

# Overview of the ImageCLEF 2007 Medical Retrieval and Annotation Tasks

Henning Müller<sup>1</sup>, Thomas Deselaers<sup>2</sup>, Eugene Kim<sup>3</sup>, Jayashree Kalpathy–Cramer<sup>3</sup>,  
Thomas M. Deserno<sup>4</sup>, William Hersh<sup>3</sup>

<sup>1</sup> Medical Informatics, University and Hospitals of Geneva, Switzerland

<sup>2</sup> Computer Science Dep., RWTH Aachen University, Germany

<sup>3</sup> Oregon Health and Science University (OHSU), Portland, OR, USA

<sup>4</sup> Medical Informatics, RWTH Aachen University, Germany

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 news data sets for a total of over 66'000 images. Tasks were derived from a log file of the Pubmed biomedical literature search system, creating realistic information needs with a clear user model in mind.

The medical annotation task was in 2007 organised in a new format as a hierarchical classification had to be performed and classification could be stopped at any confidence 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.

## Categories and Subject Descriptors

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

## General Terms

Measurement, Performance, Experimentation

## Keywords

Image Retrieval, Performance Evaluation, Image Classification, Medical Imaging

## 1 Introduction

ImageCLEF<sup>1</sup> [3, 2] started within CLEF<sup>2</sup> (Cross Language Evaluation Forum [15]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. A medical image

---

<sup>1</sup><http://ir.shef.ac.uk/imageclef/>

<sup>2</sup><http://www.clef-campaign.org/>

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 were both part of ImageCLEF [12].

The enthusiastic participation in CLEF and particularly for ImageCLEF has shown the need for benchmarks and their usefulness to the research community. Again in 2007, a total of 48 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.

Other important benchmarks in the field of visual information retrieval include TRECVID<sup>3</sup> on the evaluation of video retrieval systems [18], ImagEval<sup>4</sup>, mainly on visual retrieval of images and image classification, and INEX<sup>5</sup> (INiative for the Evaluation of XML retrieval) concentrating on retrieval of multimedia based on structured data. Close contact exists with these initiatives to develop complementary evaluation strategies.

This article focuses on the two medical tasks of ImageCLEF 2007, whereas two other papers [7, 4] describe the new object classification task and the new photographic retrieval task. More detailed information can also be found on the task web pages for ImageCLEFmed<sup>6</sup> and the medical annotation task<sup>7</sup>. A 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 many research groups registering for CLEF. In total 31 groups from all continents and 25 countries registered. A total of 13 groups 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*<sup>8</sup> dataset was made available to participants [13], containing almost 9'000 images of 2'000 cases [14]. Images present in Casimage include mostly radiology modalities, but also photographs, PowerPoint slides and illustrations. Cases are mainly in French, with around 20% being in English and 5% without annotation. We also used the *PEIR*<sup>9</sup> (Pathology Education Instructional Resource) database with annotation based on the *HEAL*<sup>10</sup> project (Health Education Assets Library, mainly Pathology images [1]). This dataset contains 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<sup>11</sup> [22], was also made available to us for ImageCLEFmed. This dataset contains over 2'000 images mainly from nuclear medicine with annotations provided per case and in English. Finally, the PathoPic<sup>12</sup>

---

<sup>3</sup><http://www-nlpir.nist.gov/projects/t01v/>

<sup>4</sup><http://www.imageval.org/>

<sup>5</sup><http://inex.is.informatik.uni-duisburg.de/2006/>

<sup>6</sup><http://ir.ohsu.edu/image>

<sup>7</sup><http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/medicalaat.html>

<sup>8</sup><http://www.casimage.com/>

<sup>9</sup><http://peir.path.uab.edu/>

<sup>10</sup><http://www.healcentral.com/>

<sup>11</sup><http://gamma.wustl.edu/home.html>

<sup>12</sup><http://alf3.urz.unibas.ch/pathopic/intro.htm>

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

collection (Pathology images [6]) was included into our dataset. It contains 9'000 images with extensive annotation on a per image basis in German. A short part of the German annotation is translated into English.

