Overview of the ImageCLEFmed 2007 Medical Retrieval and Medical Annotation Tasks

Henning Müller[1,2], Thomas Deselaers[3], Thomas M. Deserno[4], Jayashree Kalpathy–Cramer[5], Eugene Kim[5], and William Hersh[5]

[1] Medical Informatics, University and Hospitals of Geneva, Switzerland
[2] Business Information Systems, University of Applied Sciences Sierre, Switzerland
[3] Computer Science Dep., RWTH Aachen University, Germany
[4] Dept. of Medical Informatics, RWTH Aachen University, Germany
[5] Oregon Health and Science University (OHSU), Portland, OR, USA
henning.mueller@sim.hcuge.ch

Abstract. This paper describes the medical image retrieval and medical image annotation tasks of ImageCLEF 2007. Separate sections describe each of the two tasks, with the participation and an evaluation of major findings from the results of each given. A total of 13 groups participated in the medical retrieval task and 10 in the medical annotation task.

The medical retrieval task added two new data sets for a total of over 66'000 images. Topics were derived from a log file of the Pubmed biomedical literature search system, creating realistic information needs with a clear user model.

The medical annotation task was in 2007 organized in a new format as a hierarchical classification had to be performed and classification could be stopped at any hierarchy level. This required algorithms to change significantly and to integrate a confidence level into their decisions to be able to judge where to stop classification to avoid making mistakes in the hierarchy. Scoring took into account errors and unclassified parts.

1 Introduction

ImageCLEF[1] [1,2] started within CLEF[2] (Cross Language Evaluation Forum [3]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. A medical image retrieval task was added in 2004 to explore domain–specific multilingual information retrieval and also multi–modal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task have both been part of ImageCLEF [4].

The important participation in CLEF and particularly ImageCLEF has shown the need for benchmarks, and their usefulness to the research community. In 2007, a total of 50 groups registered for ImageCLEF to get access to the data sets and tasks. Among these, 13 participated in the medical retrieval task and 10 in the medical automatic annotation task.

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

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 472–491, 2008.

Other important benchmarks in the field of visual information retrieval include TRECVID[3] on the evaluation of video retrieval systems [5], ImagEval[4], mainly on visual retrieval of images and image classification, and INEX[5] (INiative for the Evaluation of XML retrieval) concentrating on retrieval of multimedia based on structured data. Close contact with these initiatives exists to develop complementary evaluation strategies.

This article focuses on the two medical tasks of ImageCLEF 2007, whereas two other papers [6,7] describe the new object classification task and the photographic retrieval task. More detailed information can also be found on the task web pages. An even more detailed analysis of the 2005 medical image retrieval task and its outcomes is also available in [8].

2 The Medical Image Retrieval Task

The medical image retrieval task has been run for four consecutive years. In 2007, two new databases were added for a total of more than 66'000 images in the collection. For the generation of realistic topics or information needs, log files of the medical literature search system Pubmed were used.

2.1 General Overview

Again and as in previous years, the medical retrieval task showed to be popular among research groups registering for CLEF in 2007. In total 31 groups from all continents and 25 countries registered. A total of 13 groups finally submitted 149 runs that were used for the pooling required for the relevance judgments.

2.2 Databases

In 2007, the same four datasets were used as in 2005 and 2006 and two new datasets were added. The *Casimage* dataset was made available to participants [9], containing almost 9'000 images of 2'000 cases [10]. Images present in Casimage included mostly radiology modalities, but also photographs, PowerPoint slides, and illustrations. Cases were mainly in French, with around 20% being in English and 5% without any annotation. We also used the *PEIR*[6] (Pathology Education Instructional Resource) database with annotation based on the *HEAL*[7] project (Health Education Assets Library, mainly Pathology images [11]). This dataset contained over 33'000 images with English annotations, with the annotation being on a per image and not a per case basis as in Casimage. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[8] [12], was also

[3] http://www-nlpir.nist.gov/projects/t01v/
[4] http://www.imageval.org/
[5] http://inex.is.informatik.uni-duisburg.de/2006/
[6] http://peir.path.uab.edu/
[7] http://www.healcentral.com/
[8] http://gamma.wustl.edu/home.html

made available. This dataset contained over 2'000 images mainly from nuclear medicine with annotations provided per case and in English. The PathoPic[9] collection (Pathology images [13]) was included in our dataset containing about 7'800 images, with extensive annotation on a per image basis in German. Part of the German annotation was translated into English.

In 2007, we added two new datasets. The first was the *myPACS*[10] dataset of 15'140 images and 3'577 cases, all in English and containing mainly radiology images. The second was the Clinical Outcomes Research Initiative (*CORI*[11]) Endoscopic image database containing 1'496 images with an English annotation per image and not per case. The latter database extended the spectrum of the total dataset since there were previously only a few endoscopic images in the dataset. An overview of all datasets is shown in Table 1.

Table 1. The databases used in ImageCLEFmed 2007

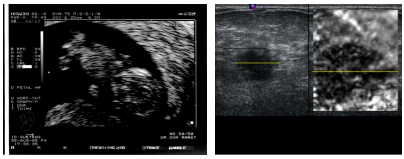
Collection Name	Cases	Images	Annotations	Annotations by Language
Casimage	2076	8725	2076	French – 1899, English – 177
MIR	407	1177	407	English – 407
PEIR	32319	32319	32319	English – 32319
PathoPIC	7805	7805	15610	German – 7805, English – 7805
myPACS	3577	15140	3577	English – 3577
Endoscopic	1496	1496	1496	English – 1496
Total	47680	66662	55485	French – 1899, English – 45781, German – 7805

2.3 Registration and Participation

In 2007, 31 groups from all 6 continents and 25 countries registered for the ImageCLEFmed retrieval task, underlining the strong interest in this evaluation campaign. As in previous years, about half of the registered groups submitted results, with those not submitting results blaming a lack of time. The feedback from the non–submitting groups remains positive as they report that the data is a very useful resource. The following groups submitted results:

- CINDI group, Concordia University, Montreal, Canada;
- Dokuz Eylul University, Izmir, Turkey;
- IPAL/CNRS joint lab, Singapore, Singapore;
- IRIT–Toulouse, Toulouse, France;
- MedGIFT group, University and Hospitals of Geneva, Switzerland;
- Microsoft Research Asia, Beijing, China;
- MIRACLE, Spanish University Consortium, Madrid, Spain;
- MRIM–LIG, Grenoble, France;
- OHSU, Oregon Health & Science University, Portland, OR, USA;

[9] `http://alf3.urz.unibas.ch/pathopic/intro.htm`
[10] `http://www.mypacs.net/`
[11] `http://www.cori.org`



Ultrasound with rectangular sensor.
Ultraschallbild mit rechteckigem Sensor.
Ultrason avec capteur rectangulaire.

