Overview of the CLEF 2010 medical image retrieval track

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

[1]Geneva 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. The seventh edition of the ImageCLEF medical retrieval task was organized in 2010. As in 2008 and 2009, the collection in 2010 uses images and captions from the *Radiology* and *Radiographics* journals published by RSNA (Radiological Society of North America). Three sub–tasks were conducted within the auspices of the medical task: modality detection, image–based retrieval and case–based retrieval.

The goal of the modality detection task was to detect the acquisition modality of the images in the collection using visual, textual or mixed methods. The goal of the image–based retrieval task was to retrieve an ordered set of images from the collection that best met the information need specified as a textual statement and a set of sample images, while the goal of the case–based retrieval task was to return an ordered set of articles (rather than images) that best met the information need provided as a description of a "case".

The number of registrations to the medical task increased to 51 research groups. However, groups submitting runs have remained stable at 16, with the number of submitted runs increasing to 155. Of these, 61 were ad–hoc runs, 48 were case–based runs while the remaining 46 were modality classification runs.

The best results for the ad–hoc retrieval topics were obtained using mixed methods with textual methods also performing well. Textual methods were clearly superior for the case–based topics. For the modality detection task, although textual and visual methods alone were relatively successful, combining these techniques proved most effective.

1 Introduction

ImageCLEF[1] [1–3] started in 2003 as part of the Cross Language Evaluation Forum (CLEF[2], [4]). A medical image retrieval task was added in 2004 and has been held every year since [3, 5]. The main goal of ImageCLEF continues to be

[1] http://www.imageclef.org/

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

promoting multi–modal information retrieval by combining a variety of media including text and images for more effective information retrieval.

In 2010, the format of CLEF was changed from a workshop at the European Conference on Digital Libraries (ECDL) to an independent conference on multilingual and multimedia retrieval evaluation[3] which includes several organized evaluation tasks now called labs.

2 Participation, Data Sets, Tasks, Ground Truth

This section describes the details concerning the set–up and the participation in the medical retrieval task in 2010.

2.1 Participation

In 2010, a new record of 112 research groups registered for the four sub–tasks of ImageCLEF down from seven sub tasks in 2009. For the medical retrieval task the number of registrations also reached a new maximum with 51. 16 of the participants submitted results to the tasks, essentially the same number as in previous years. The following groups submitted at least one run:

- AUEB (Greece);
- Bioingenium (Columbia)*;
- Computer Aided Medical Diagnoses (Edu??),*;
- Gigabioinforamtics (Belgium)*;
- IRIT (France);
- ISSR (Egypt);
- ITI, NIH (USA);
- MedGIFT (Switzerland);
- OHSU (USA);
- RitsMIP (Japan)*;
- Sierre, HES–SO (Switzerland);
- SINAI (Spain);
- UAIC (Romania)*;
- UESTC (China)*;
- UIUC–IBM (USA)*;
- Xerox (France)*.

Participants marked with a star had never before participated in the medical retrieval task, indicating that the number of first–time participants was fairly high with eight among the 16 participants.

A total of 155 valid runs were submitted, 46 of which were submitted for modality detection, 61 for the image–based topics and 48 for the case–based topics. The number of runs per group was limited to ten per subtask and case–based and image–based topics were seen as separate subtasks in this view.

[3] http://www.clef2010.org/

2.2 Datasets

The database used in 2009 was again made accessible by the Radiological Society of North America (RSNA[4]). The database contained a total of 77,506 images, and was the largest collection to ever have been used for ImageCLEFmed. All images in the collection originated from the journals Radiology and Radiographics, published by the RSNA. A similar database is also available via the Goldminer[5] interface. This collection constitutes an important body of medical knowledge from the peer–reviewed scientific literature including high quality images with textual annotations. Images are associated with journal articles, and can also be part of a larger figure. Figure captions were made available to participants, as well as the sub–caption concerning a particular subfigure (if available). This high–quality set of textual annotations enabled textual searching in addition to content–based retrieval. Furthermore, the PubMed IDs of each figure's originating article were also made available, allowing participants to access the MeSH (Medical Subject Headings) index terms assigned by the National Library of Medicine for MEDLINE[6].

2.3 Modality Classification

Previous research [6] has demonstrated the utility of classifying images by modality in order to improve the precision of the search. The modality classification task was conceived as the first step for the medical image retrieval task whereby participants use the modality classifier created in this step to improve their performance for the retrieval task. For this task, 2390 images were provided as a training set where each image was classified as belonging to one of 8 classes (CT, GX, MR, NM, PET, PX, US, XR). One of the authors (JKC) had manually, but somewhat cursorily, verified the assigned modality of all images. 2620 test images were provided for the task. Each of these images were to be assigned a modality using visual, textual or mixed techniques. Participants were also requested to provide a classification for all images in the collection. A majority vote classification for all images in the collection was made available upon request to participants of the task after the evaluation.