In 2007, we added two new datasets. The first was the *myPACS*<sup>13</sup> 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*<sup>14</sup>) Endoscopic image database contains 1'496 images with an English annotation per image and not per case. This database extends the spectrum of the total dataset as so far there were only few endoscopic images in the dataset. An overview of the datasets can be seen in Table 1

As such, we were able to use more than 66'000 images, with annotations in three different languages. Through an agreement with the copyright holders, we were able to distribute these images to the participating research groups. The myPACS database required an additional copyright agreement making the process slightly more complex than in previous years.

### 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, only about half of the registered groups finally submitted results, often blaming a lack of time for this. The feedback of these groups remains positive as they say to use the data for their research as a very useful resource.

The following groups finally also submitted results for the medical image retrieval task:

- 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;

<sup>13</sup><http://www.mypacs.net/>

<sup>14</sup><http://www.corri.org>



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

Figure 1: Example for a visual topic.

- MRIM-LIG, Grenoble, France;
- OHSU, Oregon Health & Science University, Portland, OR, USA;
- 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, Bogotá, Colombia;

In total, 149 runs were submitted, with the maximum being 36 of a single group and the minimum a single run per group. Several runs had incorrect formats. These runs were corrected by the organisers 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. All groups have the possibility to describe further runs in their working notes papers after the format corrections as the `qrels` files were made available to all.

## 2.4 Query Topics

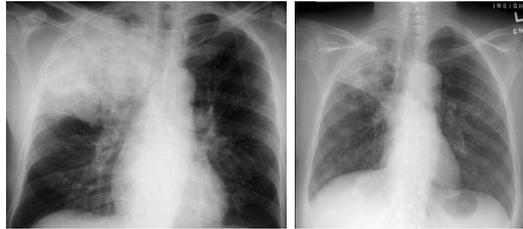
Query topics for 2007 were generated based on a log file of Pubmed<sup>15</sup>. The log file of 24 hours contained a total of 77'895 queries. In general, the search terms were fairly vague and did not contain many image-related topics, so we filtered out words such as image, video, and terms relating 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 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 are in the database and once this was finished, 30 topics were selected.

All topics were categorised 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 approach groups, ten 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 part of the topic. Topic images were searched for on the Internet and were not part of the database. This made visual retrieval significantly harder as most images were taken with different collections compared to those in the database and had changes in the grey level or colour values.

Figure 1 shows a visual topic, Figure 2 a topic that should be retrieved well with a mixed approach and Figure 3 a topics with very different images in the results sets that should be well-suited for textual retrieval, only.

<sup>15</sup><http://www.pubmed.gov/>



Lung xray tuberculosis.  
Röntgenbild Lunge Tuberkulose.  
Radio pulmonal de tuberculose.

Figure 2: Example for a mixed topic.



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

Figure 3: Example for a semantic topic.

## 2.5 Relevance Judgements

Relevance judgments in ImageCLEFmed were performed by physicians and other 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.

## 2.6 Submissions and Techniques

This section quickly summarises 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. The unit for retrieval and relevance was the image and not the case but several groups submitted case IDs that we had to replace with the first image of the case. Other problems include the change of upper/lower case for the image IDs and the change of the database names that also changed the image IDs. Some groups reused the 2006 datasets that were corrected before 2007 and also ended up with invalid IDs.

### **2.6.1 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 combine the two outcomes in a linear fashion.

### **2.6.2 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.

### **2.6.3 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.

### **2.6.4 IRIT**

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

### **2.6.5 MedGIFT**

The *MedGIFT* group submitted a total of 13 runs. For visual retrieval the GIFT (GNU Image Finding Tool) was used to create a sort of 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.

### **2.6.6 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.

### **2.6.7 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.

### **2.6.8 OHSU**

The *OHSU* group submitted 10 textual and mixed runs, using Fire as a visual system. Their mixed runs had good performance as well as the best early precision.

### **2.6.9 RWTH**

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

### 2.6.10 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.

### 2.6.11 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.

### 2.6.12 UNAL

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

### 2.6.13 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. The combinations were linear and surprisingly the results are significantly worse than the results of single techniques of the participants.

## 2.7 Results

For the first time in 2007, the best overall 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.

