multimedia Modelling Group, LIG, Grenoble, France;
- Natural Language Processing at UNED. Madrid, Spain;
- Miracle group, Spain;
- Oregon Health and Science University (OHSU), Portland, OR, USA;
- IRIT Toulouse, France;
- University of Jaen, Spain;
- Tel Aviv University, Israel;
- National University of Bogota, Colombia;
- TextMess group, University of Alicante, Spain.

Thus, a total of 15 groups from eight countries and four continents submitted results that are presented in the following chapters.

2.2 Database

The database used for the task in 2008 was made available by the Radiological Society of North America (RSNA). The database contains in total slightly more than 66,000 images taken from the radiological journals *Radiology* and *Radiographics*. The images are original figures used in published articles. The collection is a subset of a larger database that is available via the Goldminer[3] image search engine. For each image, the text of the figure caption was supplied as free text. However, this caption was sometimes associated with a multi-part image. In over 90% of the images the part of the caption actually referring to this sub–image was also provided. Additionally, links to HTML versions of the full–text articles were provided along with the relevant PubMed accession ID numbers. Both the full–size images as well as thumbnails were available to the participants. All text was in English.

The contents of this database represent a broad and significant body of medical knowledge, which makes this year's competition a potentially realistic scenario for how clinicians might use image retrieval systems in the future.

2.3 Query topics

The query topics in 2008 were a selection of 30 topics from the previous three years of ImageCLEFmed [4]. Training data in the form of the 2005-2007 database with images, annotations, topics, sample query images and qrel files was made available to participants. All topics were supposed to cover at least two of the following axes:

- Anatomic region shown in the image;
- Image modality (x–ray, CT, MRI, gross pathology, ...);
- View (frontal, sagittal,...);
- Pathology or disease shown in the image;
- abnormal visual observation (eg. enlarged heart).

From the 85 possible topics of past years, similar topics were removed to cover a wide range of different modalities and anatomic regions. A visual and textual check was then performed to make sure that at least a few relevant images exist in the dataset. Since the databases of 2008 and 2007 were very different, we wanted to ensure that each topic had more than one relevant image exist.

Each query topic consists of the information need in three languages (English, French, German) and at least two example images. Groups could decide which language and media to use for the query processing and also which part of the text to use.

[3] `http://goldminer.arrs.org/`

2.4 Relevance judgments

A new system for relevance judgments was introduced in 2008 building on a Ruby for Rails framework and allowing for simple judgments via a web interface for all judges. The first 35 images of every run were combined into "pools" with an average size of around 900 images. Such pooling is necessary to reduce the amount of data to judge, and the bias can be regarded as very limited [7]. Medical Doctors who are also students of biomedical informatics at OHSU were hired for the judgment process and paid by the hour for the judgments.

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 controlled during the judgment process.

During the judging, the new system exhibited a minor problem that resulted in certain images losing their judgments. This resulted in a short delay in the judging process, after which the affected images were re–judged by the same persons.

3 Submissions and results

This section details the submissions for the tasks and a first brief evaluation. A more detailed evaluation of the techniques will follow in the final proceedings when more details on the techniques used for the submissions will be known. Unfortunately, information on the techniques used in the submissions is not always made available by the participants well ahead of time and in great detail.

Trec_eval was used for the evaluation process with most of its performance measures.

3.1 Submissions

A total of 130 runs were submitted via the electronic submission system. Scripts to check the validity of the runs were made available to participants ahead of the submission phase, but even so, almost half of the submitted runs contained errors in either content or format and required changes. Common mistakes included a wrong trec_eval format, use of only a subset of the topics and incorrect image identifiers. In collaboration with the authors a large number of runs were repaired, resulting in 111 valid runs taken into account for the pools.

In total, only seven runs were "manual" or "interactive". There were also fewer "visual–only" runs than in all previous years, with only 8 such runs being submitted. The large majority were text–only runs, with 65 submissions. Mixed automatic runs had 31 submissions.

Groups subsequently had the chance to evaluate additional runs themselves as the qrels were made available to participants two weeks ahead of the submission deadline for these working notes.

