Overview of the CLEF 2011 medical image classification and retrieval tasks

Jayashree Kalpathy–Cramer[1], Henning Müller[2,3], Steven Bedrick[1], Ivan Eggel[2], Alba G. Seco de Herrera[2], Theodora Tsikrika[2]

[1]Oregon Health and Science University (OHSU), Portland, OR, USA
[2]University of Applied Sciences Western Switzerland, Sierre, Switzerland
[3]Medical Informatics, University of Geneva, Switzerland
henning.mueller@hevs.ch

Abstract. The eighth edition of the ImageCLEF medical retrieval task was organized in 2011. A subset of the open access collection of PubMed Central was used as the database in 2011. This database contains 231,000 images and is substantially larger than previously used collections. Additionally, there was a larger fraction of non–clinical images such as graphs and charts. As in 2010, we had three subtasks: modality classification, image–based and case–based retrieval.
A new, simple hierarchy for article figures was created. Our belief is that the use of the detected modality should help filter out non–relevant images, thereby improving precision. The goal of the image–based retrieval task was to retrieve an ordered set of images from the collection that best meet 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 meet the information need provided as a description of a "case".
The number of registrations to the medical task increased to 55 research groups. However, groups submitting runs have remained stable at 17, with the number of submitted runs increasing to 207. Of these, 130 were image–based retrieval runs, 43 were case–based runs while the remaining 34 were modality classification runs. Combining textual and visual cues most often led to best results, but results fusion needs to be used with care.

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[5–7]. The main goal of ImageCLEF continues to be promoting multi–modal information retrieval by combining a variety of media including text and images for more effective information retrieval. Each year

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

new domains of image retrieval are being added to develop new challenges in multimodal 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. In 2011, this format was continued and an increased number of visual retrieval sessions are included into the conference program, including more time for the ImageCLEF lab.

This article presents the main results of the medical image retrieval task and compares results between the various participating groups and the techniques employed.

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 2011, a new record of 130 research groups registered for the four sub–tasks of ImageCLEF down from seven sub tasks in 2009 but the same as in 2010. For the medical retrieval task the number of registrations also reached a new maximum with 55. 17 of the participants submitted results to the tasks, essentially the same number as in previous years. The following groups submitted at least one run:

- BUAA AUDR (BeiHang University, Beijing, China)*;
- CEB (National Library of Medicine, USA);
- DAEDALUS UPM (Universidad Politecnica de Madrid, Spain);
- DEMIR (Dokuz Eylul University, Turkey);
- HITEC (Ghent University, Belgium)*;
- IPL (Athens University of Economics and Business, Greece);
- IRIT (Institut de Recherche en Informatique Toulouse, France);
- LABERINTO (Universidad de Huelva, Spain)*;
- SFSU (San Francisco State University, USA)*;
- medGIFT (University of Applied Sciences Western Switzerland, Switzerland);
- MRIM (Laboratoire d'Informatique de Grenoble, France);
- Recod (Universidade Estadual de Campinas, Brazil);
- SINAI (University of Jaen, Spain);
- UESTC (University of Electronic Science and Technology, China)*;
- UNED (Universidad Nacional de Educacion a Distancia, Spain);
- UNT (University of North Texas, USA);
- XRCE (Xerox Research Centre Europe, France).

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

Participants marked with a star had never before participated in the medical retrieval task, indicating that the number of first–time participants was relatively low with five among the 17 participants.

A total of 207 valid runs were submitted, 34 of which were submitted for modality detection, 130 for the image–based topics and 43 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.

2.2 Datasets

A new database was created for the use in ImageCLEF 2011 to allow for new challenges. The database is a subset of 231,000 images from the PubMed Central database containing in total over one million images. This set of articles contains all articles in PubMed that are open access but the exact copyright for redistribution varies among the journals. The subset chosen includes all journals of BioMed Central, as these allow redistribution of the data. A set of imaging oriented journals that also allow redistribution were taken in addition to this. Two main challenges of the data set are that (1) there is a large variety of journals, not only radiology, meaning that rigor in figure legends is different and the variety of images is much larger and that (2) the data set contains a majority of images that are not or little important for retrieval (such as tables, flow charts, graphs, etc.).

2.3 Modality Classification

Previous research has demonstrated the utility of classifying images by modality in order to improve the precision of the search. A simple ad–hoc hierarchy with 18 classes in the sections radiology, microscopy, photography, graphics, other was created based on the existing data set:

- 3D : 3d reconstruction
- AN : angiography
- CM : compound figure (more than one type of image)
- CT : computed tomography
- DM : dermatology
- DR : drawing
- EM : electronMicroscopy
- EN : endoscopic imaging
- FL : fluoresence
- GL : gel
- GX : graphs
- GR : gross pathology
- HX : histopathology
- MR : magnetic resonance imaging
- PX : general photo
- RN : retinograph

– US : ultrasound
– XR : x–ray

For this hierarchy 1,000 training images and 1,000 test images were provided. Participants were requested to also classify all 231,000 images to be able to perform further analysis on the data and potentially annotate a larger part of the data. Currently, a more detailed hierarchy based on a larger data set is being elaborated.