2.4 Image–Based Topics

The topics for the image-based retrieval task were created using methods similar to previous years where realistic search topics were identified by surveying actual user needs. The starting point for the 2010 topics was a user study [7] conducted at Oregon Health & Science University (OHSU) in early 2009. Using qualitative methods, this study was conducted with medical practitioners and was focused on understanding their needs, both met and unmet, regarding medical image

[4] `http://www.rsna.org/`

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

[6] `http://www.pubmed.gov/`

retrieval. The first part of the study was dedicated to the investigation of the demographics and characteristics of participants, a population served by medical image retrieval systems (e.g., their background, searching habits, etc.). After a demonstration of state–of–the–art image retrieval systems, the second part of the study was devoted to learning about the motivation and tasks for which the intended audience uses medical image retrieval systems (e.g., contexts in which they seek medical images, types of useful images, numbers of desired answers, etc.). In the third and last part, the participants were asked to use the demonstrated systems to try to solve challenging queries, and provide responses to questions investigating how likely they would be to use such systems, aspects they did and did not like, and missing features they would like to see added. In total, the 37 participants utilized the demonstrated systems to perform a total of 95 searches using textual queries in English. We randomly selected 25 candidate queries from the 95 searches to create the topics for ImageCLEFmed 2009. Similarly, this year, we randomly selected another 25 queries from the remaining queries. From these, using the OHSU image retrieval system which was indexed using the 2009 ImageCLEF collection, we finally selected 16 topics for which at least one relevant image was retrieved by the system.

We added 2 to 4 sample images to each query from the previous collections of ImageCLEFmed. Then, for each topic, we provided a French and a German translation of the original textual description provided by the participants. Finally, the resulting set of topics was categorized into three groups: 3 visual topics, 9 mixed topics, and 4 semantic topics. This categorization of topics was based on the organizers' prior experience with how amenable certain types of search topics are to visual, textual or mixed search techniques. However, this is not an exact science and was merely provided for guidance. The entire set of topics was finally approved by a physician.

2.5 Case–Based Topics

Case–based topics were made available for the first time in 2009, and in 2010 the number of case–based topics was increased from 5 to 14, roughly half of all topics. The goal was to move image retrieval potentially closer to clinical routine by simulating the use case of a clinician who is in the process of diagnosing a difficult case. Providing this clinician with articles from the literature that discuss cases similar[7] to the case (s)he is working on can be a valuable aid to choosing a good diagnosis or treatment.

The topics were created based on cases from the teaching file Casimage [8]. This teaching file contains cases (including images) from radiological practice that clinicians document mainly for using them in teaching. 20 cases were pre–selected and a search with the diagnosis was performed in the ImageCLEF data set to make sure that there were at least a few matching articles. Fourteen topics were finally chosen. The diagnosis and all information on the chosen treatment was then removed from the cases so as to simulate the situation of the clinician

[7] "Similar" in terms of images and other clinical data on the patient.

who has to diagnose the patient. In order to make the judging more consistent, the relevance judges were provided with the original diagnosis for each case.

2.6 Relevance Judgements

The relevance judgements were performed with the same on–line system as in 2008 and 2009 for the image–based topics as well as case–based topics. The system had been adapted in 2009 for the case–based topics, displaying the article title and several images appearing in the text (currently the first six, but this can be configured). Judges were provided with a protocol for the process with specific details on what should be regarded as relevant versus non–relevant. A ternary judgement scheme was used again, 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 accurately 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 verified during the judgement process. As in previous years, judges were recruited by sending out an e–mail to current and former students at OHSU's Department of Medical Informatics and Clinical Epidemiology. Judges, primarily clinicians, were paid a small stipend for their services. Many topics were judged by two or more judges to explore inter–rater agreements and its effects on the robustness of the rankings of the systems.

3 Results

This section describes the results of ImageCLEF 2010. Runs are ordered based on the techniques used (visual, textual, mixed) and the interaction used (automatic, manual). Case–based topics and image–based topics are separated but compared in the same sections. `Trec_eval` was used for the evaluation process, and we made use of most of its performance measures.

3.1 Submissions

The numbers of submitting teams was slightly lower in 2010 than in 2009 with 16 instead of 17. The numbers of runs increased from 124 to 155. The distribution among the three run types of modality detection, image–based retrieval and case–based retrieval showed that all three types reached almost the same number of 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 the working notes.