### 2.7.1 Automatic Retrieval

As always, the vast 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, thus hiding much information about which technique performed well for what kind of tasks.

**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 themselves.

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 results tables of each category. 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 actually 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 optimised feature weighting may result in a large improvement!

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

Run	Relevant	MAP	R-prec	P10	P30	P100
RWTH-FIRE-ME-NT-tr0506	1613	0.2328	0.2701	0.4867	0.4333	0.2823
RWTH-FIRE-ME-NT-tr06	1601	0.2227	0.2630	0.4867	0.4256	0.2763
CINDLIMG_FUSION	630	0.0333	0.0532	0.1267	0.1222	0.0777
RWTH-FIRE-NT-emp	584	0.0284	0.0511	0.1067	0.0856	0.0590
RWTH-FIRE-NT-emp2	562	0.0280	0.0493	0.1067	0.0811	0.0587
miracleVisG	532	0.0186	0.0396	0.0833	0.0833	0.0470
miracleVisGFANDmm	165	0.0102	0.0255	0.0667	0.0500	0.0347
miracleVisGFANDavg	165	0.0087	0.0214	0.0467	0.0556	0.0343
UNALCO-nmi_FeatComb	644	0.0082	0.0149	0.0200	0.0144	0.0143
miracleVisGFANDmin	165	0.0081	0.0225	0.0367	0.0478	0.0333
UNALCO-nmi_RGBHisto	530	0.0080	0.0186	0.0267	0.0156	0.0153
UNALCO-svmRBF_RGBHisto	368	0.0050	0.0103	0.0133	0.0100	0.0093
UNALCO-svmRBF_Tamura	375	0.0048	0.0109	0.0067	0.0100	0.0100
GE_4-8.treceval	292	0.0041	0.0192	0.0400	0.0322	0.0203
GE-GE_GIFT8	292	0.0041	0.0194	0.0400	0.0322	0.0203
GE-GE_GIFT4	290	0.0040	0.0192	0.0400	0.0322	0.0203
DEU_CS-DEU_R2	277	0.0028	0.0052	0.0067	0.0022	0.0033
DEU_CS-DEU_R3	260	0.0018	0.0053	0.0100	0.0056	0.0057
DEU_CS-DEU_R4	238	0.0018	0.0074	0.0033	0.0056	0.0057
DEU_CS-DEU_R5	249	0.0014	0.0062	0.0000	0.0078	0.0077

**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 ImageCLEF this year for the first time. Early precision (P10) was only slightly better than the best purely visual runs and the best mixed runs had a very high early precision whereas the highest P10 was actually a purely textual system where the MAP was situated significantly lower. (Despite its name, MAP is more of a recall-oriented measure.)

**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 summarises 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.

## 2.8 Manual and Interactive retrieval

Only three runs in 2007 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 seen as a very important area to improve retrieval performance, research groups in ImageCLEF 2007 did not make use of these categories.