Fig. 1. Example for a visual topic

- RWTH Aachen Pattern Recognition group. Aachen, Germany;
- SINAI group, University of Jaen Intelligent Systems, Jaen, Spain;
- State University New York (SUNY) at Buffalo, NY, USA;
- UNAL group, Universidad Nacional Colombia, Bogotà, Colombia;

In total, 149 runs were submitted, with individual groups submitting anywhere from 1 to 36 runs. Several submitted runs had incorrect formats. These runs were corrected by the organizers whenever possible but a few runs were finally omitted from the pooling process and the final evaluation because trec_eval could not parse the results even after our modifications. Groups were able to re-score these runs as the qrels files were made available.

2.4 Query Topics

Query topics for 2007 were generated based on a log file of Pubmed[12]. The log file of 24 hours contained a total of 77'895 queries. In general, the search terms were fairly vague and did not contain many image–related topics, so we filtered queries that had words such as image, video, and terms relating to modalities such as x–ray, CT, MRI, endoscopy, etc. We also aimed for the resulting terms to cover at least two or more of the following axes: modality, anatomic region, pathology, and visual observation (e.g., enlarged heart).

A total of 50 candidate topics were taken from these and sometimes an additional axis such as modality was added. From these topics we checked whether at least a few relevant images were in the database and from this, 30 topics were selected.

All topics were categorized with respect to the retrieval approach expected to perform best: visual topics, textual (semantic) topics and mixed topics. This was performed by an experienced image retrieval system developer. For each of the three retrieval approaches, 10 topics were selected for a total of 30 query topics that were distributed among the participants. Each topic consisted of the query itself in three languages (English, German, French) and 2–3 example images for the visual retrieval. Topic images were obtained from the Internet and were not part of the database. This made visual retrieval hard as most images were taken from different collections than those in the database and had changes in the gray level or color values.

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

Pulmonary embolism all modalities.
Lungenembolie alle Modalitäten.
Embolie pulmonaire, toutes les formes.

Fig. 2. Example for a semantic topic

Figure 1 shows a visual topic, and Figure 2 a topic with very different images in the results sets that should be well–suited for textual retrieval, only.

2.5 Relevance Judgments

Relevance judgments were performed by physicians who were students in the OHSU biomedical informatics graduate program. All were paid an hourly rate for their work. The pools for relevance judging were created by selecting the top ranking images from all submitted runs. The actual number selected from each run has varied by year. In 2007, it was 35 images per run, with the goal of having pools of about 800-1200 images in size for judging. The average pool size in 2007 was 890 images. Judges were instructed to rate images in the pools as definitely relevant (DR), partially relevant (PR), or not relevant (NR). Judges were instructed to use the partially relevant designation only in case they could not determine whether the image in question was relevant.

One of the problems was that all judges were English speakers but that the collection had a fairly large number of French and German documents. If the judgment required reading the text, judges had more difficulty ascertaining relevance. This could create a bias towards relevance for documents with English annotation. We also realized that several judges were not correctly taking into account modality information given in the queries. For this reason we manually reviewed qrels and selected some topics for rejudging. This led to results in these proceedings that are slightly different from the original working notes results. Techniques using modality detection generally performed slightly better with the revised relevance judgments. As we discovered an error in using treceval, that does not take into account rank information but only the similarity score, we also calculated a new MAP for all runs taking into account only the rank information. This is the same for many runs but a few runs become significantly better.

2.6 Submissions and Techniques

This section summarizes the main techniques used by the participants for retrieval and the sort of runs that they submitted. We had for the first time several problems with the submissions although we sent out a script to check runs for correctness before submission. In 2006, this script was part of the submission web site, but performance problems had us change this setup.

CINDI. The *CINDI* group submitted a total of 4 valid runs, two feedback runs and two automatic runs, each time one with mixed media and a purely visual run. Text retrieval uses a simple tf/idf weighting model and uses English, only. For visual retrieval a fusion model of a variety of features and image representations is used. The mixed media run simply combines the outcomes in a linear way.

DEU. *Dokuz Eylul University* submitted 5 runs, 4 visual and one textual run. The text runs is a simple bag of words approach and for visual retrieval several strategies were used containing color layout, color structure, dominant color and an edge histogram. Each run contained only one single technique.

IPAL. *IPAL* submitted 6 runs, all of them text retrieval runs. After having had the best performance for two years, the results are now only in the middle of the performance scale.

IRIT. The *IRIT* group submitted a single valid text retrieval run.

MedGIFT. The *MedGIFT* group submitted a total of 13 runs. For visual retrieval the GIFT (GNU Image Finding Tool) was used to create a baseline run as this system had been used in the same configuration since the beginning of ImageCLEF. Multilingual text retrieval was performed with EasyIR and a mapping of the text in the three languages towards MeSH (Medical Subject Headings) to search in semantic terms and avoid language problems.

MIRACLE. *MIRACLE* submitted 36 runs in total and thus most runs of all groups. The text retrieval runs were among the best, whereas visual retrieval was in the midfield. The combined runs were worse than text alone and also only in the midfield.