3.2 Visual retrieval

The number of visual runs in 2008 was much lower than in previous years, and the evolution is not as fast as with textual retrieval techniques. Five groups submitted a total of eight runs in 2008. Performance as measured in MAP is very low for all these runs, reaching a maximum of 0.04 for the best run. Early precision averaged over all topics reaches around 0.2, which is absolutely acceptable. When taking into account only the visual topics these results are much better, whereas the purely semantic topics obtained extremely poor results.

Table 1 shows the results and particularly the large differences between the runs. Some runs managed to retrieve a larger part of the relevant images (809) but with a fairly low MAP, whereas some runs with a higher MAP only found a very small number of relevant images in the first 1000 results. A higher bpref in this context can mean that a larger number of images from these runs were not judged for relevance. This might also be due to the fact that only very few visual runs were submitted and thus only few visually retrieved documents were finally judged.

Table 1: Results of the automatic runs using only visual information.

Run	run_type	MAP	bpref	P5	P10	P30	num_rel
TAU_MIPLAB-TAU_norm	Visual Automatic	0.04	0.09	0.22	0.17	0.15	568
UNAL-W+QE+JS	Visual Automatic	0.04	0.06	0.13	0.13	0.11	297
GE_GIFT8	Visual Automatic	0.03	0.09	0.17	0.17	0.15	809
MIPLAB-TAU_orig	Visual Automatic	0.03	0.08	0.16	0.14	0.11	519
etfbl-max11111	Visual Automatic	0.03	0.04	0.15	0.13	0.11	212
etfbl-sum11111	Visual Automatic	0.03	0.04	0.12	0.10	0.12	194
GE_GIFT16	Visual Automatic	0.03	0.07	0.13	0.13	0.11	670
LSI_UNED	Visual Automatic	0.02	0.03	0.11	0.11	0.08	94
CEB_Image	Visual Automatic	0.01	0.04	0.03	0.04	0.05	390

Results of GIFT were available to the all the participants for combinations of visual and textual runs.

3.3 Textual retrieval

Purely automatic textual retrieval had by far the largest number of runs in 2008 with 65, more than half of all submitted runs. Table 2 shows the results for all submitted automatic text runs, ordered by MAP. Most performance measures such as bpref and early precision are similar in order. Only early precision sometimes has significant differences from the ranking with MAP.

Runs from the University of Alicante (Textmess), University of Jaen (SINAI), and LIG Grenoble teams obtained the best results, mainly by using ontologies such as MeSH (Medical Subject Headings) to code the documents. A MAP of 0.29 could be obtained and several systems have a high score very close to this. A more detailed analysis is required with the exact techniques applied for each of the runs.

3.3.1 Using various languages for the retrieval

Unfortunately, very little information was available on which languages the groups used for the retrieval. It can be assumed that most groups used English as this promises the best results. It was also possible to use all three query languages together, for example, for extracting MeSH terms. While this multi–lingual approach is not necessarily a realistic scenario, it can lead to interesting results.

The HUG group used the same techniques with several languages and showed that English obtained by far the best results, better than either French or German. The technique they applied was to map of MeSH terms form the text and queries in various languages. Through the PMIDs, the officially (manually) assigned MeSH terms of the articles were also available. The MeSH terms extracted from the article and query text performed worse for retrieval than the officially assigned terms.

3.3.2 Additional resources used for the retrieval

Groups could also state which additional resources were used for retrieval. The goal of this was to assemble a collection of available resources that could potentially be shared among participants to improve performance in future challenges. A large variety of resources were used, in large part for the combination of visual and textual runs, but also for purely textual runs. Many of the best runs used the ImageCLEFmed 2005-2007 data for training. Official MeSH terms manually assigned by the National Library of Medicine could be used through the PMIDs of the articles.

The most commonly used resources were the training data sets of ImageCLEF 2005-2007. There were numerous challenges with this approach, as the database used from 2005-07 differed greatly from the 2008 database. The annotations in the '05-07 database were of much poorer quality than in the 2008 database, and the two databases were made up of very different types

Table 2: Results of the automatic runs using only text.

