Overview of the CLEF 2009 medical image retrieval track

Henning Müller[1,2], Jayashree Kalpathy–Cramer[3], Ivan Eggel[1], Steven Bedrick[3], Saïd Radhouani[3], Brian Bakke[3], Charles E. Kahn Jr.[4], William Hersh[3]
[1]University of Applied Sciences Western Switzerland (HES–SO), Sierre, Switzerland
[2]University Hospitals and University of Geneva, 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

2009 was the sixth year for the ImageCLEF medical retrieval task. Participation was strong again with 38 registered research groups. 17 groups submitted runs and thus participated actively in the tasks. The database in 2009 was similar to the one used in 2008, containing scientific articles from two radiology journals, Radiology and Radiographics. The size of the database was increased to a total of 74,902 images. For each image, captions and access to the full text article through the Medline PMID (PubMed Identifier) were provided. An article's PMID could be used to obtain the officially assigned MeSH (Medical Subject Headings) terms. The collection was entirely in English. However, the topics were, as in previous years, supplied in German, French, and English. Twenty–five image–based topics were provided, of which ten each were visual and mixed and five were textual. In addition, for the first time, 5 case–based topics were provided as an exploratory task. Here the unit of retrieval was intended to be the article and not the image. Case–based topics are designed to be a step closer to the clinical workflow. Clinicians often seek information about patient cases with incomplete information consisting of symptoms, findings, and a set of images. Supplying cases to a clinician from the scientific literature that are similar to the case (s)he is treating can be an important application of image retrieval in the future.

As in previous years, most groups concentrated on fully automatic retrieval. However, four groups submitted a total of seven manual or interactive runs. The interactive runs submitted this year performed quite well compared to previous years but did not show a substantial increase in performance over the automatic approaches. In previous years, multimodal combinations were the most frequent submissions. However, this year, as in 2008 only about half as many mixed runs as purely textual runs were submitted. Very few fully visual runs were submitted, and again, the ones submitted performed poorly. The best mean average precisions (MAP) were obtained using automatic textual methods. There were mixed feedback runs that had high MAP. The best early precision was also obtained using automatic textual methods, with a few mixed automatic runs also doing well. We had the opportunity to perform multiple judgments on some topics. The kappas used as the metric for inter–rater agreement were mostly quite high (¿0.7). However, one of our judges consistently had low kappas as he was significantly more lenient the colleagues. We evaluated the overall performance of groups using strict and lenient judges and found that there was high correlation even though the absolute values for the metrics were different.

We also introduced a lung nodule detection task in 2009. This task used the CT slices from the Lung Imaging Data Consortium (LIDC) which included ground truth

in the form of manual annotations. The goal of the task was to create algorithms to automatically detect lung nodules. Although there seemed to be significant interest in the task as evidenced by the substantial number of registrations, only two groups submitted results with a proprietary software from a industry participant achieving impressive results.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Medical image retrieval, image retrieval, multimodal retrieval

1 Introduction

ImageCLEF[1] [1, 2, 4] started in 2003 as part of the Cross Language Evaluation Forum (CLEF[2], [9]). A medical image retrieval task was added in 2004 and has been held every year since [4, 7]. The main goal of ImageCLEF in the past has been to promote multi–modal information retrieval by combining a variety of media including text and images for more effective information retrieval. As such, it has always contained visual, textual and mixed tasks and sub–tracks. The medical image retrieval track began in 2004 as a primarily visual information retrieval task with a teaching database of 8,000 images. Since then, it progressed to a collection of over 66,000 images from several teaching collections with topics that were best suited for textual, visual and mixed methods. In 2008, images from the medical literature were used for the first time, moving the task one step closer towards applications that can be of interest in clinical scenarios. Several user studies have been performed to study the image searching behaviour of clinicians [5, 6, 3]. These studies have been used to create the task and the topics over the years. This year, for the first time, we introduced a case–based retrieval task as we continue to strive for scenarios that more closely resemble actual clinical work–flows.

