The CLEF 2005 Cross–Language Image Retrieval Track

Paul Clough[1], Henning Müller[2], Thomas Deselaers[3], Michael Grubinger[4], Thomas M. Lehmann[5], Jeffery Jensen[6], and William Hersh[6]

[1] Department of Information Studies, Sheffield University, Sheffield, UK
p.d.clough@sheffield.ac.uk
[2] Medical Informatics Service, Geneva University and Hospitals, Geneva Switzerland
henning.mueller@sim.hcuge.ch
[3] Lehrstuhl für Informatik VI, RWTH Aachen, Germany
deselaers@cs.rwth-aachen.de
[4] School of Computer Science and Mathematics, Victoria University, Australia
michael.grubinger@research.vu.edu.au
[5] Department of Medical Informatics, Medical Faculty, RWTH Aachen, Germany
lehmann@computer.org
[6] Biomedical Informatics, Oregon Health and Science University, Portland, OR, USA
hersh@ohsu.edu, jensejef@ohsu.edu

Abstract. This paper outlines efforts from the 2005 CLEF cross–language image retrieval campaign (ImageCLEF). Aim of the CLEF track is to explore the use of both text and content–based retrieval methods for cross–language image retrieval. Four tasks were offered in ImageCLEF: ad–hoc retrieval from an historic photographic collection, ad–hoc retrieval from a medical collection, an automatic image annotation task, and a user–centered (interactive) evaluation task. 24 research groups from a variety of backgrounds and nationalities (14 countries) participated in ImageCLEF. This paper presents the ImageCLEF tasks, submissions from participating groups and a summary of the main findings.

1 Introduction

ImageCLEF[7] conducts evaluation of cross–language image retrieval and is run as part of the Cross Language Evaluation Forum (CLEF) campaign. The ImageCLEF retrieval benchmark was previously run in 2003 [1] and 2004 [2] with the aim of evaluating image retrieval from multilingual document collections. Images by their very nature are language independent, but often they are accompanied by texts semantically related to the image (e.g. textual captions or metadata). Images can then be retrieved using primitive features based on pixels which form the contents of an image (e.g. using a visual exemplar), abstracted features expressed through text, or a combination of both. The language used to express the associated texts or textual queries should not affect retrieval, i.e.

[7] See http://ir.shef.ac.uk/imageclef/

an image with a caption written in English should be searchable in languages other than English.

ImageCLEF 2005 provided tasks for system–centered evaluation of retrieval systems in two domains: historic photographs and medical images. These domains offer realistic (and different) scenarios in which to test the performance of image retrieval systems and offer different challenges and problems to participants. A user–centered search task was also run using the same historic photographs, and is further described in the interactive CLEF (iCLEF) overview [3]. A major goal of ImageCLEF is to investigate the effectiveness of combining text and image for retrieval and promote the exchange of ideas which may help improve the performance of future image retrieval systems.

ImageCLEF has already seen participation from both academic and commercial research groups worldwide from communities including: Cross–Language Information Retrieval (CLIR), Content–Based Image Retrieval (CBIR), medical information retrieval and user interaction. We provide participants with the following: image collections, representative search requests (expressed by both image and text) and relevance judgements indicating which images are relevant to each search request. Campaigns such as CLEF and TREC have proven invaluable in providing standardised resources for comparative evaluation for a range of retrieval tasks and ImageCLEF aims to provide the research community with similar resources for image retrieval. In the following sections of this paper we describe separately each search task: Section 2 describes ad–hoc retrieval from historic photographs, Section 3 ad–hoc retrieval from medical images, and Section 4 the automatic annotation of medical images. For each we briefly describe the test collections, the search tasks, participating research groups, results and a summary of the main findings.

2 Ad–hoc Retrieval from Historic Photographs

Similar to previous years (see, e.g. [2]), the goal of this task is: given multilingual text queries, retrieve as many relevant images as possible from the provided image collection (the St. Andrews collection of historic photographs[8]). Queries for images based on abstract concepts rather than visual features are predominant in this task, thereby limiting the success of using visual retrieval methods alone. Either these concepts cannot be extracted using visual features and require extra external semantic knowledge (e.g. the name of the photographer), or images with different visual properties may be relevant to a search request (e.g. different views of a city). However based on feedback from participants in 2004, search tasks for 2005 were chosen to reflect more visually–based queries.

2.1 Data and Search Tasks

The St. Andrews collection consists of 28,133 images, all of which have associated structured captions written in British English (the target language). The cap-

[8] `http://specialcollections.st-and.ac.uk/photcol.htm`

Short title: Rev William Swan.
Long title: Rev William Swan.
Location: Fife, Scotland
Description: Seated, 3/ 4 face studio portrait of a man.
Date: ca.1850
Photographer: Thomas Rodger
Categories: `[ ministers ][ identified male ][ dress - clerical ]`
Notes: ALB6-85-2 jf/ pcBIOG: Rev William Swan ( ) ADD: Former owners of album: A Govan then J J? Lowson. Individuals and other subjects indicative of St Andrews provenance. By T. R. as identified by Karen A. Johnstone " Thomas Rodger 1832-1883. A biography and catalogue of selected works".

Fig. 1. An example image and caption from the St. Andrews collection.

tions consist of 8 fields (shown in Figure 1), and further examples can be found in [4] and the St. Andrews University Library[9]. Participants were given 28 topics, the main themes based on the analysis of log files from a web server at St. Andrews university, knowledge of the collection and discussions with maintainers of the image collection. After identifying main themes, queries were modified to test various aspects of cross-language and visual search. A custom–built IR system was used to identify suitable topics (in particular those topics with an estimated 20 and above relevant images). A complexity score was developed by the authors to categorise topics with respect to linguistic complexity [5].

Each topic consisted a title (a short sentence or phrase describing the search request in a few words), and a narrative (a description of what constitutes a relevant or non–relevant image for that search request). Two example images per topic were also provided, the envisaged uses being to test relevance feedback (both manual and automatic) and query–by–example searches[10]. Both topic title and narratives were translated into the following languages: German, French, Italian, Spanish (European), Spanish (Latin American), Chinese (Simplified), Chinese (Traditional) and Japanese. Translations of title only were also generated for 25 languages including: Russian, Croatian, Bulgarian, Hebrew and Norwegian. All translations were provided by native speakers and verified by at least one other native speaker.

2.2 Relevance Assessments

Relevance assessments were performed by staff at the University of Sheffield in a manner similar to previous years (see [1, 2]). The top 50 results from all submitted runs were used to create image pools giving an average of 1,376 (max: 2,193 and min: 760) images to judge per topic. The authors judged all topics to create a "gold standard" and at least two further assessments were obtained for each topic. Assessors used a custom–built tool to make judgements accessible on–line enabling them to log in when and where convenient. Assessors were asked to judge every image in the topic pool, but also to use interactive search and

[9] `http://www-library.st-andrews.ac.uk/`
[10] See `http://ir.shef.ac.uk/imageclef2005/adhoc.htm` for an example

judge: searching the collection using their own queries to supplement the image pools with further relevant images.