2.4 Image–Based Topics

The topics for the image–based retrieval task were a selection of topics that had been used in the past based on [8, 9]. Ten topics each for visual, textual and mixed retrieval were chosen to allow for the evaluation of a large variety of techniques. The reuse of existing topics allows for the comparison of the difficulty of these topics with a different database and limits the effort needed to survey clinicians and develop new topics. This also means that participats have in principal a different database available for training their systems, which can potentially increase performance.

2.5 Case–Based Topics

The case–based topics were reused from previous years. 10 topics were created based on cases from the teaching file Casimage [10]. This teaching file contains cases (including images) from radiological practice that clinicians document mainly for using them in teaching. The diagnosis and all information on the chosen treatment was then removed from the cases so as to simulate the 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.

2.6 Relevance Judgements

The relevance judgements were performed with the same on–line system as in 2008, 2009, and 2010 for the image–based topics as well as case–based topics. For the case–based topics, the system displays 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 email 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 2011. Runs are ordered based on the tasks (modality classification, image–based and case–based retrieval) and the techniques used (visual, textual, mixed). Very few manual runs were submitted.

3.1 Submissions

17 teams submitted at least one run in 2011, slightly more than in 2010. The numbers of runs increased from 155 to 207. There were more submissions on image–based retrieval task (130) than in the other two tasks modality classification (34) and case–based retrieval (43).

3.2 Modality Classification Results

The results of the modality classification task are measured in classification accuracy. With a higher number of classes, this task was more complex than in 2010. As seen in Table 1, the best results were obtained by combining visual and textual methods (86%) as in 2010. The best run using visual methods (85%) had a slightly worse accuracy than the best run using mixed methods. The best run using textual methods alone obtained a much lower accuracy (70%).

Best overall results were obtained by Xerox research but with less than a 1% difference HITEC had also very high mixed modality results. Best visual results were also obtained by Xerox, with a significant difference over the Recod group. Only one single group submitted text–based results that performed worse than visual and mixed runs.

Techniques Used for Visual Classification The Xerox team, which obtained the best results, uses a Fisher Vector representation of the images built on low level features such as Scale Invariant Feature Transform (SIFT), Local Orientation Histograms (ORH) and local RGB statistics [11]. A variety of image processing techniques were explored by the rest of the participants. The visual features used include visual descriptors for color, shape, texture or spatial information. Tools such as the GIFT[4] (GNU Image Finding Tool) [12] were employed as well as techniques such as the Color Layout Descriptor (CLD) of MPEG–7, the Color and Edge Directivity Descriptor (CEDD), the Fuzzy Color and Texture Histogram (FCTH) using the Lucene image retrieval (LIRE) library[5] [13],

[4] http://www.gnu.org/software/gift/
[5] http://freshmeat.net/projects/lirecbir/

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

Run	Group	Run Type	Classification Accuracy
CE_all_MIX_semiLM.txt	XRCE	Mixed	**0.8691**
XRCE_Testset_MIX_semiL50.txt	XRCE	Mixed	0.8642
2011.06.10-02.38.40.test.prediction.trec	HITEC	Mixed	0.8603
2011.06.09-18.36.25.test.prediction.trec	HITEC	Mixed	0.8564
XRCE_Testset_MIX_semiL25.txt	XRCE	Mixed	0.8593
2011.06.08-19.58.41.test.prediction.trec	HITEC	Mixed	0.8515
2011.06.10-00.01.26.test.prediction.trec	HITEC	Mixed	0.7685
image_text_test_result_multilevel.dat	CEB	Mixed	0.7412
2011.06.10-03.25.40.test.prediction.trec	HITEC	Mixed	0.7412
image_text_test_result_sum_ext.dat	CEB	Mixed	0.6025
image_text_test_result_CV.dat	CEB	Mixed	0.5966
image_text_test_result_multilevel_ext.dat	CEB	Mixed	0.5917
image_text_test_result_sum.dat	CEB	Mixed	0.5917
image_text_test_result_CV_ext.dat	CEB	Mixed	0.5820
image_text_test_result_original.dat	CEB	Mixed	0.5820
image_text_test_result_ext.dat	CEB	Mixed	0.5439
ICLEF2011_MED_MODALITY_09062011_1500.txt	IPL	Textual	**0.7041**
ICLEF2011_MED_MODALITY_09062011_1600.txt	IPL	Textual	0.4765
XRCE_all_VIS_semiL25.txt	XRCE	Visual	**0.8359**
XRCE_Testset_VIS_semi20_CBIR.txt	XRCE	Visual	0.8349
XRCE_all_VIS_semi20_CBIR.txt	XRCE	Visual	0.8339
recod_imageclefmed_ModCla_357l	Recod	Visual	0.6972
recod_imageclefmed_ModCla_Vl	Recod	Visual	0.6943
recod_imageclefmed_ModCla_VlNoR	Recod	Visual	0.6904
recod_imageclefmed_ModCla_VsNoR	Recod	Visual	0.6835
recod_imageclefmed_ModCla_Vs	Recod	Visual	0.6806
recod_imageclefmed_ModCla_343s	Recod	Visual	0.6787
recod_imageclefmed_ModCla_370l	Recod	Visual	0.6787
recod_imageclefmed_ModCla_370s	Recod	Visual	0.6767
recod_imageclefmed_ModCla_343l	Recod	Visual	0.6748
recod_imageclefmed_ModCla_357s	Recod	Visual	0.6669
classificationResults_GIFT.txt	medGIFT	Visual	0.6220
image_test_result_original.dat	CEB	Visual	0.5712
image_test_result_ext.dat	CEB	Visual	0.4853

the SIFT [14], as well as various combinations of these. Classifiers employed ranged from simple k–Nearest Neighbors (kNN) [12, 15] or k–means [13] to Genetic Programming (GP) [15] or Support Vector Machines (SVMs) [13].