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

Run	Relevant	MAP	R-prec	P10	P30	P100
LIG-MRIM-LIG_MU_A	2347	0.3962	0.4146	0.5067	0.4600	0.3593
LIG-MRIM-LIG_GM_A	2341	0.3947	0.4134	0.5000	0.4678	0.3617
LIG-MRIM-LIG_GM_L	2360	0.3733	0.3904	0.5200	0.4667	0.3330
SinaiC100T100	2449	0.3668	0.3942	0.5467	0.5044	0.3457
LIG-MRIM-LIG_MU_L	2363	0.3643	0.3784	0.5033	0.4422	0.3183
miracleTxtENN	2294	0.3518	0.3890	0.5800	0.4556	0.3600
SinaiC040T100	2401	0.3507	0.3737	0.5533	0.5122	0.3490
OHSU_as_out_1000rev1_c	2306	0.3453	0.3842	0.5300	0.4433	0.3033
OHSU-oshu_as_is_1000	2304	0.3453	0.3842	0.5300	0.4433	0.3033
SinaiC030T100	2345	0.3340	0.3433	0.5100	0.4889	0.3363
ohsu_text_e4_out_rev1	1850	0.3321	0.3814	0.5867	0.4878	0.2893
UB-NLM-UBTextBL1	2244	0.3182	0.3306	0.5300	0.4756	0.3190
OHSU-OHSU_txt_exp2	1433	0.3135	0.3775	0.5867	0.4878	0.2893
IPAL-IPAL1_TXT_BAY_ISA0	1895	0.3057	0.3320	0.4767	0.4044	0.3163
IPAL-IPAL_TXT_BAY_ALLREL2	1896	0.3042	0.3330	0.4633	0.4067	0.3127
IPAL-IPAL3_TXT_BAY_ISA0	1852	0.2996	0.3212	0.4733	0.3989	0.3140
miracleTxtXN	2252	0.2990	0.3540	0.4067	0.3756	0.2943
SinaiC020T100	2028	0.2950	0.3138	0.4400	0.4389	0.2980
IPAL-IPAL4_TXT_BAY_ISA0	1831	0.2935	0.3177	0.4733	0.3978	0.3073
GE_EN	2170	0.2714	0.2989	0.3900	0.3356	0.2467
UB-NLM-UBTextBL2	2084	0.2629	0.2873	0.4033	0.3644	0.2543
GE_MIX	2123	0.2416	0.2583	0.3500	0.3133	0.2243
DEU_CS-DEU_R1	891	0.1694	0.2191	0.3967	0.3622	0.2533
GE_DE	1364	0.1631	0.1770	0.2200	0.1789	0.1333
GE_FR	1306	0.1557	0.1781	0.1933	0.2067	0.1520
UB-NLM-UBTextFR	1503	0.1184	0.1336	0.2033	0.1767	0.1320
miracleTxtDET	694	0.0991	0.0991	0.2300	0.1222	0.0837
miracleTxtDEN	724	0.0932	0.1096	0.1800	0.1356	0.0970
IRIT_RunMed1	1418	0.0660	0.0996	0.0833	0.1100	0.1023

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

Run	Relevant	MAP	R-prec	P10	P30	P100
SinaiC100T80	2433	0.3719	0.4050	0.5667	0.5122	0.3517
SinaiC100T70	2405	0.3598	0.3925	0.5500	0.4878	0.3453
ohsu_m2_rev1_c	2164	0.3461	0.3892	0.5567	0.4622	0.3287
UB-NLM-UBTL1	2237	0.3230	0.3443	0.5167	0.4911	0.3317
UB-NLM-UBTL3	2253	0.3228	0.3388	0.5367	0.4767	0.3270
RWTH-FIRE-ME-tr0506	1920	0.3044	0.3409	0.5267	0.4644	0.3410
RWTH-FIRE-ME-tr06	1916	0.3022	0.3370	0.5300	0.4611	0.3363
miracleMixGENTRIGHTmin	2002	0.2740	0.2876	0.4500	0.3822	0.2697
UB-NLM-UBmixedMulti2	2076	0.2734	0.2995	0.4167	0.3767	0.2693
RWTH-FIRE-emp2	1813	0.2537	0.3085	0.4533	0.4467	0.3017
miracleMixGENTRIGHTmax	2045	0.2502	0.2821	0.3767	0.3500	0.2900
miracleMixGENTRIGHTmm	2045	0.2486	0.2817	0.3733	0.3578	0.2890
RWTH-FIRE-emp	1809	0.2457	0.3123	0.4567	0.4467	0.3020
GE_VT1_4	2123	0.2425	0.2596	0.3533	0.3133	0.2253
GE_VT1_8	2123	0.2425	0.2596	0.3533	0.3133	0.2253
SinaiC030T50	2313	0.2371	0.2594	0.4600	0.3756	0.2700
SinaiC020T50	1973	0.2148	0.2500	0.4033	0.3422	0.2403
OHSU-ohsu_m1	652	0.2117	0.2618	0.5200	0.4578	0.2173
GE_VT10_4	1402	0.1938	0.2249	0.3600	0.3133	0.2160
GE_VT10_8	1407	0.1937	0.2247	0.3600	0.3133	0.2157
CINDI_TXT_IMAGE_LINEAR	1053	0.1659	0.2196	0.3867	0.3300	0.2270
miracleMixGFANDminENTORmm	1972	0.1427	0.1439	0.2200	0.2000	0.1793
miracleMixGFANDminENTORmax	1972	0.1419	0.1424	0.2067	0.1911	0.1770
UB-NLM-UBmixedFR	1308	0.1201	0.1607	0.2100	0.2022	0.1567
OHSU-oshu_c_e_f_q	598	0.1129	0.1307	0.2000	0.1544	0.0837
ohsu_fire_ef_wt2_rev1_c	542	0.0586	0.0914	0.2000	0.1211	0.0760
3fire-7ohsu	2222	0.0344	0.0164	0.0100	0.0078	0.0113
3gift-3fire-4ohsu	2070	0.0334	0.0235	0.0067	0.0111	0.0137
5gift-5ohsu	1627	0.0188	0.0075	0.0033	0.0044	0.0070
7gift-3ohsu	1629	0.0181	0.0060	0.0033	0.0044	0.0073
miracleMixGFANDminENTLEFTmm	165	0.0099	0.0240	0.0533	0.0544	0.0363
miracleMixGFANDminENTLEFTmax	165	0.0081	0.0225	0.0367	0.0478	0.0333