Assessments were based on a ternary classification scheme: (1) relevant, (2) partially relevant and (3) not relevant. Based on these judgements, various combinations were used to create the set of relevant images (qrels). As in previous years we used the `pisec-total` set: those images judged as relevant or partially–relevant by the topic creator and at least one other assessor.

2.3 Participating Groups

In total, 19 groups registered for this task and 11 submitted results (including 5 new groups compared to last year) giving a total of 349 runs (all of which were evaluated). Participants were given queries and relevance judgements from 2004 as training data and access to a CBIR system (GIFT/Viper). Descriptions of individual techniques used can be found in descriptions by the participants:

- CEA from France [6]
- National Institute of Informatics (NII) from Japan [7]
- University of Alicante (Computer Science) from Spain [8]
- Chinese University of Hong Kong (CUHK) [9]
- Dublin City University (DCU - Computer Science) from Ireland [10]
- University Hospitals Geneva from Switzerland [11]
- University of Indonesia (Computer Science) [12]
- Daedalus and Madrid University from Spain (Miracle) [13]
- National Taiwan University (NTU) from Taiwan [14]
- University of Jaén (Intelligent Systems) from Spain [15]
- UNED from Spain [16]

In summary, five groups experimented with combining both text and visual runs [6, 9, 10, 12, 14]. Groups experimented with merging visual and textual runs [10, 12, 14], and using visual runs to reorder the text runs [6, 9]. Purely visual runs were submitted by University Hospitals Geneva [11] and NTU [14] and provide a visual baseline against which to compare mixed approaches.

Most groups made use of relevance feedback (in the form of pseudo relevance feedback) to perform query expansion and improve subsequent runs. Of particular interest are: NII who used a learned word association model to improve a language model [7], Alicante who used an ontology created automatically created from the St. Andrews collection to relate a query with several image categories [8] and UNED who experimented with creating structured queries based on identifying named entities in the caption fields [16].

Some groups focused on dealing with specific languages (e.g. Chinese [14], Japanese [7], Spanish [16] and Indonesian [12]); others used generic tools (e.g. freely available MT systems) to tackle larger numbers of languages [8, 13]. A voting-based strategy was developed joining three different systems of participating universities: University of Alicante, University of Jaén and UNED [8].

Participants were asked to categorise their submissions by the following dimensions: query language, type (automatic or manual), use of feedback (typically

relevance feedback is used for automatic query expansion), modality (text only, image only or combined) and the initial query (visual only, title only, narrative only or a combination). A summary of submissions by these dimensions is shown in Table 3. No manual runs were submitted, and a large proportion of text runs used only information from the titles. Table 1 provides a summary of submissions by query language. At least one group submitted for each language [13], the most popular (non-English)being French, German and Spanish (European).

Table 1. Ad–hoc experiments listed by query language

Query Language	#Runs	#Participants
English	69	9
Spanish (Latinamerican)	36	4
German	29	5
Spanish (European)	28	6
Chinese (simplified)	21	4
Italian	19	4
French	17	5
Japanese	16	4
Dutch	15	4
Russian	15	4
Portuguese	12	3
Greek	9	3
Indonesian	9	1
Chinese (traditional)	8	2
Swedish	7	2
Filipino	5	1
Norwegian	5	1
Polish	5	1
Romanian	5	1
Turkish	5	1
Visual	4	2
Bulgarian	2	1
Croatian	2	1
Czech	2	1
Finnish	2	1
Hungarian	2	1

2.4 Results

Results for submitted runs were computed using the latest version of trec_eval[11] from NIST (v7.3). Submissions were evaluated using uninterpolated Mean Average Precision (MAP), Precision at rank 10 (P10), and the number of relevant images retrieved (RelRetr) from which we compute recall (the proportion of relevant retrieved). Table 2 summarises the top performing systems in the ad–hoc task by language based on MAP. The highest English (monolingual) retrieval score is 0.4135, with a P10 of 0.5500 and recall of 0.8434. The relatively high recall score, but low MAP and P10 scores indicate that relevant images are being retrieved at lower rank positions. The highest monolingual score is obtained using combined visual and text retrieval and relevance feedback (see [9]).

[11] http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz

Table 2. Systems with highest MAP for each language in the ad–hoc retrieval task.

Language	MAP	Recall	Group	Run ID	Init. Query	Feedback	Modality
English	0.4135	0.5500	CUHK	CUHK-ad-eng-tv-kl-jm2	title+img	with	text+img
Chinese (Trad)	0.3993	0.7526	NTU	NTU-CE-TN-WEprf-Ponly	title+narr	with	text+img
Spanish (Lat)	0.3447	0.7891	Alicante, Jaen	R2D2vot2SpL	title	with	text
Dutch	0.3435	0.4821	Alicante, Jaen	R2D2vot2Du	title	with	text
German	0.3375	0.4929	Alicante, Jaen	R2D2vot2Ge	title	with	text
Spanish (Euro)	0.3175	0.8048	UNED	unedESENent	title	with	text
Portuguese	0.3073	0.7542	Miracle	imirt0attrpt	title	without	text
Greek	0.3024	0.6383	DCU	DCUFbTGR	title	with	text
French	0.2864	0.7322	Jaen	SinaiFrTitleNarrFBSystran	title+narr	with	text
Japanese	0.2811	0.7333	Alicante	AlCimg05Exp3Jp	title	with	text
Russian	0.2798	0.6879	DCU	DCUFbTRU	title	with	text
Italian	0.2468	0.6227	Miracle	imirt0attrit	title	without	text
Chinese (Sim)	0.2305	0.6153	Alicante	AlCimg05Exp3ChS	title	with	text
Indonesian	0.2290	0.6566	Indonesia	UI-T-IMG	title	without	text+img
Turkish	0.2225	0.6320	Miracle	imirt0allftk	title	without	text
Swedish	0.2074	0.5647	Jaen	SinaiSweTitleNarrFBWordlingo	title	without	text
Norwegian	0.1610	0.4530	Miracle	imirt0attrno	title	without	text
Filipino	0.1486	0.3695	Miracle	imirt0allffl	title	without	text
Polish	0.1558	0.5073	Miracle	imirt0attrpo	title	without	text
Romanian	0.1429	0.3747	Miracle	imirt0attrro	title	without	text
Bulgarian	0.1293	0.5694	Miracle	imirt0allfbu	title	without	text
Czech	0.1219	0.5310	Miracle	imirt0allfcz	title	without	text
Croatian	0.1187	0.4362	Miracle	imirt0attrcr	title	without	text
Finnish	0.1114	0.3257	Miracle	imirt0attrfi	title	without	text
Hungarian	0.0968	0.3789	Miracle	imirt0allfhu	title	without	text
Visual	0.0829	0.2834	Geneva	GE_A_88	visual	without	img