LIG. *MRIM–LIG* submitted 6 runs, all of them textual runs. Besides the best textual results, this was also the best overall result in 2007.

OHSU. *OHSU* submitted 10 textual and mixed runs, using Fire as a visual system. Their mixed runs had good performance as well as best early precision. Their modality detection run was the best performing mixed run.

RWTH. The human language technology and pattern recognition group from the RWTH Aachen University, Germany, submitted 10 runs using the FIRE system. The runs are based on a wide variety of 8 visual descriptors including image thumbnails, patch histograms, and various texture features. For the runs using text, a text retrieval system is used in the same way as in the last years. The weights for features are trained with a maximum entropy training method using the qrels of the 2005 and 2006 queries.

SINAI. The *SINAI* group submitted 30 runs in total, all of them textual or mixed. For text retrieval, the terms of the query are mapped onto MeSH, and then, the query is expanded with these MeSH terms.

SUNY. *SUNY* submitted 7 runs, all of which are mixed runs using Fire as visual system. One of the runs is among the best mixed runs.

UNAL. The *UNAL* group submitted 8 visual runs. The runs use a single visual feature and range towards the lower end of the performance spectrum.

MIXED. The combination of runs from *RWTH, OHSU, MedGIFT* resulted in 13 submissions, all of which were automatic and all used visual and textual information. These runs obtained a significantly better result when taking into account rank information for treceval.

2.7 Results

For the first time in 2007, the best overall official system used only text for the retrieval. Up until now the best systems always used a mix of visual and textual information. Nothing can really be said on the outcome of manual and relevance feedback submissions as there were too few submitted runs.

It became clear that most research groups participating had a single specialty, usually either visual or textual retrieval. By supplying visual and textual results as example, we gave groups the possibility to work on multi–modal retrieval as well.

Automatic Retrieval. As always, the majority of results were automatic and without any interaction. There were 146 runs in this category, with 27 visual runs, 80 mixed runs and 39 textual submissions, making automatic mixed media runs the most popular category. The results shown in the following tables are averaged over all 30 topics.

Visual Retrieval. Purely visual retrieval was performed in 27 runs and by six groups. Results from GIFT and FIRE (Flexible Image Retrieval Engine) were made available for research groups not having access to a visual retrieval engine. New MAP is the MAP calculated when taking into account rank information with treceval.

To make the tables shorter and to not bias results shown towards groups with many submissions, only the best two and the worst two runs of every group are shown in the tables. Table 2 shows the results for the visual runs. Most runs had an extremely low MAP (<3% MAP), which had been the case during the previous years as well. The overall results were lower than in preceding years, indicating that tasks might have become harder. On the other hand, two runs had good results and rivaled, at least for early precision, the best textual results. These two runs used data from 2005 and 2006 that was somewhat similar to the tasks in 2007 to train the system for optimal feature selection. This showed that an optimized feature weighting may result in a large improvement!

Textual Retrieval. A total of 39 submissions were purely textual and came from nine research groups. Table 3 shows the best and worst two results of every group for purely textual retrieval. The best overall runs were from LIG and were purely textual, which happened for the first time in ImageCLEF. LIG participated in

Table 2. Automatic runs using visual information (best/worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
RWTH-FIRE-ME-NT-tr0506	1376	0.2427	0.2426	0.283	0.48	0.45	0.3756
RWTH-FIRE-ME-NT-tr06	1368	0.23	0.2300	0.2696	0.48	0.4467	0.3722
CINDI_IMG_FUSION	567	0.0355	0.0354	0.0751	0.1533	0.1233	0.1122
RWTH-FIRE-NT-emp	506	0.0264	0.0264	0.056	0.0933	0.0933	0.0744
RWTH-FIRE-NT-emp2	474	0.0255	0.0255	0.0535	0.1067	0.0933	0.0656
miracleVisG	496	0.0182	0.0182	0.0448	0.0933	0.08	0.0767
miracleVisGFANDmm	156	0.01	0.01	0.0221	0.0667	0.0667	0.05
miracleVisGFANDavg	156	0.0085	0.0085	0.0185	0.0467	0.0467	0.0556
miracleVisGFANDmin	156	0.0079	0.0079	0.0184	0.04	0.0367	0.0478
UNALCO-nni_Sobel	433	0.0072	0.0076	0.0668	0.02	0.02	0.0133
UNALCO-nni_FeatComb	531	0.0066	0.0205	0.0825	0.0133	0.02	0.0122
DEU_CS-DEU_R2	239	0.0062	0.0111	0.0433	0.0133	0.0067	0.0022
UNALCO-svmRBF_RGBHis	329	0.0048	0.0135	0.0481	0.0133	0.0133	0.0089
UNALCO-svmRBF_Tamura	341	0.0046	0.0055	0.0536	0.0133	0.0067	0.01
GE_4_8	245	0.0035	0.0035	0.0241	0.04	0.0333	0.0233
GE-GE_GIFT4	244	0.0035	0.0035	0.024	0.04	0.0333	0.0233
GE-GE_GIFT8	245	0.0035	0.0035	0.024	0.04	0.0333	0.0233
DEU_CS-DEU_R4	199	0.0017	0.0035	0.04	0.0067	0.0033	0.0056
DEU_CS-DEU_R3	216	0.0016	0.0079	0.0442	0.0067	0.01	0.0056
DEU_CS-DEU_R5	195	0.0013	0.0038	0.0351	0	0	0.0078

ImageCLEF this year for the first time. Early precision (P5) was similar to the best purely visual runs and the best mixed runs had a very high early precision. The highest P10 was a mixed system where the MAP was situated lower. Despite its name, MAP is more of a recall–oriented measure. Re–scoring of the results with treceval basing the order of documents on the rank results in a few runs becoming significantly better but does not change many of the other runs.

Mixed Retrieval. Mixed automatic retrieval had the highest number of submissions of all categories. There were 80 runs submitted by 8 participating groups.