3.2 Modality Detection Results

A variety of commonly used image processing techniques were explored by the participants. Features used included local binary patterns (LBP) [9], Tamura texture features [10], Gabor features [11], the GIFT (GNU Image Finding Tool), the Color Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD) from MPEG–7, Color and Edge Directivity Descriptor (CEDD) and Fuzzy Color and Texture Histogram (FCTH) using the Lucene image retrieval (LIRE) library, Scale Invariant Feature Transform (SIFT) [12] as well as various combinations of these. Classifiers ranged from simple k–nearest neighbors (kNN) to Ada–Boost, multilayer perceptrons and Support Vector Machines (SVMs) as well as a variety of techniques to combine the output from multiple classifiers including those derived from Bayes theory such as product, sum, maximum and mean rules

The results of the modality detection tasks are given in Table 1below. As seen in the table, the best results were obtained using mixed methods (94%) for the modality classification task. The best run using textual methods (90%) had a slightly better accuracy than the best run using visual methods (87%). However, for groups that submitted runs using different methods, the best results were obtained when they combined visual and textual methods.

3.3 Image–Based Retrieval Results

The best results for the ad–hoc retrieval topics were obtained using mixed methods. Textual methods, as in previous years also performed well. However, visual methods by themselves, were not very effective for this collection.

Visual Retrieval As in previous years, only 8 of the 61 submitted runs used purely visual techniques. As discussed previously, this collection, with extremely well annotated textual captions and images that are primarily from radiology, does not lend itself to purely visual techniques. However, as seen from the results of the mixed runs, the use of the visual information contained in the image can improve the search performance over that of a purely textual system.

An analysis of the results shows that most techniques are in a very similar range and only a single run had a significantly better result in terms of MAP. The baseline system GIFT (GNU Image Finding Tool) is in the upper half of the performance.

Textual Retrieval Participants explored a variety of information retrieval techniques from the use of stop word removal and stemming to utilizing Lucene or Lemur, commonly used toolkits to techniques using Latent Semantic Indexing, database searches using full–text Boolean queries, query expansion with external sources such as MeSH terms (manually or automatically assigned), UMLS concepts (using MetaMap) or wikipedia, to modality filtration to more complex language models that incorporate phrases (not just words) or paragraphs,

Table 1. Results of the runs of modality classification task.

Run	Group	Run Type	Classification Accuracy
XRCE_MODCLS_COMB_testset.txt	XRCE	Mixed	0.94
XRCE_MODCLS_COMB_allset.txt	XRCE	Mixed	0.94
Modality_combined.txt	RitsMIP	Mixed	0.93
result_text_image_combined.dat	ITI	Mixed	0.92
result_text_image_comb_Max.dat	ITI	Mixed	0.91
result_text_image_comb_Prod.dat	ITI	Mixed	0.91
gigabioinformatics-both.txt	GIGABIOINFORMATICS	Mixed	0.90
result_text_image_comb_CV.dat	ITI	Mixed	0.89
result_text_image_comb_Sum.dat	ITI	Mixed	0.87
Modalityall_Mix.txt	RitsMIP	Mixed	0.78
XRCE_MODCLS_TXT_allset.txt	XRCE	Textual	0.90
result_text_titile_caption_mod_Mesh.dat	ITI	Textual	0.89
entire_text_based_modality_class.dat	ITI	Textual	0.86
gigabioinformatics-text.txt	GIGABIOINFORMATICS	Textual	0.85
Modality_text.txt	RitsMIP	Textual	0.85
Modalityall_Text.txt	RitsMIP	Textual	0.85
ipl_aueb_rhcpp_full_CT.txt	AUEB	Textual	0.74
ipl_aueb_rhcpp_full_CTM.txt	AUEB	Textual	0.71
ipl_aueb_rhcpp_full_CTMA.txt	AUEB	Textual	0.53
ipl_aueb_svm_full_CT.txt	AUEB	Textual	0.53
ipl_aueb_svm_full_CTM.txt	AUEB	Textual	0.49
ipl_aueb_svm_full_CTMA.txt	AUEB	Textual	0.32
ipl_aueb_rhcpp_test_only_CTM.txt	AUEB	Textual	0.13
ipl_aueb_rhcpp_test_only_CT.txt	AUEB	Textual	0.13
ipl_aueb_rhcpp_test_only_CTMA.txt	AUEB	Textual	0.12
XRCE_MODCLS_IMG_allset.txt	XRCE	Visual	0.87
UESTC_modality_boosting	UESTC	Visual	0.82
UESTC_modality_svm	UESTC	Visual	0.80
result_image_comb_sum.dat	ITI	Visual	0.80
result_image_comb_CV.dat	ITI	Visual	0.80
entire_result_image_comb_CV.dat	ITI	Visual	0.80
entire_result_image_comb_CV.dat	ITI	Visual	0.80
result_image_combined.dat	ITI	Visual	0.79
entire_result_image_combined.dat	ITI	Visual	0.79
result_image_comb_Max.dat	ITI	Visual	0.76
entire_result_image_comb_max.dat	ITI	Visual	0.76
gigabioinformatics-visual.txt	GIGABIOINFORMATICS	Visual	0.75
result8.txt	CAMD	Visual	0.33
result7.txt	CAMD	Visual	0.32
result1.txt	CAMD	Visual	0.31
result2.txt	CAMD	Visual	0.30
result3.txt	CAMD	Visual	0.30
result4.txt	CAMD	Visual	0.30
result5.txt	CAMD	Visual	0.29
result6.txt	CAMD	Visual	0.29
entire_result_image_comb_sum.dat	ITI	Visual	0.22