The highest cross–language MAP is Chinese (traditional) for the NTU submission which is 97% of highest monolingual score. Retrieval performance is variable across language with some performing poorly, e.g. Romanian, Bulgarian, Czech, Croatian, Finnish and Hungarian. Although these languages did not have translated narratives available for retrieval, it is more likely low performance results from limited availability of translation and language processing resources and difficult language structure (e.g. results from CLEF 2004 showed Finnish to be a very challenging language due to its complex morphology). Hungarian performs the worst at 23% of monolingual, however it is encouraging to see participation in CLEF for these languages. On average, MAP for English is 0.2840 (P10=0.3933 and Recall=0.6454) and across all languages MAP is 0.2027 (P10=0.2985 and Recall=0.5737) – see Table 3. Using the Mann-Whitney U test for two-independent samples, this difference is significant (at $p < 0.05$).

Table 3 shows the average MAP score averaged across all submissions by query dimension. We also include standard deviation (SD), median and highest MAP scores because the arithmetic mean is distorted by outliers in the data distribution. There is also a wide variation in counts for each dimension, therefore results are only an indication of effects on performance for each dimension.

From Table 3, it would appear that runs using some kind of feedback (e.g. query expansion) perform approximately 14.8% better than those without. From Figure 3 this appears true for individual topics also and mean differences are significant at $p < 0.05$. Also from Table 3 it appears that combined text and visual runs perform on average 31.5% better than text runs alone (based on average MAP). However, low retrieval scores due to translation draw the text–only results down. If we compare text–only scores for the 5 groups who submitted text and visual runs, the MAP score is 0.2723, approximately 12.1% lower than

Table 3. MAP results for each query dimension.

Dimension	type	#Runs	#Groups	Mean Average Precision (MAP) Mean (SD)	Median	Highest
Language	English	69	9	0.2840 (0.1441)	0.3574	0.4135
	non-English	277	10	0.2027 (0.0784)	0.2143	0.3993
Feedback	yes	142	9	0.2399 (0.1119)	0.2482	0.4135
	no	207	10	0.2043 (0.0887)	0.2069	0.4030
Modality	image	3	2	0.0749 (0.0130)	0.0819	0.0829
	text	318	11	0.2121 (0.0976)	0.2170	0.4115
	text+image	28	5	0.3098 (0.0782)	0.3023	0.4135
Initial Query	image only	4	3	0.1418 (0.1342)	0.0824	0.3425
	title only	274	11	0.2140 (0.0975)	0.2246	0.4115
	narr only	6	2	0.1313 (0.0555)	0.1298	0.1981
	title+narr	57	6	0.2314 (0.0929)	0.2024	0.4083
	title+image	4	1	0.4016 (0.0126)	0.4024	0.4135
	title+narr+image	4	1	0.3953 (0.0153)	0.3953	0.4118

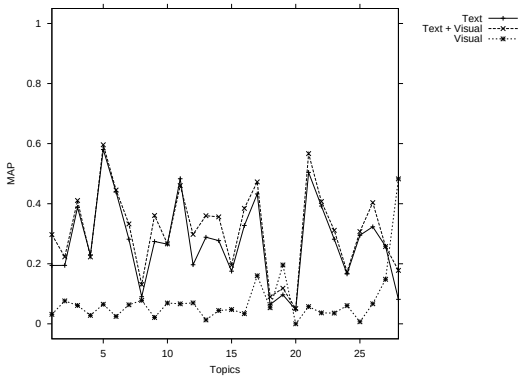


Fig. 2. Comparison between average MAP for visual and text runs from 5 groups using text and visual methods.

the combined runs. This difference is significant at $p < 0.05$ using the Mann-Whitney U test. As expected, visual–only runs perform poorly for this task.

2.5 Discussion

The variety of submissions in the ad–hoc task this year has been pleasing with six groups experimenting with both visual and text-based retrieval methods and five groups combining the two (although the number of runs submitted as combined is lower than 2004). As in 2004, a combination of text and visual approaches appears to give highest retrieval effectiveness (based on MAP) indicating this is still an area for research.

Considering individual topics, Figure 2 shows improvements for 19 topics based on comparing text–only and text+visual results for the 5 groups who submitted combined runs. In particular we observe clear improvements for topics such as “aircraft on the ground” and “portrait views of mixed sex groups” where

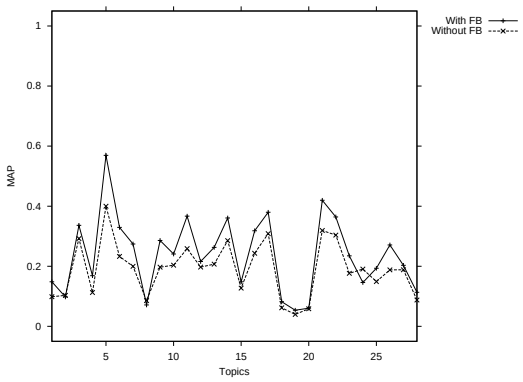


Fig. 3. Comparison between average MAP for runs with/without feedback (FB).

a combination of using visual features and semantic knowledge gained from the associated text caption improves over using a single approach. In addition, certain topics do seem better suited to a visual–only approach including topics 28 ("colour pictures of woodland scenes around St. Andrews") and 19 ("composite postcards of Northern Ireland") which obtain the highest MAP results. This begins to indicate the kinds of topics that are likely to perform well and for which visual cues are likely effective for retrieval (i.e. the set of relevant images are themselves visually similar).

Figure 2 also show that results vary widely across topic, and as expected some are much "harder" than others. For example, topics 8 ("building covered in snow"), 18 ("woman in white dress") and 20 ("royal visits to Scotland (not Fife)") are consistently the lowest scoring topics (based on average and highest MAP scores). The "easiest" topics appear to be topics 5 ("animal statue") and 21 ("monument to poet Robert Burns"). This requires further investigation and we have started analysis based on a measure of topic difficulty [5].

We wanted to offer a wider range of languages in 2005, of which 13 of these obtained runs from at least two groups (compared to 10 in 2004). It would seem that the focus for many groups in 2005 has been translation (and query expansion) with more use made of both title and narrative than 2004. However, it is interesting to see languages such as Chinese (traditional) and Spanish (Latin American) perform above European languages such as French, German and Spanish (European) which performed best in 2004.

