The CLEF 2005 Cross-Language Image Retrieval Track

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

Abstract

The purpose of this paper is to outline efforts from the 2005 CLEF cross-language image retrieval campaign (ImageCLEF). The aim of this 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 the ImageCLEF track: a 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 that is explained in the iCLEF summary. 24 research groups from a variety of backgrounds and nationalities (14 countries) participated in ImageCLEF. In this paper we describe the ImageCLEF tasks, submissions from participating groups and summarise the main findings.

Categories and Subject Descriptors

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

General Terms

Image retrieval, evaluation, visual retrieval

Keywords

visual image retrieval, multimodal retrieval, medical image retrieval, automatic annotation

1 Introduction

ImageCLEF[1] 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 established in 2003 [4] and run again in 2004 [3] 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. an image with a caption written in English should be searchable in languages other than English.

ImageCLEF provides tasks for both system–centered and user–centered retrieval evaluation within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections. These domains offer realistic scenarios in which to test the performance of image retrieval systems, offering different challenges and problems to participating research groups. 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, section sec:annotation the automatic annotation of medical images and. 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

2.1 Aims and Objectives

This is a bilingual ad–hoc retrieval task in which a system is expected to match a user's one–time query against a more or less static collection (i.e. the set of documents to be searched is known prior to retrieval, but the search requests are not). Similar to the task run in previous years (see, e.g. [3]), 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). Queries for images based on abstract concepts rather than visual features are predominant in this task. This limits the effectiveness of using visual retrieval methods alone as 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 Rome). However, based on feedback from participants in 2004, the search tasks for 2005 are aimed to reflect more visually–based queries.

2.2 Data and Search Tasks

The St. Andrews collection consists of 28,133 images, all of which have associated textual captions written in British English (the target language). The captions consist of 8 fields including title,

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

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".

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

photographer, location, date and one or more pre–defined categories (all manually assigned by domain experts). For example, see Fig. 1. Further examples can be found in [5] and the St. Andrews University Library[2]. We provided participants with 28 topics (titles shown in Table 11 and an example image shown in Fig. 5), the main themes based on analysis of log files from a web server at St. Andrews university, knowledge of the image collection and discussions with maintainers of the image collection. After identifying these main themes, we modified queries to test various aspects of cross-language and visual search and used a custom–built IR system 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 [8].

Each topic consists of 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). In addition to the text description for each topic, we also provided two example images which we envisage could be used for relevance feedback (both manual and automatic) and query–by–example searches[3]. Both topic title and narratives have been translated into the following languages: German, French, Italian, Spanish (European), Spanish (Latin American), Chinese (Simplified), Chinese (Traditional) and Japanese. Translations have also been produced for the titles only and these are available in 25 languages including: Russian, Croatian, Bulgarian, Hebrew and Norwegian. All translations have been provided by native speakers and verified by at least one other native speaker.

2.3 Creating Relevance Assessments

Relevance assessments were performed by staff at the University of Sheffield (the majority unfamiliar with the St. Andrews collection but given training and access to the collection through our IR system). The top 50 results from all submitted runs (349) 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. We asked assessors to judge every image in the topic pool, but also to use interactive search and judge: searching the collection using their own queries to supplement the image pools with further relevant.

The assessment of images in this ImageCLEF task is based on using a ternary classification scheme: (1) relevant, (2) partially relevant and (3) not relevant. The aim of the ternary scheme is to help assessors in making their relevance judgements more accurate (e.g. an image is definitely relevant in some way, but maybe the query object is not directly in the foreground: it is therefore considered partially relevant). Relevance assessment for the more general topics are based entirely on the visual content of images (e.g. "aircraft on the ground"). However, certain topics also require the use of the caption to make a confident decision (e.g. "pictures of North Street St Andrews"). What constitutes a relevant image is a subjective decision, but typically a relevant image will have

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

the subject of the topic in the foreground, the image will not be too dark in contrast, and maybe the caption confirms the judge's decision.

Based on these judgements, various combinations are used to create the set of relevant images and as in previous years, we used the `pisec-total` set: those images judges as relevant or partially–relevant by the topic creator and at least one other assessor. These are then used to evaluate system performance and compare submissions. The size of pools and number of relevant images is shown in Table 11 (the %max indicating the pool size compared to the maximum possible pool size, i.e. if all top 50 images from each submission were unique).

2.4 Participating Groups

In total, 19 groups registered for this task and 11 ended up submitting (including 5 new groups compared to last year) 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 default CBIR system (GIFT/Viper). Submissions from participants are briefly described in the following.

CEA: CEA from France, submitted 9 runs. Experimented with 4 languages, title and title+narrative, and merging between modalities (text and image). This is simply based on normalised scores obtained by each search and is conservative (results obtained using visual topics and CBIR system are used only to reorder results obtained using textual topics)

NII: National Institute of Informatics from Japan, submitted 16 runs with 3 languages. These experiments were aimed to see if the inclusion of learned word association model - the model which represents how words are related - can help finding relevant images in adhoc CLIR setting. To do this, basic unigram language models were combined with differently estimated word association models that performs soft word-expansion. Also, combining simple keyword matching-like language models to above mentioned soft word-expansion language models at the model-output level. All runs were text only.