This paper reports on the medical retrieval task. Additionally, other papers within ImageCLEF describe the other five tasks of ImageCLEF 2009. More information on the tasks and on how to participate in CLEF can also be found on the ImageCLEF web pages.

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 2009. A new management system for participation in ImageCLEF was created to better manage the increasing number of registrations and submissions to the ImageCLEF benchmark in a fully electronic fashion. The interface allowed registrations for particular tasks, provided the links to the description of the data sets available, allowed submission of the results and enabled the final evaluation.

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

2.1 Participation

In 2009, a new record of 85 research groups registered for the seven sub–tasks of ImageCLEF. For the medical retrieval task the participation remained similar to the previous year with 37 registrations. 17 of the participants submitted results to the tasks, a slight increase from 15 in 2008. The following groups submitted at least one run:

- NIH (USA);
- Liris (France);
- ISSR (Egypt)*;
- UIIP Minsk (Belarus)*;
- MedGIFT (Switzerland);
- Sierre (Switzerland)*;
- SINAI (Spain);
- Miracle (Spain);
- BiTeM (Switzerland);
- York University (Canada)*;
- AUEB (Greece);
- University of Milwaukee (USA)*;
- University of Alicante (Spain);
- University of North Texas (USA)*;
- OHSU (USA);
- University of Fresno (USA);
- DEU (Turkey).

Participants marked with a star had never participated in the past in a medical retrieval task, indicating that the number of first–time participants is fairly high with six among the 17 participants.

A total of 124 valid runs were submitted, 106 of which were submitted for the image–based topics while 18 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. This was an increase compared to the 111 runs submitted last year.

2.2 Datasets

The database in 2009 was again made accessible by the Radiological Society of North America (RSNA[3]). The database contained a total of 74'902 images, the largest collection yet. All images are taken from the journals Radiology and Radiographics of the RSNA. A similar database is also available via the Goldminer[4] interface. This collection constitutes an important body of medical knowledge from the peer–reviewed scientific literature including high quality images with annotations. Images are associated with journal articles and can be part of a figure. Figure captions are made available to participants as well as the part concerning a particular subfigure if

[3] http://www.rsna.org/
[4] http://goldminer.arrs.org/

available. This creates high–quality textual annotations enabling textual searching in addition to content–based retrieval. As the PubMed IDs were also made available, participants could access the MeSH (Medical Subject Headings) terms created by the National Library of Medicine for PubMed[5].

2.3 Image–Based Topics

The image-based topics were created using methods similar to previous years where realistic search topics were identified by surveying actual user needs. The starting point for this year's topics was a user study [8] conducted at Oregon Health & Science University (OHSU) during early 2009. Based on qualitative methods, this study was conducted with 37 medical practitioners in order to understand the needs, both met and unmet, in medical image retrieval. The first part of the study was dedicated to the investigation of the characteristics of a large portion of the 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, trying to solve challenging queries, and provide responses to them in terms of how likely they would be to use them, which aspects they did and did not like, and which missing features they would like to see added. In total, the 37 participants used 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. We added to each candidate query 2 to 4 sample images 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 the topics was categorized into three groups: 10 visual topics, 10 mixed topics, and 5 semantic topics. The entire set of topics was finally approved by a physician.

2.4 Case–Based Topics

Case–based topics were made available for the first time in 2009. 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 treat cases similar to the case (s)he is working on ("similar" based on images and other clinical data on the patient) can be a valuable aide to choosing a good diagnosis or treatment.

The topics were cerated based on cases from the teaching file Casimage. This teaching file contains cases including images from radiological practice. 10 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. Five topics were finally chosen. The diagnosis and all information on the chosen treatment was then removed from the cases to simulate a situation of the clinician 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.

2.5 Relevance Judgements

The relevance judgements were performed with the same on–line system as in 2008 for the image–based topics. The system was adapted for the case–based topics showing the article title and several images appearing in the text (currently the first six, but this can be configured). Besides a short description for the judgements, a full document was prepared to describe the judging process, including 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",

[5] http://www.pubmed.gov/

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 verified during the judgement process.