Although topics were designed to be more suited to visual retrieval methods (based on comments from participants in 2004), the topics are still dominated by semantics and background knowledge; pure visual similarity still plays a less significant role. The current ad-hoc task is not well-suited to purely visual retrieval because colour information, which typically plays an important role in CBIR, is ineffective due to the nature of the St. Andrews collection (historic photographs). Also unlike typical CBIR benchmarks, the images in the St. Andrews collection are very complex containing both objects in the foreground

and background which prove indistinguishable to CBIR methods. Finally, the relevant image set is visually different for some queries (e.g. different views of a city) making visual retrieval methods ineffective. This highlights the importance of using either text-based IR methods on associated metadata alone, or combined with visual features. Relevance feedback (in the form of automatic query expansion) still plays an important role in retrieval as also demonstrated by submissions in 2004: a 17% increase in 2005 and 48% in 2004 (see Figure 3).

We are aware that research in the ad-hoc task using the St. Andrews collection has probably reached a plateau. There are obvious limitations with the existing collection: mainly black and white images, domain-specific vocabulary used in associated captions, restricted retrieval scenario (i.e. searches for historic photographs) and experiments with limited target language (English) are only possible (i.e. cannot test further bilingual pairs). To address these and widen the image collections available to ImageCLEF participants, we have been provided with access to a new collection of images from a personal photographic collection with associated textual descriptions in German and Spanish (as well as English). This is planned for use in the ImageCLEF 2006 ad-hoc task.

3 Ad–hoc Retrieval from Medical Image Collections

Domain–specific information retrieval is increasingly important, and this holds especially true for the medical field, where patients, clinicians, and researchers have their particular information needs [17]. Whereas information needs and retrieval methods for textual documents have been well researched, there has been little investigation of information needs and search system use for images and other multimedia data [18], even less so in the medical domain. ImageCLEFmed is creating resources to evaluate information retrieval tasks on medical image collections. This process includes the creation of image collections, query tasks, and the definition of correct retrieval results for these tasks for system evaluation. Some of the tasks have been based on surveys of medical professionals and how they use images [19].

Much of the basic structure is similar to the non–medical ad–hoc task, such as the general outline, the evaluation procedure and the relevance assessment tool used. These similarities will not be described in detail in this section.

3.1 Data Sets Used and Query Topics

In 2004, only the Casimage[12] dataset was made available to participants [20], containing almost 9.000 images of 2.000 cases, 26 query topics, and relevance judgements by three medical experts [21]. Casimage is also part of the 2005 collection. 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. For 2005, we were also given permission to use the PEIR[13] (Pathology Education Instructional

[12] http://www.casimage.com/

[13] http://peir.path.uab.edu/

Resource) database using annotation based on the HEAL[14] project (Health Education Assets Library, mainly Pathology images [22]). 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[15] [23], 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[16] collection (Pathology images [24]) was included into our dataset. It contains 9.000 images with extensive annotation on a per image basis in German. Part of the German annotation is translated into English. As such, we were able to use a total of more than 50.000 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 image topics were based on a small survey administered to clinicians, researchers, educators, students, and librarians at Oregon Health & Science University (OHSU)[19]. Based on this survey, topics for ImageCLEFmed were developed along the following axes:

- Anatomic region shown in the image;
- Image modality (x–ray, CT, MRI, gross pathology, ...);
- Pathology or disease shown in the image;
- abnormal visual observation (eg. enlarged heart).

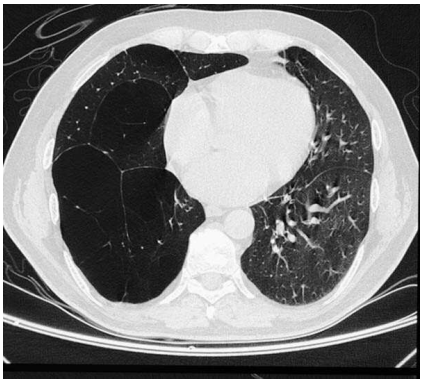
As the goal was to accommodate both visual and textual research groups, we developed a set of 25 topics containing three different groups of topics: those expected to work most effectively with a visual retrieval system (topics 1–12), those where both text and visual features were expected to perform well (topics 13–23), and semantic topics, where visual features were not expected to improve results (topics 24–25). All query topics were of a higher semantic level than the 2004 ImageCLEF medical topics because the 2005 automatic annotation task provided a testbed for purely visual retrieval/classification. All 25 topics contained one to three images, with one having an image as negative feedback. The topic text was provided with the images in the three languages present in the collections: English, German, and French. An example for a visual query of the first category can be seen in Figure 4.

A query topic requiring more than purely visual features is shown in Figure 5.

3.2 Relevance Judgements

The relevance assessments were performed by graduate students who were also physicians in the OHSU biomedical informatics program. A simple interface was used from previous ImageCLEF relevance assessments. Nine judges, all medical doctors except for one image processing specialist with medical knowledge,

[14] http://www.healcentral.com/
[15] http://gamma.wustl.edu/home.html
[16] http://alf3.urz.unibas.ch/pathopic/intro.htm



Show me chest CT images with emphysema.
Zeige mir Lungen CTs mit einem Emphysem.
Montre–moi des CTs pulmonaires avec un emphysème.

Fig. 4. An example of a query that is at least partly solvable visually, using the image and the text as query. Still, use of annotation can augment retrieval quality. The query text is presented in three languages.

Show me all x–ray images showing fractures.
Zeige mir Röntgenbilder mit Brüchen.
Montres–moi des radiographies avec des fractures.

Fig. 5. A query requiring more than visual retrieval but visual features can deliver hints to good results.

performed the relevance judgements. Half of the images for most of topics were judged in duplicate.

To create the pools for the judgements, the first 40 images of each submitted run were used to create pools with an average size of 892 images. The largest pool size was 1.167 and the smallest one 470. It took the judges an average of about three hours to judge the images for a single topic. Compared to the purely visual topics from 2004 (around one hour of judgement per topic containing an average of 950 images), the judgement process took much longer per image. This was most likely due to the semantic topics requiring the judges to verify the text and/or an enlarged version of the images. The longer time might also be due to the fact that in 2004, all images were pre–marked as irrelevant, and only relevant images required a change, whereas this year we did not have anything pre–marked. Still, this process was generally faster than most text research judgements, and a large number of irrelevant images could be sorted out quickly.

We use a ternary judgement scheme including relevant, partially–relevant, and non–relevant. For the official qrels, we only used images judged as relevant (and not those judged partially relevant). For the topics judged by two persons, we only used the first judgements for the official relevance. (Later we plan to analyse the duplicate judgements and their effect on the results of runs.)