Alicante: University of Alicante (Computer Science) from Spain, submitted 62 runs (including 10 joint runs with UNED and Jaen). They experimented with 13 languages using title, automatic query expansion and text only. Their system combines probabilistic information with ontological information and a feedback technique. Several information streams are created using different sources: stems, words and stem bigrams, the final result obtained by combining them. An ontology has been created automatically from the St. Andrews collection to relate a query with several image categories. Four experiments were carried out to analyse how different features contribute to retrieval results. Moreover, a voting-based strategy was developed joining three different systems of participating universities: University of Alicante, University of Jaén and UNED.

CUHK: Chinese University of Hong Kong, submitted 36 runs for English and Chinese (simplified). CUHK experimented with title, title+narrative and using visual methods to rerank search results (visual features are composed of two parts: DCT coefficients and Colour moments with a dimension of 9). Various IR models used for retrieval (trained on 2004 data), together with query expansion. LDC Chinese segmentor is used to extract words from Chinese queries and translated into English using a dictionary.

DCU: Dublin City University (Computer Science) from Ireland, submitted 33 runs for 11 languages. All runs were automatic using title only. Standard OKAPI used incorporating stop word removal, suffix stripping and query expansion using pseudo relevance feedback. Their main focus of participation was to explore an alternative approach to combining text and image retrieval in an attempt to make use of information provided by the query image. Separate ranked lists returned using text retrieval without feedback and image retrieval based on standard low-level colour, edge and texture features, were investigated to find documents returned by both methods. These documents were then assumed to be relevant and used for text based pseudo relevance feedback and retrieval as in our standard method.

Geneva: University Hospitals Geneva from Switzerland, submitted 2 runs based on visual retrieval only (automatic and no feedback).

Table 1: Ad hoc experiments listed by query dimension.

Dimension	type	#Runs (%)
Language	non-English	230 (66%)
Run type	Automatic	349 (100%)
Feedback (QE)	yes	142 (41%)
Modality	image	4 (1%)
	text	318 (91%)
	text+image	27 (8%)
Initial Query	image only	4 (1%)
	title only	274 (79%)
	narr only	6 (2%)
	title+narr	57 (16%)
	title+image	4 (1%)
	title+narr+image	4 (1%)

Indonesia: University of Indonesia (Computer Science), submitted 9 runs using Indonesian queries only. They experimented with using title and title+narrative, with and without query expansion and combining text and image retrieval (all runs automatic).

MIRACLE: Daedalus and Madrid University from Spain, submitted 106 runs for 23 languages. All runs were automatic, using title only, no feedback and text-based only.

NTU: National Taiwan University from Taiwan, submitted 7 runs for Chinese (traditional) and English (also included a visual-only run). All runs are automatic and NTU experimented with using query expansion, using title and title+narrative and combining visual and text retrieval.

Jaen: University of Jaén (Intelligent Systems) from Spain, submitted 64 runs in 9 languages (all automatic). Jaen experimented with title and title+narrative, with and without feedback and combining both text and visual retrieval. Jaén experimented with both term weighting and the use of pseudo relevance feedback.

UNED: UNED from Spain, submitted 5 runs for Spanish (both Latin American and European) and English. All runs were automatic, title, text only and with feedback. UNED experimented with three different approaches: i) a naive baseline using a word by word translation of the title topics; ii) a strong baseline based on Pirkola's work; and iii) a structured query using the named entities with field search operators and Pirkola's approach.

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 1. No manual runs have been submitted this year, and a large proportion are text only using just the title. Together with 41% of submissions using query expansion, this co–incides with the large number of query languages offered this year and the focus on query translation by participating groups (although 6 groups submitted runs involving CBIR). An interesting submission this year was the combined efforts of Jaen, UNED and Alicante to create an approach based on voting for images. Table 2 provides a summary of submissions by query language. At least one group submitted for each language, the most popular (non-English) being French, German and Spanish (European).

Table 2: Ad hoc experiments listed by query language.

Query Language	#Runs	#Participants
English	70	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	3	2
Bulgarian	2	1
Croatian	2	1
Czech	2	1
Finnish	2	1
Hungarian	2	1

2.5 Results

Results for submitted runs were computed using the latest version of trec_eval [4] from NIST (v7.3). From the scores output, four chosen to evaluate submissions are Mean Average Precision (MAP), precision at result 10 (P10), precision at result 100 (P100) and the number of relevant images retrieved (RelRet) from which we compute recall (the proportion of relevant retrieved). Table 3 summarises the top performing systems in the ad-hoc task based on MAP. Whether MAP is the best score to rank image retrieval systems is debatable, hence our inclusion of P10 and P100 scores. The highest English (monolingual) retrieval score is 0.4135, with a P10 of 0.5500 and P100 of 0.3197. On average recall is high (0.8434), but low MAP and P10 indicating that relevant images are likely retrieved at lower rank positions. The highest monolingual score is obtained using combined visual and text retrieval and relevance feedback.

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 CLEF2004 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 at CLEF for these languages. On average, MAP for English is 0.2084 (0.3933 P10; 0.6454 recall) and across all languages is 0.2009 (0.2985 P10; 0.5737 recall) – see Table 4.