We had the opportunity to perform multiple judgements on many topics, both image–based and case–based. Inter–rater agreement was assessed using the kappa metric, given as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the observed agreement between judges and $P(E)$ is the expected (random) agreement. These are calculated using a 2x2 table for the relevances of images or articles. These were calculated for both lenient where a "partly relevant" is considered relevant, and strict judgments where "partly relevant" is considered not–relevant. It is generally accepted that a kappa < 0.7 is good and sufficient for an evaluation. In general the agreement between the judges was fairly high with few exceptions and the overall average κ is similar to other evaluation campaigns. Regarding the case–based topics it seems necessary to take longer for the judges but we did not receive much feedback, so all judges seemed to be satisfied with the written description on the judgments that were supplied.

3 Results

This section describes the results of ImageCLEF 2009. 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.

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 sufficient detail.

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 higher in 2009 than in 2008 with 17. The numbers of runs increased from 111 to 124. This was partly due to the fact that with the case–based topics and the image–based topics there were two more run categories.

A total of 124 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, so only few 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 participants, a large number of runs were quickly repaired, resulting in 122 valid runs taken into account for the pools.

In total, only 13 runs were "manual" or "interactive." There were only 16 "visual–only". The large majority were "text–only runs", with 59 submissions. There were 30 mixed runs

Groups subsequently had the chance to evaluate additional runs themselves as the qrels were made available to participants 2 weeks ahead of the submission deadline for the working notes.

3.2 Image–Based Results

3.2.1 Visual Retrieval

The number of visual runs in 2009 was small, and the improvement in the results is not as fast as with textual retrieval techniques. 5 groups submitted a total of 16 runs in 2009, one of which was

Table 1: Results of the visual runs for the medical image retrieval task.

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
CBIR_FUSION_MERGE	Visual Automatic	0.01	0.04	0.08	0.07	0.05	295
medGIFT_sep_max	Visual Automatic	0.01	0.03	0.09	0.08	0.06	266
medGIFT_sum_withAR	Visual Automatic	0.01	0.03	0.09	0.07	0.05	262
CBIR_FUSION_CV_MERGE	Visual Automatic	0.01	0.03	0.09	0.08	0.05	289
medGIFT_sep_sum	Visual Automatic	0.01	0.03	0.05	0.05	0.06	259
medGIFT_sep_max_withAR	Visual Automatic	0.01	0.03	0.08	0.08	0.05	253
CBIR_FUSION_CATEGORY	Visual Automatic	0.01	0.03	0.06	0.06	0.04	315
medGIFT_sum_withNegImg	Visual Automatic	0.01	0.03	0.06	0.04	0.05	210
medGIFT_max_withNegImg.txt	Visual Automatic	0.01	0.02	0.06	0.04	0.04	201
clef2009	Visual Automatic	0.00	0.02	0.02	0.03	0.02	242
UIIPMinsk_visual_1	Visual Automatic	0.00	0.01	0.03	0.02	0.01	80
UIIPMinsk_visual_2	Visual Automatic	0.00	0.01	0.02	0.02	0.01	91
CSUFresno_visual_CEDD	Visual Automatic	0.00	0.03	0.04	0.02	0.02	162
CSUFresno_visual_CEDD	Visual Automatic	0.00	0.01	0.04	0.02	0.02	13
CSUFresno_visual_CEDD	Visual Automatic	0.00	0.02	0.02	0.01	0.01	89

feedback. Performance as measured in MAP is very low for all these runs, reaching a maximum of 0.0136 for the best run. Both early precision and recall were quite low for the visual runs when compared to the textual runs but there were significant differences between the visual runs. The University of Minsk only submitted runs to a subset of the topics and this made their average performance look much worse than than the other runs. A more detailed per topics analysis seems necessary to really compare the systems.