3.3 Participants

The medical retrieval task had 12 participants in 2004 when it was purely visual task and 13 in 2005 as a mixture of visual and non-visual retrieval. Only 13 of the 28 registered groups ended up submitting results, which was likely due to the short time span between delivery of the images and the deadline for results submission. Another reason was that several groups registered very late, as they did not have information about ImageCLEF beforehand, but were still interested in the datasets also for future participations. As the registration to the task was free, they could simply register to get this access.

The following groups registered but were finally not able to submit results for a variety of reasons:

- University of Alicante, Spain
- National Library of Medicine, Bethesda, MD, USA
- University of Montreal, Canada
- University of Science and Medical Informatics, Innsbruck, Austria
- University of Amsterdam, Informatics department, The Netherlands
- UNED, LSI, Valencia, Spain
- Central University, Caracas, Venezuela
- Temple University, Computer science, USA
- Imperial College, Computing lab, UK
- Dublin City University, Computer science, Ireland
- CLIPS Grenoble, France
- University of Sheffield, UK

– Chinese University of Hong Kong, China

In the end, 13 groups (two from the same laboratory but different groups in Singapore) submitted results for the medical retrieval task, including a total of 134 runs. Only 6 manual runs were submitted. Here is a list of their participation including a description of submitted runs:

National Chiao Tung University, Taiwan: submitted 16 runs in total, all automatic. 6 runs were visual only and 10 mixed runs. They use simple visual features (color histogram, coherence matrix, layout features) as well as text retrieval using a vector–space model with word expansion using Wordnet.

State University of New York (SUNY), Buffalo, USA: submitted a total of 6 runs, one visual and five mixed runs. GIFT was used as visual retrieval system and SMART as textual retrieval system, while mapping the text to UMLS.

University and Hospitals of Geneva, Switzerland: submitted a total of 19 runs, all automatic runs. This includes two textual and two visual runs plus 15 mixed runs. The retrieval relied mainly on the GIFT and easyIR retrieval systems.

RWTH Aachen, Computer science, Germany: submitted 10 runs, two being manual mixed retrieval, two automatic textual retrieval, three automatic visual retrieval and three automatic mixed retrieval. Fire was used with varied visual features and a text search engine using English and mixed–language retrieval.

Daedalus and Madrid University, Spain: submitted 14 runs, all automatic. 4 runs were visual only and 10 were mixed runs; They mainly used semantic word expansions with EuroWordNet.

Oregon Health and Science University (OHSU), Portland, OR, USA: submitted three runs in total, two of which were manual. One of the manual runs combined the output from a visual run using the GIFT engine. For text retrieval, the Lucene system was used.

University of Jaen, Spain: had a total of 42 runs, all automatic. 6 runs were textual, only, and 36 were mixed. GIFT is used as a visual query system and the LEMUR system is used for text in a variety of configurations to achieve multilingual retrieval.

Institute for Infocomm research, Singapore: submitted 7 runs, all of them automatic visual runs; For their runs they first manually selected visually similar images to train the features. These runs should probably have been classified as a manual runs. Then, they use a two–step approach for visual retrieval.

Institute for Infocomm research – second group , Singapore: submitted a total of 3 runs, all visual with one being automatic and two manual runs The main technique applied is the connection of medical terms and concepts to visual appearances.

RWTH Aachen – medical informatics, Germany: submitted two visual only runs with several visual features and classification methods of the IRMA project.

CEA, France: submitted five runs, all automatic with two being visual, only and three mixed runs. The techniques used include the PIRIA visual retrieval system and a simple frequency–based text retrieval system.

IPAL CNRS/ I2R, France/Singapore: submitted a total of 6 runs, all automatic with two being text only and the other a combination of textual and visual features. For textual retrieval they map the text onto single axes of the MeSH ontology. They also use negative weight query expansion and mix visual and textual results for optimal results.

University of Concordia, Canada: submitted one visual run containing a query only for the first image of every topic using only visual features. The technique applied is an association model between low–level visual features and high–level concepts mainly relying on texture, edge and shape features.

In Table 4 an overview of the submitted runs can be seen with the query dimensions.

Table 4. Query dimensions of the submissions for the medical retrieval task.

Dimension	type	#Runs (%)
Run type	Automatic	128 (95.52%)
Modality	image	28 (20.90%)
	text	14 (10.45%)
	text+image	86 (64.18%)
Run type	Manual	6 (4.48%)
Modality	image	3 (2.24%)
	text	1 (0.75%)
	text+image	2 (1.5%)

3.4 Results

This section gives an overview of the best results of the various categories and performs some more analyses. Table 5 shows all the manual runs that were submitted with a classification of the techniquae used for retrieval

Table 6 gives the best 5 results for textual retrieval only and the best ten results for visual and for mixed retrieval. The results for individual topics varied widely, and further analysis will attempt to explore why this was so. If we calculate the average over the best system for each query we would be much closer to 0.5 than to what the best system actually achieved, 0.2821. So far, non of the systems optimised the feature selection based on the query input.

Table 5. Overview of the best manual retrieval results.

Run identifier	visual	textual	MAP	P10
OHSUmanual.txt		x	0.2116	0.4560
OHSUmanvis.txt	x		0.1601	0.5000
i2r-vk-avg.txt	x		0.0921	0.2760
i2r-vk-sem.txt	x		0.06	0.2320
i6-vistex-rfb1.clef	x	x	0.0855	0.3320
i6-vistex-rfb2.clef	x	x	0.077	0.2680

Table 6. Overview of the best automatic retrieval results.

Run identifier	visual	textual	MAP	P10
IPALI2R_Tn		x	0.2084	0.4480
IPALI2R_T		x	0.2075	0.4480
i6-En.clef		x	0.2065	0.4000
UBimed_en-fr.T.BI2		x	0.1746	0.3640
SinaiEn_okapi_nofb		x	0.091	0.2920
I2Rfus.txt	x		0.1455	0.3600
I2RcPBcf.txt	x		0.1188	0.2640
I2RcPBnf.txt	x		0.1114	0.2480
I2RbPBcf.txt	x		0.1068	0.3560
I2RbPBnf.txt	x		0.1067	0.3560
mirabase.qtop(GIFT)	x		0.0942	0.3040
mirarf5.1.qtop	x		0.0942	0.2880
GE_M_4g.txt	x		0.0941	0.3040
mirarf5.qtop	x		0.0941	0.2960
mirarf5.2.qtop	x		0.0934	0.2880
IPALI2R_TIan	x	x	0.2821	0.6160
IPALI2R_TIa	x	x	0.2819	0.6200
nctu_visual+text_auto_4	x	x	0.2389	0.5280
UBimed_en-fr.TI.1	x	x	0.2358	0.5520
IPALI2R_TImn	x	x	0.2325	0.5000
nctu_visual+text_auto_8	x	x	0.2324	0.5000
nctu_visual+text_auto_6	x	x	0.2318	0.4960
IPALI2R_TIm	x	x	0.2312	0.5000
nctu_visual+text_auto_3	x	x	0.2286	0.5320
nctu_visual+text_auto_1	x	x	0.2276	0.5400

3.5 Discussion

The results show a few clear trends. Very few groups performed manual submissions using relevance feedback, which was most likely due to the need for more resources for such evaluations. Still, relevance feedback has shown to be extremely useful in many retrieval tasks and the evaluation of it seems extremely necessary, as well. Surprisingly, in the submitted results, relevance feedback did not seem to give a much superior performance compared to the automatic runs. In the 2004 tasks, the relevance feedback runs were often significantly better than without feedback.