Table 4 shows the average MAP score averaged across all submissions by query dimension. There is a wide variation in counts for each dimension and type, therefore results are only an indication of effects on performance for each dimension. On average, it would appear that submissions with feedback (e.g. query expansion) performed better than without, submissions based on a combination of image and text retrieval appear to give higher performance (modality, although the NTU visual–only runs also perform well giving this type a high MAP score) and using both the image and text for the initial query (title+image) gives highest average MAP score (although

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

Table 3: Systems with highest MAP for each language in the ad–hoc retrieval task.

Query Language	MAP	P10	P100	Recall	Group	Run ID	Init. Query	Feedback	Modality
English	0.4135	0.5500	0.3197	0.8434	CUHK	CUHK-ad-eng-tv-kl-jm2	title+img	with	text+im
Chinese (trad.)	0.3993	0.5893	0.3211	0.7526	NTU	NTU-CE-TN-WEprf-Ponly	title+narr	with	text+im
Spanish (LA)	0.3447	0.4857	0.2839	0.7891	Alicante, Jaen	R2D2vot2SpL	title	with	text
Dutch	0.3435	0.4821	0.2575	0.7891	Alicante, Jaen	R2D2vot2Du	title	with	text
Visual	0.3425	0.5821	0.2650	0.7009	NTU	NTU-adhoc05-EX-prf	visual	with	image
German	0.3375	0.4929	0.2514	0.6383	Alicante, Jaen	R2D2vot2Ge	title	with	text
Spanish (Euro)	0.3175	0.4536	0.2804	0.8048	UNED	unedESENent	title	with	text
Portuguese	0.3073	0.4250	0.2436	0.7542	Miracle	imirt0attrpt	title	without	text
Greek	0.3024	0.4321	0.2389	0.6383	DCU	DCUFbTGR	title	with	text
French	0.2864	0.4036	0.2582	0.7322	Jaen	SinaiFrTitleNarrFBSystran	title+narr	with	text
Japanese	0.2811	0.3679	0.2086	0.7333	Alicante	AlCimg05Exp3Jp	title	with	text
Russian	0.2798	0.3571	0.2136	0.6879	DCU	DCUFbTRU	title	with	text
Italian	0.2468	0.3536	0.2054	0.6227	Miracle	imirt0attrit	title	without	text
Chinese (simp)	0.2305	0.3179	0.1732	0.6153	Alicante	AlCimg05Exp3ChS	title	with	text
Indonesian	0.2290	0.4179	0.2068	0.6566	Indonesia	UI-T-IMG	title	without	text+im
Turkish	0.2225	0.3036	0.1929	0.6320	Miracle	imirt0allftk	title	without	text
Swedish	0.2074	0.3393	0.1664	0.5647	Jaen	SinaiSweTitleNarrFBWordlingo	title	without	text
Norwegian	0.1610	0.1964	0.1425	0.4530	Miracle	imirt0attrno	title	without	text
Filipino	0.1486	0.1571	0.1229	0.3695	Miracle	imirt0allffl	title	without	text
Polish	0.1558	0.2643	0.1239	0.5073	Miracle	imirt0attrpo	title	without	text
Romanian	0.1429	0.2214	0.1218	0.3747	Miracle	imirt0attrro	title	without	text
Bulgarian	0.1293	0.2250	0.1196	0.5694	Miracle	imirt0allfbu	title	without	text
Czech	0.1219	0.1929	0.1343	0.5310	Miracle	imirt0allfcz	title	without	text
Croatian	0.1187	0.1679	0.1075	0.4362	Miracle	imirt0attrcr	title	without	text
Finnish	0.1114	0.1321	0.1211	0.3257	Miracle	imirt0attrfi	title	without	text
Hungarian	0.0968	0.1321	0.0768	0.3789	Miracle	imirt0allfhu	title	without	text

Table 4: MAP for ad–hoc averaged across all submissions by query dimension.

Dimension	type	#Runs	Average MAP
Language	English	119	0.2084
	non-English	230	0.2009
Feedback	yes	142	0.2399
	no	207	0.2043
Modality	image	4	0.3322
	text	318	0.2121
	text+image	27	0.3086
Initial Query	image only	4	0.1418
	title only	274	0.2140
	narr only	6	0.1313
	title+narr	57	0.2314
	title+image	4	0.4016
	title+narr+image	4	0.3953

again small counts for this dimension type).

Table 11 shows the highest MAP, P10, P100 and RelRet scores obtained from submissions for each topic. Results vary across topic as expected; some topics are harder than others. In this initial evaluation, we find that 18 topics have a recall of 1, 18 topics a P10 of 1, and 12 topics with a maximum MAP score greater than 0.7. The highest performing topic (easiest) is 11 "Swiss mountain scenery" and the lowest is topic 18 "woman in white dress". In addition, 18 topics have a maximum Relretr=relevant (i.e. a recall of 1) indicating that all relevant images have been retrieved in the top 1000 results.

2.6 Discussion

The variety of submissions in the ad-hoc task this year has been pleasing with a number of groups experimenting with both visual and text-based retrieval methods and combining the two (although the number of runs submitted as combined is much lower than 2004). As in 2004, the combination of text and visual retrieval appears to give highest retrieval effectiveness (based on MAP) indicating this is still an area for research. We aimed to offer a wider range of languages of which 13 have submissions from at least two groups (compared to 10 in 2004). It would seem that the focus for many groups in 2005 has been translation 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.