Table 2. Results of the visual runs for the medical image retrieval task.

Run Name	Retrieval Type	Run Type	Group	MAP	bPref	P10
fusion_cv_merge_mean.dat	Visual	Automatic	ITI	0.0091	0.0179	0.0125
XRCE_IMG_max.trec	Visual	Automatic	XRCE	0.0029	0.0069	0.0063
fusion_cv_merge_max.dat	Visual	Automatic	ITI	0.0026	0.0075	0.0063
GE_GIFT8.treceval	Visual	Automatic	medGIFT	0.0023	0.006	0.0125
NMFAsymmetricDCT5000_k2_7	Visual	Automatic	Bioingenium	0.0018	0.011	0.0063
fusion_cat_merge_max.dat	Visual	Automatic	ITI	0.0018	0.0057	0.0063
NMFAsymmetricDCT2000_k2_5	Visual	Automatic	Bioingenium	0.0015	0.0079	0.0063
NMFAsymmetricDCT5000_k2_5	Visual	Automatic	Bioingenium	0.0014	0.0076	0.0063

sentence selection and query translation, as well as techniques such as pseudo relevance feedback. Many participants found the use of the manually assigned MeSH terms to be most useful. Modality filtration, using either text–based or image–based modality detection techniques was found to be useful by some participants while others found only minimal benefit using the modality.

Table 3. Results of the textual runs for the medical image retrieval task.

Run Name	Retrieval Type	Run Type	Group	MAP	bPref	P10
WIKI_AX_MOD_late.trec	Textual	Not applicable	XRCE	0.338	0.3828	0.5062
ipl_aueb_AdHoc_default_TC.txt	Textual	Automatic	AUEB	0.3235	0.3109	0.4687
ipl_aueb_adhoq_default_TCg.txt	Textual	Automatic	AUEB	0.3225	0.3087	0.4562
ipl_aueb_adhoq_default_TCM.txt	Textual	Automatic	AUEB	0.3209	0.3063	0.4687
ipl_aueb_AdHoc_pivoting_TC.txt	Textual	Automatic	AUEB	0.3155	0.2998	0.45
ipl_aueb_adhoq_Pivoting_TCg.txt	Textual	Automatic	AUEB	0.3145	0.2993	0.45
ipl_aueb_adhoq_Pivoting_TCM.txt	Textual	Automatic	AUEB	0.3102	0.3005	0.4375
OHSU_pm_all_all_mod.txt	Textual	Automatic	OHSU	0.3029	0.344	0.4313
OHSU_pm_major_all_mod.txt	Textual	Automatic	OHSU	0.3004	0.3404	0.4375
OHSU_all_mh_major_jaykc_mod.txt	Textual	Automatic	OHSU	0.2983	0.3428	0.4188
XRCE_CHI2AX_MOD_late.trec	Textual	Feedback	XRCE	0.2925	0.3027	0.4125
ipl_aueb_AdHoc_Modality_Filtering	Textual	Automatic	AUEB	0.2787	0.296	0.375
I_CT_T.lst	Textual	Not applicable	SINAI	0.2764	0.2693	0.425
UESTC_image_pNw.txt	Textual	Automatic	UESTC	0.2751	0.3028	0.3438
XRCE_DIR_TXT.trec	Textual	Feedback	XRCE	0.2722	0.2837	0.4
UESTC_image_pBasic.txt	Textual	Automatic	UESTC	0.2713	0.2963	0.3438
I_CT_TMDM.lst	Textual	Feedback	SINAI	0.2672	0.2683	0.4125
OHSU_high_recall_with_tit_mod_reord	Textual	Automatic	OHSU	0.2623	0.2754	0.3875
I_CT_TM.lst	Textual	Feedback	SINAI	0.2616	0.2638	0.4062
OHSU_high_recall_with_titles.txt	Textual	Automatic	OHSU	0.2592	0.2714	0.3875
issr_CT.txt	Textual	Automatic	ISSR	0.2583	0.2667	0.3187
hes-so-vs_image-based_captions	Textual	Automatic	HES-SO VS	0.2568	0.278	0.35
OHSU_all_mh_major_jaykc_mod_reord	Textual	Automatic	OHSU	0.256	0.2533	0.3813
OHSU_mm_all_mod.txt	Textual	Automatic	OHSU	0.2476	0.2594	0.4125
WIKI_LGD_IMG_MOD_late.trec	Textual	Not applicable	XRCE	0.2343	0.2463	0.3937
issr_CTS.txt	Textual	Automatic	ISSR	0.231	0.2367	0.2812
issr_CTP.txt	Textual	Automatic	ISSR	0.2199	0.2424	0.325
ad_hoc_QE_0.1_Citations_All_Im_Txt	Textual	Automatic	ITI	0.188	0.2158	0.375
ad_hoc_queries_terms_0.1_Cit_All_Txt	Textual	Automatic	ITI	0.1589	0.1785	0.325
issr_CT_T_MT.txt	Textual	Automatic	ISSR	0.1472	0.1516	0.2571
issr_CT_T_dic.txt	Textual	Manual	ISSR	0.1394	0.1117	0.1929
hes-so-vs_image_fulltext	Textual	Automatic	HES-SO VS	0.1312	0.1684	0.1813
NMFText_k2_11	Textual	Automatic	Bioingenium	0.1005	0.1289	0.1875
issr_CT_T_MeSh.txt	Textual	Automatic	ISSR	0.0985	0.1205	0.15