Table 5: The only three runs not using automatic retrieval.

Run	Relevant	MAP	R-prec	P10	P30	P100	media	interaction
CINDI_TXT_IMG_RF_LIN	860	0.08	0.12	0.38	0.27	0.14	mixed	feedback
CINDI_IMG_FUSION_RF	690	0.04	0.05	0.14	0.13	0.08	visual	feedback
OHSU-oshu_man2	2245	0.34	0.37	0.54	0.44	0.3	textual	manual

## 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 still perform well using visual techniques. Extensive learning of feature selection and weighting can have enormous gain in performance as shown by the FIRE runs.

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. 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.

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 close 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 too big to assemble sufficient training data to create suitable classifiers. To address this issue, a hierarchical class structure such as the IRMA code [9] 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 & 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 released 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. 3.1.1). 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 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 4.



1121-120-200-700

T: x-ray, plain radiography, analog, overview image  
 D: coronal, anteroposterior (AP, coronal), unspecified  
 A: cranium, unspecified, unspecified  
 B: musculoskeletal system, unspecified, unspecified



1121-120-310-700

T: x-ray, plain radiography, analog, overview image  
 D: coronal, anteroposterior (AP, coronal), unspecified  
 A: spine, cervical spine, unspecified  
 B: musculoskeletal system, unspecified, unspecified



1121-127-700-500

T: x-ray, plain radiography, analog, overview image  
 D: coronal, anteroposterior (AP, coronal), supine  
 A: abdomen, unspecified, unspecified  
 B: uropoietic system, unspecified, unspecified



1123-211-500-000

T: x-ray, plain radiography, analog, high beam energy  
 D: sagittal, lateral, right-left, inspiration  
 A: chest, unspecified, unspecified  
 B: unspecified, unspecified, unspecified

Figure 4: Example images from the medical annotation task with full IRMA-code and its textual representation.

