Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks

Paul Clough[1], Michael Grubinger[2], Thomas Deselaers[3],
Allan Hanbury[4], Henning Müller[5]

[1] Sheffield University, Sheffield, UK
[2] Victoria University, Melbourne, Australia
[3] Computer Science Department, RWTH Aachen University, Germany
[4] Vienna University, Vienna, Austria
[5] University and Hospitals of Geneva, Switzerland

Abstract

This paper describes the general photographic retrieval and object annotation tasks of the ImageCLEF 2006 evaluation campaign. These tasks provide both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information systems for image retrieval and automatic image annotation. Both tasks offer something new for 2006 and attracted a large number of submissions: 12 groups participating in ImageCLEFphoto and 3 in the automatic annotation task. This paper summarises components used in the benchmark, including the collections, the search and annotation tasks, the submissions from participating groups, and results.

The general photographic retrieval task, ImageCLEFphoto, used a new collection – the IAPR-TC12 Benchmark – of 20,000 colour photographs with semi-structured captions in English and German. This new collection replaces the St Andrews collection of historic photographs used for the previous three years. For ImageCLEFphoto groups submitted mainly text-only runs. However, 31% of runs involved some kind of visual retrieval technique, typically combined with text through the merging of image and text retrieval results. Bilingual text retrieval was performed using two target languages: English and German, with 59% of runs bilingual. Highest monolingual of English was shown to be 74% for Portuguese-English and 39% of German for English-German. Combined text and retrieval approaches were seen to give, on average, higher retrieval results (+54%) than using text (or image) retrieval alone. Similar to previous years, the use of relevance feedback (most commonly in the form of pseudo relevance feedback) to enable query expansion was seen to improve the text-based submissions by an average of 39%. Topics have been categorised and analysed with respect to various attributes including an estimation of their "visualness" and linguistic complexity.

The general automatic object annotation task used a hand collected dataset of 81,211 images from 268 classes provided by LTUtech. Given training data, participants were required to classify previously unseen images. The error rate of submissions for this task was high (ranging from 77.3% to 93.2%) resulting in a large proportion of test images being misclassified by any of the proposed classification methods. The task can therefore be said to have been very challenging for participants.

Categories and Subject Descriptors

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

Management]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, image classification, performance evaluation

1 Introduction

The evaluation of text information retrieval has benefited from the use of standardised benchmarks and evaluation events, performed since the 1960s [2]. With TREC[1] (Text REtrieval Conference [11]) a standard was set that has been used as the model for evaluation events in related fields. One such event is CLEF[2] (Cross Language Evaluation Forum) and within CLEF, the retrieval of images from multilingual collections: ImageCLEF. Over the past 2-3 years, ImageCLEF has expanded to deal with multiple domains (most noticeably the retrieval of medical images) and aspects of retrieval such as the automatic annotation of images with text descriptors. In this paper, we describe three tasks at ImageCLEF 2006: the general photographic retrieval task (ImageCLEFphoto), a general visual retrieval task, and a general image annotation (or classification) task. Section 2 describes the first general retrieval task, section 3 the visual retrieval task aimed more specifically at evaluating purely visual retrieval systems, and section 4 describes the automatic annotation task.

2 The ImageCLEFphoto photographic retrieval task

2.1 General Overview

This task is similar to the classic TREC ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (i.e. topics are not known to the system in advance). The goal of ImageCLEFphoto 2006 is: given a multilingual statement describing a user information need, find as many relevant images as possible from the given document collection. After three years of image retrieval evaluation using the St. Andrews database [3], a new database was used in this year's task: the *IAPR TC-12 Benchmark* [5], created under Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR[3]). This collection differs from the St Andrews collection used in previous campaigns in two major ways: (1) it contains mainly colour photographs (the St Andrews collection was primarily black and white) and (2) it contains semi-structured captions in English *and* German (the St Andrews collection used only English).

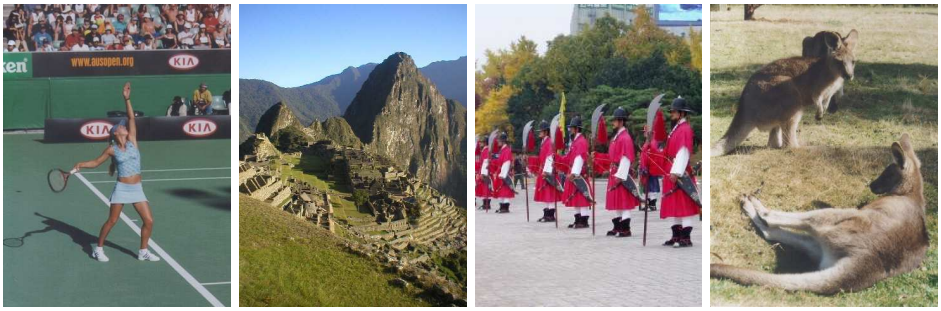
2.2 Document Collection

The IAPR TC-12 Benchmark contains 20,000 photos taken from locations around the world and comprises a varying cross-section of still natural images. Figure 1 illustrates a number of sample images from a selection of categories. The majority of images have been provided by *viventura*[4], an independent travel company that organises adventure and language trips to South-America. Travel guides accompany the tourists and maintain a daily online diary including photographs of

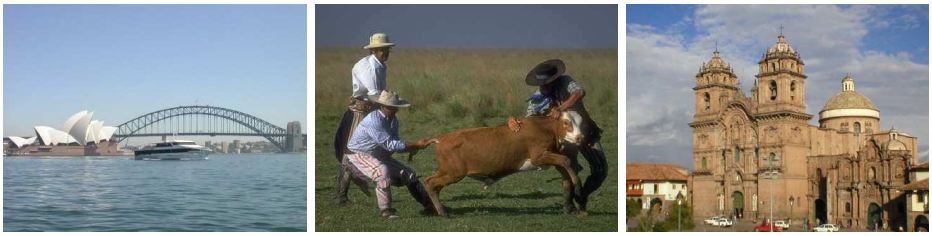
[1] `http://trec.nist.gov/`
[2] `http://www-clef-campaign.org/`
[3] `http://www.iapr.org/`
[4] `http://www.viventura.de`

trips made and general pictures of each location including accommodation, facilities and ongoing social projects. The collection contains many different images of similar visual content, but varying illumination, viewing angle and background. This makes it a challenge for the successful application of visual analysis techniques.

Figure 1: Sample images from the IAPR TC-12 collection.



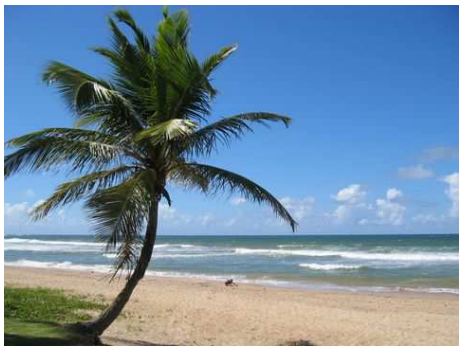
Sports. Landscapes. People. Animals.



Cities. Actions. Architecture.