Best results were obtained by the University of Athens and Xerox. The baseline runs using Lucene with the captions and full text and without any optimization performed in the lower half.

Multimodal Retrieval This year, the run with the highest MAP utilized a multimodal approach to retrieval. However, many groups that performed a pure fusion of the text–based and image–based runs found a significant deterioration in performance as the visual runs had very poor performance. This year's results again emphasize the previously noted observations that although the use of visual information can improve the search results over purely textual methods, the process of effectively combining the information from the captions and image itself can be quite complex and are often not robust. Simple approaches of fusing visual and textual runs rarely lead to optimized performance.

Table 4. Results of the multimodal runs for the medical image retrieval task.

Run Name	Retrieval Type	Run Type	Group	MAP	bPref	P10
XRCE_AX_rerank_comb.trec	Mixed	Automatic	XRCE	0.3572	0.3841	0.4375
XRCE_CHI2_LOGIT_IMG_MOD_late.trec	Mixed	Automatic	XRCE	0.3167	0.361	0.3812
XRCE_AF_LGD_IMG_late.trec	Mixed	Automatic	XRCE	0.3119	0.3201	0.4375
WIKI_AX_IMG_MOD_late.trec	Mixed	Automatic	XRCE	0.2818	0.3279	0.3875
OHSU_all_mh_major_all_mod_reorder.txt	Mixed	Automatic	OHSU	0.256	0.2533	0.3813
OHSU_high_recall.txt	Mixed	Automatic	OHSU	0.2386	0.2533	0.3625
queries_terms_0.1_Modalities.trec	Mixed	Automatic	ITI	0.1067	0.1376	0.2812
XRCE_AX_rerank.trec	Mixed	Automatic	XRCE	0.0732	0.1025	0.1063
Exp_Queries_Cit_CBIR_CV_MERGE_MAXt	Mixed	Automatic	ITI	0.0641	0.0962	0.1438
runMixt.txt	Mixed	Automatic	UAIC2010	0.0623	0.0666	0.1313
Exp_Queries_Cit_CBIR_CAT_MERGE_MAX	Mixed	Automatic	ITI	0.0616	0.0975	0.1375
Queries_Citations_CBIR_CV_MERGE_MAX	Mixed	Automatic	ITI	0.0583	0.0783	0.125
Multimodal-Rerank-ROI-QE-Merge	Mixed	Automatic	ITI	0.0486	0.0803	0.1
NMFAsymmetricMixed_k2_11	Mixed	Automatic	Bioingenium	0.0395	0.047	0.0438
GE_Fusion_img_fulltext_Vis0.2.run	Mixed	Automatic	medGIFT	0.0245	0.0718	0.0375
GE_Fusion_img_captions_Vis0.2.run	Mixed	Automatic	medGIFT	0.0208	0.0753	0.0375

Interactive Retrieval This year, as in previous years, interactive retrieval was only used by a very small number of participants. The results were not substantially better than automatic runs. This continues to be an area where we would like to see improved participation but little success in doing so. For this reason the manual and interactive runs are not shown in separate tables.