Techniques Used for Classification Based on Text Only one team (IPL) submitted runs for the modality classification using text. The system is based on the Lucene[6] search engine [16].

Techniques Used for Multimodal Classification The Xerox team obtained the best results also on multimodal classification by averaging text and image classification scores separately [11].

Besides the techniques mentioned in the visual classification section, the participants used visual techniques such as pixel mean pixels [17], skin detection [17], mutual information or innovative similarities metrics based on visual concepts [13].These techniques were then combined with text methods to improve the results considering often binary features [11, 17, 13] from image captions, a mapping of the free text onto MeSH (Medical Subject Headings), but also the title, abstract, etc. of the text.

3.3 Image–Based Retrieval Results

As in most previous years, the best results for the image–based retrieval topics were obtained using multimodal methods. Most of the runs submitted to this task use textual methods that perform well. 26 visual runs were submitted but the results were still much lower than the textual and multimodal techniques.

Visual Retrieval 26 of the 130 submitted runs used purely visual techniques. In addition to the techniques used in the modality classification task, participants used visual features such as Edge Histogram descriptors (EHD) or the brightness texture histogram (BTDH) and applied feature weighting using SVM accuracy [13]. The IPL system that uses LIRE [16] obtained the four best results, most of them automatically, but also with one feedback run reaching a MAP of 0.0338. Then, the GIFT baseline was the second best group with a MAP of 0.0274. In terms of P10, both systems are almost the same and in terms of bPref GIFT has the best results (0.0807 compared to 0.0716), indicating that GIFT had more non–judged images in its relevance set.

Textual Retrieval Participants explored a variety of retrieval techniques, with many using Lucene [12, 16, 13, 18, 19]. Simple boolean text queries were used and query expansion was applied by exploiting external sources such as MeSH terms (manually or automatically assigned)[18, 12, 20], the Unified Medical Language System (UMLS) [16, 21, 22, 20], concepts using MetaMap [20, 23] and even

[6] `http://lucene.apache.org/`

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

Run Name	Group	Run Type	MAP	P10	bPref
IPL2011Visual-DECFc	IPL	Automatic	**0.0338**	**0.1500**	0.0717
IPL2011Visual-DEFC	IPL	Automatic	0.0322	0.1467	0.0715
IPL2011Visual-DEC	IPL	Automatic	0.0312	0.1433	0.0716
ILP2011Visual-DEF	IPL	Feedback	0.0283	0.1367	0.0703
gift_visual_ib	medGIFT	Automatic	0.0274	0.1467	**0.0807**
ILP2011Visual-DTG	IPL	Automatic	0.0253	0.1333	0.0715
visual_ib	medGIFT	Automatic	0.0252	0.1267	0.0752
iti-lucene-image	CEB	Automatic	0.0245	0.1333	0.0627
image_fusion_category_weight_filter	CEB	Automatic	0.0221	0.1167	0.0651
image_fusion_category_weight_filter_merge	CEB	Automatic	0.0201	0.1000	0.0629
image_fusion_category_weight_merge	CEB	Automatic	0.0193	0.0933	0.0620
cedd_norm_min_1	DEMIR	Automatic	0.0174	0.1067	0.0602
Daedalus_BasImgC_MC	DAEDALUS UPM	Feedback	0.0147	0.0967	0.0582
Daedalus_ImgC_MCCI	DAEDALUS UPM	Automatic	0.0147	0.0967	0.0582
bovw_visual_ib	medGIFT	Automatic	0.0126	0.0867	0.0437
Daedalus_BasImg	DAEDALUS UPM	Automatic	0.0125	0.0733	0.0528
AUDR_VISUAL_CEDD	BUAA AUDR	Automatic	0.0082	0.0400	0.0427
bovw_s2_visual_ib	medGIFT	Automatic	0.0076	0.0900	0.0279
okada_lab_cosine_tfidf	lambdasfsu	Automatic	0.0007	0.0167	0.0084
okada_lab_cosine_reg	lambdasfsu	Automatic	0.0005	0.0067	0.0094
okada_lab_emd_L1_tfidf	lambdasfsu	Automatic	0.0004	0.0100	0.0036
okada_lab_diffusion	lambdasfsu	Automatic	0.0003	0.0000	0.0065
okada_lab_emd	lambdasfsu	Automatic	0.0003	0.0100	0.0051
okada_lab_emd_L1	lambdasfsu	Automatic	0.0003	0.0100	0.0051
okada_lab_emd_tfidf	lambdasfsu	Automatic	0.0003	0.0067	0.0042
okada_lab_diffusion_tfidf	lambdasfsu	Automatic	0.0002	0.0000	0.0063