Each image in the collection has a corresponding semi-structured caption consisting of the following seven fields: (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual contents of the image, (4) notes for additional information, (5) the provider of the photo and fields describing (6) where and (7) when the photo was taken. These fields exist in English and German, with a Spanish version currently being verified. Figure 2 shows a sample image with its corresponding English annotation.

Figure 2: Sample image caption.



```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>Flamingo Beach</TITLE>
<DESCRIPTION>a photo of a brown sandy beach; the dark
  blue sea with small breaking waves behind it; a dark
  green palm tree in the foreground on the left; a blue
  sky with clouds on the horizon in the background;
</DESCRIPTION>
<NOTES>Original name in Portuguese: "Praia do Flamengo";
  Flamingo Beach is considered as one of the most
  beautiful beaches of Brazil;</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>2 October 2004</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

These annotations are stored in a database allowing subsets of the collection to be created for benchmarking based on specifying particular parameters (e.g. which caption fields to use).

One of these parameters is *annotation quality*: in order to provide a more realistic scenario, the annotation files have been generated with a varying degree of annotation "completeness":

- 70% of the annotations contain title, description, notes, location and date.
- 10% of the annotations contain title, location and date.
- 10% of the annotations contain location and date.
- 10% of the images are not annotated (or have empty tags respectively).

2.3 Query Topics

Participants were given 60 topics, created using a custom-built topic creation and administration system to "achieve a natural, balanced topic set accurately reflecting real word user statements of information needs" [9] (pp.1069). The following information was considered in the topic creation process:

Number of topics. In order to increase the reliability of results, a total of 60 topics was provided to participants.

Log file Analysis. To make the task realistic, topics were derived from analysing a log file[5] from a web-based interface to the IAPR TC-12 collection which is used by employees and customers of viventura. A total of 40 topics were taken directly from the log file (semantically equivalent but perhaps with slight syntactic modification, e.g. "lighthouse sea" to "lighthouses at the sea") and 10 topics derived from entries in the log file (e.g. "straight roads in Argentina" changed to "straight roads in the USA"). The remaining 10 topics were not taken directly from the log file but created to test various aspects of text and image retrieval (e.g. "black and white photos of Russia").

Geographic Constraints. Corresponding to the findings from previous log file analyses (see, e.g. [12]), many search requests exhibit geographic constraints and this was found to be similar with the IAPR TC-12 collection. Thus, 24 of the topics were created with a geographic constraint (e.g. "tourist accommodation near Lake Titicaca" specifies a location and spatial operator *near*); 20 of the topics specifying a geographic feature or a permanent man-made object (e.g. "group standing in salt pan") and the remaining topics having no geography (e.g. "photos of female guides").

Visual Features. All topics were classified according to how "visual" they were considered to be. An average rating between 1-5[6] was obtained for each topic from three experts in the field of image analysis, and the retrieval score from a baseline content-based image retrieval (CBIR) system[7]. A total of 30 topics are classed as "semantic" (levels 1 and 2) for which visual approaches are highly unlikely to improve results; 20 topics are "neutral" (level 3) for which visual approaches may or may not improve results and 10 are "visual" topics for which content-based approaches are most likely to improve retrieval results.

Topic Difficulty. A topic complexity measure was used to categorise topics according to their linguistic complexity [6]. A total of 31 topics were chosen to be rather easy topics (levels 1 and 2), 25 topics were medium–hard topics (level 3), and 4 topics were difficult (levels 4 and 5).

Size of Target Set. Topic creators aimed for a target set size between 20 and 100 relevant images and thus had to further modify some of the topics (broadening or narrowing the concepts). The minimum was chosen in order to be able to use P(20) as a performance measure, whereas the upper limit of relevant images should limit the retrieval of relevant images by chance and to keep the relevance judgment pools to a manageable size.

Annotation Quality. Another dimension considered was the distribution of the topics in regards to the level of annotation quality of relevant images for the particular queries. In other

[5] Log file taken between 1st February and 15th April 2006 containing 980 unique queries.

[6] We asked experts in the field to rate these topics according to the following scheme: (1) CBIR will produce very bad or random results, (2) bad results, (3) average results, (4) good results and (5) very good results.

[7] The FIRE system was used based on using all query images.

words, 18 topics were provided in which all relevant images have complete annotations, 10 topics with 80% - 100% of the relevant images having complete annotations, further 19 topics with 60% - 80% of the relevant images with complete annotations, and 13 topics with less than 60% of the relevant images with complete annotations.

Attributes of Text Retrieval. Various aspects of text retrieval on a more semantic level were considered too, concentrating on vocabulary mismatches, general versus specific concepts, word disambiguation and abbreviations.

Participant Feedback. In last year's break-out session, participants suggested we provide groups of similar topics in order to facilitate the analysis of weak performing queries. This year saw groups of up to five similar topics (e.g. "tourist groups / destinations / Machu Picchu in bad weather").

Each original topic comprised 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 each request). In addition, three image examples were provided with each topic in order to test relevance feedback (both manual and automatic) and query-by-example searches. The topic titles were then translated into 15 languages including German, French, Spanish, Italian, Portuguese, Dutch, Russian, Japanese, and Simplified and Traditional Chinese. All translations were provided by at least one native speaker and verified by at least another native speaker. Unlike in past campaigns, however, the topic narratives were neither translated nor evaluated this year. A list of all topics can be found in Table 5.

Figure 3: Topic with three sample images.