3.4 Case–based Retrieval Results

In terms of case–based retrieval almost all groups focused on using textual retrieval techniques as combining visual retrieval on a case basis is a difficult approach. Best results were obtained with a textual retrieval approach when using relevance feedback.

Visual Retrieval The performance of the single visual run submitted (see Table 5) shows that the results are much lower than the text–based techniques. Still, compared with the image–based retrieval only a single image–based run had a higher MAP, meaning that also case–based retrieval is possible with purely visual retrieval techniques and can be used as a complement to the text approaches.

Textual Retrieval The vast majority of submissions was in the category of textual retrieval (see Table6). Best results were obtained by a collaboration of IBM and UIUC in the textual part. Surprisingly the baseline text result of using

Table 5. Results of the visual runs for the medical image retrieval task (case–based topics).

Run Name	Retrieval Type	Run Type	Group	MAP	bPref	P10
1276253404955_GE_GIFT8_case	Visual	Automatic	medGIFT	0.0358	0.0612	0.0929

Lucene with the full text articles and with absolutely no optimization has the third best result and is within the limit of statistical significance of the best run. The first three runs are basically very close and then the performance slowly drops of. In general results are slightly lower than for the image–based topics. The baseline run using the image captions and then combining results of the single images obtains a much lower performance.

For the first time in several years there was actually a substantial number of feedback runs, although only two groups submitted feedback runs (see Table 7). These runs show that relevance feedback can improve results, although the improvement is fairly low compared with the automatic run. All but one of the feedback runs has very good results, showing that the techniques work in a stable manner.

Multimodal Retrieval Only two participants actually submitted a mixed case–based result, and the performance of these runs is fairly low highlighting the difficulty in combining the textual and visual results properly. Much more research on the visual and combined retrieval seems necessary as the current techniques in this field do not seem to work in a satisfying way. For this reason an information fusion task using ImageCLEF 2009 data was organized at ICPR 2010, showing an enormous increase in performance when good fusion techniques are applied even when the base results have very strong variations in performance [13]. Very few of these runs using more sophisticated fusion techniques had a degradation in performance over the best single run.

3.5 Relevance Judgement Analysis

A number of topics, both image–based and case–based, were judged by two or even three judges. Seven topics were judged by two judges while two additional topics were judged by three judges. There were significant variations in the kappa metric used to evaluate the inter–rater agreement. Kappa for these topics ranged from 0 to 1. The average kappa was 0.47. However, there were 4 topics where the kappa was zero as one judge had assessed no images as being relevant while the other had said that 1–11 images were relevant. On the other hand, there was a topic where both judges agreed that only a single image was relevant. Topics with low number of relevant images (¡10) can cause difficulties in evaluation as difference in opinions between judges one a single image can result in large differences in performance metrics for that topic. Without these topics, the average kappa was 0.657, a more acceptable figure.

Table 6. Results of the textual runs for the medical image retrieval task (Case–Based Topics).