Table 4 summarizes the best two and the worst two mixed runs of every group. For some groups the results for mixed runs were better than the best text runs but for others this was not the case. This underlines the fact that combinations between visual and textual features have to be done with care. Another interesting fact is that some systems with only a mediocre MAP performed extremely well with respect to early precision. All early precision values (P5, P10, P30) had their best results with mixed submissions.

Another interesting fact could be observeredafter correctly rescroting the results as the best mixed run is in this case much better than the best textual run. All combination runs of gift, fire, and ohsu obtain extremely much better results bringing them up the performing runs.

Table 3. Automatic runs using only text (best and worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
LIG-MRIM-LIG_MU_A	1904	0.3538	0.3533	0.3954	0.42	0.43	0.3844
LIG-MRIM-LIG_GM_A	1898	0.3517	0.3513	0.395	0.42	0.4233	0.3922
miracleTxtENN	1842	0.3385	0.3427	0.406	0.4933	0.4567	0.3578
LIG-MRIM-LIG_GM_L	1909	0.3345	0.3338	0.3855	0.4467	0.4433	0.3856
ohsu_text_e4_out_rev1	1459	0.3317	0.3467	0.3957	0.46	0.4733	0.3956
LIG-MRIM-LIG_MU_L	1912	0.3269	0.3263	0.3802	0.44	0.4333	0.3656
OHSU-OHSU_txt_exp2	1162	0.3192	0.3339	0.3688	0.46	0.4733	0.3956
SinaiC100T100	1985	0.2944	0.3052	0.3505	0.3933	0.4367	0.3967
UB-NLM-UBTextBL1	1825	0.2897	0.2897	0.3279	0.3867	0.41	0.3678
SinaiC040T100	1937	0.2838	0.2978	0.3269	0.4067	0.4533	0.4033
IPAL1_TXT_BAY_ISA0.2	1515	0.2784	0.2781	0.323	0.42	0.39	0.31
IPAL1_TXT_BAY_ISA0.1	1517	0.2783	0.278	0.3233	0.4133	0.39	0.3122
OHSU-as_out_1000rev1_c	1871	0.2754	0.2799	0.3346	0.44	0.4367	0.36
OHSU-oshu_as_is_1000	1871	0.2754	0.2816	0.3345	0.44	0.4367	0.36
IPAL_TXT_BAY_ALLREL2	1520	0.275	0.2746	0.3215	0.4067	0.3767	0.3122
IPAL4_TXT_BAY_ISA0.4	1468	0.2711	0.2708	0.3218	0.3933	0.3867	0.3078
SinaiC030T100	1910	0.271	0.2748	0.3126	0.42	0.41	0.3822
miracleTxtXN	1784	0.2647	0.2659	0.3711	0.3267	0.3367	0.3167
UB-NLM-UBTextBL2	1666	0.2436	0.2437	0.2921	0.3133	0.3033	0.2811
GE_EN	1839	0.2369	0.2373	0.2867	0.2867	0.3333	0.2678
SinaiC020T100	1589	0.2356	0.2366	0.2665	0.34	0.3467	0.3422
GE_MIX	1806	0.2186	0.2192	0.2566	0.3133	0.2967	0.2622
DEU_CS-DEU_R1	727	0.1611	0.1618	0.1876	0.3067	0.32	0.3033
GE_DE	1166	0.1433	0.1441	0.209	0.2267	0.2	0.15
UB-NLM-UBTextFR	1248	0.1414	0.1413	0.2931	0.2	0.1933	0.1533
GE_FR	1139	0.115	0.115	0.1503	0.1	0.1267	0.1289
miracleTxtFRT	906	0.0863	0.085	0.1195	0.1733	0.1733	0.15
miracleTxtFRN	815	0.0846	0.0822	0.1221	0.26	0.18	0.1367
IRIT_RunMed1	1163	0.0486	0.1201	0.1682	0.0533	0.05	0.0756

2.8 Manual and Interactive Retrieval

Only three runs were in the manual or interactive sections, making any real comparison impossible. Table 5 lists these runs and their performance Although information retrieval with relevance feedback or manual query modifications are thought to be a very important area to improve performance, research groups in ImageCLEF 2007 did not make use of it.

2.9 Conclusions

Visual retrieval without learning had very low results for MAP and even for early precision (although with a smaller difference from text retrieval). Visual topics perform well using visual techniques. Extensive learning of feature selection and weighting can have enormous gain in performance as shown by FIRE.