```
<top>
<num> Number: 14 </num>
<title> scenes of footballers in action </title>
<narr> Relevant images will show football (soccer)
players in a game situation during a match. Images with
footballers that are not playing (e.g. players posing for
a group photo, warming up before the game, celebrating
after a game, sitting on the bench, and during the half-
time break) are not relevant. Images with people not
playing football (soccer) but a different code (American
Football, Australian Football, Rugby Union, Rugby League,
Gaelic Football, Canadian Football, International Rules
Football, etc.) or some other sport are not relevant.
</narr>
<image> images/31/31609.jpg </image>
<image> images/31/31673.jpg </image>
<image> images/32/32467.jpg </image>
</top>
```

In addition, 30 purely visual topics were provided in a *visual subtask* to attract more visual groups. These visual topics are, in fact, a modified subset of the 60 original topics in which non-visual features like geographic constraints or proper names were removed. Only three example images and no textual information like topic titles or narrative descriptions were provided. Section 3 provides more details about this task.

2.4 Relevance Assessments

Relevance assessments were carried out by the two topic creators[8] using a custom-built online tool. The top 40 results from all submitted runs were used to create image pools giving an average of 1,045 images (max: 1468; min: 575) to judge per topic. The topic creators judged all images in the topic pools and also used interactive search and judge (ISJ) to supplement the pools with further relevant images. The ISJ was based on purely text searches. The assessments were based on a ternary classification scheme: (1) relevant, (2) partially relevant, and (3) not relevant. Based on these judgments, only those images judged relevant by both assessors were considered for the set of relevant images (qrels).

2.5 Participating Groups and Methods

A record number of 36 groups registered for ImageCLEFphoto this year, with exactly one third of them submitting a total of 157 runs (all of which were evaluated). This is similar to last year (11 groups in 2005), although fewer runs (349 in 2005). Table 2 shows an overview of these participating groups and the number of runs submitted. New groups submitting in 2006 include Berkeley, RWTH, CINDI, TUC and CELI. All groups (with the exception of RWTH) submitted a monolingual English run with the most popular languages appearing as Italian, Japanese and Simplified Chinese.

Table 1: Participating groups.

Group ID	Institution	Runs
Berkeley	University of California, Berkeley, USA	7
CEA-LIC2M	Fontenay aux Roses Cedex, France	5
CELI	CELI srl, Torino, Italy	9
CINDI	Concordia University, Montreal, Canada	3
DCU	Dublin City University, Dublin, Ireland	40
IPAL	IPAL, Singapore	9(+4)
NII	National Institute of Informatics, Tokyo, Japan	6
Miracle	Daedalus University, Madrid, Spain	30
NTU	National Taiwan University, Taipei, Taiwan	30
RWTH	RWTH Aachen University, Aachen, Germany	2(+2)
SINAI	University of Jaén, Jaén, Spain	12
TUC	Technische Universität Chemnitz, Germany	4

A brief description of the methods of the submitted runs is provided for each group (listed alphabetically by their group ID). Participants were also asked to categorise their submissions according to the following: query language, annotation language (English or German), type (automatic or manual), use of feedback or automatic query expansion, and modality (text only, image only or combined). Table 4 shows the overall results according to runs categorised by these dimensions. Most submissions made use of the image metadata, with 8 groups submitting bilingual runs and 11 groups monolingual runs. For many participants, the main focus of their submission was combining visual and text features (11 groups text-only and 7 groups combined text and image) and/or using some kind of relevance feedback to provide query expansion (8 groups using some kind of feedback).

Berkeley. The School of Information Management and Systems of the University of California in Berkeley, USA, submitted seven runs. All runs were text only: 4 monolingual English, 2 monolingual German, one bilingual English-German. Berkeley submitted 3 runs using feedback and 3 runs using title + narrative. The retrieval algorithm used was a form of logistic regression

[8] One of the topic generators is part of the viventura travel company.

as used in TREC2 with blind relevance feedback method (10 highest weighting terms from top 10 documents). Translation was using Babelfish and expanding queries using the metadata of relevant images was found to work well. An interesting result was that using query expansion *without* any translation of terms worked surprisingly well for the bilingual run.

CEA-LIC2M. The CEA-LIC2M group from Fontenay aux Roses Cedex in France submitted five runs without using feedback or query expansion techniques. The group submitted 2 visual, 2 text, 1 mixed, 2 monolingual English and 1 bilingual French-English run. Separate initial queries were performed using the text and visual components of the topics, and then merged a-posteriori. Documents and queries are processed using a linguistic analyser to extract "concepts". Performing visual retrieval on each query image and merging results appeared to provide better results than visual retrieval with all three example images simultaneously.

CELI. The participants from CELI srl of Torino, Italy, submitted 9 text-only, automatic runs without feedback: 1 monolingual English, 8 bilingual, Italian-English and 6 runs with different query expansion techniques. Translation is achieved using bilingual dictionaries and a disambiguation approach based on Latent Semantic Analysis was implemented. Using a Boolean AND operator of the translations was found to provide higher results than using an OR operator. Results for P10 and P20 were shown to give similar results across runs compared to a more variable MAP result. The use of query expansion was shown to increase retrieval effectiveness to bridge the gap between the uncontrolled language of the query and the controlled language of the metadata.

CINDI. The CINDI group from Concordia University in Montreal, Canada, submitted 3 monolingual English runs, 2 text only, 1 mixed, 2 automatic, 1 manual, 2 with feedback (manual), 1 without feedback, 2 with query expansion and 1 without query expansion. The use of manual relevance feedback and the integration of text and image achieved the best performance for this group.

DCU. Dublin City University in Dublin, Ireland, submitted 40 automatic runs, 14 mixed, 26 text-only, 27 with feedback and 13 without feedback. DCU submitted 6 monolingual and 34 bilingual runs exploring 10 different query languages and both annotation languages. Text retrieval is performed using the BM25 weighting operator, and visual features matched using the Jeffrey Divergence function. Image retrieval on individual images was performed and merged using the CombMAX operator. Text and visual runs were fused using the weighted CombSUM operator. The results showed that fused text and image retrieval consistently outperformed text-only methods. The use of pseudo relevance feedback was also shown to improve the effectiveness of the text retrieval model.