Run	Retrieval Type	Run Type	Group	MAP	bPref	P10
baselinefbWMR_10_0.2sub	Textual	Automatic	UIUCIBM	0.2902	0.3049	0.4429
baselinefbWsub	Textual	Automatic	UIUCIBM	0.2808	0.2816	0.4429
hes-so-vs_case-based_fulltext.txt	Textual	Automatic	HES-SO VS	0.2796	0.2699	0.4214
baselinefbsub	Textual	Automatic	UIUCIBM	0.2754	0.2856	0.4286
baselinefbWMD_25_0.2sub	Textual	Automatic	UIUCIBM	0.2626	0.2731	0.4
C_TA_T.lst	Textual	Not applicable	SINAI	0.2555	0.2518	0.3714
IRIT_SemAnnotator-2.0_BM25_N28.res	Textual	Automatic	IRIT	0.2265	0.2351	0.3429
C_TA_TM.lst	Textual	Not applicable	SINAI	0.2201	0.2307	0.3643
IRIT_SemAnnotator-2.0_BM25_N28_1.res	Textual	Automatic	IRIT	0.2193	0.2139	0.3286
IRIT_SemAnnotator-1.5.2_BM25_N34.res	Textual	Automatic	IRIT	0.2182	0.2267	0.3571
IRIT-run-bl.res	Textual	Automatic	IRIT	0.2103	0.1885	0.2786
IRIT_SemAnnotator-1.5.2_BM25_N34_1.res	Textual	Automatic	IRIT	0.2085	0.2083	0.3143
IRIT_SemAnnotator-2.0_BM25_N34_1.res	Textual	Automatic	IRIT	0.2085	0.2083	0.3143
ISSR_cb_cts.txt	Textual	Automatic	ISSR	0.1986	0.1883	0.3071
ISSR_cp_ctp.txt	Textual	Automatic	ISSR	0.1977	0.1873	0.3
ISSR_CB_CT.txt	Textual	Automatic	ISSR	0.1977	0.1873	0.3
ipl_aueb_CaseBased_CTM_0.2.txt	Textual	Automatic	IPL	0.1874	0.1927	0.3214
ipl_aueb_CaseBased_CTM_0.1.txt	Textual	Automatic	IPL	0.186	0.1897	0.3214
ipl_aueb_CaseBased_CT.txt	Textual	Automatic	IPL	0.1841	0.1803	0.3143
ipl_aueb_CaseBased_CTM_0.3.txt	Textual	Automatic	IPL	0.1833	0.1919	0.3143
ipl_aueb_CaseBased_CTM_0.4.txt	Textual	Automatic	IPL	0.1809	0.1895	0.3143
ipl_aueb_CaseBased_CTM_0.4.txt	Textual	Automatic	IPL	0.1809	0.1895	0.3143
ipl_aueb_CaseBased_CTM_0.5.txt	Textual	Automatic	IPL	0.1716	0.1811	0.3429
UESTC_case_pBasic.txt	Textual	Automatic	UESTC	0.1692	0.184	0.2643
UESTC_case_pQE.txt	Textual	Automatic	UESTC	0.1677	0.1852	0.2786
UESTC_case_pNw.txt	Textual	Automatic	UESTC	0.1522	0.1725	0.2714
case_based_expanded_queries_backoff_0.1.trec	Textual	Automatic	ITI	0.1501	0.1749	0.2929
case_based_queries_backoff_0.1.trec	Textual	Automatic	ITI	0.128	0.1525	0.2357
hes-so-vs_case-based_nodoubles_captions.txt	Textual	Automatic	HES-SO VS	0.1273	0.1375	0.25
case_based_expanded_queries_types_0.1.trec	Textual	Automatic	ITI	0.1217	0.1502	0.2929
C_TAbs_TM.lst	Textual	Not applicable	SINAI	0.1146	0.1661	0.2643
case_based_queries_pico_MA_0.1.trec	Textual	Automatic	ITI	0.1145	0.1439	0.2
C_TAbs_T.lst	Textual	Not applicable	SINAI	0.1076	0.166	0.2571
case_based_queries_types_0.1.trec	Textual	Automatic	ITI	0.0996	0.1346	0.2286
case_based_queries_terms_0.1.trec	Textual	Automatic	ITI	0.0522	0.07	0.0857
GE_GIFT8_case.treceval	Visual	Automatic	medGIFT	0.0358	0.0612	0.0929

Table 7. Results of the textual interactive and feedback runs for the medical image retrieval task (Case–Based Topics).

Run	Retrieval Type	Run Type	Group	MAP	bPref	P10
PhybaselineRelfbWMR_10_0.2sub	Textual	Feedback	UIUCIBM	0.3059	0.3348	0.4571
PhybaselineRelfbWMD_25_0.2sub	Textual	Feedback	UIUCIBM	0.2837	0.3127	0.4571
PhybaselineRelFbWMR_10_0.2_top20sub	Textual	Feedback	UIUCIBM	0.2713	0.2897	0.4286
case_based_queries_pico_backoff_0.1.trec	Textual	Feedback	ITI	0.1386	0.1666	0.2
PhybaselinefbWMR_10_0.2sub	Textual	Manual	UIUCIBM	0.3551	0.3714	0.4714
PhybaselinefbWsub	Textual	Manual	UIUCIBM	0.3441	0.348	0.4714
PhybaselinefbWMD_25_0.2sub	Textual	Manual	UIUCIBM	0.3441	0.348	0.4714
case_based_expanded_queries_terms_0.1.trec	Textual	Manual	ITI	0.0601	0.0825	0.0857

Table 8. Results of the multimodal runs for the medical image retrieval task (Case–Based Topics).