Run	run_type	MAP	bpref	P5	P10	P30	num_rel
EXPPRFNegativaMesh	Text Automatic	0.29	0.35	0.49	0.46	0.41	2165
sinai_CT_Mesh	Text Automatic	0.28	0.33	0.44	0.41	0.37	2106
LIG_COS0506_MPTT_Emi	Text Automatic	0.28	0.34	0.51	0.47	0.43	2224
LIG-LIG_MPTT_Emix	Text Automatic	0.28	0.34	0.43	0.45	0.43	2138
TEXTMESSmeshType_CT	Text Automatic	0.28	0.32	0.43	0.41	0.37	2106
IRn2baseline	Text Automatic	0.28	0.33	0.48	0.42	0.35	1986
IRn2ExpNeg	Text Automatic	0.28	0.33	0.45	0.40	0.34	2006
LIG_RET_MPTT_Emix	Text Automatic	0.27	0.34	0.46	0.45	0.41	2129
LIG_COS_MPTT_Emix	Text Automatic	0.27	0.33	0.47	0.47	0.43	2275
LIG_CR_MPTT_Emix	Text Automatic	0.27	0.33	0.48	0.47	0.41	2265
IRn2ExpNegMesh	Text Automatic	0.27	0.32	0.45	0.42	0.36	2038
MirBaselineEN	Text Automatic	0.27	0.32	0.51	0.47	0.39	1861
IRn2Explca	Text Automatic	0.26	0.33	0.45	0.41	0.35	2096
LIG_RET_MP_Emix	Text Automatic	0.26	0.32	0.47	0.45	0.42	1979
IRn2ExpPRF	Text Automatic	0.26	0.32	0.47	0.41	0.36	1980
LIG_MP_Emix	Text Automatic	0.25	0.33	0.45	0.42	0.43	2007
MirAPEN	Text Automatic	0.25	0.31	0.49	0.46	0.39	1773
sinai_CT_Base	Text Automatic	0.25	0.31	0.32	0.35	0.33	2030
MirTaxEN	Text Automatic	0.25	0.32	0.38	0.37	0.37	1867
LIG-LIG_COS_MP_Emix	Text Automatic	0.24	0.31	0.45	0.41	0.41	2120
LIG-LIG_CR_MP_Emix	Text Automatic	0.24	0.31	0.47	0.40	0.39	2108
sinai_CT_Umls	Text Automatic	0.23	0.27	0.37	0.35	0.30	1927
bp_acad_textonly	Text Automatic	0.22	0.28	0.49	0.43	0.35	1726
Ssinai_CTA_Mesh	Text Automatic	0.21	0.27	0.46	0.40	0.29	1683
ohsu_text_umls_4	Text Automatic	0.20	0.30	0.31	0.29	0.25	1973
sinai_CTA_Base	Text Automatic	0.20	0.27	0.41	0.36	0.30	1702
LIG-LIG_MPadd_Emix	Text Automatic	0.19	0.29	0.34	0.37	0.34	2032
sinai_CTS_Base	Text Automatic	0.18	0.25	0.33	0.31	0.31	1790
sinai_CTA_Umls	Text Automatic	0.18	0.25	0.35	0.32	0.32	1553
HUG-MH-EN	Text Automatic	0.18	0.24	0.34	0.30	0.22	1957
HUG-MHnOVID-EN	Text Automatic	0.18	0.24	0.34	0.30	0.22	1957
sinai_CTS_Mesh	Text Automatic	0.16	0.24	0.32	0.29	0.27	1828