IPAL. IPAL Singapore submitted 13 automatic runs (monolingual only): 6 visual, 4 mixed and 3 text only. Various indexing methods were tested and the XIOTA system used for text retrieval. The group used pseudo relevance feedback and an interesting feature of this was using feedback from one modality to influence the other (e.g. the result of image ranking used to drive query expansion through documents). Results indicate that the combination of text and image retrieval leads to better performance. They submitted a further 4 runs to the visual-only subtask.

NII. The National Institute of Informatics from Tokyo, Japan, submitted 6 text-only, automatic runs without feedback or query expansion, concentrating on all possibilities of three languages: English, German and Japanese: 1 monolingual English, 1 monolingual German and four bilingual runs. NII used the Lemur toolkit for text retrieval (unigram language modelling algorithm), Babelfish for translation, and a visual feature-based micro-clustering algorithm was trialled for the linking of near identical images annotated in different languages. This clustering approach did not improve retrieval effectiveness.

Miracle. The Miracle group of the Daedalus University in Madrid, Spain, submitted 30 automatic runs: 28 text only, 2 mixed and 10 runs involving query expansion based on Wordnet. The group used only the English annotations and generated 18 monolingual English runs and 12 bilingual runs (Russian, Polish, Japanese and simplified Chinese). A total of 8 runs used narrative descriptions only, 9 runs used both title and narratives and the remaining used the titles only. The most effective approach was shown to be the indexing of nouns from the image captions with no other processing.

NTU. The National Taiwan University from Taipei, Taiwan, submitted 30 automatic runs: 10 text only, 20 mixed, 12 with feedback and 18 without feedback. A total of 2 monolingual English, 2 monolingual German, 1 visual run and 25 bilingual runs (using English annotations only) exploring 10 different languages were submitted. NTU showed that the use of visual features could improve text-only retrieval based on the image annotations. A novel word-image ontology approach did not perform as well as retrieval with the image captions. Systran was used to provide translation and the initial query images were found to improve ad-hoc retrieval.

RWTH. The Human Language Technology and Pattern Recognition Group from the RWTH Aachen University in Aachen, Germany, submitted a total number of 4 entirely visual runs: 2 for the standard ad-hoc task, and 2 to the visual retrieval sub-task. Visual-only retrieval did not perform well in either task.

SINAI. The University of Jaén, Spain, submitted 12 automatic text-only runs, 8 runs with query expansion, using English annotations only. The group submitted 4 monolingual runs and 8 bilingual runs (Dutch, French, German, Italian, Portuguese and Spanish). A number of different MT systems were used for translation and the Lemur toolkit implementation of Okapi used as the retrieval model.

TUC. Technische Universität Chemnitz from Germany submitted four automatic monolingual English runs: 3 text only and 1 mixed; 3 with feedback (and query expansion) and 1 without. Combining/merging independent visual and text runs appear to give highest retrieval effectiveness, together with the use of text-based query expansion.

Table 2: Ad-hoc experiments listed by query and annotation language.

Query Language	Annotation	# Runs	# Participants
English	English	49	11
Italian	English	15	4
Japanese	English	10	4
Simplified Chinese	English	10	3
French	English	8	4
Russian	English	8	3
German	English	7	3
Spanish	English	7	3
Portuguese	English	7	3
Dutch	English	4	2
Traditional Chinese	English	4	1
Polish	English	3	1
Visual	English	1	1
German	German	8	4
English	German	6	3
French	German	3	1
Japanese	German	1	1
Visual	(none)	6	3
Visual Topics	(none)	6	2

2.6 Results and Discussion

2.6.1 Analysis of System Runs

Results for submitted runs were computed using the latest version of TREC_EVAL[9]. Submissions were evaluated using uninterpolated (arithmetic) Mean Average Precisions MAP and Precision at rank 20 (P20) because most online image retrieval engines like Google, Yahoo and Altavista display 20 images by default. Further measures considered include Geometric Mean Average Precision (GMAP) to test robustness [10], and the Binary Preference (bpref) measure which is a good indicator for the completeness of relevance judgments [1]. Using Kendall's Tau to compare system ranking between measures, we have found significant correlations at the 0.001 level between all measures above 0.74. This requires further investigation, but it would appear that the measure used to rank systems does affect the system ranking.

Table 3 shows the runs which achieved the highest MAP for each language pair. Of these runs, 83% use feedback of some kind (typically pseudo relevance feedback) and a similar proportion use both visual and textual features for retrieval. It is noticeable that submissions from NTU and DCU dominate the results (see participant's workshop papers for further information about their runs). It is interesting to note that English monolingual outperforms the German monolingual (19% lower) and the highest bilingual to English run was Portuguese-English which performed 74% of monolingual , but the highest bilingual to German run was English to German which performed only at only 39% of monolingual. Also, unlike previous years, the top-performing bilingual runs have involved Portuguese, traditional Chinese and Russian as the source language showing an improvement of the retrieval methods using these languages.

Table 3: System with highest MAP for each language.

Language (Annotation)	Group	Run ID	MAP	P20	GMAP	bpref
English (English)	CINDI	Cindi_Exp_RF	0.385	0.530	0.282	0.874
German (German)	NTU	DE-DE-AUTO-FB-TXTIMG-T-WEprf	0.311	0.335	0.132	0.974
Portuguese (English)	NTU	PT-EN-AUTO-FB-TXTIMG-T-WEprf	0.285	0.403	0.177	0.755
T. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG-TOnt-WEprf	0.279	0.464	0.154	0.669
Russian (English)	NTU	RU-EN-AUTO-FB-TXTIMG-T-WEprf	0.279	0.408	0.153	0.755
Spanish (English)	NTU	SP-EN-AUTO-FB-TXTIMG-T-WEprf	0.278	0.407	0.175	0.757
French (English)	NTU	FR-EN-AUTO-FB-TXTIMG-T-WEprf	0.276	0.416	0.158	0.750
Visual (English)	NTU	AUTO-FB-TXTIMG-WEprf	0.276	0.448	0.107	0.657
S. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG-T-WEprf	0.272	0.392	0.168	0.750
Japanese (English)	NTU	JA-EN-AUTO-FB-TXTIMG-T-WEprf	0.271	0.402	0.170	0.746
Italian (English)	NTU	IT-EN-AUTO-FB-TXTIMG-T-WEprf	0.262	0.398	0.143	0.722
German (English)	DCU	combTextVisual_DEENEN	0.189	0.258	0.070	0.683
Dutch (English)	DCU	combTextVisual_NLENEN	0.184	0.234	0.063	0.640
English (German)	DCU	combTextVisual_ENDEEN	0.122	0.175	0.036	0.524
Polish (English)	Miracle	miratctdplen	0.108	0.139	0.005	0.428
French (German)	DCU	combTextVisual_FRDEEN	0.104	0.147	0.002	0.245
Visual (none)	RWTHi6	RWTHi6-IFHTAM	0.063	0.182	0.022	0.366
Japanese (German)	NII	mcp.bl.jpn_tger_td.skl_dir	0.032	0.051	0.001	0.172

Table 4 shows results by different dimensions and shows that on average: monolingual results are higher than bilingual, retrieval using English annotations is higher than German, combined text and image retrieval is higher than text or image only, and retrieval with feedback gives higher results than without (we are currently determining statistical significance of these results). This trend has continued for the past three years (combined media and feedback runs performing the highest). Absolute retrieval results are lower than previous years and we attribute this to the choice of topics, a more visually challenging photographic collection and there being incomplete annotations provided with the collection. All groups have shown that combining visual features from the image and semantic knowledge derived from the captions offers optimum retrieval for many of the topics. In general, feedback (typically in the form of query expansion based on pseudo relevance feedback) also appears to work well on short captions (including results from previous years) and is likely due to the limited vocabulary exhibited by the captions.