Table 4. Automatic runs using mixed information (best and worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
ohsu_m2_rev1_c	1778	0.3415	0.4084	0.4099	0.4467	0.4333	0.37
SinaiC100T80	1976	0.2999	0.3026	0.3425	0.4	0.4567	0.4067
RWTH-FIRE-ME-tr0506	1566	0.2962	0.2962	0.3414	0.4733	0.4667	0.3978
RWTH-FIRE-ME-tr06	1566	0.296	0.296	0.3407	0.4933	0.47	0.3978
UB-NLM-UBTI_3	1833	0.2938	0.2938	0.3306	0.3867	0.4167	0.3689
UB-NLM-UBTI_1	1831	0.293	0.2928	0.335	0.3867	0.4	0.3867
SinaiC040T80	1948	0.2914	0.2949	0.3236	0.4267	0.4667	0.4133
UB-NLM-UBmixedMulti2	1666	0.2537	0.2537	0.3011	0.3467	0.3167	0.29
miracleMixGENTRIGHTmin	1608	0.248	0.2439	0.2936	0.3667	0.3533	0.3011
RWTH-FIRE-emp2	1520	0.2302	0.2302	0.2803	0.3867	0.4	0.3689
RWTH-FIRE-emp	1521	0.2261	0.2261	0.2758	0.38	0.4	0.3711
miracleMixGENTRIGHTmax	1648	0.2225	0.2259	0.2687	0.3067	0.32	0.2856
GE_VT1_4	1806	0.2195	0.2199	0.2567	0.32	0.3033	0.2622
GE_VT1_8	1806	0.2195	0.2204	0.2566	0.32	0.3033	0.2622
OHSU-ohsu_m1	509	0.2167	0.2374	0.2405	0.3867	0.3933	0.3567
CINDI_TXT_IMAGE_LINEAR	944	0.1906	0.1914	0.2425	0.34	0.3133	0.2822
SinaiC060T50	1863	0.1874	0.1882	0.2245	0.4	0.3767	0.2789
GE_VT10_4	1192	0.1828	0.1829	0.2141	0.3	0.31	0.2633
GE_VT10_8	1196	0.1828	0.1839	0.214	0.3	0.31	0.2633
SinaiC020T50.clef	1544	0.1727	0.1726	0.1967	0.3133	0.3267	0.2744
UB-NLM-UBmixedFR	997	0.1364	0.1363	0.2168	0.2133	0.2	0.1789
ohsu_comb3_ef_wt1_rev1_c	903	0.1113	0.1144	0.1525	0.2533	0.2433	0.1522
ohsu_fire_ef_wt2_rev1_c	519	0.0577	0.0608	0.0888	0.16	0.16	0.1122
3fire-7ohsu	1887	0.0303	0.2355	0.1115	0.0067	0.01	0.0067
5fire-5ohsu	1892	0.0291	0.2871	0.1012	0.0067	0.0067	0.0078
5gift-5ohsu	1317	0.0153	0.1867	0.1151	0	0.0033	0.0022
7gift-3ohsu	1319	0.0148	0.2652	0.1033	0	0.0033	0.0022
miracleGFANDminLEFTmm	156	0.0097	0.0097	0.0197	0.0533	0.0533	0.0544
miracleGFANDminLEFTmax	156	0.0079	0.0079	0.0184	0.04	0.0367	0.0478

Table 5. The only three runs not using automatic retrieval

Run	Rel.	MAP	new	bpref	P10	P30	media	interaction
CINDI_IMG_FUSION_RF	610	0.04	0.04	0.09	0.15	0.119	visual	feedback
CINDI_TXT_IMG_RF_LIN	773	0.12	0.12	0.19	0.36	0.251	mixed	feedback
OHSU-oshu_man2	1795	0.35	0.36	0.40	0.443	0.349	textual	manual

Purely textual runs had the best overall results for the first time and text retrieval was shown to work well for most topics. Mixed–media runs were the most popular category and are often better in performance than text or visual features alone. When correctly scoring all runs the best performance was actually in this category. Still, in many cases the mixed media runs did not perform as

well as text alone, showing that care needs to be taken to combine media. These runs do have the best performance for all early precision values.

Interactive and manual queries were almost absent from the evaluation and this remains an important problem. ImageCLEFmed has to put these domains more into the focus of the researchers although this requires more resources to perform the evaluation. System–oriented evaluation is an important part but only interactive retrieval can show how well a system can really help the users.

With respect to performance measures, there was less correlation between the measures than in previous years. The runs with the best early precision (P10) were not as good in MAP to the best overall systems. This needs to be investigated as MAP is indeed a good indicator for overall system performance but early precision might be much more what real users are looking for.

3 The Medical Automatic Annotation Task

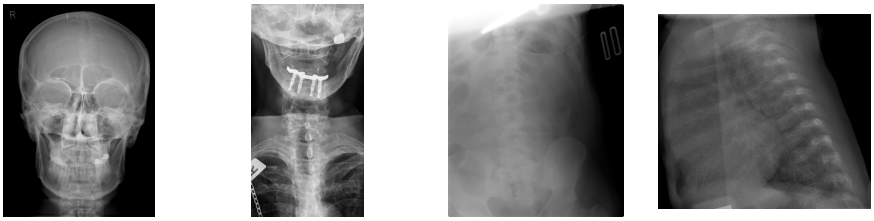
Over the last two years, automatic medical image annotation has been evolved from a simple classification task with about 60 classes to a task with about 120 classes. From the very start however, it was clear that the number of classes cannot be scaled indefinitely, and that the number of classes that are desirable to be recognised in medical applications is far to big to assemble sufficient training data to create suitable classifiers. To address this issue, a hierarchical class structure such as the IRMA code [14] can be a solution which allows to create a set of classifiers for subproblems. The classes in the last years were based on the IRMA code where created by grouping similar codes in one class. This year, the task has changed and the objective is to predict complete IRMA codes instead of simple classes.

This year's medical automatic annotation task builds on top of last year: 1,000 new images were collected and are used as test data, the training and the test data of last year was used as training and development data respectively.

3.1 Database and Task Description

The complete database consists of 12'000 fully classified medical radiographs taken randomly from medical routine at the RWTH Aachen University Hospital. 10'000 of these were release together with their classification as training data, another 1'000 were also published with their classification as validation data to allow for tuning classifiers in a standardised manner. One thousand additional images were released at a later date without classification as test data. These 1'000 images had to be classified using the 11'000 images (10'000 training + 1'000 validation) as training data.

Each of the 12'000 images is annotated with its complete IRMA code (see Sec. 3.1). In total, 116 different IRMA codes occur in the database, the codes are not uniformly distributed, but some codes have a significant larger share among the



1121-120-200-700 1121-120-310-700 1121-127-700-500 1123-211-500-000

Fig. 3. Example images from the medical annotation task with full IRMA-code. The textual representation of the IRMA codes is (from left to right):
T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), unspecified; A: cranium, unspecified, unspecified; B: musculosceletal system, unspecified, unspecified.
T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), unspecified; A: spine, cervical spine, unspecified; B: musculosceletal system, unspecified, unspecified.
T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), supine; A: abdomen, unspecified, unspecified; B: uropoietic system, unspecified, unspecified.
T: x-ray, plain radiography, analog, high beam energy; D: sagittal, lateral, right-left, inspiration; A: chest, unspecified, unspecified; B: unspecified, unspecified, unspecified.

data than others. The least frequent codes however, are represented at least 10 times in the training data to allow for learning suitable models.

Example images from the database together with textual labels and their complete code are given in Figure 3.