Table 1 shows the results and particularly the large differences between the runs. Runs retrieved between 13 and 315 of 2362 possible relevant images, which is substantially lower than the poorest performing the textual runs.

Part of the performance can be explained with the extremely well annotated database that created a much larger gap between visual and textual results. The topics in ImageCLEFmed also became harder, making even the visual topics more semantic than before. This corresponds clearly to user needs. The small number of submitted visual runs also biases the pools towards the textual runs, even further widening the gap.

3.2.2 Textual Retrieval

Purely automatic textual retrieval had by far the largest number of runs in 2009 with 52, more than 46% 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 LIRIS obtained the best results with 8 of the top 10 runs. These used conceptual language modelling with the additional use of the UMLS (Unified Medical Language System) metathesaurus. They had many runs with MAP between 0.43 and 0.41. A more detailed analysis is required with the exact techniques applied for each of the runs.

3.2.3 Multimodal 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 2009 as in 2008, however, only about half as many mixed runs as purely textual runs were submitted.

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, for some groups, the mixed–media submissions often have higher early precision than the purely textual retrieval submissions.

All runs exhibited relatively high correlation between MAP and bpref.

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

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

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
LIRIS_maxMPTT_extMPTT	Text Automatic	0.43	0.46	0.70	0.66	0.55	1814
LIRIS_maxMPTT_extMPTTEF	Text Automatic	0.42	0.45	0.67	0.62	0.53	1801
LIRIS_maxMPTT_enMPTT	Text Automatic	0.42	0.44	0.70	0.68	0.56	1689
LIRIS_maxMPTT_enMMMPTTEF	Text Automatic	0.42	0.43	0.72	0.68	0.55	1685
LIRIS_KL_maxMPTT_extMMMPTTEF	Text Automatic	0.42	0.44	0.69	0.64	0.53	1784
LIRIS_KL_maxMPTT_enMMMPTTEF	Text Automatic	0.42	0.43	0.73	0.68	0.54	1678
LIRIS_KL_maxMPTT_extMMMPTT	Text Automatic	0.42	0.44	0.68	0.62	0.53	1793
LIRIS_KL_maxMPTT_enMMMPTT	Text Automatic	0.41	0.43	0.70	0.69	0.55	1682
sinai_CTM_t	Text Automatic	0.38	0.39	0.65	0.62	0.56	1884
LIRIS_maxMPTT_frTT_tradMPTT	Text Automatic	0.38	0.41	0.58	0.58	0.47	1576