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

Table 4: MAP scores for each result dimension.

Dimension	Type	# Runs	# Groups	Mean (σ)	Median	Highest
Query Language	bilingual	93	8	0.144 (0.074)	0.143	0.285
	monolingual	57	11	0.154 (0.090)	0.145	0.385
	visual	7	3	0.074 (0.090)	0.047	0.276
Annotation	English	133	11	0.152 (0.082)	0.151	0.385
	German	18	4	0.121 (0.070)	0.114	0.311
	none	6	2	0.041 (0.016)	0.042	0.063
Modality	Text Only	108	11	0.129 (0.062)	0.136	0.375
	Text + Image	43	7	0.199 (0.077)	0.186	0.385
	Image Only	6	2	0.041 (0.016)	0.042	0.063
Feedback/Expansion	without	85	11	0.128 (0.055)	0.136	0.334
	with	72	8	0.165 (0.090)	0.171	0.385

2.6.2 Analysis of Topics

Table 5 shows the average P20 and MAP scores across all runs for each topic (together with the number of relevant images per topic). There are considerable differences between topics, e.g. "photos of radio telescopes" (topic 57) has an average MAP of 0.5161; whereas "tourist accommodation near Lake Titicaca" (topic 9) has an average MAP of 0.0027. Reasons for these differences are likely due to the discriminating power of query terms in the collection, the complexity of topics (e.g. topic 9 involves a location and fuzzy spatial operator which will not be handled appropriately unless necessary support is given for spatial queries), the level of semantic knowledge required to retrieve relevant images (this will limit the success of purely visual approaches), and translation success (e.g. whether proper names have been successfully handled). Based on all results, we find the following trends according to average MAP (standard deviation):

- **Log file Analysis**. For topics taken from the log file MAP=0.1296 (0.0928); topics derived from the log file MAP=0.1155 (0.0625) and topics not taken from the log file MAP=0.2191 (0.1604). It is likely that most topics not derived from the log file are more "visual" and perhaps therefore simpler to execute.

- **Geographic Constraints**. Topics specifying specific locations and spatial operators MAP=0.1146 (0.0872); topics specifying general locations or man-made objects MAP=0.1785 (0.1111) and topics with no geography MAP=0.1313 (0.1219). Most groups did not use geographic retrieval methods.

- **Visual Features**. For topics where it is estimated visual techniques will not improve results (levels 1 and 2) MAP=0.1179 (0.1041); for topics where visual retrieval could improve results (level 3) MAP=0.1318 (0.0940) and topics where visual techniques are expected to improve results (levels 4 and 5) MAP=0.2250 (0.1094). More visual topics are likely to perform better given many participants made use of combined visual and textual approaches.

- **Topic Difficulty**. Topics rated as linguistically easy (complexity levels 1 and 2) MAP=0.1794 (0.1191); topics rated as challenging MAP=0.1107 (0.0728) and topics rated as difficult MAP=0.0234 (0.0240).

- **Annotation Quality**. Topics with all relevant images having annotations MAP=0.1668 (0.1356); topics with 80-99% of relevant images having annotations MAP=0.1290 (0.0653); topics with 60-79% of relevant images having annotations MAP=0.1353 (0.1002) and topics with 0-59% of relevant images having complete annotations MAP=0.1198 (0.1027). The use of non-text approaches is the likely cause of successful retrieval for topics with relevant images containing incomplete annotations.

We are currently investigating the effects of various retrieval strategies (e.g. use of visual and textual features) on results for different topics which will be reported in further work. We expect

that the use of visual techniques will improve topics which can be considered "more visual" (e.g. "sunset over water" is more visual than "pictures of female guides" which one could consider more semantic) and that topics which are considered "more difficult" linguistically (e.g. "bird flying" is linguistically simpler than "pictures taken on Ayers Rock") will require more complex language processing techniques.

3 The ImageCLEFphoto visual retrieval sub-task

3.1 General Overview

The ImageCLEFphoto visual retrieval sub-task offers a challenge that is similar to the general ImageCLEFphoto task: given a user information need described by three sample images, find as many relevant images as possible from a given document collection using content-based image retrieval only.

The main goal of this task is to investigate the current status quo of CBIR as regards general photographic collections, or in other words, how well CBIR techniques can, at this stage of research, handle realistic user queries on general still-natural images (in contrast to very specific tasks); it was created to further attract more visually orientated groups to ImageCLEFphoto, which was predominated by participating groups using text-orientated approaches in previous years.

3.2 Document Collection and Query Topics

The same document collection was used as with the ImageCLEFphoto task, namely the 20,000 colour photos of the *IAPR TC-12 collection*, without the corresponding image captions.

The topic creators selected 30 topics (also from the ImageCLEFphoto task) that were as collection- independent as possible, removing geographic constraints (e.g. "black and white photos" instead of "black and white photos *from Russia*") and other, non-visual constraints (e.g. "*child* wearing baseball cap" instead of "*godson* wearing baseball cap") in order to make them more visual (narrative descriptions for the relevance assessments was adjusted accordingly). Yet, the participants were only allowed to use three images representative for the textual description of each topic[10]. These 30 topics were further classified into three evenly sized groups according to how visual they were estimated to be (the same approach as described in the **Visual Features** paragraph of section 2.3).

Based on these findings, the topics were categorized into 10 *easy* topics that should do well with CBIR techniques (level > 3), 10 *hard* topics that will be quite difficult for CBIR (level ≤ 2), and 10 *medium* topics that should lie in between these two categories ($2 <$ level ≤ 3). Table 6 displays the title of the visual queries together with the average value of the individual expert judgments and the aforementioned categorisation.

3.3 Participating Groups and Methods

Two out of 12 groups that participated at the general ImageCLEFphoto task also submitted a total of six runs for the visual subtask.

IPAL. The IPAL group from Singapore submitted four slightly different runs in which only visual similarities are used: the query images and all the images of the collection were indexed with feature reduction using Latent Semantic Indexing, and the images were then ranked according to their distances to the query images.

[10]The same three sample images as in the ImageCLEFphoto retrieval task were used.

RWTHi6. The RWTHi6 group from the RWTH University Aachen, Germany, submitted two runs to the visual sub-task: one using invariant and tamura texture feature histograms which are compared using JSD, weighing IFH twice as strong as texture features based on the assumption that colour information is more important than texture information for databases of general photographs; the other one using 2048 bin histograms of image patches in colour which are compared according to their colour and texture using JSD.