We also found that the topics developed were much more geared towards textual retrieval than visual retrieval. The best results for textual retrieval were much higher than for visual retrieval only, and a few of the poorly performing textual runs appeared to have indexing problems. When analysing the topics in more detail, a clear division becomes evident between the developed visual and textual topics. However, some of the topics marked as visual actually had better results using a textual system. Some systems performed extremely well on a few topics but then extremely poorly on other topics. No system was the best system for more than two of the topics.

The best results were clearly obtained when combining textual and visual features most likely due to the fact that there were queries for which only a combination of the feature sets works well.

4 Automatic Annotation Task

4.1 Introduction, Idea, and Objectives

Automatic image annotation is a classification task, where an image is assigned to its correspondent class from a given set of pre–defined classes. As such, it is an important step for content–based image retrieval (CBIR) and data mining [25]. The aim of the *Automatic Annotation Task* in ImageCLEFmed 2005 was to compare state–of–the–art approaches to automatic image annotation and to quantify their improvements for image retrieval. In particular, the task aims at finding out how well current techniques for image content analysis can identify the medical image modality, body orientation, body region, and biological system examined. Such an automatic classification can be used for multilingual image annotations as well as for annotation verification, e.g., to detect false information held in the header streams according to Digital Imaging and Communications in Medicine (DICOM) standard [26].

4.2 Database

The database consisted of 9.000 fully classified radiographs taken randomly from medical routine at the Aachen University Hospital. 1.000 additional radiographs for which classification labels were unavailable to the participants had to be classified into one of the 57 classes, from which the 9.000 database images come from. Although only 57 simple class numbers were provided for ImageCLEFmed 2005. The images are annotated with the complete IRMA code, a multi–axial code for image annotation [27]. The code is currently available in English and German. It is planned to use the results of such automatic image annotation tasks for further textual image retrieval tasks in the future.

Some example images together with their class number and their complete English annotation are given in Figure 6.

4.3 Participating Groups

In total 26 groups registered for participation in the automatic annotation task. All groups have downloaded the data but only 12 groups submitted runs. Each group had at least two different submissions. The maximum number of submissions per group was 7. In total, 41 runs were submitted which are briefly described in the following.

CEA: CEA from France, submitted three runs. In each run different feature vectors were used and classified using a k–Nearest Neighbour classifier (k was either 3 or 9). In the run labelled `cea/pj-3.txt` the images were projected along horizontal and vertical axes to obtain a feature histogram. For `cea/tlep-9.txt` histograms of local edge pattern features and colour features were created, and for `cea/cime-9.txt` quantified colours were used.

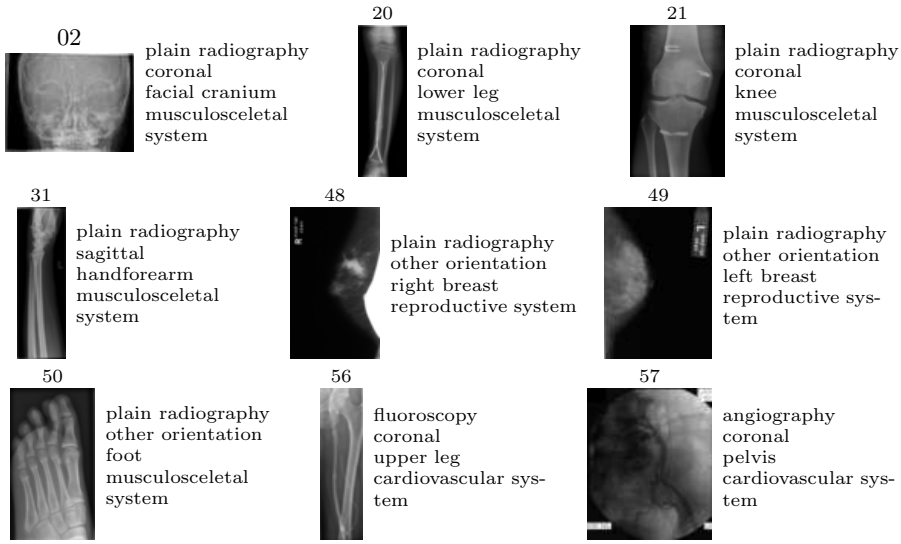


Fig. 6. Example images of the IRMA 10.000 database together with their class and annotation

CINDI: The CINDI group from Concordia University in Montreal, Canada used multi–class SVMs (one–vs–one) and a 170 dimensional feature vector consisting of colour moments, colour histograms, cooccurence texture features, shape moment, and edge histograms.

Geneva: The medGIFT group from Geneva, Switzerland used various different settings for gray levels, and Gabor filters in their medGIFT retrieval system.

Infocomm: The group from Infocomm Institute, Singapore used three kinds of 16x16 low–resolution–map–features: initial gray values, anisotropy and contrast. To avoid overfitting, for each of 57 classes, a separate training set was selected and about 6.800 training images were chosen out of the provided 9.000 images. Support Vector Machines with RBF (radial basis functions) kernels were applied to train the classifiers which were then employed to classify the test images.

Miracle: The Miracle Group from UPM Madrid, Spain used GIFT and a decision table majority classifier to calculate the relevance of each individual result in `miracle/mira20relp57.txt`. In `mira20relp58IB8.txt` additionally a k–nearest neighbour classifier with $k = 8$ and attribute normalisation is used.

