Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks

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

[1] Sheffield University, Sheffield, UK
[2] Victoria University, Melbourne, Australia
[3] Computer Science Department, RWTH Aachen University, Germany
[4] Vienna University of Technology, Vienna, Austria
[5] University and Hospitals of Geneva, Switzerland
p.d.clough@sheffield.ac.uk

Abstract. This paper describes the general photographic retrieval and object annotation tasks of the ImageCLEF 2006 evaluation campaign. These tasks provided 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 offered something new for 2006 and attracted a large number of submissions: 12 groups participated in ImageCLEFphoto and 3 groups in the automatic annotation task. This paper summarises these two tasks including collections used in the benchmark, the tasks proposed, a summary of submissions from participating groups and the main findings.

1 The Photographic Retrieval Task: ImageCLEFphoto

The ImageCLEFphoto task provides the resources for the comparison of system performance in a laboratory-style setting. This kind of evaluation is system-centred and similar to the classic TREC[1] (Text REtrieval Conference [1]) ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but search topics are not known to the system in advance. Evaluation aims to compare algorithms and systems, and not assess aspects of user interaction (iCLEF addresses this). The specific goal of ImageCLEFphoto is: given a statement describing a user information need, find as many relevant images as possible from the given document collection (with the query in a language different from that used to describe the images). After three years of evaluation using the St. Andrews database [2], a new database was used in this year's task: the *IAPR TC-12 Benchmark* [3], created under Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR[2]). This collection differs from the

[1] http://trec.nist.gov/
[2] http://www.iapr.org/

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 579–594, 2007.

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).

1.1 Document Collection

The ImageCLEFphoto collection contains 20,000 photos taken from locations around the world, comprising a varying cross-section of still natural images on a variety of topics (Fig. 1 shows some examples). The majority of images have been provided by *viventura*[3], an independent travel company organising adventure and language trips to South America. Travel guides accompanying the tourists maintain a daily online diary including photographs of the trips made and general pictures of each location. For example, pictures include accommodation, facilities, people and social projects. The remainder of the images have been collected by the second author over the past few years from personal experiences (e.g. holidays and events). The collection is publicly available for research purposes and unlike many existing photographic collections used to evaluate image retrieval systems, this collection is very general in content. 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 techniques involving visual analysis.

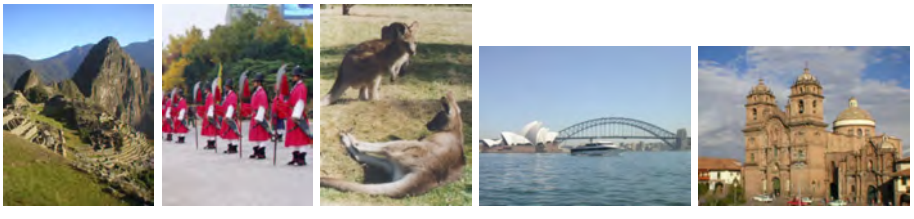

Fig. 1. Sample images from the IAPR TC-12 collection

The content of the collection is varied (and realistic) and associated descriptive annotations have been carefully created and applied in a systematic manner (e.g. all fields contain values of a similar style and format) to all images. Each image in the collection has a corresponding semi-structured caption consisting of the following seven fields (similar to the previous St Andrews collection): (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. Although consistent and careful annotations are typically not found in

[3] `http://www.viventura.de`

practice, the goal of creating this resource was to provide a general-purpose collection which could be used for a variety of research purposes. For example, this year we decided to create a more realistic scenario for participants by releasing a version of the collection with a varying degree of annotation "completeness" (i.e. with different caption fields available for indexing and retrieval). For 2006, the collection contained the following levels of annotation: (a) 70% of the annotations contain title, description, notes, location and date; (b) 10% of the annotations contain title, location and date; (c) 10% of the annotations contain location and date; and (d) 10% of the images are not annotated (or have empty tags respectively).

1.2 Query Topics

Participants were given 60 topics representing typical search requests for this document collection. The topic creation process is an important aspect of a test collection as one must aim to create a balanced and representative set of information needs. Topics for ImageCLEFphoto were derived from analysing the log file[4] from a web-based interface to the image collection which is used by employees and customers of viventura. Domain knowledge from the authors was also used. To provide an element of control over the topics, the final set given to participants were based on considering a number of parameters including: geographical constraint, "visualness" of the topic, an estimation of linguistic difficulty, degree of annotation completeness and the estimated number of relevant images. Topic creators aimed for a target set size of between approximately 20 to 100 relevant images and thus had to further modify some of the topics (broadening or narrowing the concepts).