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

3.1 Goals and objectives

Domain–specific information retrieval is getting increasingly important and this holds especially true for the medical field, where patients as well as clinicians and researchers have their particular information needs [11]. Whereas information needs and retrieval methods for textual documents have been well researched, there is only a small amount of information available on the need to search for images [17], and even less so for the use of images in the medical domain. Image-

CLEFmed is creating resources to evaluate information retrieval tasks on medical image collections. This process includes the creation of image collections, of query tasks, and the definition of correct retrieval results for these tasks for system evaluation. Part of the tasks have been based on surveys of medical professionals and how they use images [12].

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 any detail in this section.

3.2 Data sets used and query topics

In 2004, only the Casimage[5] dataset was made available to participants [18], containing almost 9.000 images of 2.000 cases [22], 26 query topics with relevance judgements of three medical experts. It is also part of the 2005 collection. Images present in the data set include mostly radiology modalities, but also photographs, powerpoint slides and illustrations. Cases are mainly in French, with around 20% being in English. We were also allowed to use the PEIR[6] (Pathology Education Instructional Resource) database using annotation from the HEAL[7] project (Health Education Assets Library, mainly Pathology images [2]). This dataset contains over 33.000 images with English annotation, with the annotation being in XML per image and not per case as casimage. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology[8] [24], was also made available to us for ImageCLEF. This dataset contains over 2.000 images mainly from nuclear medicine with annotations per case and in English. Finally, the PathoPic[9] collection (Pathology images [7]) was included into our dataset. It contains 9.000 images with an extensive annotation per image in German. Part of the German annotation is translated into English, but it is still incomplete. This means, that a total of more than 50.000 images was made available with annotations in three different languages. Two collections have case–based annotations whereas two collections have image image–based annotations. Only through the access to the data by the copyright holders, we were able to distribute these images to the participating research groups.

The image topics were based on a small survey at OHSU. Based on this survey, the topics were developed along the following main 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 clearly to accommodate both visual and textual research groups we developed a set of 25 topics containing three different groups of queries: queries that are expected to be solvable with a visual retrieval system (topics 1-12), topics where both text and visual features are expected to perform well (topics 13-23) and semantic topics, where visual features are not expected to improve results. All query topics were of a higher semantic level than the 2004 topics because the automatic annotation task provides a testbed for purely visual retrieval/classification. All 25 topics contain one to three images, one query also an image as negative feedback. The query text was given out 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 2.

A query topic that will require more than purely visual features can be seen in Figure 3.

[5] http://www.casimage.com/
[6] http://peir.path.uab.edu/
[7] http://www.healcentral.com/
[8] http://gamma.wustl.edu/home.html
[9] 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.

Figure 2: 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.

Figure 3: A query that requires more than visual retrieval but visual features can deliver some hints to good results as well.

3.3 Relevance judgements

The relevance assessments were performed at OHSU in Portland, Oregon. A simple interface was used from previous ImageCLEF relevance assessments. 9 judges, mainly medical doctors and one image processing specialist performed the relevance judgements. Due to a lack of resources, only part of the topics could be judged by more than one person.

To create the image pools for the judgements, the first 40 images of each submitted run were taken into account to create pools with an average size of 892 images. The largest pool size was 1167 and the smallest one 470. It took the judges an average of roughly three hours to judge the images for a single topic. Compared to the purely visual topics from 2004 (around one hour judgement per topic containing an average of 950 images) the judgement process took much longer per image as the semantic queries required to verify the text and often 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 is significantly faster than most text research judgements, as a large number of irrelevant images could be sorted out very quickly.

We use a ternary judgement scheme including relevant, partially–relevant, and non–relevant. For the official qrels, we only used images marked as relevant. We also had several topics judged by two persons, but still took only the first judgements for the evaluations. Further analysis will follow in the final conference proceedings when more knowledge is available on the used techniques as well.

3.4 Participants

The number of registered participants of ImageCLEF has multiplied over the last three years. ImageCLEF started with 4 participants in 2003, then in 2004 a total of 18 groups participated and in 2005 we have 36 registered groups. The medical retrieval task had 12 participants in 2004 when it was purely visual and 13 in 2005 as a mixture of visual and non-visual retrieval. A surprisingly small number of groups (13 of 28 registered groups) finally submitted results, which can be due to the short time span between delivery of the images and the deadline for results submission. Another point was the fact that several groups only registered very late as they had not had information about ImageCLEF beforehand, but they were still interested in the datasets also for future participations. As the registration to the task is 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 Honk Kong, China

Finally 13 groups (two of them 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 short list of their participation including a short description of the submitted runs:

National Chiao Tuna 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. The Fire retrieval engine 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, Portland, OR, USA: submitted three runs in total, two manual runs, one for visual and one for textual retrieval and one automatic textual run. As retrieval engines GIFT and Lucene are being 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, which should rather be classified as a manual run, then. 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.

Table 5: 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%)

Table 6: Overview of the manual retrieval results.