HUG-ltc-EN	Text Automatic	0.16	0.23	0.31	0.28	0.20	1713
HUG-mixPapers_EN	Text Automatic	0.15	0.21	0.33	0.27	0.20	1883
ohsu_text_3	Text Automatic	0.15	0.23	0.39	0.31	0.22	1786
sinai_CTS_Umls	Text Automatic	0.14	0.21	0.23	0.21	0.21	1558
TEXTMESSumlsType_CT	Text Automatic	0.14	0.17	0.33	0.32	0.25	1045
sigRunTxt	Text Automatic	0.14	0.19	0.29	0.24	0.22	858
HUG-BL_EN	Text Automatic	0.14	0.21	0.31	0.26	0.24	1615
HUG-HUG-BL HUG-BL	Text Automatic	0.14	0.21	0.31	0.26	0.24	1615
HUG-capMH_EN	Text Automatic	0.13	0.19	0.33	0.28	0.24	1499
HUG-capMH_EN	Text Automatic	0.13	0.19	0.33	0.28	0.24	1499
OHSU-text_or_1	Text Automatic	0.11	0.18	0.31	0.26	0.24	1420
HUG-ltc-FR	Text Automatic	0.11	0.18	0.19	0.20	0.16	1218
HUG-MH-FR	Text Automatic	0.11	0.17	0.19	0.17	0.16	1419
HUG-MHnOVID-FR	Text Automatic	0.11	0.17	0.19	0.17	0.16	1419
MirRF0505EN	Text Automatic	0.11	0.18	0.28	0.24	0.24	1372
HUG-MHnOVID-GE	Text Automatic	0.10	0.14	0.21	0.19	0.17	894
TEXTMESSmeshType_CTS	Text Automatic	0.10	0.18	0.23	0.23	0.15	1828
HUG-ltc-GE	Text Automatic	0.09	0.14	0.17	0.16	0.13	869
HUG-capMH_FR	Text Automatic	0.09	0.16	0.23	0.20	0.17	1364
TEXTMESSumlsType_CTS	Text Automatic	0.09	0.14	0.23	0.23	0.16	933
CEB_BaseC_QE	Text Automatic	0.08	0.14	0.33	0.29	0.23	887
CCEB_BaseC_QE	Text Automatic	0.08	0.14	0.35	0.28	0.23	887
CEB_BaseC	Text Automatic	0.08	0.14	0.31	0.28	0.22	893
MirRF1005EN	Text Automatic	0.07	0.15	0.22	0.16	0.15	1248
HUG-MH-GE	Text Automatic	0.07	0.11	0.17	0.15	0.14	866
HUG-BL-FR	Text Automatic	0.07	0.11	0.17	0.16	0.15	942
MirRFTax1005EN	Text Automatic	0.07	0.14	0.15	0.13	0.14	1260
MirRFTax1005FR	Text Automatic	0.07	0.11	0.13	0.11	0.09	823
MirRFTax1005DE	Text Automatic	0.05	0.08	0.09	0.09	0.06	461
CEB_BaseM	Text Automatic	0.04	0.09	0.20	0.17	0.15	532
HUG-BL-GE	Text Automatic	0.03	0.05	0.07	0.06	0.06	432
HUG-capMH_GE	Text Automatic	0.03	0.05	0.07	0.06	0.06	432
CEB_BaseC_QE	Text Automatic	0.02	0.03	0.06	0.05	0.04	182