Wikipedia [16, 23]. More complex language models that incorporate phrases (not just words), sentence selection and query translation were used [20], and pseudo relevance feedback [16] was also applied. Modality filtering using either text–based or image–based modality detection techniques was found to be useful by some participants. Best results (see Table 3) were obtained by the University of Huelva (LABERINTO team), Universidad Nacional de Educación a Distancia (UNED) and Athens University of Economics and Business with all results being within less than 1% difference. It is interesting to remark that these three gropus applied query expansion strategies; applying query expansion does not always lead to better results as indicated by the results obtained by other groups such as BUAA AUDR.

Multimodal Retrieval This year, the multimodal run with the highest MAP was submitted by the DEMIR team (see Table 4) obtaining better results than visual and textual techniques alone. The runs may use manual optimization as they are not marked as automatic runs. To combine visual and textual techniques, participants use filtering and re–ranking and simple fusion with linear combinations [13, 24, 22]. In the past, rank–based fusion often performed better than score–based fusion and still many groups submit rank–based rather score–based fusion runs. The best automatic runs using multimodal techniques are lower than the best automatic runs for text only retrieval. In general the average performance is lower than for purely textual retrieval underlining the importance of good fusion techniques.

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

Run Name	Group	Run Type	MAP	P10	bPref
laberinto_CTC	LABERINTO	Automatic	**0.2172**	0.3467	0.2402
Run2_Txt	UNED	Automatic	0.2158	0.3533	0.2514
IPL2011AdHocT1-C6-M0_2-R0_01-DEFAULT	IPL	Automatic	0.2145	**0.4033**	0.2434
laberinto_BC	LABERINTO	Automatic	0.2133	0.3400	0.2384
IPL2011AdHocT1-C6-M0_2-DEFAULT	IPL	Automatic	0.2130	0.3567	0.2370
Run3_Txt	UNED	Automatic	0.2125	0.3867	0.2430
IPL2011AdHocT0_113-C0_335-M0_1-DEFAULT	IPL	Automatic	0.2016	0.3733	0.2269
IVSCT5G	MRIM	Automatic	0.2008	0.3033	0.2331
IVSCT5GK	MRIM	Automatic	0.2008	0.3033	0.2331
IVPCT5GKin	MRIM	Automatic	0.1975	0.2967	0.2257
IVPCT5G	MRIM	Automatic	0.1974	0.2967	0.2256
IVPCT5GKout	MRIM	Automatic	0.1973	0.2967	0.2256
Daedalus_BasTxtC	DAEDALUS UPM	Automatic	0.1966	0.3900	0.2564
IPL2011AdHocTC0_9-M0_1-DEFAULT	IPL	Manual	0.1945	0.3700	0.2255
DEMIR_MED_1	DEMIR	Automatic	0.1942	0.3400	0.2215
laberinto_ETPCC	LABERINTO	Automatic	0.1939	0.2933	0.2198
TFIDFModel_TopicModel_ImageCaption+Article	BUAA AUDR	Automatic	0.1917	0.3400	0.2237
Daedalus_SemEC	DAEDALUS UPM	Automatic	0.1906	0.3867	**0.2690**