york.In_expB2c1.0	Text Automatic	0.37	0.38	0.61	0.60	0.51	1762
sinai_CT_t	Text Automatic	0.36	0.36	0.58	0.60	0.54	1869
ISSR_text_1	Text Automatic	0.35	0.36	0.58	0.56	0.49	1717
ceb-essie2-automatic	Text Automatic	0.35	0.40	0.65	0.62	0.54	1554
york.bm25	Text Automatic	0.35	0.36	0.60	0.57	0.48	1759
deu_run1_pivoted	Text Automatic	0.34	0.35	0.58	0.52	0.45	1742
clef2009	Text Automatic	0.34	0.36	0.67	0.60	0.50	1803
ISSR_Text_2	Text Automatic	0.33	0.35	0.58	0.52	0.44	1768
sinai_C_t	Text Automatic	0.33	0.36	0.61	0.57	0.50	1590
sinai_CTM_tM	Text Automatic	0.33	0.35	0.58	0.54	0.48	1666
BiTeM_EN	Text Automatic	0.32	0.33	0.52	0.50	0.42	1752
sinai_CM_t	Text Automatic	0.31	0.34	0.62	0.57	0.54	1582
deu_run2_simple	Text Automatic	0.31	0.34	0.61	0.53	0.45	1620
sinai_CT_tM	Text Automatic	0.31	0.32	0.51	0.53	0.44	1729
ISSR_Text_FR_1	Text Automatic	0.30	0.34	0.53	0.48	0.40	1643
BiTeM_FRtl	Text Automatic	0.29	0.32	0.48	0.44	0.38	1699
deu_simple_rrank_dtree	Text Automatic	0.29	0.32	0.59	0.51	0.46	1615
sinai_CM_tM	Text Automatic	0.28	0.33	0.54	0.56	0.47	1399
deu_run3_pivot_rrank_dtree	Text Automatic	0.28	0.32	0.59	0.52	0.42	1570
sinai_C_tM	Text Automatic	0.28	0.32	0.56	0.52	0.45	1494
BiTeM_FR	Text Automatic	0.28	0.30	0.46	0.44	0.37	1641
UNTtextrf1	Text Automatic	0.26	0.28	0.53	0.44	0.38	1762
UNTtextb1	Text Automatic	0.24	0.28	0.46	0.40	0.37	1642
BiTeM_DEtl	Text Automatic	0.23	0.27	0.41	0.37	0.34	1606
BiTeM_DE	Text Automatic	0.22	0.25	0.41	0.37	0.34	1545
BiTeM_ENsy	Text Automatic	0.20	0.24	0.40	0.38	0.31	1387
ISSR_Text_DE_1	Text Automatic	0.20	0.24	0.42	0.39	0.30	1608
OHSU_SR1	Text Automatic	0.18	0.22	0.59	0.54	0.41	801
MirEN	Text Automatic	0.17	0.23	0.62	0.55	0.39	912
MirTaxEN	Text Automatic	0.16	0.22	0.59	0.52	0.38	913
Mir	Text Automatic	0.15	0.21	0.58	0.47	0.31	842
uwmTextOnly	Text Automatic	0.13	0.18	0.44	0.40	0.31	572
Alicante-Run3	Text Automatic	0.13	0.16	0.34	0.36	0.34	996
Alicante-Run1	Text Automatic	0.13	0.16	0.38	0.40	0.35	958
MirRF0505EN	Text Automatic	0.13	0.18	0.59	0.51	0.34	567
MirTax	Text Automatic	0.13	0.19	0.50	0.40	0.29	843
ohsu_j_no_mod	Text Automatic	0.12	0.18	0.42	0.38	0.30	896
MirRFTax0505EN	Text Automatic	0.10	0.15	0.45	0.36	0.23	568
MirRF1005EN	Text Automatic	0.09	0.13	0.54	0.41	0.23	459
MirRF0505	Text Automatic	0.07	0.11	0.43	0.31	0.21	430
MirRFTax1005EN	Text Automatic	0.07	0.12	0.41	0.30	0.18	470
MirRFTax0505	Text Automatic	0.05	0.09	0.29	0.22	0.17	447