3.4 Relevance Judgments and Results

The relevance judgments were performed as described in Section 2.4: the top 40 results from the six submitted runs were used to create image pools giving an average of 171 images (max: 190; min: 83) to judge per topic. The topic creators judged all images in the topic pools and also heavily used interactive search and judge (ISJ) to supplement the pools with further relevant images.

Most runs had quite promising results for precision values at a low cut-off (P20 = 0.285 for the best run, compare the results shown in Table 7). However, it is felt that this is due to the fact that some relevant images in the database are visually very similar to the query images, rather than algorithms really understanding what one is searching for. The retrieved images at higher ranks seemed to be quite random and further relevant images were only found by chance, which is also reflected by the quite low MAP scores (0.101 for the best run) and further backs up the aforementioned assumption.

3.5 Discussion

Many image retrieval systems have recently achieved decent results in retrieval tasks of quite specific domains or in tasks which are purely tailored to the current level of CBIR. The low results of the visual sub-task, however, show that content-based image retrieval is a far cry from actually bridging the semantic gap for visual information retrieval from databases of general, real-life photographs.

It has to be further investigated with the participants why only two (out of 36 registered) groups actually submitted their results. On the one hand, some groups mentioned in their feedback that they couldn't submit due to lack of time; the generally low results for this task might have also discouraged several groups from submitting their results. On the other hand, there were twice as many groups that submitted purely content-based runs to the main ImageCLEFphoto task; the question might arise whether this visual task has been promoted sufficiently enough and it should further be discussed with participants.

4 The Object Annotation Task

After the big success of the automatic medical annotation task from last year [7], which clearly showed the need for evaluation challenges in computer vision, and several demands for a similar task in a less specific domain by participants, a plan for a non-medical automatic image classification or annotation task was created. In contrast to the medical task, images to be labeled are of everyday objects and hence do not require The aim of this newly created image annotation task is to identify objects shown in images and label the image accordingly. In contrast to the PASCAL visual object classes challenge[11] [4] where several two-class experiments are performed, i.e. independent prediction of presence or absence of various object classes, here several object classes are tackled jointly.

[11] `http://www.pascal-network.org/challenges/VOC/`

4.1 Database & Task Description

LTUtech[12] kindly provided their hand collected dataset of images from 268 classes. Each image of this dataset shows one object in a rather clean environment, i.e. the images show the object and some mostly homogeneous background.

To facilitate participation in the first year, the number of classes taken into account is considerably lowered to 21 classes. The classes 1) "*ashtrays*", 2) "*backpacks*", 3) "*balls*", 4) "*banknotes*", 5) "*benches*", 6) "*books*", 7) "*bottles*", 8) "*cans*", 9) "*calculators*", 10) "*chairs*", 11) "*clocks*", 12) "*coins*", 13) "*computer equipment*", 14) "*cups and mugs*", 15) "*hifi equipment*", 16) "*cutlery(knives, forks and spoons)*", 17) "*plates*", 18) "*sofas*", 19) "*tables*", 20) "*mobile phones*", and 21) "*wallets*" are used. Removing all images that do not belong to one of these classes leads to a database of 81211 images. To create a new set of test data, 1100 new images of objects from these classes were taken. In these images, the objects are in a more "natural setting", i.e. there is more background clutter than in the training images. To simplify the classification task, it is specified in advance that each test image belongs to only one of the 21 classes. Multiple objects of the same class may appear in an image. Objects not belonging to any of the 21 classes may appear as background clutter.

The training data was released together with 100 randomly sampled test images with known classification to allow for tuning of the systems. At a later date, the remaining 1000 test images were published without their classification as test data.

The distribution of the classes is not uniform in either of these datasets. An overview of the distribution of the classes is given in Table 8 and Figure 4 gives an example from the training data and from the test data for each of the classes. From these images it can be seen that the task is hard, as the test data contains far more clutter than the training data.

4.2 Participating Groups & Methods

In total, 20 groups registered and 3 of these submitted a total of 8 runs. Here for each group a very short description of the methods of the submitted runs is provided. The groups are listed alphabetically by their group id, which is later used in the results section to refer to the groups.

CINDI. The CINDI group from Concordia University in Montreal, Canada submitted 4 runs. For their experiments they use MPEG7 edge direction histograms and MPEG7 color layout descriptors which are classified by a nearest neighbor classifier and by different combinations of support vector machines. They expect their run SVM-Product to be their best submission.

DEU. This group from the Department of Computer Engineering of the Dokuz Eylul University in Tinaztepe, Turkey submitted 2 runs. For their experiments they use MPEG7 edge direction histograms and MPEG7 color layout descriptors respectively. For classification, a nearest prototype approach is taken.

RWTHi6. The Human Language Technology and Pattern Recognition Group from the RWTH Aachen University in Aachen, Germany submitted 2 runs. For image representation they use a bag-of-features approach and for classification a discriminatively trained maximum entropy (log-linear) model is used. The runs differ with respect to the histogram bins and vector quantization methods chosen.

MedGIFT. The medGIFT group of the University and Hospitals of Geneva submitted three runs to the medical automatic annotation task. One was entirely based on tf/idf weighting of the GNU Image Finding Tool (GIFT) and thus acted as a baseline using only collection frequencies of features with no learning on the training data supplied. The other submission is a combination of several separate runs by voting. The single results were quite different, so the combination-run

[12] `http://www.ltutech.com`

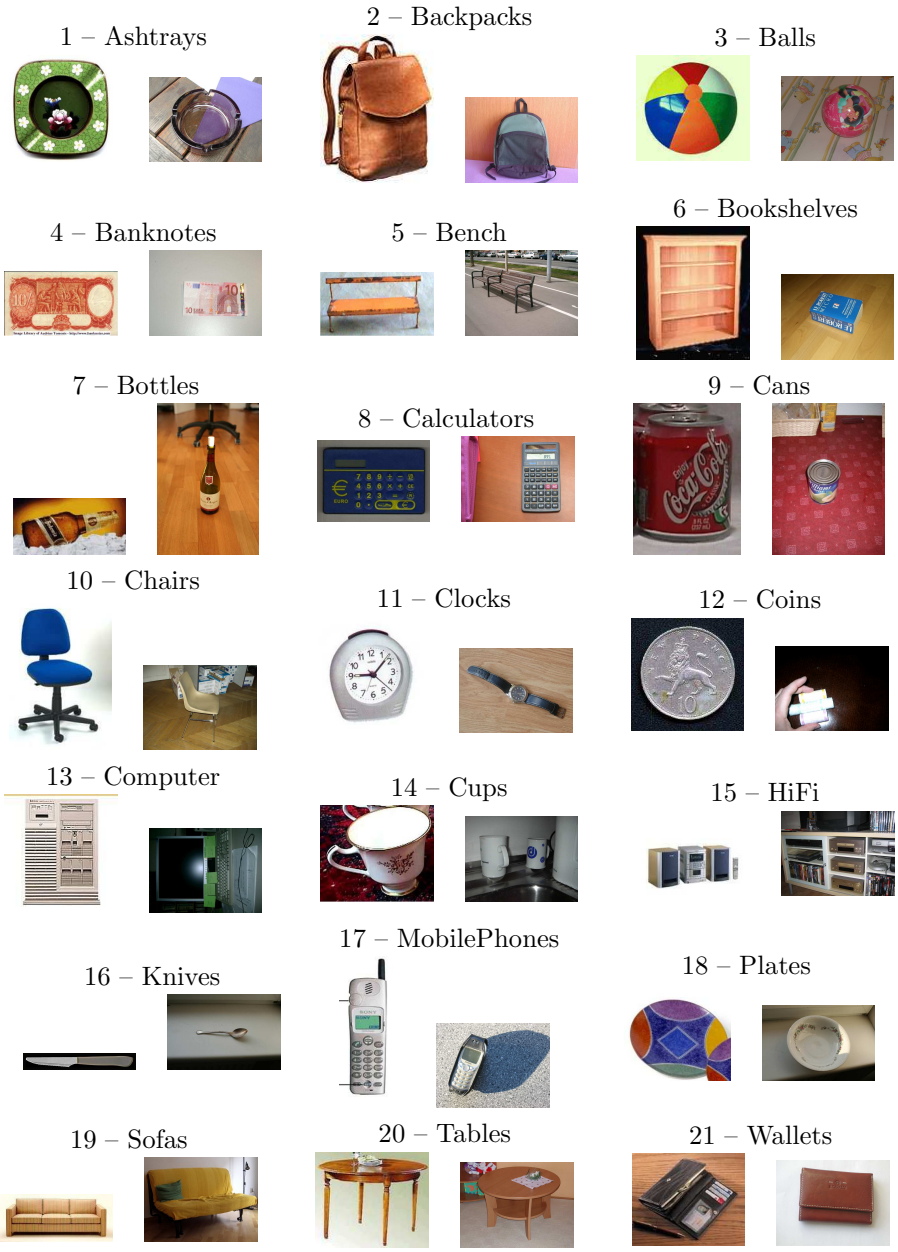


Figure 4: One image from the training (left) and the test (right) data for each of the classes of the object annotation task.

is expected to be the best submission. The runs were submitted after the evaluation ended and are thus not ranked.

4.3 Results