Montreal: The group from University of Montreal, Canada submitted 7 runs, which differ in the features used. They estimated, which classes are best represented by which features and combined appropriate features.

mtholyoke: For the submission from Mount Holyoke College, MA, USA, Gabor energy features were extracted from the images and two different cross–media relevance models were used to classify the data.

nctu–dblab: The NCTU–DBLAB group from National Chiao Tung University, Taiwan used a support vector machine (SVM) to learn image feature characteristics. Based on the SVM model, several image features were used to predict the class of the test images.

ntu: The group from National Taiwan University used mean gray values of blocks as features and different classifiers for their submissions.

rwth–i6: The Human Language Technology and Pattern Recognition group from RWTH Aachen, Germany had two submissions. One used a simple zero–order image distortion model taking into account local context. The other submission used a maximum entropy classifier and histograms of patches as features.

rwth–mi: The IRMA group from Aachen, Germany used features proposed by Tamura et al to capture global texture properties and two distance measures for downscaled representations, which preserve spatial information and are robust w.r.t. global transformations like translation, intensity variations, and local deformations. The weighting parameters for combining the single classifiers were guessed for the first submission and trained on a random 8.000 to 1.000 partitioning of the training set for the second submission.

ulg The ulg (University of Liége) method is based on random sub–windows and decision trees. During the training phase, a large number of multi–size subwindows are randomly extracted from training images. Then, a decision tree model is automatically built (using Extra Trees and/or Tree Boosting), based on normalised versions of the subwindows, and operating directly on pixel values. Classification of a new image similarly entails the random extraction of subwindows, the application of the model to these, and the aggregation of subwindows predictions.

4.4 Results

The error rates range between 12.6 % and 73.3 % (Table 7). Based on the training data, a system guessing the most frequent group for all 1.000 test images would result with 70.3 % error rate, since 297 radiographs of the test set were from class 12. A more realistic baseline of 36.8 % error rate is computed from an 1–nearest–neighbour classifier comparing downscaled 32×32 versions of the images using the Euclidean distance.

Interestingly, the classes are very different in difficulty. The average classification accuracy ranges from 6.3 % to 90.7 %, and there is a tendency that classes with less training images are more difficult. For example, images from class 2 were extremely often classified to be from class 44: on average 46% of the images from class 2 were classified to be from class 44. This is probably partly due to a much higher a–priori probability for class 44, which has 193 images in the training set while class 2 only has 32 training images. Classes 7 and 8 are often classified to be from class 6, where once again class 6 is much better represented

in the training data. Furthermore, quite a few classes (6,13,14,27,28,34,44,51,57) are often misclassified to be from class 12, which is by far the largest class in the training data. This strongly coincides with the fact that class 12 is the class with the highest classification accuracy: 90.7% of the test images from class 12 were classified correctly. The three classes with the lowest classification accuracies, that is those three classes were on the average most of the images were misclassified, together have less then 1% of the training data.

4.5 Discussion

Similar experiments have been described in the literature. However, previous experiments have been restricted to a small number of categories. For instance, several algorithms have been proposed for orientation detection of chest radiographs, where lateral and frontal orientation are distinguished by means of image content analysis [28, 29]. In a recent investigation, Pinhas and Greenspan report error rates below 1 % for automatic categorisation of 851 medical images into 8 classes [30]. In previous investigations, error rates between 5.3% and 15% were reported for experiments with 1617 of 6 [31] and 6,231 of 81 classes, respectively. Hence, error rates of 12 % for 10.000 of 57 classes are plausible.

As mentioned before, classes 6, 7, and 8 were frequently confused. All show parts of the arms and thus look extremely similar (Fig. 6). However, a reason for the common misclassification in favour of class 6 might be that there are by a factor of 5 more training images from class 6 than from classes 7 and 8 together.

Given the confidence files from all runs, classifier combination was tested using the sum and the product rule in such a manner that first the two best confidence files were combined, then the three best confidence files, and so forth. Unfortunately, the best result was 12.9%. Thus, no improvement over the current best submission was possible using simple classifier combination techniques.

Having results close to 10% error rate, classification and annotation of images might open interesting vistas for CBIR systems. Although the task considered here is more restricted than the *Medical Retrieval Task* and can thus be considered easier, techniques applied will probably be apt to be used in future CBIR applications. Therefore, it is planned to use the results of such automatic image annotation tasks for further, textual image retrieval tasks.

5 Conclusions

ImageCLEF has continued to attract researchers from a variety of global communities interested in image retrieval using both low–level image features and associated texts. This year we have improved the ad–hoc medical retrieval by enlarging the image collection and creating more semantic queries based on realistic information needs of medical professionals. The ad–hoc task has continued to attract interest and this year has seen an increase in the number of translated topics and those with translated narratives. The addition of the IRMA annotation task has provided a further challenge to the medical side of ImageCLEF

Table 7. Resulting error rates for the submitted runs

submission	error rate [%]
rwth-i6/IDMSUBMISSION	12.6
rwth_mi-ccf_idm.03.tamura.06.confidence	13.3
rwth-i6/MESUBMISSION	13.9
ulg/maree-random-subwindows-tree-boosting.res	14.1
rwth-mi/rwth_mi1.confidence	14.6
ulg/maree-random-subwindows-extra-trees.res	14.7
geneva-gift/GIFT5NN_8g.txt	20.6
infocomm/Annotation_result4_I2R_sg.dat	20.6
geneva-gift/GIFT5NN_16g.txt	20.9
infocomm/Annotation_result1_I2R_sg.dat	20.9
infocomm/Annotation_result2_I2R_sg.dat	21.0
geneva-gift/GIFT1NN_8g.txt	21.2
geneva-gift/GIFT10NN_16g.txt	21.3
miracle/mira20relp57.txt	21.4
geneva-gift/GIFT1NN_16g.txt	21.7
infocomm/Annotation_result3_I2R_sg.dat	21.7
ntu/NTU-annotate05-1NN.result	21.7
ntu/NTU-annotate05-Top2.result	21.7
geneva-gift/GIFT1NN.txt	21.8
geneva-gift/GIFT5NN.txt	22.1
miracle/mira20relp58IB8.txt	22.3
ntu/NTU-annotate05-SC.result	22.5
nctu-dblab/nctu_mc_result_1.txt	24.7
nctu-dblab/nctu_mc_result_2.txt	24.9
nctu-dblab/nctu_mc_result_4.txt	28.5
nctu-dblab/nctu_mc_result_3.txt	31.8
nctu-dblab/nctu_mc_result_5.txt	33.8
cea/pj-3.txt	36.9
mtholyoke/MHC_CQL.RESULTS	37.8
mtholyoke/MHC_CBDM.RESULTS	40.3
cea/tlep-9.txt	42.5
cindi/Result-IRMA-format.txt	43.3
cea/cime-9.txt	46.0
montreal/UMontreal_combination.txt	55.7
montreal/UMontreal_texture_coarsness_dir.txt	60.3
nctu-dblab/nctu_mc_result_gp2.txt	61.5
montreal/UMontreal_contours.txt	66.6
montreal/UMontreal_shape.txt	67.0
montreal/UMontreal_contours_centred.txt	67.3
montreal/UMontreal_shape_fourier.txt	67.4
montreal/UMontreal_texture_directionality.txt	73.3
Euclidean Distance, 32x32 images, 1-Nearest-Neighbor	36.8

and proven a popular task for participants, covering mainly the visual retrieval community. The user–centered retrieval task, however, remains with low participation, mainly due to the high level of resources required to run an interactive task. We will continue to improve tasks for ImageCLEF 2006 mainly based on feedback from participants.