Run identifier	visual	textual	results
OHSUmanual.txt		x	0.2116
OHSUmanvis.txt	x		0.1601
i2r-vk-avg.txt	x		0.0921
i2r-vk-sem.txt	x		0.06
i6-vistex-rfb1.clef	x	x	0.0855
i6-vistex-rfb2.clef	x	x	0.077

CEA, France: submitted five runs, all automatic with two being visual, only and three mixed runs. The techniques used include the 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 5 an overview of the submitted runs can be seen including the query dimensions.

3.5 Results

This section will give an overview of the best results of the various categories and will also do some more in depth analysis on a topic basis. More needs to follow based on the submissions of the papers from the participants.

Table 6 shows all the manual runs that were submitted with a classification into the technique used for the retrieval

In Table 7 are the best 5 results for textual retrieval only and the best ten results for visual and for mixed retrieval.

If we are looking at single topics it becomes clear that the systems vary extremely over the topics. 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 7: Overview of the best manual retrieval results.

Run identifier	visual	textual	results
IPALI2R_Tn		x	0.2084
IPALI2R_T		x	0.2075
i6-En.clef		x	0.2065
UBimed_en-fr.T.BI2		x	0.1746
SinaiEn_okapi_nofb		x	0.091
I2Rfus.txt	x		0.1455
I2RcPBcf.txt	x		0.1188
I2RcPBnf.txt	x		0.1114
I2RbPBcf.txt	x		0.1068
I2RbPBnf.txt	x		0.1067
mirabase.qtop(GIFT)	x		0.0942
mirarf5.1.qtop	x		0.0942
GE_M_4g.txt	x		0.0941
mirarf5.qtop	x		0.0941
mirarf5.2.qtop	x		0.0934
IPALI2R_TIan	x	x	0.2821
IPALI2R_TIa	x	x	0.2819
nctu_visual+text_auto_4	x	x	0.2389
UBimed_en-fr.TI.1	x	x	0.2358
IPALI2R_TImn	x	x	0.2325
nctu_visual+text_auto_8	x	x	0.2324
nctu_visual+text_auto_6	x	x	0.2318
IPALI2R_TIm	x	x	0.2312
nctu_visual+text_auto_3	x	x	0.2286
nctu_visual+text_auto_1	x	x	0.2276

3.6 Discussion

The results show a few clear trends. Very few groups performed manual submissions using relevance judgements, which is most likely due to the need of 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 does not seem to have a much superior performance compared to the automatic runs. In the 2004 tasks the relevance feedback runs were often significantly better than without feedback.

It also becomes clear that the topics developed were much more geared towards textual retrieval than visual retrieval. The best results for textual retrieval are much higher than for visual retrieval only, and a few of the bad textual runs seem simply to have indexing problems. When analysing the topics in more details a clear division becomes clear between the developed visual and textual topics, but also some of the topics marked as visual had actually better results using a textual system. Some systems actually perform extremely well on a few topics but then extremely bad on other topics. No system is actually 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 that either one of the feature sets would work 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 [14]. 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 [9].

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, the 9,000 database images come from. Although only 57 simple class numbers were provided for ImageCLEFmed 2005. The images are annotated with complete IRMA code, a multi-axial code for image annotation. 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.

Example images together with their class number are given in Figure 4. Table 8 gives the English textual description for each of the classes.

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` histogram of local edge patterns features and colour features were created, and for `cea/cime-9.txt` quantified colours were used.

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 image 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 over-fitting, for each of 57 classes, a separate training set was selected and about 6,800 training images were chosen out of the given 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 uses 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 used features used. They to estimated, which classes are best represented by which features and combined appropriate features.

Figure 4: Example images from the IRMA database which was used for the automatic annotation task.

Table 8: Class numbers together with their English IRMA annotation.