Table 3: Results of the automatic runs mixing text and visual information.

Run	run_type	MAP	bpref	P5	P10	P30	num_rel
sinai_CT_Mesh_Fire20	Mixed Automatic	0.29	0.33	0.45	0.43	0.40	2132
TEXTMESSmeshTypeFIREidf_CT	Mixed Automatic	0.28	0.32	0.43	0.41	0.37	2106
IRn2ExpNegRRIDF	Mixed Automatic	0.28	0.33	0.45	0.40	0.34	2006
IRn2ExpNegMeshRRIDF	Mixed Automatic	0.27	0.32	0.45	0.42	0.36	2038
ohsu_vis_mod_umls_4	Mixed Automatic	0.23	0.35	0.41	0.37	0.28	2052
ohsu_vis_mod_5	Mixed Automatic	0.23	0.33	0.41	0.38	0.29	1995
EXTMESSmeshTypeFIRE_CT	Mixed Automatic	0.22	0.27	0.30	0.30	0.30	2106
ohsu_mod_pars2_sp	Mixed Automatic	0.21	0.30	0.58	0.55	0.46	1561
OHSU_vis_mod_3	Mixed Automatic	0.15	0.25	0.41	0.32	0.24	1829
TEXTMESSumlsTypeFIREidf_CT	Mixed Automatic	0.14	0.17	0.33	0.32	0.25	1045
TEXTMESSumlsTypeFIRE_CT	Mixed Automatic	0.13	0.18	0.25	0.22	0.23	1045
TEXTMESSmeshTypeFIRE_CTS	Mixed Automatic	0.12	0.19	0.25	0.25	0.20	1828
SIG_IRIT-SigRunMixt	Mixed Automatic	0.11	0.16	0.30	0.29	0.23	859
TEXTMESSumlsTypeFIRE_CTS	Mixed Automatic	0.09	0.15	0.21	0.22	0.21	928
GE_GIFT8_EN_0.5	Mixed Automatic	0.08	0.19	0.27	0.24	0.24	1835
GE_EN_reGIFT8	Mixed Automatic	0.08	0.19	0.24	0.23	0.23	1957
GE_EN_GIFT8_mix	Mixed Automatic	0.08	0.19	0.28	0.24	0.25	1610
GE_GIFT8_EN_0.9	Mixed Automatic	0.07	0.12	0.31	0.27	0.25	812
GE_GIFT8_reEN	Mixed Automatic	0.07	0.12	0.29	0.24	0.25	812
IRn2ExpNegGiftRR	Mixed Automatic	0.05	0.11	0.13	0.12	0.11	830
IRIT-SigRunComb5	Mixed Automatic	0.05	0.10	0.28	0.24	0.17	793
IRIT-SigRunComb1	Mixed Automatic	0.05	0.10	0.28	0.25	0.17	791
IRIT-SigRunComb2	Mixed Automatic	0.05	0.10	0.28	0.24	0.16	789
IRIT-SigRunComb3	Mixed Automatic	0.05	0.10	0.27	0.24	0.16	782
IRIT-SigRunComb7	Mixed Automatic	0.04	0.09	0.25	0.22	0.16	805
IRIT-SigRunComb4	Mixed Automatic	0.04	0.09	0.25	0.22	0.16	770
IRIT-SigRunComb6	Mixed Automatic	0.04	0.09	0.25	0.22	0.16	771
IRIT-SigRunComb8	Mixed Automatic	0.04	0.09	0.24	0.22	0.16	817
CEB_IBaseC	Mixed Automatic	0.04	0.13	0.17	0.15	0.10	893
CEB_ITD3	Mixed Automatic	0.03	0.10	0.07	0.11	0.10	945
IRn2ExpNegMeshGiftRR	Mixed Automatic	0.03	0.08	0.11	0.11	0.09	662

of images. Nonetheless, the 2008 topics were a subset of those from previous years' competitions, and so the scenario was somewhat realistic with respect to the training data.

3.4 Mixed retrieval

The promotion of mixed–media retrieval has always been one of the main goals of ImageCLEF. In past years, mixed–media retrieval had the highest submission rate. In 2008, however, only half as many mixed runs were submitted than purely textual runs.

Table 3 shows the results for all submitted runs. It is clear that, for a large number of the runs, the MAP results for the mixed retrieval submissions were very similar to those from the purely textual retrieval systems. An interesting observation is that the mixed-media submissions often have higher early precision than the purely textual retrieval submissions. This confirms what has been previously observed.

The text–only runs exhibited relatively high correlation between MAP and bpref. This was not the case among the mixed–media runs. One possible explanation for this difference could be that the mixed–media runs used a wider variety of techniques than the text–only runs. Another possible explanation is that more of the mixed–media runs were submitted after the deadline for pool inclusion. If the mixed–media runs retrieved a higher proportion of non-judged images than the text–only runs, the result would be a larger MAP/bpref variance.

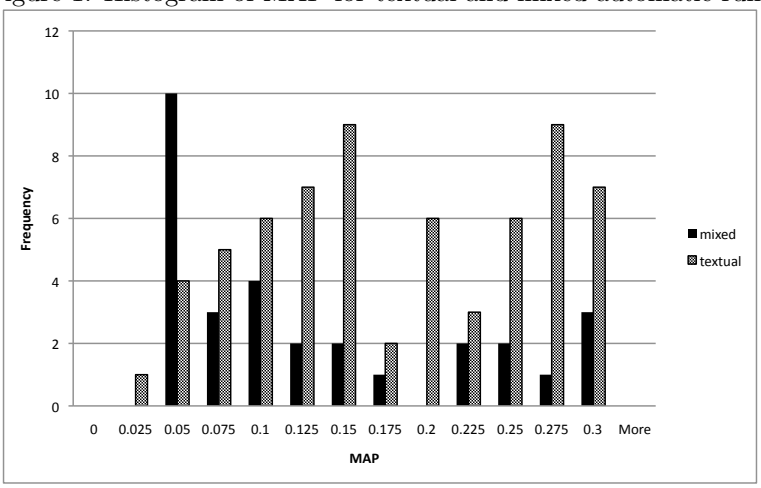
When comparing these mixed–media results with those from the text–only runs, it becomes clear that mixed retrieval can obtain very low results. From examining mixed–media runs which had corresponding text–only runs, it is particularly clear that combining good textual retrieval techniques with questionable visual retrieval techniques can negatively affect system performance. This demonstrates the difficulty of usefully integrating both textual and visual information, and

Table 4: Results of the interactive and manual runs.