SINAI-ImgCaption	SINAI	Automatic	0.1890	0.3300	0.2247
SINAI-ImgCaptionExpand1	SINAI	Automatic	0.1890	0.3300	0.2247
SINAI-ImgCaptionExpand2	SINAI	Automatic	0.1890	0.3300	0.2247
SINAI-ImgCaptionExpand3	SINAI	Automatic	0.1890	0.3300	0.2247
SINAI-ImgCaptionExpand4	SINAI	Automatic	0.1890	0.3300	0.2247
SINAI-ImgCaptionExpand5	SINAI	Automatic	0.1890	0.3300	0.2247
XRCE_RUN_TXTax_dir_spl	XRCE	Feedback	0.1870	0.3233	0.2156
Daedalus_SemAC	DAEDALUS UPM	Automatic	0.1818	0.3767	0.2496
XRCE_RUN_TXT_noMOD	XRCE	Automatic	0.1802	0.3100	0.2122
AUDR_TFIDF_CAPTION	BUAA AUDR	Automatic	0.1758	0.3133	0.2187
HES-SO-VS_IMAGE-BASED_CAPTIONS	medGIFT	Automatic	0.1742	0.3000	0.2179
UESTC_adhoc_p1	UESTC	Automatic	0.1672	0.2667	0.1946
UESTC_adhoc_p2	UESTC	Feedback	0.1672	0.2733	0.1995
UESTC_adhoc_p1QE_sw	UESTC	Automatic	0.1669	0.2833	0.1977
UESTC_adhoc_p2QE_sw_chd	UESTC	Automatic	0.1666	0.2700	0.2049
UESTC_adhoc_p1_sw	UESTC	Automatic	0.1635	0.2733	0.1908
UESTC_adhoc_p1QE	UESTC	Automatic	0.1632	0.2533	0.1994
IPL2011AdHocTCM-DEFAULT-DEFAULT	IPL	Automatic	0.1599	0.3367	0.1874
ESU_Ib_bl	UNT	Automatic	0.1594	0.2667	0.1889
UESTC_adhoc_p2QE_sw	UESTC	Automatic	0.1590	0.2567	0.1956
UESTC_adhoc_indri	UESTC	Automatic	0.1588	0.2600	0.1873
UESTC_adhoc_p2QE	UESTC	Automatic	0.1583	0.2500	0.1974
ESU_Ib_blRF	UNT	Automatic	0.1558	0.2433	0.1865
ESU_Ib_Struc	UNT	Automatic	0.1540	0.2800	0.1906
IPL2011AdHocTC0_9-M0_1-BM25F	IPL	Feedback	0.1510	0.3033	0.1909
laberinto_BIR	LABERINTO	Automatic	0.1496	0.3400	0.1992
IPL2011AdHocT1-C6-M0_2-R0_01-BM25F	IPL	Automatic	0.1492	0.3067	0.1848
IPL2011AdHocT1-C6-M0_2-BM25F	IPL	Automatic	0.1485	0.3067	0.1839
UESTC_adhoc_p1QE_sw_chd	UESTC	Automatic	0.1471	0.2100	0.1807
laberinto_CTIR	LABERINTO	Automatic	0.1466	0.3433	0.1953
textual_rerank_reindex	DEMIR	Automatic	0.1452	0.3033	0.1859
laberinto_ETPCIR	LABERINTO	Automatic	0.1411	0.3000	0.1887
ESU_Ib_StrucRF	UNT	Automatic	0.1346	0.2300	0.1874
IPL2011AdHocT0_113-C0_335-M0_1-BM25F	IPL	Automatic	0.1312	0.2767	0.1670
Run6_Txt	UNED	Automatic	0.1309	0.3433	0.1597
IPL2011AdHocTCM-BM25	IPL	Automatic	0.1289	0.2867	0.1744
Run5_Txt	UNED	Automatic	0.1270	0.3100	0.1622
iti-lucene-baseline+expanded-concepts	CEB	Automatic	0.1255	0.2733	0.1828
laberinto_BFT	LABERINTO	Automatic	0.1146	0.2533	0.1786
laberinto_CTFT	LABERINTO	Automatic	0.1101	0.2500	0.1691
laberinto_ETFT	LABERINTO	Automatic	0.1050	0.2567	0.1640
laberinto_ETPCFT	LABERINTO	Automatic	0.1014	0.2400	0.1571
iti-essie-baseline+expanded-concepts	CEB	Automatic	0.0966	0.2133	0.1556
HES-SO-VS_IMAGE-BASED_FULLTEXT	medGIFT	Automatic	0.0921	0.2167	0.1506
BM25Model_TopicModel_ImageCaption+Article	BUAA AUDR	Automatic	0.0878	0.1900	0.1250
KLDQueryExpansion_TopicModel_ImageCaption+Article	BUAA AUDR	Automatic	0.0811	0.1733	0.1191