For many of the topics, successful retrieval using text-based IR methods required the use of query analysis (e.g. expansion of query terms or logical inference). These reflected examples found in the log files, e.g. for the query "group pictures on a beach", many of the annotations would not use the term "group" but rather terms such as "men" and "women" or the names of individuals. Similarly for the query "accommodation with swimming pool" (also from the log file), the query would result in limited effectiveness unless "accommodation" was expanded to terms such as "hotel" and "B&B". Queries such as "images of typical Australian animals" required a higher level of inference and access to world knowledge (this query was not found in the log file but could be a feasible request by users of an image retrieval system).

Similar to previous analyses of search log files (see, e.g. [4]), we found many search requests to exhibit some kind of geographical constraint (e.g. specifying a location). Therefore, we created 24 topics with a geographic constraint (e.g. "tourist accommodation *near Lake Titicaca*"), 20 topics with a geographic feature or a permanent man-made object (e.g. "group standing in *salt pan*") and 16 topics with no geography (e.g. "photos of female guides"). All topics were classified regarding how "visual" they were considered to be. An average rating

[4] Log file from 1st February–15th April 2006 containing 980 unique queries.

between 1-5[5] 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[6]. A total of 30 topics were classed as "semantic" (levels 1 and 2) for which visual approaches would be highly unlikely to improve results (e.g. "cathedrals in Ecuador"); 20 topics classified as "neutral" (level 3) for which visual approaches may or may not improve results (e.g. "group pictures on a beach") and 10 were "visual" for which content-based approaches would be most likely to improve retrieval results (e.g. "sunset over water"). To consider topics from a linguistic viewpoint, a complexity measure was used to categorise topics according to their linguistic difficulty [5]. A total of 31 "easy" topics were selected (levels 1 and 2, e.g. "bird flying"), 25 "medium–hard" (level 3, e.g. "pictures taken on Ayers Rock"), and 4 "difficult" topics (levels 4 and 5, e.g. "tourist accomodation near Lake Titicaca"). Various aspects of text retrieval on a more semantic level were considered too, concentrating on vocabulary mismatches, general versus specific concepts, ambiguous terms and use of abbreviations.

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 (these images were not removed from the collection, but removed from the set of relevance judgments). 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.

1.3 Relevance Assessments

Relevance assessments were carried out by two topic creators[7] 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 (on average 25%). The ISJ was based on purely textual 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).

[5] 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.

[6] The FIRE system was used based on using all query images `http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html`

[7] One of the topic generators a member of the viventura travel company.

1.4 Participating Groups and Methods

A total 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). Of the 12 participating groups, four of them were new to ImageCLEF: Berkeley, CINDI, TUC and CELI. Table 1 summarises participating groups and the number of runs submitted by them. All groups (with the exception of RWTH) submitted a monolingual English run with the most popular languages appearing as Italian, Japanese and Simplified Chinese (see Table 2).

Table 1. Participating groups for ImageCLEFphoto

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

Overall 157 runs were submitted using a variety of approaches. Participants were asked to categorise their submissions according to the following dimensions: query language, annotation language (English or German), run 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 annotations (or metadata), with 8 groups submitting bilingual runs and 11 groups monolingual runs (many groups used MT systems for translation, e.g. Berkeley, DCU, NII, NTU and SINAI). For many participants, the main focus of their submission was combining visual and text features (11 groups submitted text-only runs; 7 used a combination of text and visual information: CEA, CINDI, DCU, IPAL, Miracle, NTU and TUC) and/or using some kind of relevance feedback to provide query expansion (8 groups using some kind of feedback: Berkeley, CINDI, DCU, IPAL, Miracle, NTU, SINAI and TUC).

Based on all submitted runs, 59% were bilingual (85% X-English; 15% X-German), 31% involved the use of image retrieval (27% using combined visual and textual features) and 46% of runs involved some kind of relevance feedback (typically in the form of query expansion). The use of query expansion was shown to increase retrieval effectiveness by bridging the gap between the languages of the query and annotations. The majority of runs were automatic (i.e. involving no human intervention); 1 run was manual. Further details of methods used in the submitted runs can be found in the workshop papers submitted by participants.