class	textual description
01	plain radiography, coronal, cranium, musculosceletal system
02	plain radiography, coronal, facial cranium, musculosceletal system
03	plain radiography, coronal, cervical spine, musculosceletal system
04	plain radiography, coronal, thoracic spine, musculosceletal system
05	plain radiography, coronal, lumbar spine, musculosceletal system
06	plain radiography, coronal, hand, musculosceletal system
07	plain radiography, coronal, radio carpal joint, musculosceletal system
08	plain radiography, coronal, handforearm, musculosceletal system
09	plain radiography, coronal, elbow, musculosceletal system
10	plain radiography, coronal, upper arm, musculosceletal system
11	plain radiography, coronal, shoulder, musculosceletal system
12	plain radiography, coronal, chest, unspecified
13	plain radiography, coronal, bones, musculosceletal system
14	plain radiography, coronal, abdomen, gastrointestinal system
15	plain radiography, coronal, abdomen, uropoietic system
16	plain radiography, coronal, upper abdomen, gastrointestinal system
17	plain radiography, coronal, pelvis, musculosceletal system
18	plain radiography, coronal, foot, musculosceletal system
19	plain radiography, coronal, ankle joint, musculosceletal system
20	plain radiography, coronal, lower leg, musculosceletal system
21	plain radiography, coronal, knee, musculosceletal system
22	plain radiography, coronal, upper leg, musculosceletal system
23	plain radiography, coronal, hip, musculosceletal system
24	plain radiography, sagittal, facial cranium, musculosceletal system
25	plain radiography, sagittal, neuro cranium, musculosceletal system
26	plain radiography, sagittal, cervical spine, musculosceletal system
27	plain radiography, sagittal, thoracic spine, musculosceletal system
28	plain radiography, sagittal, lumbar spine, musculosceletal system
29	plain radiography, sagittal, hand, musculosceletal system
30	plain radiography, sagittal, radio carpal joint, musculosceletal system
31	plain radiography, sagittal, handforearm, musculosceletal system
32	plain radiography, sagittal, elbow, musculosceletal system
33	plain radiography, sagittal, shoulder, musculosceletal system
34	plain radiography, sagittal, chest, unspecified
35	plain radiography, sagittal, foot, musculosceletal system
36	plain radiography, sagittal, ankle joint, musculosceletal system
37	plain radiography, sagittal, lower leg, musculosceletal system
38	plain radiography, sagittal, knee, musculosceletal system
39	plain radiography, sagittal, upper leg, musculosceletal system
40	plain radiography, sagittal, hip, musculosceletal system
41	plain radiography, axial, right breast, reproductive system
42	plain radiography, axial, left breast, reproductive system
43	plain radiography, axial, knee, musculosceletal system
44	plain radiography, other orientation, facial cranium, musculosceletal system
45	plain radiography, other orientation, neuro cranium, musculosceletal system
46	plain radiography, other orientation, cervical spine, musculosceletal system
47	plain radiography, other orientation, hand, musculosceletal system
48	plain radiography, other orientation, right breast, reproductive system
49	plain radiography, other orientation, left breast, reproductive system
50	plain radiography, other orientation, foot, musculosceletal system
51	fluoroscopy, coronal, hilum, respiratory system
52	fluoroscopy, coronal, upper abdomen, gastrointestinal system
53	fluoroscopy, coronal, pelvis, cardiovascular system
54	fluoroscopy, coronal, lower leg, cardiovascular system
55	fluoroscopy, coronal, knee, cardiovascular system
56	fluoroscopy, coronal, upper leg, cardiovascular system
57	angiography, coronal, pelvis, cardiovascular system

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 University, 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 down-scaled representations, which preserve spatial information and are robust w.r.t. global transformations like translation, intensity variations, and local deformations. The weighing 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.ac.be: The ULg method is based on random sub-windows and decision trees. During the training phase, a large number of multi-size sub-windows are randomly extracted from training images. Then, a decision tree model is automatically built (using Extra-Trees and/or Tree Boosting), based on size-normalised versions of the sub-windows, and operating directly on their pixel values. Classification of a new image similarly entails the random extraction of sub-windows, the application of the model to these, and the aggregation of sub-window predictions.

4.4 Results

The error rates ranges between 12.6 % and 73.3 % (Table 9). 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 (Table 10). A more realistic baseline of 36.8 % error rate is computed from an 1-nearest-neighbour classifier comparing down-scaled 32×32 versions of the images using the Euclidean distance.

For each class, a more detailed analysis including the number of training and test images as well as with respect to all 41 submitted runs, the average classification accuracy, the class most frequently misclassified, and the average percentage over all submitted runs of images being assigned to this class is given in Table 10. Obviously, the difficulty of the 57 classes diversifies. The average classification accuracy range from 6.3 % to 90.7 %, and there is a tendency that classes with less training images are more difficult. For instance for class 32, 78 images were contained in the training but only one image in the test data. In 23 runs, this test image was misclassified (43.9 %). Five times, it was labelled to be from class 25 (12.2 %). Also, it can be seen that many images of the classes 7 and 8 have been classified to be of class 6.

4.5 Discussion

Similar experiments have been described in 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 [20, 1]. For this two-class experiment, the error rates are

Table 9: 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.ac.be/maree-random-subwindows-tree-boosting.res	14.1
rwth-mi/rwth_mi1.confidence	14.6
ulg.ac.be/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

Table 10: The number of training and test images in the classes, the average, minimum, and maximum error rates.

class	train images	test images	avg. classification accuracy [%]	most mistaken class	avg. images that were classified to the most mistaken class [%]
1	336	38	84.0	25	3.9
2	32	3	18.7	44	46.3
3	215	24	69.6	5	3.0
4	102	12	57.3	3	5.9
5	225	25	75.6	3	2.6
6	576	67	66.0	12	4.4
7	77	8	27.7	6	21.0
8	48	3	6.5	6	38.2
9	69	10	21.0	21	19.8
10	32	7	6.3	6	10.8
11	108	12	26.0	6	9.8
12	2563	297	90.7	34	1.5
13	93	7	17.1	12	18.5
14	152	14	57.1	12	9.9
15	15	3	26.8	5	18.7
16	23	1	9.8	6	31.7
17	217	24	71.3	34	5.1
18	205	12	43.5	6	19.5
19	137	17	62.1	6	4.7
20	31	2	13.4	21	24.4
21	194	29	66.6	6	4.3
22	48	3	25.2	19	9.8
23	79	10	29.5	21	8.0
24	17	4	32.3	6	16.5
25	284	36	71.0	1	10.4
26	170	23	61.3	3	5.3
27	109	13	62.3	12	5.6
28	228	16	63.0	12	7.5
29	86	8	18.6	6	26.2
30	59	7	26.1	21	11.5
31	60	8	8.2	6	19.8
32	78	1	43.9	25	12.2
33	62	5	22.9	6	12.7
34	880	79	88.5	12	5.5
35	18	4	25.6	6	9.8
36	94	21	40.7	6	5.9
37	22	2	6.1	36	17.1
38	116	19	37.6	21	13.5
39	38	5	7.8	22	12.2
40	51	3	12.2	23	19.5
41	65	15	59.3	48	24.4
42	74	13	66.0	49	21.2
43	98	8	56.1	6	9.1
44	193	23	39.1	12	7.2
45	35	3	26.0	1	19.5
46	30	1	17.1	28	26.8
47	147	15	42.4	6	33.2
48	79	6	66.7	41	20.3
49	78	9	54.2	42	35.2
50	91	8	33.5	6	25.6
51	9	1	43.9	12	17.1
52	9	1	51.2	26	9.8
53	15	3	16.3	5	29.3
54	46	3	57.7	21	11.4
55	10	2	11.0	54	23.2
56	15	0	-	-	-
57	57	7	81.5	12	5.2