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

Run Name	Group	Run Type	MAP	P10	bPref
mixed_3_2_cedd_baseline_run	DEMIR	Not applicable	**0.2372**	0.3933	**0.2738**
mixed_cedd_baseline_run	DEMIR	Not applicable	0.2307	0.3967	0.2606
mixed_3_2_cedd_weighted_run	DEMIR	Not applicable	0.2014	0.3400	0.2481
mixed_3_2_cedd_rerank_reindex_run	DEMIR	Feedback	0.1983	**0.4067**	0.2428
mixed_cedd_weighted_run	DEMIR	Not applicable	0.1972	0.3367	0.2383
mixed_cedd_rerank_reindex_run	DEMIR	Not applicable	0.1853	0.3667	0.2230
DEMIR_MED2011	DEMIR	Automatic	0.1645	0.3967	0.2198
XRCE_RUN_MIX_SFLMODSc_ax_dir_spl	XRCE	Feedback	0.1643	0.3800	0.2234
XRCE_RUN_MIX_SFLMOD_ax_dir_spl	XRCE	Feedback	0.1545	0.3800	0.2053
XRCE_RUN_MIX_SFLMODFL2_ax_dir_spl	XRCE	Automatic	0.1520	0.3633	0.2049
XRCE_RUN_MIX_SFL_MOD_ax_dir_spl_lgd	XRCE	Feedback	0.1512	0.3667	0.2031
DEF-T1-C6-M0_2-BM25F-0_39-0_01	IPL	Automatic	0.1494	0.3067	0.1849
DEF-T1-C6-M0_2-R0_01-BM25F-0_39-0_01	IPL	Automatic	0.1493	0.3067	0.1849
DTG-T1-C6-M0_2-R0_01-BM25F-0_39-0_01	IPL	Automatic	0.1492	0.3067	0.1849
DTG-T1-C6-M0_2-BM25F-0_39-0_01	IPL	Automatic	0.1489	0.3067	0.1840
XRCE_RUN_MIX_SFL_noMOD_ax_dir_spl	XRCE	Automatic	0.1472	0.3433	0.1874
XRCE_RUN_MIX_SFL_noMOD_ax_dir_spl_lgd	XRCE	Feedback	0.1429	0.3367	0.1860
iti-lucene-baseline+expanded-concepts+image	CEB	Automatic	0.1356	0.2833	0.1970
Run6_TxtImg_OwaOr03	UNED	Automatic	0.1346	0.3467	0.1604
Run6_TxtImg_PtPi	UNED	Automatic	0.1311	0.3333	0.1557
Run5_TxtImg_OwaOr03	UNED	Automatic	0.1299	0.3200	0.1641
mixed_GIFT_Lucene_captions_ib	medGIFT	Automatic	0.1230	0.3133	0.1733
Run5_TxtImg_PtPi	UNED	Feedback	0.1176	0.2800	0.1614
DEF-T1-C6-M0_2-BM25F	IPL	Automatic	0.0952	0.2967	0.1610
DTG-T1-C6-M0_2-BM25F	IPL	Automatic	0.0945	0.2700	0.1613
DTG-T0_113-C0_335-M0_1-BM25F	IPL	Automatic	0.0924	0.2733	0.1600
DEF-T0_113-C0_335-M0_1-BM25F	IPL	Automatic	0.0911	0.2733	0.1583
Multimodal_Rerank_Filter_Merge	CEB	Automatic	0.0910	0.2867	0.1572
Multimodal_Rerank_Merge	CEB	Automatic	0.0903	0.2833	0.1547
Run6_TxtImg_Pi	UNED	Automatic	0.0891	0.2400	0.1288
mixed_GIFT_Lucene_full_ib	medGIFT	Automatic	0.0857	0.2900	0.1308
iti-essie-baseline+expanded-concepts+image	CEB	Automatic	0.0843	0.2167	0.1331
Run5_TxtImg_Pi	UNED	Automatic	0.0699	0.1667	0.1394
AUDR_MIXED_CEDD_TFIDFModel	BUAA AUDR	Automatic	0.0556	0.1767	0.1018
AUDR_MIXED_CEDD_KLD_QE	BUAA AUDR	Automatic	0.0341	0.1633	0.0720
AUDR_MIXED_CEDD_BM25Model	BUAA AUDR	Automatic	0.0328	0.1500	0.0719
Daedalus_CombSemA#C_MC#SC_MCCI	DAEDALUS UPM	Automatic	0.0232	0.1133	0.0851
Daedalus_CombSemE#C_MC#C_MCCI	DAEDALUS UPM	Automatic	0.0218	0.1167	0.0858
Daedalus_CombSemA#C_MC#C_MCCI	DAEDALUS UPM	Automatic	0.0211	0.1167	0.0845
Daedalus_CombBas#C_MC#C_MC	DAEDALUS UPM	Automatic	0.0204	0.1167	0.0827

3.4 Case–based Retrieval Results

As in 2010, almost all teams used textual retrieval techniques in the case–based retrieval task. Only medGIFT submitted visual case–based retrieval runs. MedGIFT and CEB submitted runs to the multimodal task. Best results were obtained with a textual retrieval approach by the University of Electronic Science and Technology of China (UESTC) and the University of Applied Sciences Western Switzerland (medGIFT). Multimodal fusion runs do not perform as well as text retrieval runs.

Visual Retrieval Table 5 shows that the results using visual retrieval are lower than most of text–based techniques. Still, the difference in performance is lower than for the image–based retrieval tasks. The only visual run is described in more detail in [12]. Fusion of the single–image runs for visual case–based retrieval is more complex than text–based retrieval of cases, where the image full text can be taken.

Table 5. Results of the **visual** runs for the medical case–based retrieval task.

Run	Group	Run Type	MAP	P10	bPref
gift_visual	medGIFT	Automatic	**0.0204**	0.0444	**0.0292**
bovw_visual_cb	medGIFT	Automatic	0.0164	**0.0556**	0.0267
visual_ib	medGIFT	Automatic	0.0150	0.0444	0.0228
bovw_s2_visual_cb	medGIFT	Automatic	0.0082	0.0333	0.0113

Textual Retrieval 35 runs for the case–based task out of 43 use textual retrieval techniques (see Table 6). The methods used are similar to the techniques of the image–based retrieval. In addition, the participants use tools such as the Terrier IR platform[7] for indexing documents or query expansion techniques based on Rocchio's method [25].

A simple application of Lucene on the article full text obtains second best retrieval results, with the difference not being statistically significant. Using the image captions, which delivered best results for the image–based retrieval task had a much lower performance for the same Lucene setup (MAP of 0.044 vs. 0.129).

Multimodal Retrieval Only two groups submitted multimodal case–based runs. Best results were obtained by medGIFT although it is still low compared to textual approaches (see Table 7). Again, good fusion techniques should be able to improve these mixed results significantly but currently only few groups seem to stress this and rather concentrate on optimizing text–based retrieval.

[7] `http://terrier.org/`

Table 6. Results of the **textual** runs for the medical case–based retrieval task.