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

1.5 Results and Discussion

Analysis of System Runs. Results for submitted runs were computed using version 7.3 of `trec_eval`[8]. Submissions were evaluated using uninterpolated (arithmetic) Mean Average Precision (MAP) and Precision at rank 20 (P20). We also considered Geometric Mean Average Precision (GMAP) to test robustness [6]. Using Kendall's Tau to compare system ranking between measures, significant correlations existed at $p \leq 0.01$ between all measures above 0.74. This implies that the measure used to rank systems *does* affect system ranking and requires further investigation.

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 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 involved Portuguese, traditional Chinese and Russian as the source language showing an improvement of the retrieval methods using these languages.

Table 4 shows results by different dimensions and indicates that, on average, combining visual features from the image and semantic information from the annotations gives a 54% improvement over text alone, using some kind of feedback (visual and textual) gives a 39% improvement, bilingual retrieval performs 7% lower than monolingual and that results for English as the target language are 26% higher than those for German (differences are statistically significant at $p \leq 0.05$ using a Student's t-Test). The differences are less impressive, however,

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

Table 3. Systems with highest MAP for each query language (ranked by descending order of MAP scores)

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

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

if the *highest* MAP score is considered: the highest bilingual run performs 35% lower than the monolingual, combining text and image gives only 3% increase and using feedback of some kind gives a 15% increase.

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.

Analysis of Topics. There are considerable differences between the retrieval effectiveness of individual topics. Possible causes for these differences include: the discriminating power of query terms in the collection, the complexity of topics

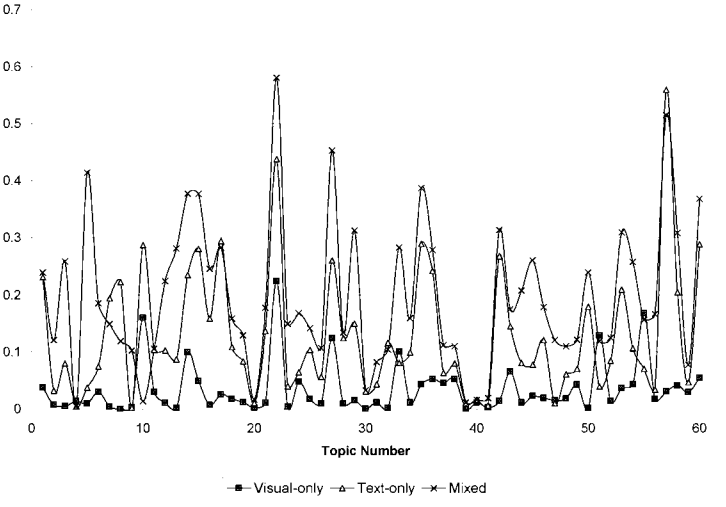

Fig. 2. MAP by topic based on modality

(e.g. topic 9 "Tourist accommodation near Lake Titicaca" involves a location and fuzzy spatial operator which would not be handled appropriately unless support for spatial queries was provided), the level of semantic knowledge required to retrieve relevant images (this limits the success of purely visual approaches), and translation success (e.g. whether proper names had been successfully handled). Fig. 2 shows mean average precision across (all) system runs for each topic based on modality. Many topics clearly show improvement through the use of combining textual and visual features (mixed) than any single modality alone. Part of this is likely to be attributed to the availability of visual examples with the topics which could be used in the mixed runs (and that these examples were also present in the collection).

We are still investigating the effects of various retrieval strategies (e.g. use of visual and textual features and relevance feedback) on the results for different topics. 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.

1.6 The ImageCLEFphoto Visual Retrieval Sub-task

To investigate further the success of visual techniques, thirty topics from the ImageCLEFphoto task were selected and modified to reduce semantic information and make better suited to visual retrieval techniques. For example removing

Table 5. The visual results

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

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"). We wanted to attract more visually orientated groups to ImageCLEFphoto which to date has been dominated by groups using textual approaches.

The same document collection was used as with the ImageCLEFphoto task, but without the corresponding image captions. Participants were given three example images to describe each topic and were required to perform query-by-visual-example retrieval to begin the search. Two out of 12 groups that participated at the general ImageCLEFphoto task also submitted a total of six runs for the visual subtask: IPAL and RWTH.

Relevance judgments were performed as described in Section 1.3: 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 used (text-based) 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 5). 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 is being searched for. The retrieved images at higher ranks appeared random and further relevant images were only found by chance. This is also reflected by the low MAP scores (0.101 for the best run). Image retrieval systems typically achieve good results for tasks based on specific domains, or in tasks which are well-suited to the current level of CBIR. The low results of the visual sub-task highlight the fact that the successful application of visual techniques in applications which involve more general (and less domain-specific) pictures is still requiring much investigation.