below 1 % [15]. In a recent investigation, Pinhas and Greenspan report error rates below 1 % for automatic categorisation of 851 medical images into 8 classes [21]. In previous investigations of the IRMA group, error rates between 5.3% and 15% were reported for experiments with 1617 of 6 [13] and 6,231 of 81 classes [10], 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. 4). 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 results was 12.9%. Thus, no improvement over the current best submission was possible using simple classifier combination techniques.

Having some 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 thus can be considered easier, techniques applied here will most probably be apt to be used in future CBIR applications, too. 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 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 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, there is 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 at the CLEF workshop to discuss the presented techniques.

6 Acknowledgements

This work has been funded in part by the EU Sixth Framework Programme (FP6) within the Bricks project (IST contract number 507457) as well as the SemanticMining project (IST NoE 507505). The establishment of the IRMA database was funded by the German Research Community DFG under grand Le 1108/4. We also acknowledge the generous support of National Science Foundation (NSF) grant ITR-0325160.

References

[1] Boone JM, Seshagiri S, Steiner RM. Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. Journal of Digital

Imaging 1992; 5(3): 190-193.

[2] C. S. Candler, S. H. Uijtdehaage, and S. E. Dennis. Introducing HEAL: The health education assets library. *Academic Medicine*, 78(3):249–253, 2003.

[3] Clough, P., Müller, H. and Sanderson, M. The CLEF 2004 Cross Language Image Retrieval Track, In Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign, Eds (Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M. and Magnini, B.), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2005, Volume 3491/2005, 597-613.

[4] Clough, P.D. and Sanderson, M.(2003), The CLEF 2003 cross language image retrieval track, In Proceedings of Cross Language Evaluation Forum (CLEF) 2003 Workshop, Trondheim, Norway.

[5] Paul Clough, Mark Sanderson, and Henning Müller. A proposal for the CLEF cross language image retrieval track (ImageCLEF) 2004. In *The Challenge of Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, July 2004. Springer LNCS 3115.

[6] . J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. Proceedings of the 13th International Conference on Pattern Recognition, 3:361–369, 1996.

[7] K. Glatz-Krieger, D. Glatz, M. Gysel, M. Dittler, and M. J. Mihatsch. Webbasierte Lernwerkzeuge für die Pathologie – web–based learning tools for pathology. *Pathologe*, 24:394–399, 2003.

[8] Grubinger, M., Leung, C. and Clough, P.D. Towards a Topic Complexity Measure for Cross-Language Image Retrieval, In Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria

[9] Güld MO, Kohnen M, Keysers D, Schubert H, Wein B, Bredno J, Lehmann TM. Quality of DICOM header information for image categorization. Procs SPIE 2002; 4685: 280-287.

[10] Güld MO, Keysers D, Leisten M, Schubert H, Lehmann TM. Comparison of global features for categorization of medical images. Procs SPIE 2004; 5371: 211-222.

[11] W. R. Hersh and D. H. Hickam. How well do physicians use electronic information retrieval systems? *Journal of the American Medical Association*, 280(15):1347–1352, 1998.

[12] W. Hersh, H. Müller, P. Gorman, and J. Jensen. 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, June 2005.

[13] Keysers D, Gollan C., Ney H. Classification of Medical Images using Non-linear Distortion Models. Proc. Bildverarbeitung fr die Medizin 2004 : 366-370

[14] Lehmann TM, Gld MO, Deselaers T, Keysers D, Schubert H, Spitzer K, Ney H, Wein BB. Automatic categorisation of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics 2005; 29(2): 143-155.

[15] Lehmann TM, Güld MO, Keysers D, Schubert H, Kohnen M, Wein BB. Determining the view position of chest radiographs. Journal of Digital Imaging 2003; 16(3): 280-291.

[16] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. Procs SPIE 2003; 5033: 440-451.

Table 11: Topics used in ImageCLEF and maximum MAP, P10, P100 and RelRetr scores.