IRMA Code. Existing medical terminologies such as the MeSH thesaurus are poly-hierarchical, i.e., a code entity can be reached over several paths. However, in the field of content-based image retrieval, we frequently find class-subclass relations. The mono-hierarchical multi-axial IRMA code strictly relies on such part-of hierarchies and, therefore, avoids ambiguities in textual classification [14]. In particular, the IRMA code is composed from four axes having three to four positions, each in $\{0, \ldots 9, a, \ldots z\}$, where "'0"' denotes "'not further specified"'. More precisely,

- the technical code (T) describes the imaging modality;
- the directional code (D) models body orientations;
- the anatomical code (A) refers to the body region examined; and
- the biological code (B) describes the biological system examined.

This results in a string of 13 characters (IRMA: TTTT – DDD – AAA – BBB). For instance, the body region (anatomy, three code positions) is defined as follows:

```
AAA
000 not further specified
...
400 upper extrimity (arm)
410 upper extrimity (arm); hand
411 upper extrimity (arm); hand; finger
412 upper extrimity (arm); hand; middle hand
413 upper extrimity (arm); hand; carpal bones
420 upper extrimity (arm); radio carpal joint
430 upper extrimity (arm); forearm
431 upper extrimity (arm); forearm; distal forearm
432 upper extrimity (arm); forearm; proximal forearm
440 upper extrimity (arm); ellbow
...
```

The IRMA code can be easily extended by introducing characters in a certain code position, e.g., if new imaging modalities are introduced. Based on the hierarchy, the more code position differ from "'0"', the more detailed is the description.

Hierarchical Classification. To define a evaluation scheme for hierarchical classification, we can consider the 4 axes to be uncorrelated. Hence, we assume the axes independently and just sum up the errors for each axis independently.

Hierarchical classification is a well-known topic in different field. For example the classification of documents often is done using an ontology-based class hierarchy [15] and in information extraction similar techniques are applied [16]. In our case, however we developed a novel evaluation scheme to account for the particularities of the IRMA code which considers errors that are made early in a hierarchy to be worse than errors that are made at a fine level, and it is explicitly possible to predict a code partially, i.e. to predict a code up to a certain position and put wild-cards for the remaining positions, which is penalised but only with half the penalty a misclassification is penalised.

Our evaluation scheme is described in the following, where we only consider one axis. The same scheme is applied to each axis individually.

Let $l_1^I = l_1, l_2, \ldots, l_i, \ldots, l_I$ be the *correct* code (for one axis) of an image, i.e. if a classifier predicts this code for an image, the classification is perfect. Further, let $\hat{l}_1^I = \hat{l}_1, \hat{l}_2, \ldots, \hat{l}_i, \ldots, \hat{l}_I$ be the *predicted* code (for one axis) of an image.

The correct code is specified completely: l_i is specified for each position. The classifiers however, are allowed to specify codes only up to a certain level, and predict "*don't know*" (encoded by $*$) for the remaining levels of this axis.

Given an incorrect classification at position $\hat{l}_i$ we consider all succeeding decisions to be wrong and given a not specified position, we consider all succeeding decisions to be not specified.

We want to penalise wrong decisions that are easy (fewer possible choices at that node) over wrong decisions that are difficult (many possible choices at that node), we can say, a decision at position l_i is correct by chance with a probability

of $\frac{1}{b_i}$ if b_i is the number of possible labels for position i. This assumes equal priors for each class at each position.

Furthermore, we want to penalise wrong decisions at an early stage in the code (higher up in the hierarchy) over wrong decisions at a later stage in the code (lower down on the hierarchy) (i.e. l_i is more important than l_{i+1}).

Assembling the ideas from above straight forwardly leads to the following equation:

$$\sum_{i=1}^{I} \underbrace{\frac{1}{b_i}}_{(a)} \underbrace{\frac{1}{i}}_{(b)} \underbrace{\delta(l_i, \hat{l_i})}_{(c)}$$

with

$$\delta(l_i, \hat{l_i}) = \begin{cases} 0 & \text{if } l_j = \hat{l_j} \quad \forall j \leq i \\ 0.5 & \text{if } l_j = * \quad \exists j \leq i \\ 1 & \text{if } l_j \neq \hat{l_j} \quad \exists j \leq i \end{cases}$$

where the parts of the equation account for

(a) accounts for difficulty of the decision at position i (branching factor)
(b) accounts for the level in the hierarchy (position in the string)
(c) correct/not specified/wrong, respectively.

In addition, for every code, the maximal possible error is calculated and the errors are normed such that a fully incorrect decision (i.e. all positions wrong) gets an error count of 1.0 and an image classified correctly in all positions has an error of 0.0.

Table 6 shows examples for a correct code with different predicted codes. Predicting the completely correct code leads to an error measure of 0.0, predicting all positions incorrectly leads to an error measure of 1.0. The examples in Table 6 demonstrate that a classification error in a position at the back of the code results in a lower error measure than a position in one of the first positions. The last column of the table show the effect of the branching factor b. In this column we assumed the branching factor of the code is $b = 2$ in each node of the hierarchy. It can be observed that the errors for the later positions have more weight compared to the real errors in the real hierarchy.

Table 6. Example scores for hierarchical classification, based on the correct code IRMA TTTT = 318a and assuming the branching factor would be 2 in each node of the hie

classified	error measure	error measure (b=2)
318a	0.000	0.000
318*	0.024	0.060
3187	0.049	0.120
31*a	0.082	0.140
31**	0.082	0.140
3177	0.165	0.280
3***	0.343	0.260
32**	0.687	0.520
1000	1.000	1.000

3.2 Participating Groups and Methods

In the medical automatic annotation task, 29 groups registered of which 10 groups participated, submitting a total of 68 runs. The group with the highest number of submissions had 30 runs in total.

In the following, groups are listed alphabetically and their methods are described shortly.