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

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
deu_imaged_vsm	Mixed Automatic	0.37	0.39	0.63	0.54	0.48	1754
york.BO1.EdgeHistogram0.2	Mixed Automatic	0.36	0.37	0.58	0.58	0.54	1724
york.BO1.Tamura0.2	Mixed Automatic	0.35	0.37	0.62	0.57	0.51	1722
BM25b=0.75k_1=1.2k_3=8.0_ICLEFPProcess_3	Mixed Automatic	0.35	0.37	0.60	0.59	0.50	1763
York.BO1.colorHistogram0.2	Mixed Automatic	0.34	0.36	0.59	0.57	0.50	1719
BM25b=0.75k_1=1.2k_3=8.0_ICLEFPProcess_1	Mixed Automatic	0.33	0.35	0.59	0.57	0.47	1757
medGIFT0.3_withNegImg_EN	Mixed Automatic	0.29	0.32	0.63	0.60	0.52	1176
Multimodal_Text_Rerank	Mixed Automatic	0.27	0.40	0.49	0.52	0.45	1553
UNTMixed Automatic1	Mixed Automatic	0.24	0.28	0.46	0.40	0.37	1659
medGIFT0.5_EN	Mixed Automatic	0.21	0.25	0.70	0.59	0.43	848
UNTMixed Automaticrf1	Mixed Automatic	0.19	0.24	0.50	0.42	0.37	1197
ohsu_j_umls	Mixed Automatic	0.18	0.21	0.71	0.66	0.42	591
ohsu_j_mod1	Mixed Automatic	0.17	0.22	0.59	0.55	0.38	943
OHSU_SR6	Mixed Automatic	0.16	0.20	0.68	0.61	0.43	543
OHSU_SR2	Mixed Automatic	0.16	0.21	0.62	0.54	0.39	801
OHSU_SR3	Mixed Automatic	0.15	0.20	0.61	0.52	0.37	801
clef2009	Mixed Automatic	0.15	0.21	0.38	0.33	0.26	1381
medGIFT_mix_0.5vis_withNegImg	Mixed Automatic	0.14	0.17	0.56	0.49	0.33	547
Alicante-Run4	Mixed Automatic	0.13	0.17	0.31	0.34	0.33	992
uwmTextAndModality	Mixed Automatic	0.13	0.17	0.49	0.46	0.38	521
Alicante-Run5	Mixed Automatic	0.13	0.16	0.33	0.35	0.33	982
OHSU_SR4	Mixed Automatic	0.11	0.15	0.60	0.48	0.31	381
OHSU_SR5	Mixed Automatic	0.11	0.15	0.58	0.52	0.31	514
uwmTextAndImageDistance	Mixed Automatic	0.07	0.09	0.44	0.40	0.27	204
medGIFT_sum_withNegImg	Mixed	0.01	0.03	0.06	0.04	0.05	210

Table 4: Results of the interactive and manual runs for the medical image retrieval task.

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
ceb-interactive-with-pad	Mixed Interactive	0.38	0.43	0.74	0.72	0.55	1545
In_expB2c1.0_Bo1bfree	Mixed Interactive	0.37	0.39	0.62	0.56	0.51	1803
TEXT_MANUAL_CBIR_RF	Mixed Interactive	0.04	0.09	0.28	0.22	0.14	496
BM25b=0.75k_1=1.2k_3=8.0_Bo1bfree	Text Interactive	0.37	0.38	0.61	0.55	0.50	1810
ISSR_Text_FR_2	Text Interactive	0.31	0.34	0.47	0.49	0.44	1811
ISSR_Text_1rfb	Text Interactive	0.28	0.29	0.41	0.43	0.39	1604
ISSR_Text_5	Text Interactive	0.27	0.29	0.38	0.43	0.37	1738
ISSR_Text_4	Text Interactive	0.27	0.29	0.34	0.42	0.37	1734
ISSR_Text_DE_2	Text Interactive	0.20	0.22	0.25	0.25	0.27	1438
Alicante-Run2	Text Interactive	0.14	0.17	0.32	0.35	0.34	994
CBIR_RF	Visual Interactive	0.01	0.03	0.06	0.05	0.05	306
york.BO1.MeSH.TamuraHistogram0.2	Mixed Manual	0.35	0.36	0.62	0.58	0.48	1760
Multimodal_Text_QE_CBIR	Mixed Manual	0.04	0.09	0.27	0.19	0.13	456

both textual and visual information, and the fragility that such combinations can introduce into retrieval systems. The distribution of MAP for the textual runs was higher than that for the mixed runs. A significant mode exists around a MAP of 0.2 for the mixed runs, while the mode for the textual runs is at around 0.3.

3.2.4 Interactive Retrieval

This year, as in previous years, interactive retrieval was only used by a very small number of participants. However, the manual and interactive runs submitted this year performed relatively well with one of the runs achieving the highest overall early precision (P5 and P10). Table 4 shows the results of all manual and interactive runs submitted.

There is definitly a need to promote interactive an manual retrieval further as the potential of this does not seem to have been exploited well, so far.

3.3 Case–based results

A total of six groups participated in this introductory task, submitting at total of 18 runs. The results were quite promising with one group achieving a relatively high MAP of 0.33. As with the image–based retrieval, automatic textual results achieved the best results with poor results

Table 5: Results of the visual runs for the medical image retrieval task (Case–Based Topics).