A large number of participants only registered but finally did not submit results. This means that the resources are very valuable and already access to the resources is a reason to register. Still, only if we have participants submitting results with different techniques, is there really the possibility to compare retrieval systems and developed better retrieval for the future. So for 2006 we hope to receive much feedback for tasks and many people who register, submit results and participate in the CLEF workshop to discuss the presented techniques. Further information can be found in [32, 33].

6 Acknowledgements

This work has been funded by the EU Sixth Framework Programme within the Bricks project (IST contract number 507457) and the SemanticMining project (IST NoE 507505). The establishment of the IRMA database was funded by the German Research Community DFG under grant Le 1108/4. We also acknowledge the support of National Science Foundation (NSF) grant ITR–0325160.

References

1. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval track. In Peters, C., Gonzalo, J., Braschler, M., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: Results of the Fourth CLEF Evaluation Campaign, Lecture Notes in Computer Science (LNCS), Springer, Volume 3237 (2004) 581–593
2. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross language image retrieval track. In Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Lecture Notes in Computer Science (LNCS), Springer, Volume 3491 (2005) 597–613
3. Gonzalo, J., Paul, C., Vallin, A.: Overview of the CLEF 2005 interactive track. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
4. Clough, P., Sanderson, M., Müller, H.: A proposal for the CLEF cross language image retrieval track (ImageCLEF) 2004. In: The Challenge of Image and Video Retrieval (CIVR 2004), Dublin, Ireland, Springer LNCS 3115 (2004)
5. Grubinger, M., Leung, C., Clough, P.: Towards a topic complexity measure for cross–language image retrieval. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
6. Besançon, R., Millet, C.: Data fusion of retrieval results from different media: Experiments at ImageCLEF 2005. In: Proceedings of Cross–Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)

7. Inoue, M.: Easing erroneous translations in cross–language image retrieval using word associations. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
8. Izquierdo-Beviá, R., Tomás, D., Saiz-Noeda, M., Vicedo, J.L.: University of Alicante in ImageCLEF2005. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
9. Hoi, S.C.H., Zhu, J., Lyu, M.R.: CUHK at ImageCLEF 2005: Cross–language and cross–media image retrieval. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
10. Jones, G.J., McDonald, K.: Dublin city university at CLEF 2005: Experiments with the ImageCLEF St. Andrews collection. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
11. Müller, H., Geissbühler, A., Marty, J., Lovis, C., Ruch, P.: The use of MedGIFT and EasyIR for imageCLEF 2005. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
12. Adriani, M., Arnely, F.: Retrieving images using cross–lingual text and image features. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
13. Martínez-Fernández, J., Villena Román, J., García-Serrano, A.M., González-Cristóbal, J.C.: Combining textual and visual features for image retrieval. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
14. Chang, Y.C., Lin, W.C., Chen, H.H.: A corpus–based relevance feedback approach to cross–language image retrieval. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
15. Martín-Valdivia, M., Garcí-Cumbreras, M., Dí-Galiano, M., Ureña López, L., Montejo-Raez, A.: The university of jaén at ImageCLEF 2005: Adhoc and medical taskseasing erroneous translations in cross–language image retrieval using word associations. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
16. Peinado, V., López-Ostenero, F., Gonzalo, J., Verdejo, F.: UNED at ImageCLEF 2005: Automatically structured queries with named entities over metadata. In: Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria. (2005)
17. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? Journal of the American Medical Association **280** (1998) 1347–1352
18. Markkula, M., Sormunen, E.: Searching for photos – journalists' practices in pictorial IR. In Eakins, J.P., Harper, D.J., Jose, J., eds.: The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval. Electronic Workshops in Computing, Newcastle upon Tyne, The British Computer Society (1998)
19. Hersh, W., Müller, H., Gorman, P., Jensen, J.: Task analysis for evaluating image retrieval systems in the ImageCLEF biomedical image retrieval task. In: Slice of Life conference on Multimedia in Medical Education (SOL 2005), Portland, OR, USA (2005)
20. 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** (2004) 295–305
21. 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** (2004) 1–6

22. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. Academic Medicine **78** (2003) 249–253
23. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet–based nuclear medicine teaching file. Journal of Nuclear Medicine **36** (1995) 1520–1527
24. Glatz-Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Web-basierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. Pathologe **24** (2003) 394–399
25. Lehmann, T.M., Güld, M.O., Deselaers, T., Schubert, H., Spitzer, K., Ney, H., Wein, B.B.: Automatic categorization of medical images for content–based retrieval and data mining. Computerized Medical Imaging and Graphics **29** (2005) 143–155
26. Güld, M., Kohnen, M., Keysers, D., Schubert, H., Wein, B., Bredno, J., Lehmann, T.: Quality of DICOM header information for image categorization. In: Proceedings of the SPIE conference Photonics West, Electronic Imaging, special session on benchmarking image retrieval systems. (2002) 280–287
27. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: Medical Imaging. Volume 5033 of SPIE Proceedings., San Diego, California, USA (2003)
28. Pietka, E., Huang, H.: Orientation correction for chest images. Journal of Digital Imaging **5** (1992) 185–189
29. Boone, J.M., Seshagiri, S., Steiner, R.: Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. Journal of Digital Imaging **5(3)** (1992) 190–193
30. Güld, M.O., Keysers, D., Deselaers, T., Leisten, M., Schubert, H., Ney, H., Lehmann, T.M.: Comparison of global features for categorization of medical images. In: Medical Imaging 2004. Volume 5371 of SPIE Proceedings. (2004)
31. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non–linear distortion models. In: Proceedings Bildverarbeitung für die Medizin, Berlin, Germany (2004) 366–370
32. Müller, H., Clough, P., Hersh, W., Deselaers, T., Lehmann, T.M., Geissbuhler, A.: Using heterogeneous annotation and visual information for the benchmarking of image retrieval systems. In: Proceedings of the SPIE conference Photonics West, Electronic Imaging, special session on benchmarking image retrieval systems, San Diego, USA (2006)
33. Müller, H., Clough, P., Hersh, W., Deselaers, T., Lehmann, T., Geissbuhler, A.: Axes for the evaluation of medical image retrieval systems - the ImageCLEF experience. In: Proceedings of the of ACM Multimedia 2005 (Brave New Topics track), 6–12 November, Singapore (2005) 1014–1022