Run	run_type	MAP	bpref	P5	P10	P30	num_rel
ohsu_int_2	Mixed Interactive	0.22	0.31	0.57	0.49	0.39	1580
ohsu_sdb_full_interactive	Mixed Interactive	0.18	0.29	0.53	0.46	0.33	1626
ohsu_sdb_lsa	Mixed Interactive	0.10	0.20	0.27	0.27	0.27	1601
CEB_ITD_ALL	Mixed Manual	0.03	0.11	0.08	0.11	0.11	964
CEB_IBaseM	Mixed Manual	0.02	0.10	0.08	0.08	0.06	532
CEB_TD_ALL	Text Manual	0.08	0.16	0.24	0.27	0.25	1198
CEB_TD3	Text Manual	0.08	0.16	0.24	0.27	0.25	1189

the fragility that such combinations can introduce into retrieval systems. As seen in 1, the distribution of MAP for the textual runs was higher than that for the mixed runs. A significant mode exists around a MAP of 0.05 for the mixed runs, while the modes for the textual runs are at 0.15 and 0.28.

Figure 1: Histogram of MAP for textual and mixed automatic runs



3.5 Interactive retrieval

This year, as in previous years, interactive retrieval was only used by a very small number of participants. Interactive retrieval is extremely important, and it is a pity so few groups chose to attempt anything other than purely automated systems.

Table 4 shows the results of all manual and interactive runs submitted. Two runs from OHSU had fairly good results; the other runs were competitive in neither the MAP nor the early precision categories when compared to the fully automatic runs. In general, MAP and early precision were well-correlated ($R^2 = 0.82$ for textual runs, 0.68 for mixed runs); these two runs, however, had higher early precision than their MAP would predict.

3.6 Topic Analysis

Overall, most groups performed significantly better on the semantic topics than on the mixed or visual topics, as can be seen in the table below. Topics 6 and 11–18 were quite difficult for many participants. Table 5 gives an overview of the best and average perform per topic. Some topics with a small number of relevant images have a particularly low performance.

The fact that many of the visual topics obtained poorer performance than the semantic topics also shows that groups have much more experience working on semantic topics, and that visual retrieval currently has much more difficulty obtaining good results. That said, visual retrieval

Table 5: Best results and average for all topics, showing the significant differences between topics.

Topic	Topic	Ave. MAP	Max. MAP	no. rel.
1.	Show me photographs of benign or malignant skin lesions.	0.04	0.29	2
2.	Show me images containing one or several full-body scintigraphies.	0.02	0.61	10
3.	Show me Doppler ultrasound images (colored).	0.24	0.50	284
4.	Show me photographs showing an entire fetus.	0.04	0.26	5
5.	Show me chest CT images with emphysema.	0.16	0.58	69
6.	Show me images of a frontal head MRI.	0.01	0.08	27
7.	Show me images of a knee x-ray.	0.07	0.40	137
8.	Show me x-ray images of a hip joint with prosthesis.	0.07	0.38	28
9.	Show me images of PowerPoint slides.	0.32	1.00	17
10.	mediastinal CT	0.23	0.52	358
11.	Show me abdominal CT images showing liver blood vessels.	0.05	0.21	331
12.	Show me microscopic pathology images of the kidney.	0.04	0.47	51
13.	Show me gross pathologies of myocardial infarction.	0.08	0.35	10
14.	Show me chest CT images showing micro nodules.	0.06	0.22	71
15.	Show me chest x-ray images of cases with tuberculosis.	0.07	0.33	204
16.	Show me all x-ray images containing one or more fractures.	0.04	0.27	218
17.	Show me MRI images of the brain with a blood clot.	0.01	0.09	11
18.	gastrointestinal endoscopy with polyp	0.08	0.35	46
19.	CT liver abscess	0.24	0.76	101
20.	MRI or CT of colonoscopy	0.20	0.60	306
21.	Show me photographs of tumours.	0.11	0.39	334
22.	Show me images of muscle cells.	0.13	0.50	90
23.	Show me x-ray images of bone cysts .	0.05	0.29	17
24.	Show me images containing a Budd-Chiari malformation.	0.38	0.94	74
25.	Merkel cell carcinoma	0.40	1.00	24
26.	gastrointestinal neoplasm	0.13	0.37	279
27.	tuberous sclerosis	0.34	0.77	52
28.	mitral valve prolapse	0.14	0.53	3
29.	pulmonary embolism all modalities	0.26	0.55	237
30.	microscopic giant cell	0.13	0.50	39