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
medGIFT_case_bySimilarity_vis_maxWithAR	Visual Automatic	0.02	0.03	0.04	0.04	0.05	41
medGIFT_case_bySimilarity_vis_sumWithAR	Visual Automatic	0.02	0.03	0.04	0.06	0.05	42
clef2009	Visual Automatic	0.01	0.00	0.00	0.00	0.01	39
medGIFT_case_byfreq_vis_sumWithAR	Visual Automatic	0.00	0.00	0.00	0.00	0.01	26
medGIFT_case_byfreq_vis_maxWithAR	Visual Automatic	0.00	0.00	0.00	0.00	0.01	26

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

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
ceb-cases-essie2-automatic	Textual Automatic	0.34	0.28	0.32	0.34	0.23	74
sinai_TA_cbt	Textual Automatic	0.26	0.23	0.32	0.34	0.23	89
sinai_TA_cbtM	Textual Automatic	0.26	0.22	0.32	0.30	0.25	89
clef2009	Textual Automatic	0.19	0.13	0.32	0.24	0.19	93
HES-SO-VS_txt_case	Textual Automatic	0.19	0.15	0.32	0.32	0.20	71
Alicante-CaseBased-Run5	Textual Automatic	0.07	0.07	0.16	0.10	0.09	61
Alicante-CaseBased-Run2	Textual Automatic	0.05	0.04	0.08	0.08	0.07	58
Alicante-CaseBased-Run4	Textual Automatic	0.05	0.04	0.08	0.08	0.07	58
Alicante-CaseBased-Run3	Textual Automatic	0.05	0.04	0.08	0.08	0.07	59
Alicante-CaseBased-Run1	Textual Automatic	0.05	0.04	0.08	0.08	0.07	59

being obtained by visual methods. Results were quite varied however with the MAP varying from 0.0025 to 0.335

3.3.1 Visual Retrieval

Purely visual methods were not able to achieve good performance as seen in the Table 5 below. This is not entirely surprising as the set of sample images provided for each topic were quite varied in visual appearance and it needs to be explored how this information can be used well.

3.3.2 Textual Retrieval

Textual methods were more effective in retrieving relevant articles as seen in the Table 6 below. Interestingly, the early precision for the best runs were not significantly higher than the MAP, unlike in the image–based topics where the early precision was substantially higher than the MAP for many runs.

3.3.3 Multimodal Retrieval

Unlike the image–based topics, here the multimodal runs performed quite poorly as seen in Table 7. This could be due to the variety of reasons including the diversity of sample images, poor visual performance of runs for case–based topics and the fact that the two best groups did not submit runs in this category.

3.4 Relevance Judgement Analysis

A number of topics, both image–based and case–based were judged by two or even three judges. There were significant variations in the kappa metric used to evaluate the inter–rater agreement. The kappas for the image–based topics are given below in Table 8. The kappas are usually reasonably high except when involving some judges. As seen in the table, judge 12 was extremely

Table 7: Results of the multimodal runs for the medical image retrieval task (Case–Based Topics).

Run	Run Type	MAP	bpref	P5	P10	P30	rel_ret
medGIFT0.5_BySimilarity_EN	Mixed Automatic	0.07	0.05	0.12	0.14	0.09	74
clef2009	Mixed Automatic	0.02	0.00	0.00	0.00	0.02	57

Table 8: Kappas for Image–Based Topics.

Topic	Judge 1	Judge 2	Kappa
1	3	4	0.341
3	3	11	0.715
7	4	6	0.302
8	6	15	0.639
10	4	12	0.15
13	7	12	0.021
14	11	12	0.0298
15	6	7	0.885
17	3	4	0.821
18	4	15	0.884
20	7	12	0.0388

Table 9: Kappas for Case–Based Topics.