BIOMOD: University of Liege, Belgium. The Bioinformatics and Modelling group from the University Liege in Belgium submitted four runs. The approach is based on an object recognition framework using extremely randomised trees and randomly extracted sub-windows [17]. All runs use the same technique but differ how the code is assembled.

BLOOM: IDIAP, Switzerland. The Blanceflor-om2-toMed group from IDIAP in Martigny, Switzerland submitted 7 runs. All runs use support vector machines (either in one-against-one or one-against-the-rest manner). Features used are downscaled versions of the images, SIFT features extracted from sub-images, and combinations of these [18].

Geneva: medGIFT Group, Switzerland. The medGIFT group from Geneva, Switzerland submitted 3 runs, each of the runs uses the GIFT image retrieval system. The runs differ in the way, the IRMA-codes of the top-ranked images are combined [19].

CYU: Information Management AI lab, Taiwan. The Information Management AI lab from the Ching Yun University of Jung-Li, Taiwan submitted one run using a nearest neighbour classifier using different global and local image features which are particularly robust with respect to lighting changes.

MIRACLE: Madrid, Spain. The Miracle group from Madrid, Spain submitted 30 runs. The classification was done using a 10-nearest neighbour classifier and the features used are gray-value histograms, Tamura texture features, global texture features, and Gabor features, which were extracted using FIRE. The runs differ which features were used and how the prediction of the code was done.

Oregon Health State University, Portland, OR, USA. The Department of Medical Informatics and Clinical Epidemiology of the Oregon Health and Science University in Portland, Oregon submitted two runs using neural networks and GIST descriptors. One of the runs uses a support vector machine as a second level classifier to help discriminating the two most difficult classes.

RWTHi6: RWTH Aachen University, Aachen, Germany. The Human Language Technology and Pattern Recognition group of the RWTH Aachen University in Aachen, Germany submitted 6 runs, all are based on sparse histograms of image patches which were obtained by extracting patches at each position in the image [20]. One run is a combination of 4 normal runs, and one run does the classification axis-wise.

IRMA: RWTH Aachen University, Medical Informatics, Aachen, Germany. The IRMA group from the RWTH Aachen University Hospital in

Aachen, Germany submitted three baseline runs using weighted combinations of nearest neighbour classifiers using texture histograms, image cross correlations, and the image deformation model. The parameters used are exactly the same as used in previous years. The runs differ in the way in which the codes of the five nearest neighbours are used to assemble the final predicted code.

UFR: University of Freiburg, Computer Science Dep., Freiburg, Germany. The Pattern Recognition and Image Processing group from the University Freiburg, Germany, submitted four runs using relational features calculated around interest points which are later combined to form cluster cooccurrence matrices [21]. Three different classification methods were used.

UNIBAS: University of Basel, Switzerland. The Databases and Information Systems group from the University Basel, Switzerland submitted 14 runs using a pseudo two-dimensional hidden Markov model to model image deformation in the images which were scaled down keeping the aspect ratio such that the longer side has a length of 32 pixels [23].

3.3 Results

An overview of the results of the evaluation is given in Table 7. For each group, the number of submissions, the best and the worst rank, the minimal and the maximal score, the mean and the median score, the best and the worst error rate, the mean and the median error rate are given.

The method which had the best result last year is now at rank 8, which gives an impression how much improvement in this field was achieved over the last year.

Looking at the results for individual images, we noted, that only one image was classified correctly by all submitted runs (top left image in Fig. 3). No image was misclassified by all runs.

3.4 Discussion

Analysing the results, it can be observed that the top-performing runs do not consider the hierarchical structure of the given task, but rather use each individual code as one class and train a 116 classes classifier. This approach seems to work better given the currently limited amount of codes, but obviously would not scale up infinitely and would probably lead to a very high demand for appropriate training data if a much larger amount of classes is to be distinguished. The best run using the code is on rank 6, builds on top of the other runs from the same group and uses the hierarchy only in a second stage to combine the four runs.

Furthermore, it can be seen that a method that is applied once accounting for the hierarchy/axis structure of the code and once using the straight forward classification into 116 classes approach, the one which does not know about the hierarchy clearly outperforms the other one (runs on ranks 11 and 13/7 and 14,16).

Table 7. Results of the evaluation by participating group. For each group, the number of submitted runs, the rank of the best and worst run, and the minimum, maximum, mean, and medium error count and error rate are given.

		rank		score				ER			
group	# sub	min	max	min	max	mean	median	min	max	mean	median
BIOMOD	4	30	35	73.82	95.25	80.90	77.26	22.90	36.00	29.28	29.10
BLOOM	7	1	29	26.85	72.41	40.44	29.46	10.30	20.80	13.77	11.50
Geneva	3	63	65	375.72	391.02	385.68	390.29	99.00	99.70	99.33	99.30
CYU	1	33	33	79.30	79.30	79.30	79.30	25.30	25.30	25.30	25.30
MIRACLE	30	36	68	158.82	505.62	237.42	196.18	49.30	89.00	62.09	55.50
OHSU	2	26	27	67.81	67.98	67.89	67.89	22.70	22.70	22.70	22.70
RWTHi6	6	6	13	30.93	44.56	35.16	33.88	11.90	17.80	13.38	12.55
IRMA	3	17	34	51.34	80.47	61.45	52.54	18.00	45.90	27.97	20.00
UFR	5	7	16	31.44	48.41	41.29	45.48	12.10	17.90	15.36	16.80
UNIBAS	7	19	25	58.15	65.09	61.64	61.41	20.20	23.20	22.26	22.50

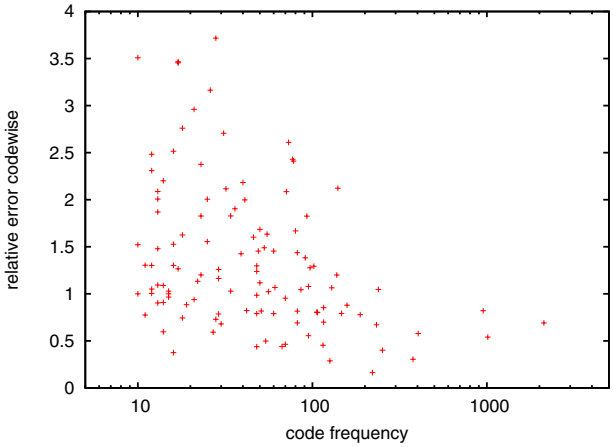


Fig. 4. Code–wise relative error as a function of the frequency of this code in the training data

Another clear observation is that methods using local image descriptors outperform methods using global image descriptors. In particular, the top 16 runs are all using either local image features alone or local image features in combination with a global descriptor.