Table 6: Inter–rater reliability

Topic	Judge 1	Judge 2	Strict Kappa	Lenient kappa
3.	User 3	User 4	0.91	0.95
5.	User 5	User 7	0.70	0.79
6.	User 3	User 5	0.48	0.48
25.	User 7	User 10	0.69	0.70

can have an important positive influence, and it seems necessary to promote it further by having potentially a larger number of visual topics to push groups towards using visual techniques.

3.7 Inter-rater agreement

Four topics were each judged by two judges. We performed tests of inter-rater agreements using kappa statistics, as seen in table 6. In 3 of the four cases, the inter-rater agreement was quite good. In the last case, one judge interpreted the query more strictly than the other.

4 Conclusions

The focus of many participants in this year's ImageCLEF 2008 has been text–based retrieval. The increasingly semantic topics combined with a database containing high–quality annotations in 2008 may have resulted in less impact of using visual techniques as compared to previous years. This tendency is also seen when looking at the performance by topic where visual topics had significantly lower results than the semantic topics. Our goal in the upcoming ImageCLEF medical retrieval task is to increase the number of visual runs submitted. We hope to modify the task to favor more integrated approaches. Another important aspect is that interactive retrieval has always had a poor participation and definitely needs to be regarded more strongly. Relevance feedback and query modifications have a potential to significantly improve results, but of course research favors laboratory style evaluations.

Visual runs were rare and had no single run with a very convincing performance as was the case in 2007, where the best visual runs had an extremely good performance. Mixed–media runs were very similar in performance to textual runs when looking at MAP. The only difference was that mixed–media runs obtained better early precision in general. Several mixed–media runs were also broken, resulting in a very poor performance. This highlights that the combination is still not very stable.

A per–topic analysis shows that visual topics obtained lower average results than semantic topics. The analysis also shows that several runs with very few relevant images have a very low average performance, whereas topics with a larger number seem to perform better.

Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. The images for the 2008 ImageCLEFmed challenge were contributed by the Radiological Society of North America (RSNA). This work was partially funded by the Swiss National Science Foundation (FNS) under contract 205321–109304/1, the American National Science Foundation (NSF) with grant ITR–0325160, and by the University of Applied Sciences Western Switzerland (HES SO) in the context of the BeMeVIS project.

References

[1] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF 2006 Proceedings*, volume 4730 of *Springer Lecture Notes in Computer Science*, pages 579–594, 2007.

[2] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Michael Jones, Gareth J. F.and Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pages 597–613, Bath, UK, 2005. Springer.

[3] William Hersh, Henning Müller, Jeffery Jensen, Jianji Yang, Paul Gorman, and Patrick Ruch. Advancing biomedical image retrieval: Development and analysis of a test collection. *Journal of the American Medical Informatics Association*, September/October:488–496, 2006.

[4] William Hersh, Henning Müller, and Jayashree Kalpathy-Cramer. The imageclefmed medical image retrieval task test collection. *Journal of Digital Imaging*, 2008.

[5] Henning Müller, Thomas Deselaers, Eugene Kim, Jayashree Kalpathy-Cramer, Thomas M. Deserno, Paul Clough, and William Hersh. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In *CLEF 2007 Proceedings*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, Budapest, Hungary, 2008. Springer.

[6] Jacques Savoy. Report on CLEF–2001 experiments. In *Report on the CLEF Conference 2001 (Cross Language Evaluation Forum)*, pages 27–43, Darmstadt, Germany, 2002. Springer LNCS 2406.

[7] Justin Zobel. How reliable are the results of large–scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.