Topic	judge1	judge2	Kappa
26	4	7	0.06
27	4	7	-0.10
28	4	11	0.37
29	4	7	-0.25
29	4	11	0.13
29	7	11	0.28
30	7	11	0.56

lenient compared to all other judges leading to extremely low kappas for any pairwise comparison involving this judge. For instance, on topic 13, judge 12 evaluated 342 images as being relevant while judge 7 (our most strict judge) only evaluated 7 images as being relevant. We discovered this during the judging process and did not use the judgements from judge 12 in creating the official qrels for any of the topics. We performed extensive evaluation of the effect of the judge's strictness in establishing relevance and found that overall, the results obtained using strict judges and those obtained lenient judges correlated well. However, results for a particular topic could be affected by the judge's parsimony in the evaluation of relevance.

For the case–based topics, the kappa values were generally lower as seen in Table 9. The 2x2 tables indicated that judge 4 was the most lenient and judge 7 was the strictest.

3.5 Lung Nodule Detection Task

We also introduced a lung nodule detection task in 2009. This task used the CT (Computed Tomography) slices from the Lung Imaging Data Consortium (LIDC). This collection consisted for 100–200 slices per study and were manually annotated by 4 clinicians. Although more than 25 groups had registered for the task and more than a dozen had downloaded the datasets, only two groups submitted runs. A commercial proprietary software package performed quite well in the task of detecting the nodules.

4 Conclusions

The focus of many participants in this year's ImageCLEF has been text–based retrieval. The increasingly semantic topics combined with a database containing high–quality annotations in 2009 may have resulted in less impact of using visual techniques as compared to previous years. Visual runs were rare and generally poor in performance. Mixed–media runs were very similar in performance to textual runs when looking at MAP. 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.

Case–based topics were introduced for the first time and only a few groups participated with results being slightly lower than for the image–based topics.

A kappa analysis between several relevance judgements for the same topics shows that there are differences between judges but that agreement is generally high. A few judges can nevertheless have disagreeing results with all other judges, something that we need to investigate further.

For future campaign it seems important that more research on visual techniques including massive learning should be done as currently techniques do not perform well. Interactive and manual retrieval do also seem to have room for improvements and should be put forward to participants who generally prefer automatic text–based approaches.

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, the TrebleCLEF project and Google. We would like to thank the RSNA for supplying the images of their journals Radiology and Radiographics for the ImageCLEF campaign.

References

[1] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 cross–language image retrieval track. In *Cross Language Evaluation Forum (CLEF 2005)*, Springer Lecture Notes in Computer Science, pages 535–557, September 2006.

[2] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael 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, Jeffery Jensen, Henning Müller, Paul Gorman, and Patrick Ruch. A qualitative task analysis for developing an image retrieval test collection. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, pages 11–16, Vienna, Austria, 2005.

[4] 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)*, pages 473–491, Budapest, Hungary, 2008. Springer.

[5] Henning Müller, Christelle Despont-Gros, William Hersh, Jeffery Jensen, Christian Lovis, and Antoine Geissbuhler. Health care professionals' image use and search behaviour. In *Proceedings of the Medical Informatics Europe Conference (MIE 2006)*, IOS Press, Studies in Health Technology and Informatics, pages 24–32, Maastricht, The Netherlands, August 2006.

[6] Henning Müller, Jayashree Kalpathy-Cramer, William Hersh, and Antoine Geissbuhler. Using Medline queries to generate image retrieval tasks for benchmarking. In *Medical Informatics Europe (MIE2008)*, pages 523–528, Gothenburg, Sweden, May 2008. IOS press.

[7] Henning Müller, Antoine Rosset, Jean-Paul Vallée, Francois Terrier, and Antoine Geissbuhler. A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, 28(6):295–305, 2004.

[8] Saïd Radhouani, William Hersh, Jayashree Kalpathy-Cramer, and Steven Bedrick. Understanding and improving image retrieval in medicine. Technical report, Oregon Health and Science University, 2009.

[9] 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.