Number	Title	Pool size (% max)	Relevant	MAP	P10	P100	Rel retr
1	Aircraft on the ground	1690 (9.7%)	85	0.8259	1.0000	0.7600	83
2	People gathered at bandstand	2420 (13.9%)	27	0.6899	0.9000	0.2600	27
3	Dog in sitting position	763 (4.4%)	34	0.5723	1.0000	0.2700	34
4	Steam ship docked	1797 (10.3%)	76	0.4316	0.9000	0.4000	72
5	Animal statue	861 (4.9%)	37	0.8349	1.0000	0.3400	37
6	Small sailing boat	1447 (8.3%)	122	0.6975	1.0000	0.7200	118
7	Fishermen in boat	1182 (6.8%)	32	0.5151	0.9000	0.3000	32
8	Building covered in snow	1329 (7.6%)	38	0.4177	0.7000	0.2800	35
9	Horse pulling cart or carriage	1435 (8.2%)	108	0.5972	1.0000	0.6200	108
10	Sun pictures & Scotland	1553 (8.9%)	203	0.7139	1.0000	0.9300	197
11	Swiss mountain scenery	1460 (8.4%)	83	0.9660	1.0000	0.8000	83
12	Postcards from Iona & Scotland	1665 (9.5%)	34	0.7493	1.0000	0.3400	34
13	Stone viaduct with several arches	1567 (9.0%)	184	0.5587	1.0000	0.7000	174
14	People at the marketplace	1203 (6.9%)	55	0.8207	1.0000	0.5100	55
15	Golfer putting on green	1367 (7.8%)	48	0.5652	0.9000	0.3700	48
16	Waves breaking on beach	1544 (8.8%)	71	0.5281	1.0000	0.4100	68
17	Man or woman reading	1074 (6.2%)	13	0.8156	0.8000	0.1300	13
18	Woman in white dress	1112 (6.4%)	40	0.2696	0.5000	0.2600	39
19	Composite postcards of Northern Ireland	1943 (11.1%)	50	0.5017	1.0000	0.5000	50
20	Royal visit to Scotland (not Fife)	1359 (7.8%)	13	0.7820	0.9000	0.1300	13
21	Monument to poet Robert Burns	875 (5.0%)	35	0.7349	1.0000	0.3300	35
22	Building with waving flag	1221 (7.0%)	56	0.6475	1.0000	0.4800	56
23	Tomb inside church or cathedral	1706 (9.8%)	62	0.7653	1.0000	0.5500	62
24	Close-up picture of bird	1414 (8.1%)	33	0.6353	1.0000	0.2700	29
25	Arched gateway	2037 (11.7%)	235	0.5857	1.0000	0.8700	208
26	Portrait pictures of mixed sex group	1410 (8.1%)	30	0.7618	0.9000	0.2900	30
27	Woman or girl carrying basket	1000 (5.7%)	14	0.5011	0.6000	0.1400	14
28	Colour pictures of woodland scenes around St Andrews	2263 (13.0%)	98	0.8200	1.0000	0.6700	98

[17] M. Markkula and E. Sormunen. Searching for photos – journalists' practices in pictorial IR. In J. P. Eakins, D. J. Harper, and J. Jose, editors, *The Challenge of Image Retrieval, A Workshop and Symposium on Image Retrieval*, Electronic Workshops in Computing, Newcastle upon Tyne, 5–6 February 1998. The British Computer Society.

[18] H. Müller, A. Rosset, J.-P. Vallée, F. Terrier, and A. Geissbuhler. A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, 28:295–305, 2004.

[19] Petrelli, D. and Clough, P.D. Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation, In Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria

[20] Pietka E, Huang HK. Orientation correction for chest images. Journal of Digital Imaging 1992; 5(3): 185-189.

[21] Pinhas A, Greenspan H. A continuous and probabilistic framework for medical image representation and categorization. Procs SPIE 2003; 5371: 230-238.

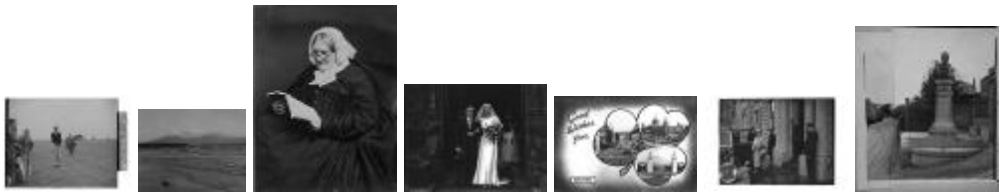
[22] A. Rosset, H. Müller, M. Martins, N. Dfouni, J.-P. Vallée, and O. Ratib. Casimage project – a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[23] Villena-Román, R., Crespo-García, R.M., and González-Cristóbal, J.C. Boolean Operators in Interactive Search, In Proceedings of Cross Language Evaluation Forum (CLEF) 2005 Workshop, Vienna, Austria

[24] J. W. Wallis, M. M. Miller, T. R. Miller, and T. H. Vreeland. An internet–based nuclear medicine teaching file. *Journal of Nuclear Medicine*, 36(8):1520–1527, 1995.

5.1 5.2 5.3 5.4 5.5 5.6 5.7

5.8 5.9 5.10 5.11 5.12 5.13 5.14



5.15 5.16 5.17 5.18 5.19 5.20 5.21

5.22 5.23 5.24 5.25 5.26 5.27 5.28

Figure 5: Example images given to participants for the ad-hoc retrieval task (1 of 2 images).