Run	Group	Run Type	MAP	P10	bPref
UESTC_full_indri	UESTC	Automatic	**0.1297**	0.1889	**0.1212**
HES-SO-VS_CASE_BASED_FULLTEXT	medGIFT	Automatic	0.1293	**0.2000**	0.1122
UESTC_full_p2QE	UESTC	Automatic	0.1199	0.1556	0.1082
UESTC_full_p2	UESTC	Automatic	0.1179	0.1889	0.1162
MRIM_KJ_A_VM_Sop_T4G	MRIM	Automatic	0.1114	0.1444	0.1064
IRIT_LGDc1.0_KLbfree_d_20_t_20_1	IRIT	Automatic	0.1030	0.1556	0.0930
IRIT_CombSUMc1.0_KLbfree_d_20_t_20_1	IRIT	Automatic	0.0947	0.1333	0.0862
iti-essie-manual	CEB	Manual	0.0941	0.1667	0.1162
IRIT_LGDc1.0_KLbfree_d_20_t_20_1_ignore_low_idf	IRIT	Automatic	0.0937	0.1111	0.0716
MRIM_KJ_A_VM_Pos_T4G	MRIM	Automatic	0.0911	0.1111	0.0938
UESTC_full_okapi	UESTC	Automatic	0.0907	0.1444	0.0970
IRIT_CombSUMc1.0_KLbfree_d_20_t_20_2	IRIT	Automatic	0.0874	0.1111	0.0710
IRIT_LGDc1.0	IRIT	Automatic	0.0872	0.1111	0.0722
IRIT_CombSUMc1.0_3	IRIT	Automatic	0.0859	0.1444	0.0783
UESTC_ac_okapi	UESTC	Automatic	0.0835	0.1222	0.0734
IRIT_In_expB2c1.0_KLbfree_d_20_t_20_0_ignore_low_idf	IRIT	Automatic	0.0793	0.1444	0.0707
IRIT_In_expB2c1.0_KLbfree_d_20_t_20_0	IRIT	Automatic	0.0772	0.1000	0.0675
UESTC_ac_indri	UESTC	Automatic	0.0767	0.1111	0.0669
iti-lucene-baseline	CEB	Automatic	0.0762	0.1444	0.0737
UESTC_full_okapi_fb	UESTC	Automatic	0.0762	0.1333	0.0841
IRIT_In_expB2c1.0_1	IRIT	Automatic	0.0743	0.1111	0.0730
UESTC_ac_p2	UESTC	Automatic	0.0722	0.1222	0.0628
IRIT_CombSUMc1.0_2_ignore_low_idf	IRIT	Automatic	0.0721	0.1333	0.0683
UESTC_ac_p2QE	UESTC	Automatic	0.0677	0.1000	0.0633
UESTC_ac_okapi_fb	UESTC	Automatic	0.0500	0.0778	0.0484
IPL2011CaseBasedT1-C6-M0_2-RO_01-BM25F-AVG	IPL	Automatic	0.0463	0.0889	0.0588
IPL2011CaseBasedT1-C6-M0_2-BM25F-AVG	IPL	Automatic	0.0461	0.0889	0.0588
HES-SO-VS_CASE_BASED_CAPTIONS	medGIFT	Automatic	0.0437	0.1111	0.0540
iti-lucene-baseline+expanded-concepts	CEB	Automatic	0.0264	0.0333	0.0252
iti-lucene-baseline+expanded-concepts+cases	CEB	Automatic	0.0249	0.0333	0.0230
iti-lucene-expanded-concepts	CEB	Automatic	0.0243	0.0333	0.0249
IPL2011CaseBasedT1-C6-M0_2-BM25F-SUM	IPL	Automatic	0.0201	0.0333	0.0176
IPL2011CaseBasedT1-C6-M0_2-RO_01-BM25F-SUM	IPL	Automatic	0.0201	0.0333	0.0174
iti-essie-frames	CEB	Automatic	0.0174	0.0667	0.0333
iti-lucene-frames	CEB	Automatic	0.0141	0.0667	0.0239

Table 7. Results of the **multimodal** runs for the medical case retrieval task.

Run	Group	Run Type	MAP	P10	bPref
mixed_GIFT_Lucene_fulltext_cb	medGIFT	Automatic	**0.0754**	**0.1667**	**0.0958**
iti-lucene-baseline+expanded-concepts+image	CEB	Automatic	0.0269	0.0333	0.0252
iti-lucene-baseline+expanded-concepts+image+cases	CEB	Automatic	0.0255	0.0333	0.0230
iti-lucene-expanded-concepts+image	CEB	Automatic	0.0247	0.0333	0.0249

3.5 Relevance Judgement Analysis

Ten of the topics had relevance judgements performed by two or more judges in order to asses inter–rater agreement. In most cases, the inter–rater agreement, as calculated using the kappa metric was in the substantial range with a few in the moderate or almost perfect range. The average kappa was 0.67 with a maximum of 0.93 and a minimum of 0.36.