It has to be further investigated with the participants why only two (out of 36 registered) groups actually submitted visual-only results. On the one hand, some groups mentioned in their feedback that they could not 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.

2 The Object Annotation Task

After the success of the automatic medical annotation task in 2005 [7] that clearly showed the need for public evaluation challenges in computer vision, and several calls for a more general annotation task from ImageCLEF 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 were of everyday objects and hence did not require specialised domain knowledge. The aim of this new annotation task was to identify objects shown in test images and label the image accordingly. In contrast to the PASCAL visual object classes challenge[9] [8] where several two-class experiments are performed, i.e. independent prediction of presence or absence of various object classes, in this task several object classes were tackled jointly.

2.1 Database and Task Description

LTUtech[10] 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 were reduced to 21. 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*" were used. Removing all images that did not belong to one of these classes lead to a database of 13,963 images. To create a new set of test data, 1,100 new images of objects from these classes were taken. In these images, the objects were in a more "natural setting", i.e. there was more background clutter than in the training images. To simplify the classification task, it was specified in advance that each test image belonged 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 classifications to allow for the tuning of system parameters. Following this, the remaining 1000 images were published unclassified as the test data.

The distribution of classes in both the training and test data was non–uniform. Table 6 summaries the distributions and Figure 3 shows examples from the training and test data for each of the classes. It can be seen from the images that the task is hard. This is because the test data contains far more clutter than the training data.

[9] `http://www.pascal-network.org/challenges/VOC/`
[10] `http://www.ltutech.com`

Table 6. Overview of the data of the object annotation task

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

2.2 Participating Groups and Methods

20 groups registered for the general annotation task and 3 of these submitted a total of 8 runs. Details about each of the participating groups and their submissions is provided in the following text (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 used MPEG7 edge direction histograms and MPEG7 color layout descriptors, classified by a nearest neighbour classifier and by different combinations of support vector machines. They expected their run `SVM-Product` to be their best submission.
- *DEU.* The DEU group from the Department of Computer Engineering of the Dokuz Eylul University in Tinaztepe, Turkey, submitted 2 runs. For their experiments they used MPEG7 edge direction histograms and MPEG7 colour layout descriptors respectively. For classification, a nearest prototype approach was 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 used a bag-of-features approach and for classification a discriminatively trained maximum entropy (log-linear) model was used. Their runs differed with respect to the histogram bins and vector quantization methods chosen.
- *MedGIFT.* The MedGIFT group of the University and Hospitals of Geneva, Switzerland, submitted two runs. One run was entirely based on tf/idf weighting of the GNU Image Finding Tool (GIFT). This acted as a baseline using only collection frequencies of features with no learning on the training data