### 3.1.1 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 [9]. In particular, the IRMA code is composed from four axes having three to four positions, each in  $\{0, \dots, 9, a, \dots, z\}$ , 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 extremity (arm)
410 upper extremity (arm); hand
411 upper extremity (arm); hand; finger
412 upper extremity (arm); hand; middle hand
413 upper extremity (arm); hand; carpal bones
```

420 upper extremity (arm); radio carpal joint  
430 upper extremity (arm); forearm  
431 upper extremity (arm); forearm; distal forearm  
432 upper extremity (arm); forearm; proximal forearm  
440 upper extremity (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.

### 3.1.2 Hierarchical Classification

To define a evaluation scheme for hierarchical classification, we can consider the 4 axes to be independent, such that we can consider 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 a ontology based class hierarchy [20] and in information extraction similar techniques are applied [11]. 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 very 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  $l_1^I = l_1, l_2, \dots, l_i, \dots, l_I$  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  $\hat{l}_1^I = \hat{l}_1, \hat{l}_2, \dots, \hat{l}_i, \dots, \hat{l}_I$  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 in a straight forward way 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

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 hierarchy

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

- (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 completely wrong decision (i.e. all positions wrong) gets an error count of 1.0 and a completely correctly classified image has an error of 0.0.

Table 7 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 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 shows the effect of the branching factor. In this column we assumed the branching factor of the code is 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.

## 3.2 Participating Groups & 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.

### 3.2.1 BIOMOD: University of Liege, Belgium

The Bioinformatics and Modelling group from the University Liege<sup>16</sup> in Belgium submitted four runs. The approach is based on an object recognition framework using extremely randomised trees and randomly extracted sub-windows [10]. The runs all use the same technique and differ how the code is assembled. One run predicts the full code, one run predicts each axis independently and the other two runs are combinations of the first ones.

### 3.2.2 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 [21].

<sup>16</sup><http://www.montefiore.ulg.ac.be/services/stochastic/biomod>

### 3.2.3 Geneva: medGIFT Group, Switzerland

The medGIFT group<sup>17</sup> 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 [23].

### 3.2.4 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.

### 3.2.5 MIRACLE: Madrid, Spain

The Miracle group from Madrid Spain<sup>18</sup> 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, how the prediction was done (predicting the full code, axis-wise prediction, different subsets of axes jointly), and whether the features were normalised or not.

### 3.2.6 Oregon Health State University, Portland, OR, USA

The Department of Medical Informatics and Clinical Epidemiology<sup>19</sup> 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.

### 3.2.7 RWTHi6: RWTH Aachen University, Aachen, Germany

The Human Language Technology and Pattern Recognition group<sup>20</sup> 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. The histograms have 65536 or 4096 bins [5]. The runs differ in the resolution of the images. One run is a combination of 4 normal runs, and one run does the classification axis-wise, the other runs, directly predict the full code.

### 3.2.8 IRMA: RWTH Aachen University, Medical Informatics, Aachen, Germany

The IRMA group from the RWTH Aachen University Hospital<sup>21</sup>, 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.