Run	Retrieval Type	Run Type	Group	MAP	bPref	P10
case_based_queries_cbir_with_case_backoff	Mixed	Automatic	ITI	0.0353	0.0509	0.0429
case_based_queries_cbir_without_case_backoff	Mixed	Automatic	ITI	0.0308	0.0506	0.0214
GE_Fusion_case_captions_Vis0.2	Mixed	Automatic	medGIFT	0.0143	0.0657	0.0357
GE_Fusion_case_fulltext_Vis0.2	Mixed	Automatic	medGIFT	0.0115	0.0786	0.0357

3.6 Robustness of Rankings

We briefly explored the variability in the rankings of the various runs caused by using different judges for a topic, especially on topics that had very few relevant images. Topics 2 and 8 had a kappa of zero as one judge had not found any relevant images in the pool with the other found 1 and 9 relevant images respectively. Both judges had found one relevant image for topic 7. We explored the changes in ranking caused by eliminating these topics from the evaluation. Most runs had a none to substantial improvement in bpref with three runs demonstrating a substantial improvement in rankings without these topics. However, four runs had a drop in bpref as these runs had performed quite well on topic 7 and extremely well on topic 8. The relative rankings of the groups were vastly unchanged with using the assessment of different judges aside from topics with low number of relevant images.

4 Conclusions

As in 2009, the largest number of runs for the image–based and case–based tasks used textual techniques. The semantic topics combined with a database containing high–quality annotations lend themselves to textual methods. However, unlike in 2009, the best runs were those that effectively combined visual and textual methods. Visual runs continue to be rare and generally poor in performance.

Case–based topics had an increased participation over last year. As may be expected based on the nature of the task, case–based retrieval is more easily accomplished using textual techniques. Unlike in the ad–hoc runs, combining visual image severely degraded the performance for case–based topics, meaning that much more care needs to be taken with these combinations. More focus has to be put on the combinations to increase performance. Maybe a pure fusion task of results could be an additional challenge for the coming years.

A kappa analysis between several relevance judgements for the same topics shows that, although there are differences between judges, there was moderate agreement on topics that have more than 10 relevant images. As a result topics with very few relevant images could be removed or a more thorough testing could already remove them during the topic creation process.

For future campaign it seems important to explore how to effectively combine visual techniques with the text–based methods. As has been stated at previous ImageCLEFs, we strongly believe that interactive and manual retrieval are important and we strive to improve participation in these. This year's results show that even simple feedback can significantly improve results.

5 Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. This work was partially funded by the Swiss National Science Foundation (FNS) under contracts 205321–109304/1 and PBGE22–121204, the American National Science Foundation (NSF) with grant ITR–0325160, Google, the National Library of MEdicine grant K99LM009889, and the EU FP7 projects Khresmoi and Promise. We would like to thank the RSNA for supplying the images of their journals Radiology and Radiographics for the ImageCLEF campaign.

References

1. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Cross Language Evaluation Forum (CLEF 2005). Springer Lecture Notes in Computer Science (September 2006) 535–557
2. Clough, P., Müller, H., Sanderson, M.: The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign. Volume 3491 of Lecture Notes in Computer Science (LNCS)., Bath, UK, Springer (2005) 597–613
3. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: 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, Springer (2008) 473–491
4. Savoy, J.: Report on CLEF–2001 experiments. In: Report on the CLEF Conference 2001 (Cross Language Evaluation Forum), Darmstadt, Germany, Springer LNCS 2406 (2002) 27–43
5. Müller, H., Rosset, A., Vallée, J.P., Terrier, F., Geissbuhler, A.: A reference data set for the evaluation of medical image retrieval systems. Computerized Medical Imaging and Graphics **28**(6) (2004) 295–305
6. Kalpathy-Cramer, J., Hersh, W.: Multimodal medical image retrieval: image categorization to improve search precision. In: MIR '10: Proceedings of the international conference on Multimedia information retrieval, New York, NY, USA, ACM (2010) 165–174
7. Radhouani, S., Hersh, W., Kalpathy-Cramer, J., Bedrick, S.: Understanding and improving image retrieval in medicine. Technical report, Oregon Health and Science University (2009)
8. Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallée, J.P., Ratib, O.: Casimage project — a digital teaching files authoring environment. Journal of Thoracic Imaging **19**(2) (2004) 1–6

9. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray–scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7) (July 2002) 971–987
10. Tamura, H., Mori, S., Yamawaki, T.: Texture features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics **8**(6) (1978) 460–472
11. Ma, W., Manjunath, B.: Texture features and learning similarity. In: Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96), San Francisco, California (June 1996) 425–430
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
13. Müller, H., Kalpathy-Cramer, J.: The ImageCLEF medical retrieval task at icpr 2010 — information fusion to combine viusal and textual information. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2010). Lecture Notes in Computer Science (LNCS), Istanbul, Turkey, Springer (August 2010) in press.