4 Conclusions

As in previous years, the largest number of runs submitted was 130 for the image–based retrieval task. Text retrieval was still dominant in terms of retrieval performance but for the modality classification visual retrieval obtained better results than text retrieval alone. Best results were in most cases obtained using visual features and text together. Only for the case–based retrieval task the mixed runs had worse results than the best text retrieval run. Only few groups submitted for the case–based retrieval task, particular for the visual and mixed categories.

For the modality classification task only few text–only runs were submitted and these had a limited performance, lower than visual or mixed runs. This can be due to lower quality of the image captions compared to radiology journals, where captions are often strictly controlled. Still, some of the mixed runs might use better text runs than those that were submitted in the competition, so submitting all base runs would be good to really compare results and better understand the underlying techniques.

For the case–based retrieval results a simple Lucene on the full text had almost the best results, showing also that caption information is in this case of only limited interest. More visual groups need to be motivated to submit for the case–based task but such combinations would require good fusion techniques.

For image–based retrieval combination of text and visual runs had best techniques. Visual runs had a low performance compared to purely textual runs. Information from the captions clearly led to best result compared to using the full text of the articles.

Again and as in past years fusion of text and visual results might be the key to improving the performance of current systems. Interactive and feedback runs are also seen as good possibilities to increase performance of existing systems. To highlight the importance of interactive retrieval a demo session demonstrating the iterfaces of participating systems will be organized at ImageCLEF 2011 in Amsterdam.

5 Acknowledgements

We would like to thank the EU FP7 projects Khresmoi (257528), Promise (258191) and Chorus+ (249008) for their support as well as the Swiss national science foundation with the MANY project (number 205321–130046). Jayashree

Kalpathy–Cramer was supported in part by a National Library of Medicine grant 5K99 LM009889.

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). Lecture Notes in Computer Science (LNCS), Springer (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., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Said, R., Bakke, B., Jr., C.E.K., Hersh, W.: Overview of the CLEF 2009 medical image retrieval track. In: Working Notes of CLEF 2009, Corfu, Greece (September 2009)
6. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Said, R., Bakke, B., Jr., C.E.K., Hersh, W.: Overview of the CLEF 2010 medical image retrieval track. In: Working Notes of CLEF 2010 (Cross Language Evaluation Forum). (September 2010)
7. 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) (September 2004) 295–305
8. Hersh, W., Müller, H., Kalpathy-Cramer, J., Kim, E., Zhou, X.: The consolidated ImageCLEFmed medical image retrieval task test collection. Journal of Digital Imaging **22**(6) (2009) 648–655
9. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: 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, Maastricht, The Netherlands (August 2006) 24–32
10. 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
11. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In: Working Notes of CLEF 2011. (2011)
12. Markonis, D., Eggel, I., G.Seco de Herrera, A., Müller, H.: The medGIFT group in ImageCLEFmed 2011. In: Working Notes of CLEF 2011. (2011)
13. Simpson, M., Rahman, M.M., Phadnis, S., Apostolova, E., Demmer-Fushman, D., Antani, S., Thoma, G.: Text– and content–based approaches to image modality classification and retrieval for the ImageCLEF 2011 medical retrieval track. In: Working Notes of CLEF 2011. (2011)

14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
15. Faria, F.A., Calumby, R.T., Torres, R.d.S.: RECOD at ImageCLEF 2011: Medical modality classification using genetic programming. In: Working Notes of CLEF 2011. (2011)
16. Gkoufas, Y., Morou, A., Kalamboukis, T.: IPL at ImageCLEF 2011 medical retrieval task. In: Working Notes of CLEF 2011. (2011)
17. Gál, V., Solt, I.: Multi–disciplinary modality classification for medical images. In: Working Notes of CLEF 2011. (2011)
18. Mata, J., Crespo, M., Maña, M.J.: LABERINTO at ImageCLEF 2011 medical image retrieval task. In: Working Notes of CLEF 2011. (2011)
19. Castellanos, Á., Benavent, X., Benavent, J., García-Serrano, A.: UNED–UV at medical retrieval task of ImageCLEF 2011. In: Working Notes of CLEF 2011. (2011)
20. Wu, H., Tian, C.: UESTC at ImageCLEF 2011 medical retrieval task. In: Working Notes of CLEF 2011. (2011)
21. Abdulahhad, K., Chevallet, J.P., Berrut, C.: Multi–facet document representation and retrieval. In: Working Notes of CLEF 2011. (2011)
22. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.C.: DAELUS at imageCLEF medical retrieval 2011. textual, visual and multimodal experiments. In: Working Notes of CLEF 2011. (2011)
23. Ruiz, M.E., Leong, C.W., Hassan, S.: UNT at ImageCLEF 2011: Relevance models and salient semantic analysis for image retrieval. In: Working Notes of CLEF 2011. (2011)
24. Alpokocak, A., Ozturkmenoglu, O., Berber, T., Vahid, A.H., Hamed, R.G.: DEMIR at ImageCLEFmed 2011: Evaluation of fusion techniques for multimodal content–based medical image retrieval. In: Working Notes of CLEF 2011. (2011)
25. Dinh, D., Tamine, L.: IRIT at ImageCLEF 2011: medical retrieval task. In: Working Notes of CLEF 2011. (2011)