### 3.2.9 UFR: University of Freiburg, Computer Science Dep., Freiburg, Germany

The Pattern Recognition and Image Processing group from the University Freiburg<sup>22</sup>, Germany, submitted four runs using relational features calculated around interest points which are later combined to form cluster cooccurrence matrices [17]. Three different classification methods were used: a flat classification scheme using all of the 116 classes, an axiswise-flat classification scheme

---

<sup>17</sup><http://www.sim.hcuge.ch/medgift/>

<sup>18</sup><http://www.mat.upm.es/miracle/introduction.html>

<sup>19</sup><http://www.ohsu.edu/dmice/>

<sup>20</sup><http://www-i6.informatik.rwth-aachen.de>

<sup>21</sup><http://www.irma-project.org>

<sup>22</sup><http://lmb.informatik.uni-freiburg.de/>

(i.e. 4 multi-class classifiers), and a binary classification tree (BCT) based scheme. The BCT based approach is much faster to train and classify, but this comes at a slight performance penalty. The tree was generated as described in [16].

### 3.2.10 UNIBAS: University of Basel, Switzerland

The Databases and Information Systems group from the University Basel<sup>23</sup>, 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 [19]. The runs differ in the features (pixels, Sobel features) that were used to determine the deformation and in the k-parameter for the k-nearest neighbour.

## 3.3 Results

The results of the evaluation are given in Table 7. For each run, the run-id, the score as described above and additionally, the error rate, which was used in the last years to evaluate the submissions to this task 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. 4). 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).

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 wrt. to the applied hierarchical measure and the ranking wrt. 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 wrt. to 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.

---

<sup>23</sup><http://dbis.cs.unibas.ch/>

Table 7: Results of the medical image annotation task. Score is the hierarchical evaluation score, and ER is the error rate in % that was used last year to evaluate the annotation results.

rank	run id	score	ER
1	BLOOM-BLOOM_MCK_oa	26.8	10.3
2	BLOOM-BLOOM_MCK_oo	27.5	11.0
3	BLOOM-BLOOM_SIFT_oo	28.7	11.6
4	BLOOM-BLOOM_SIFT_oa	29.5	11.5
5	BLOOM-BLOOM_DAS	29.9	11.1
6	RWTHi6-4RUN-MV3	30.9	13.2
7	UFR-UFR_cooc_flat	31.4	12.1
8	RWTHi6-SH65536-SC025-ME	33.0	11.9
9	UFR-UFR_cooc_flat2	33.2	13.1
10	RWTHi6-SH65536-SC05-ME	33.2	12.3
11	RWTHi6-SH4096-SC025-ME	34.6	12.7
12	RWTHi6-SH4096-SC05-ME	34.7	12.4
13	RWTHi6-SH4096-SC025-AXISWISE	44.6	17.8
14	UFR-UFR_cooc_codewise	45.5	17.9
15	UFR-UFR_cooc_tree2	47.9	16.9
16	UFR-UFR_cooc_tree	48.4	16.8
17	rwth_mi_k1_tn9.187879e-05_common.run	51.3	20.0
18	rwth_mi_k5_majority.run	52.5	18.0
19	UNIBAS-DBIS-IDM_HMM_W3_H3_C	58.1	22.4
20	UNIBAS-DBIS-IDM_HMM2_4812_K3	59.8	20.2
21	UNIBAS-DBIS-IDM_HMM2_4812_K3_C	60.7	23.2
22	UNIBAS-DBIS-IDM_HMM2_4812_K5_C	61.4	23.1
23	UNIBAS-DBIS-IDM_HMM2_369_K3_C	62.8	22.5
24	UNIBAS-DBIS-IDM_HMM2_369_K3	63.4	21.5
25	UNIBAS-DBIS-IDM_HMM2_369_K5_C	65.1	22.9
26	OHSU-OHSU_2	67.8	22.7
27	OHSU-gist_pca	68.0	22.7
28	BLOOM-BLOOM_PIXEL_oa	68.2	20.1
29	BLOOM-BLOOM_PIXEL_oo	72.4	20.8
30	BIOMOD-full	73.8	22.9
31	BIOMOD-correction	75.8	25.3
32	BIOMOD-safe	78.7	36.0
33	im.cyu.tw-cyu_w1i6t8	79.3	25.3
34	rwth_mi_k5_common.run	80.5	45.9
35	BIOMOD-independant	95.3	32.9
36	miracle-miracleAAn	158.8	50.3
37	miracle-miracleVAn	159.5	49.6
38	miracle-miracleAATDABn	160.2	49.9
39	miracle-miracleAATABDn	162.2	50.1
40-62	runs from miracle group	-	-
63	GE-GE_GIFT10_0.5ve	375.7	99.7
64	GE-GE_GIFT10_0.15vs	390.3	99.3
65	GE-GE_GIFT10_0.66vd	391.0	99.0
66	miracle-miracleVATDAB	419.7	84.4
67	miracle-miracleVn	490.7	82.6
68	miracle-miracleV	505.6	86.8

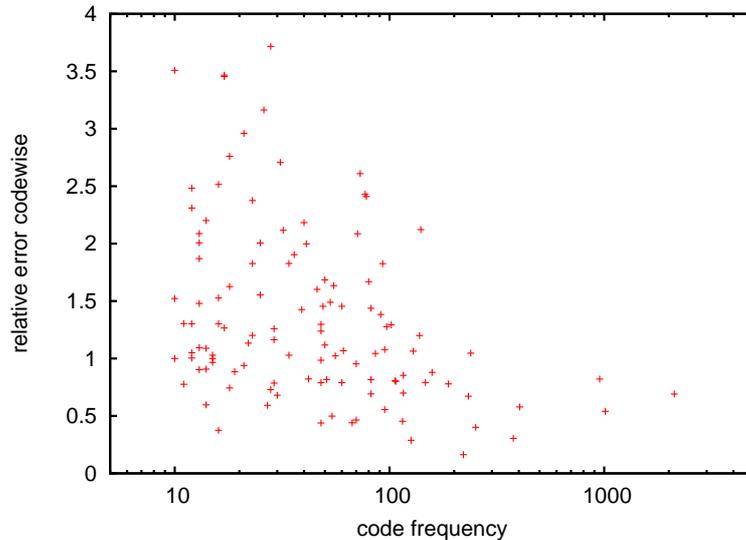


Figure 5: Code-wise relative error as a function of the frequency of this code in the training data.

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 is 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 fairly 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. 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 unique code as a unique class which gives the impression that for the current number of 116 unique codes the training data is sufficient to train a joint classifier. As opposed to the retrieval task, none of the groups used any interaction although this might allow for a big performance gain.

## Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. We would like to thank all the 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] Chris. S. Candler, Sebastian. H. Uijtdehaage, and Sharon. E. Dennis. Introducing HEAL: The health education assets library. *Academic Medicine*, 78(3):249–253, 2003.
- [2] Paul Clough, Henning Müller, and Mark Sanderson. The CLEF 2004 cross language image retrieval track. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, pages 597–613. Lecture Notes in Computer Science (LNCS), Springer, Volume 3491, 2005.
- [3] Paul Clough, Henning Müller, and Mark Sanderson. Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul D. Clough, Gareth J. F. Jones, Julio Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, Lecture Notes in Computer Science, Bath, England, 2005. Springer-Verlag.
- [4] Thomas Deselaers, Allan Hanbury, and et al. Overview of the ImageCLEF 2007 object retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
- [5] Thomas Deselaers, Andre Hegerath, Daniel Keysers, and Hermann Ney. Sparse patch-histograms for object classification in cluttered images. In *DAGM 2006, Pattern Recognition, 26th DAGM Symposium*, volume 4174 of *Lecture Notes in Computer Science*, pages 202–211, Berlin, Germany, September 2006.
- [6] K. Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Webbasierte Lernwerkzeuge für die Pathologie – web-based learning tools for pathology. *Pathologie*, 24:394–399, 2003.
- [7] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
- [8] William Hersh, Henning Müller, Jeffery Jensen, Jianji Yang, Paul Gorman, and Patrick Ruch. Imageclefmed: A text collection to advance biomedical image retrieval. *Journal of the American Medical Informatics Association*, September/October, 2006.
- [9] Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohlen, and Bertold B. Wein. The IRMA code for unique classification of medical images. In *SPIE 2003*, volume 5033, pages 440–451, 2003.
- [10] Raphaël Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 34–40. IEEE, June 2005.
- [11] Diana Maynard, Wim Peters, and Yaoyong Li. Metrics for evaluation of ontology-based information extraction. In *Evaluation of Ontologies for the Web (EON 2006)*, Edinburgh, UK, 2006.

- [12] Henning Müller, Thomas Deselaers, Thomas M. Lehmann, Paul Clough, and William Hersh. Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.
- [13] 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*, 28:295–305, 2004.
- [14] Antoine Rosset, Henning Müller, Martina Martins, Natalia Dfouni, Jean-Paul Vallée, and Osman Ratib. Casimage project — a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.
- [15] 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.
- [16] Lokesh Setia and Hans Burkhardt. Learning taxonomies in large image databases. In *ACM SIGIR Workshop on Multimedia Information Retrieval*, Amsterdam, Holland, 2007.
- [17] Lokesh Setia, Alexandra Teynor, Alaa Halawani, and Hans Burkhardt. 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.
- [18] Alan F. Smeaton, Paul Over, and W. Kraaij. TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In *Proceedings of the international ACM conference on Multimedia 2004 (ACM MM 2004)*, pages 652–655, New York City, NY, USA, October 2004.
- [19] Michael Springmann, Andreas Dander, and Heiko Schuldt. T.b.a. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
- [20] Aixin Sun and Ee-Peng Lim. Hierarchical text classification and evaluation. In *IEEE International Conference on Data Mining (ICDM 2001)*, pages 521–528, San Jose, CA, USA, November 2001.
- [21] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. CLEF2007 Image Annotation Task: an SVM–based Cue Integration Approach. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
- [22] Jerold. W. Wallis, Michelle. M. Miller, Tom. R. Miller, and Thomas. H. Vreeland. An internet–based nuclear medicine teaching file. *Journal of Nuclear Medicine*, 36(8):1520–1527, 1995.
- [23] Xin Zhou, Julien Gobeill, Patrick Ruch, and Henning Müller. University and Hospitals of Geneva at ImageCLEF 2007. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.