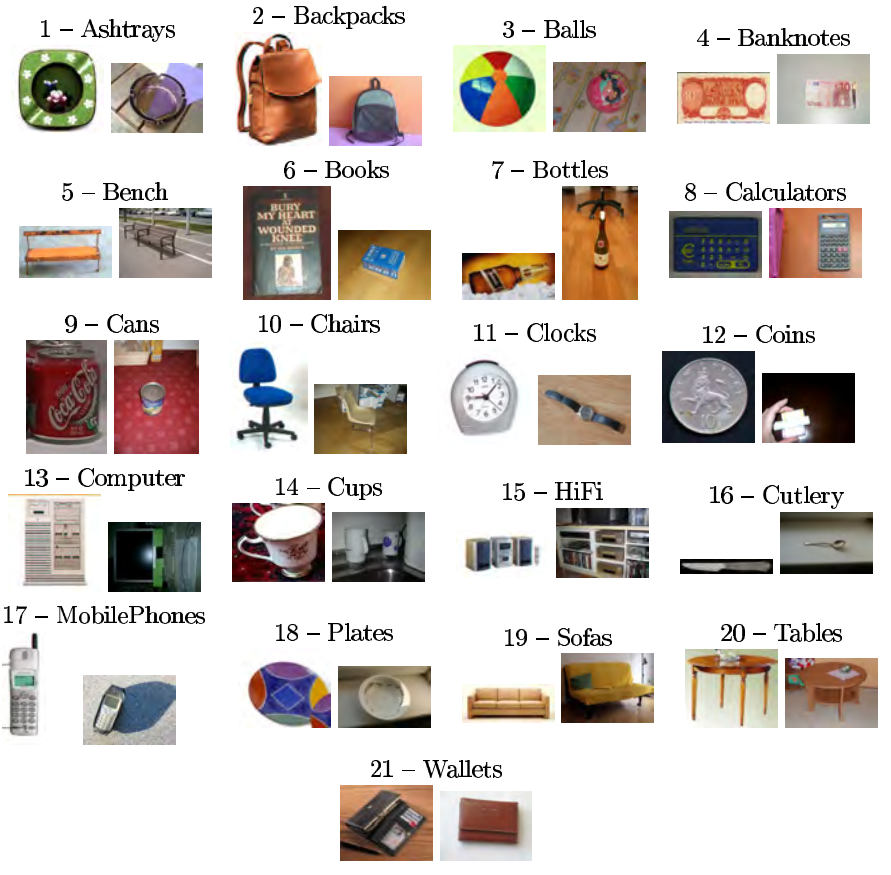

Fig. 3. One image from training (left) and test (right) data for each of the classes

supplied. The other submission was a combination of several separate runs by voting. The single results were quite different, so the combination-run was expected to be the best submission. The runs were submitted after the evaluation ended and therefore not evaluated (or ranked).

2.3 Results

The results of the evaluation are shown in Table 7. The runs are sorted by error rate. Overall, the error rates are very high due to the difficulty of the task. Scores range from 77.3% to 93.2%, indicating that many of the test images could not be classified correctly using any method. Table 7 provides details on the number of correctly classified images. None of the test images were classified correctly by all classifiers; 411 images were misclassified by all submitted runs and 301 images could be classified correctly by only one classifier.

Table 7. 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 8. The number of test images correctly classified by the 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

We have found that a combination of classifiers can improve results. Using the first two methods and summing up normalized confidences leads to an error rate of 76.7%, and using the three best submissions leads to an error rate of 75.8%. Adding further submissions could not improve the performance further. Combining all submissions lead to an error rate of 78.8%.

2.4 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 task was very challenging. One contributing factor was that the training images generally contained very little clutter and obscuring features, whereas the test images showed objects in their "natural" (and more realistic) environment. None of the participating groups specifically addressed this issue, although it would be exepcted to lead to improvement in classification accuracy. Furthermore, the results show that discriminatively trained methods outperform other methods (similar to results from the medical automatic annotation task), although only by a small amount that is probably not statistically significant.

Although the object annotation task and the medical automatic annotation tasks of ImageCLEF 2006 [9] were similar, they also differed in four important ways:

- Both tasks provided 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 were multi-class/one object per image classification tasks. Here they differ from the PASCAL visual classes challenge which has addressed 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 were from the same distribution for the medical task, whereas for the object task the training images are rather clutter-free and the test images contained 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.

3 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 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[11]. 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[12].

Like in previous years, the ImageCLEFphoto task has shown the usefulness of combining visual and textual features derived from the images themselves

[11] One of the biggest factors influencing what collections are used and provided by ImageCLEF is copyright.

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

and associated image captions (although to a lesser degree this year). 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.

For the automatic annotation/object classification task the addition of the LTU dataset has provided a more general challenge to researchers than medical images. 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 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, the EU SemanticMining project (IST NoE 507505) and the EU MUSCLE NoE.

References

1. Voorhees, E.M., Harmann, D.: Overview of the seventh Text REtrieval Conference (TREC-7). In: The Seventh Text Retrieval Conference, Gaithersburg, MD, USA, pp. 1–23 (1998)
2. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)

3. Grubinger, M., Clough, P., Müller, H., Deselears, T.: 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, Genoa, Italy, pp. 13–23 (2006)
4. Zhang, V., Rey, B., Stipp, E., Jones, R.: Geomodification in query rewriting. In: GIR '06: Proceedings of the Workshop on Geographic Information Retrieval, SIGIR (2006)
5. Grubinger, M., Leung, C., Clough, P.D.: Linguistic estimation of topic difficulty in cross-language image retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 558–566. Springer, Heidelberg (2006)
6. Voorhees, E.M.: The trec robust retrieval track. SIGIR Forum 39, 11–20 (2005)
7. Müller, H., Geissbuhler, A., Marty, J., Lovis, C., Ruch, P.: The Use of medGIFT and easyIR for ImageCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J.D.R., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 pascal visual object classes challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 117–176. Springer, Heidelberg (2006)
9. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: CLEF working notes, Alicante, Spain (2006)