It is also observed that images where a large amount of training data is available are more far more likely to be classified correctly.

Considering the ranking according to the applied hierarchical measure and the ranking according to the error rate it can clearly be seen that there are hardly any differences. Most of the differences are clearly due to use of the code (mostly inserting of wildcard characters) which can lead to an improvement for the hierarchical evaluation scheme, but will always lead to a deterioration of the error rate.

3.5 Conclusion

The success of the medical automatic annotation task could be continued, the number of participants is pretty constant, but a clear performance improvement of the best method could be observed. Although only few groups actively tried to exploit the hierarchical class structure many of the participants told us that they consider this an important research topic and that a further investigation is desired.

Our goal for future tasks is to motivate more groups to participate and to increase the database size such that it is necessary to use the hierarchical class structure actively.

4 Overall Conclusions

The two medical tasks of ImageCLEF again attracted a very large number of registrations and participation. This underlines the importance of such evaluation campaigns giving researchers the opportunity to evaluate their systems without the tedious task of creating databases and topics. In domains such as medical retrieval this is particularly important as data access if often difficult.

In the medical retrieval task, visual retrieval without any learning only obtained good results for a small subset of topics. With learning this can change strongly and deliver even for purely visual retrieval very good results. Mixed–media retrieval was the most popular category and results were often better for mixed–media than textual runs of the same groups. This shows that mixed–media retrieval requires much work and more needs to be learned on such combinations. The best systems concerning early precision were mixed media runs. Interactive retrieval and manual query modification were only used in 3 out of the 149 submitted runs. This shows that research groups prefer submitting automatic runs, although interactive retrieval is important and still must be addressed by researchers.

For the annotation task, it was observed that techniques that rely heavily on recent developments in machine learning and build on modern image descriptors clearly outperform other methods. The class hierarchy that was provided could only lead to improvements for a few groups. Overall, the runs that use the class hierarchy perform worse than those, which consider every code as a unique class giving the impression that for the current number of 116 unique codes the training data is sufficient to train a joint classifier.

Acknowledgements

We thank CLEF for supporting ImageCLEF. We also thank all organizations who provided images and annotations for this year's task, including myPACS.net (Rex Jakobovits) and the OHSU CORI project (Judith Logan).

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contracts Ne-572/6 and Le-1108/4, the Swiss National Science Foundation (FNS) under contract 205321-109304/1, the American National Science Foundation (NSF) with grant ITR–0325160, and the EU Sixth Framework Program with the SemanticMining project (IST NoE 507505) and the MUSCLE NoE.

References

1. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 243–251. Springer, Heidelberg (2005)
2. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
3. Savoy, J.: Report on CLEF–2001 experiments. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 27–43. Springer, Heidelberg (2002)
4. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: CLEF working notes, Alicante, Spain (September 2006)
5. Smeaton, A.F., Over, P., Kraaij, W.: TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In: Proceedings of the international ACM conference on Multimedia 2004 (ACM MM 2004), New York City, NY, USA, October 2004, pp. 652–655 (2004)
6. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
7. Deselaers, T., Hanbury, A., et al.: Overview of the ImageCLEF 2007 object retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
8. Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: Imageclefmed: A text collection to advance biomedical image retrieval. Journal of the American Medical Informatics Association (September/October 2006)
9. 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, 295–305 (2004)
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), 1–6 (2004)
11. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. Academic Medicine 78(3), 249–253 (2003)
12. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet–based nuclear medicine teaching file. Journal of Nuclear Medicine 36(8), 1520–1527 (1995)
13. Glatz-Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Webbasierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. Pathologe 24, 394–399 (2003)

14. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: SPIE 2003, vol. 5033, pp. 440–451 (2003)
15. Sun, A., Lim, E.P.: Hierarchical text classification and evaluation. In: IEEE International Conference on Data Mining (ICDM 2001), San Jose, CA, USA, November 2001, pp. 521–528 (2001)
16. Maynard, D., Peters, W., Li, Y.: Metrics for evaluation of ontology–based information extraction. In: Evaluation of Ontologies for the Web (EON 2006), Edinburgh, UK (2006)
17. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random subwindows for robust image classification. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005), June 2005, vol. 1, pp. 34–40. IEEE, Los Alamitos (2005)
18. Tommasi, T., Orabona, F., Caputo, B.: CLEF2007 Image Annotation Task: an SVM–based Cue Integration Approach. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
19. Zhou, X., Gobeill, J., Ruch, P., Müller, H.: University and Hospitals of Geneva at ImageCLEF 2007. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
20. Deselaers, T., Hegerath, A., Keysers, D., Ney, H.: Sparse patch–histograms for object classification in cluttered images. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) DAGM 2006. LNCS, vol. 4174, pp. 202–211. Springer, Heidelberg (2006)
21. Setia, L., Teynor, A., Halawani, A., Burkhardt, H.: Image classification using cluster-cooccurrence matrices of local relational features. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, Santa Barbara, CA, USA (2006)
22. Setia, L., Burkhardt, H.: Learning taxonomies in large image databases. In: ACM SIGIR Workshop on Multimedia Information Retrieval, Amsterdam, Holland (2007)
23. Springmann, M., Schuldt, H.: Speeding up idm without degradation of retrieval quality. In: Nardi, A., Peters, C. (eds.) Working Notes of the CLEF Workshop 2007 (2007)