The results of the evaluation are given in Table 9: the runs are sorted by the error rate. Overall, the error rates are very high due to the very hard task: they range from 77.3% to 93.2%, i.e. a large part of the test images could not be classified correctly by any of the methods. Table 9 gives details how many images could be classified correctly by how many classifiers. There is no test image that was classified correctly by all classifiers, but 411 images were misclassified by all submitted runs and 301 images could be classified correctly by only one classifier.

Here too, a combination of classifiers can improve the results: Combining the first two methods by summing up normalized confidences leads to an error rate of 76.7%. Combining the three best submissions leads to an error rate of 75.8%. Adding further submissions could not improve the performance further, and combining all submissions leads to an error rate of 78.8%.

4.4 Discussion

Considering that the error rates of the submitted runs are high and that nearly half of these images could not be classified correctly by any of the submitted methods, it can be said that the the task was very challenging. One aspect that contributes to this outcome is certainly that the training images mainly contain very little clutter and that the test images are images of the objects in their "natural" environment. None of the groups specially addressed this issue although it would be exepcted to lead to improvements. Furthermore, the results show that discriminatively trained methods outperform other methods as in the medical automatic annotation task (although only a small improvement is seen and is probably not statistically significant).

The object annotation task and the medical automatic annotation task of ImageCLEF 06 [8] are very similar, but differ in some critical aspects:

- Both tasks provide a relatively large training set and a disjunct test set. Thus, in both cases it is possible to learn a relatively reliable model for the training data (this is somewhat proven for the medical annotation task)
- Both tasks are multi-class/one object per image classification tasks. Here they differ from the PASCAL visual classes challenge which addresses a set of object vs. non object tasks where several objects (of equal or unequal type) may be contained in an image.
- The medical annotation task has only gray scale images, whereas the object task has mainly color images. This is probably most relevant for the selection of descriptors.
- The images from the test and the training set are from the same distribution for the medical task, whereas for the object task, the training images are rather clutter-free and the test images contain a significant amount of clutter. This is probably relevant and should be addressed when developing methods for the non-medical task. Unfortunately, the participating methods did not address this issue which probably has a significant impact on the results.

5 Conclusions

ImageCLEF continues to provide resources to the retrieval and computational vision communities to facilitate standardised laboratory-style testing of (predominately text-based) image retrieval systems. The main division of effort thus far in ImageCLEF has been between medical and non-medical information systems. These fields have helped to attract different groups to ImageCLEF (and CLEF) over the past 2-3 years and thereby broaden the audience of this evaluation campaign. For the retrieval task, the first 2 evaluation events were based on cross-language retrieval from a cultural heritage collection: the St Andrews historic collection of photographic images. This

provided certain challenges for both the text and visual retrieval communities, most noticeably the style of language used in the captions and the types of pictures in the collection: mainly black-and-white of varying levels of quality and visual degradation. For the automatic annotation/object classification task the addition of the LTU dataset has provided a more general challenge to researchers than medical images.

For 2006, the retrieval task moved to a new collection based on feedback from ImageCLEF participants in 2005-2006 and the availability of the IAPR-TC12 Benchmark[13]. Designed specifically as a benchmark collection, it is well-suited for use in ImageCLEF with captions in multiple languages and high-quality colour photographs covering a range of topics. This type of collection - personal photographs - is likely to become of increasing interest to researchers with the growth of the desktop search market and popularity of tools such as FlickR[14].

Like in previous years, the ImageCLEFphoto task has shown the usefulness of combining visual and textual features derived from the images themselves and associated image captions. It is noticeable that, although some topics are more "visual" than others and likely to benefit more from visual techniques, the majority of topics seem to benefit from a combination of text and visual approaches and participants continue to deal with issues involved in combining this evidence. In addition, the use of relevance feedback to facilitate, for example, query expansion in text retrieval continues to improve the results of many topics in collections used so far, likely due to the nature of the text associated with images: typically a controlled vocabulary that lends itself to blind relevance feedback.

The object annotation task has shown that current approaches to image classification and/or annotation have problems with test data that is not from the same distribution as the provided training data. Given the current high interest in object recognition and annotation in the computer vision community it is to be expected that big improvements are achievable in the area of automatic image annotation in the near future. It is planned to use image annotation techniques as a preprocessing step for a multi-modal information retrieval system: given an image, create an annotation and use the image and the generated annotation to query a multi-modal information retrieval system, which is likely to improve the results given the much better performance of combined runs in the photographic retrieval task.

Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. Furthermore, special thanks go to viventura, the IAPR and LTUtech for providing their image databases for this years' tasks, and to Tobias Weyand for creating the web interface for submissions.

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

References

[1] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM Press.

[2] Cyril W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, Aslib Cranfield Research Project, Cranfield, USA, September 1962.

[13] One of the biggest factors influencing what collections are used and provided by ImageCLEF is copyright.
[14] `http://www.flickr.com`

[3] Paul Clough, Henning Müller, and Mark Sanderson. Overview of the CLEF cross–language image retrieval track (ImageCLEF) 2004. In Carol Peters, Paul D. Clough, Gareth J. F. Jones, Julio Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Result of the fifth CLEF evaluation campaign*, Lecture Notes in Computer Science, Bath, England, 2005. Springer–Verlag.

[4] Mark Everingham, Andrew Zisserman, Christopher K. I. Williams, Luc van Gool, Moray Allan, Christopher M. Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorko, Stefan Duffner, Jan Eichhorn, Jason D. R. Farquhar, Mario Fritz, Christophe Garcia, Tom Griffiths, Frederic Jurie, Daniel Keysers, Markus Koskela, Jorma Laaksonen, Diane Larlus, Bastian Leibe, Hongying Meng, Hermann Ney, Bernt Schiele, Cordelia Schmid, Edgar Seemann, John Shawe-Taylor, Amos Storkey, Sandor Szedmak, Bill Triggs, Ilkay Ulusoy, Ville Viitaniemi, and Jianguo Zhang. The 2005 pascal visual object classes challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment (PASCAL Workshop 05)*, number 3944 in Lecture Notes in Artificial Intelligence, pages 117–176, Southampton, UK, 2006.

[5] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselears. The IAPR-TC12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06*, pages 13–23, Genoa, Italy, May 2006.

[6] Michael Grubinger, Clement Leung, and Paul Clough. Linguistic estimation of topic difficulty in cross-language image retrieval. In *CLEF 2005: Overview of the Cross Language Evaluation Forum 2005*, page to appear, September 2006.

[7] Henning Müller, Antoine Geissbuhler, Johan Marty, Christian Lovis, and Patrick Ruch. The Use of medGIFT and easyIR for ImageCLEF 2005. In *Proceedings of the Cross Language Evaluation Forum 2005*, LNCS, page in press, Vienna, Austria, September 2006.

[8] Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, and William Hersh. Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In *CLEF working notes*, Alicante, Spain, September 2006.

[9] C. Peters and M. Braschler. Cross language system evaluation: The clef campaigns. *Journal of the American Society for Information Science and Technology*, 22(12):1067–1072, 2001.

[10] Ellen M. Voorhees. The trec robust retrieval track. *SIGIR Forum*, 39(1):11–20, 2005.

[11] Ellen M. Voorhees and Donna Harmann. Overview of the seventh Text REtrieval Conference (TREC–7). In *The Seventh Text Retrieval Conference*, pages 1–23, Gaithersburg, MD, USA, November 1998.

[12] Vivian Zhang, Benjamin Rey, Eugene Stipp, and Rosie Jones. Geomodification in query rewriting. In *GIR '06: Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2006*, page to appear, August 10 2006.

Table 5: ImageCLEFphoto topics and average score across all submissions.

ID	Topic Title	#Rel	AVG P20	AVG MAP
1	accommodation with swimming pool	35	0.3157	0.2208
2	church with more than two towers	27	0.0451	0.0550
3	religious statue in the foreground	32	0.1466	0.0812
4	group standing in front of mountain landscape in Patagonia	68	0.0136	0.0070
5	animal swimming	64	0.1340	0.0537
6	straight road in the USA	84	0.1380	0.1104
7	group standing in salt pan	49	0.3472	0.2083
8	host families posing for a photo	74	0.3198	0.2174
9	tourist accommodation near Lake Titicaca	13	0.0031	0.0027
10	destinations in Venezuela	36	0.3043	0.3013
11	black and white photos of Russia	65	0.1386	0.1252
12	people observing football match	31	0.1201	0.1097
13	exterior view of school building	72	0.2037	0.0907
14	scenes of footballers in action	34	0.2929	0.2629
15	night shots of cathedrals	23	0.3432	0.2924
16	people in San Francisco	54	0.2235	0.1714
17	lighthouses at the sea	27	0.3420	0.2751
18	sport stadium outside Australia	49	0.1870	0.1178
19	exterior view of sport stadia	57	0.1636	0.0922
20	close-up photograph of an animal	73	0.0559	0.0115
21	accommodation provided by host families	70	0.1963	0.1386
22	tennis player during rally	92	0.4377	0.4589
23	sport photos from California	75	0.1525	0.0662
24	snowcapped buildings in Europe	62	0.1068	0.0901
25	people with a flag	63	0.1861	0.1086
26	godson with baseball cap	79	0.1256	0.0664
27	motorcyclists racing at the Australian Motorcycle Grand Prix	30	0.2827	0.3025
28	cathedrals in Ecuador	41	0.2599	0.1195
29	views of Sydney's world-famous landmarks	40	0.1741	0.1837
30	room with more than two beds	25	0.0312	0.0290
31	volcanos around Quito	58	0.1491	0.0519
32	photos of female guides	26	0.1401	0.1065
33	people on surfboards	50	0.2000	0.1330
34	group pictures on a beach	77	0.2608	0.1092
35	bird flying	88	0.5704	0.3001
36	photos with Machu Picchu in the background	105	0.3765	0.2393
37	sights along the Inka-Trail	92	0.1910	0.0738
38	Machu Picchu and Huayna Picchu in bad weather	23	0.1077	0.0852
39	people in bad weather	72	0.0333	0.0097
40	tourist destinations in bad weather	104	0.0623	0.0157
41	winter landscape in South America	135	0.0367	0.0090
42	pictures taken on Ayers Rock	45	0.2478	0.2622
43	sunset over water	40	0.2210	0.1472
44	mountains on mainland Australia	160	0.1750	0.1093
45	South American meat dishes	41	0.2096	0.1222
46	Asian women and/or girls	41	0.2710	0.1291
47	photos of heavy traffic in Asia	35	0.0645	0.0392
48	vehicle in South Korea	33	0.0750	0.0704
49	images of typical Australian animals	99	0.1123	0.0810
50	indoor photos of churches or cathedrals	36	0.2988	0.1866
51	photos of goddaughters from Brazil	29	0.0355	0.0634
52	sports people with prizes	29	0.1392	0.0901
53	views of walls with unsymmetric stones	44	0.2941	0.2257
54	famous television (and telecommunication) towers	18	0.1210	0.1418
55	drawings in Peruvian deserts	81	0.2361	0.0958
56	photos of oxidised vehicles	28	0.0877	0.0676
57	photos of radio telescopes	10	0.3006	0.5161
58	seals near water	56	0.4216	0.2222
59	creative group pictures in Uyuni	24	0.0627	0.0532
60	salt heaps in salt pan	28	0.4040	0.2952

Table 6: The visual topics and the three categories: easy, medium and hard.

ID	Topic Title	Level
82	sunset over water	4.75
66	black and white photos	4.25
88	drawings in deserts	4.00
71	tennis player on tennis court	3.75
78	bird flying	3.25
85	photos of dark-skinned girls	3.25
86	views of walls with asymmetric stones	3.25
68	night shots of cathedrals	3.25
64	straight road	3.25
72	snowcapped buildings	3.25
67	scenes of footballers in action	3.00
74	motorcyclists riding on racing track	3.00
76	people on surfboards	3.00
63	animal swimming	2.75
69	lighthouses at the sea	2.75
77	group pictures on a beach	2.75
81	winter landscape	2.75
90	salt heaps in salt pan	2.75
65	group standing in salt pan	2.50
84	indoor photos of churches or cathedrals	2.25
79	photos with Machu Picchu in the background	2.00
80	Machu Picchu and Huayna Picchu in bad weather	2.00
62	group in front of mountain landscape	2.00
70	close-up photograph of an animal	2.00
83	images of typical Australian animals	1.75
87	television and telecommunication towers	1.75
89	photos of oxidised vehicles	1.75
73	child wearing baseball cap	1.50
75	exterior view of churches or cathedrals	1.50
61	church with more than two towers	1.25

Table 7: The visual results.

RK	RUN ID	MAP	P20	BPREF	GMAP
1	RWTHi6-IFHTAM	0.1010	0.2850	0.4307	0.0453
2	RWTHi6-PatchHisto	0.0706	0.2217	0.3831	0.0317
3	IPAL-LSA3-VisualTopics	0.0596	0.1717	0.3360	0.0281
4	IPAL-LSA2-VisualTopics	0.0501	0.1800	0.3093	0.0218
5	IPAL-LSA1-VisualTopics	0.0501	0.1650	0.3123	0.0236
6	IPAL-MF-VisualTopics	0.0291	0.1417	0.2374	0.0119

	class	train	dev	test
1	Ashtrays	300	1	24
2	Backpacks	300	3	28
3	Balls	320	3	10
4	Banknotes	306	4	45
5	Bench	300	1	44
6	Books	604	5	65
7	Bottles	306	9	95
8	Calculators	301	1	14
9	Cans	300	0	20
10	Chairs	320	10	132
11	Clocks	1833	2	47
12	Coins	310	0	26
13	Computing equipment	3923	10	79
14	Cups	600	12	108
15	HiFi	1506	2	24
16	Cutlery	912	12	86
17	Mobile Phones	300	6	39
18	Plates	302	9	52
19	Sofas	310	3	22
20	Tables	310	2	23
21	Wallets	300	5	17
	sum	13963	100	1000

Table 8: Overview of the data of the object annotation task.

Table 9: Results from the object annotation task sorted by error rate.

rank	Group ID	Runtag	Error rate
1	RWTHi6	SHME	77.3
2	RWTHi6	PatchHisto	80.2
3	cindi	Cindi-SVM-Product	83.2
4	cindi	Cindi-SVM-EHD	85.0
5	cindi	Cindi-SVM-SUM	85.2
6	cindi	Cindi-Fusion-knn	87.1
7	DEU-CS	edgehistogr-centroid	88.2
-	medGIFT	fw-bwpruned	90.5
-	medGIFT	baseline	91.7
8	DEU-CS	colorlayout-centroid	93.2

Table 10: The number of test images that were correctly classified by the specified number of runs

number of images	number of runs in which correctly classified
411	0
301	1
120	2
69	3
54	4
30	5
13	6
2	7
0	8