Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task

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

[1] Victoria University, Melbourne, Australia
[2] Sheffield University, Sheffield, United Kingdom
[3] Vienna University of Technology, Vienna, Austria
[4] University and Hospitals of Geneva, Switzerland
[5] University of Applied Sciences, Sierre, Switzerland
michael.grubinger@research.vu.edu.au

Abstract. The general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign is described. This task provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval from generic photographic collections. In 2007, the evaluation objective concentrated on retrieval of lightly annotated images, a new challenge that attracted a large number of submissions: a total of 20 participating groups submitted 616 system runs. This paper summarises the components used in the benchmark, including the document collection and the search tasks, and presents an analysis of the submissions and the results.

1 Introduction

ImageCLEFphoto 2007 provides a system-centered evaluation for multilingual visual information retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections). The evaluation scenario is similar to the classic TREC[1] 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) [1]. The goal of the simulation is: given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from the given collection (with the query language either being identical or different from that used to describe the images).

The objective of *ImageCLEFphoto 2007* comprised the evaluation of multilingual visual information retrieval from a generic collection of lightly annotated photographs (*i.e.* containing only short captions such as the title, location, date or additional notes, but without a semantic description of the photograph). This new challenge allows for the investigation of the following research questions:

[1] http://trec.nist.gov/

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 433–444, 2008.

- Are traditional text retrieval methods still applicable for such short captions?
- How significant is the choice of the retrieval language?
- How does the retrieval performance compare to retrieval from collections containing fully annotated images (*ImageCLEFphoto 2006*)?

One major goal of *ImageCLEFphoto 2007* was to attract more content-based image retrieval approaches, as most of the retrieval approaches in previous years had predominately been concept-based. The reduced alphanumeric semantic information provided with the image collection should support this goal as content-based retrieval techniques become more significant with increasingly reduced image captions.

2 Methods

Similar to *ImageCLEFphoto 2006* [2], we generated a subset of the *IAPR TC-12 Benchmark* to provide the evaluation resources for *ImageCLEFphoto 2007*. This section provides more information on these individual components: the document collection, the query topics, relevance judgments and performance indicators.

2.1 Document Collection

The document collection of *IAPR TC-12 Benchmark* contains 20,000 colour photos taken from locations around the world and comprises a varying cross-section of still natural images. More information on the design and implementation of test collection, created under *Technical Committee 12* (TC-12) of the *International Association of Pattern Recognition* (IAPR[2]), can be found in [3].

```
<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>
```

Fig. 1. Sample image caption

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

[2] http://www.iapr.org/

(6) where and (7) when the photo was taken. Figure 2.1 shows a sample image with its corresponding English annotation.

These annotations are stored in a database, allowing the creation of collection subsets with respect to a variety of particular parameters (*e.g.* which caption fields to use). Based on the feedback from participants of previous evaluation tasks, the following was provided for *ImageCLEFphoto 2007*:

- **Annotation language:** four sets of annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images.
- **Caption fields:** only the fields for the *title*, *location*, *date* and additional *notes* were provided. Unlike 2006, the *description* field was not made available for retrieval to provide a more realistic evaluation scenario and to attract more visually oriented retrieval approaches.
- **Annotation completeness:** each image caption exhibited the same level of annotation completeness - there were no images without annotations as in 2006.

2.2 Query Topics

The participants were given 60 query topics (see Table 1) representing typical search requests for the generic photographic collection of the *IAPR TC-12 Benchmark*.

These topics had already been used in 2006, and we decided to reuse them to facilitate the objective comparison of retrieval from a generic collection of fully annotated (2006) and lightly annotated (2007) photographs. The creation of these topics is based on several factors (see [4] for detailed information), including:

- the analysis of a log file from online-access to the image collection;
- knowledge of the contents of the image collection;
- various types of linguistic and pictorial attributes;
- the use of geographical constraints;
- the estimated difficulty of the topic.

Similar to TREC, the query topics were provided as structured statements of user needs which consist of a title (a short sentence or phrase describing the search request in a few words) and three sample images that are relevant to that search request. These images were removed from the test collection and did not form part of the ground-truth in 2007.

The topic titles were offered in 16 languages including English, German, Spanish, Italian, French, Portuguese, Chinese, Japanese, Russian, Polish, Swedish, Finnish, Norwegian, Danish, and Dutch, whereby all translations had been provided by at least one native speaker and verified by at least another native speaker. The participants only received the topic titles, but not the narrative descriptions to avoid misunderstandings as they had been misinterpreted by participants in the past (they only serve to unambiguously define what constitutes a relevant image or not).

Table 1. *ImageCLEFphoto 2007* topics

ID	Topic Title	ID	Topic Title
1	accommodation with swimming pool	31	volcanos around Quito
2	church with more than two towers	32	photos of female guides
3	religious statue in the foreground	33	people on surfboards
4	group standing in front of mountain landscape in Patagonia	34	group pictures on a beach
5	animal swimming	35	bird flying
6	straight road in the USA	36	photos with Machu Picchu in the background
7	group standing in salt pan	37	sights along the Inca-Trail
8	host families posing for a photo	38	Machu Picchu and Huayna Picchu in bad weather
9	tourist accommodation near Lake Titicaca	39	people in bad weather
10	destinations in Venezuela	40	tourist destinations in bad weather
11	black and white photos of Russia	41	winter landscape in South America
12	people observing football match	42	pictures taken on Ayers Rock
13	exterior view of school building	43	sunset over water
14	scenes of footballers in action	44	mountains on mainland Australia
15	night shots of cathedrals	45	South American meat dishes
16	people in San Francisco	46	Asian women and/or girls
17	lighthouses at the sea	47	photos of heavy traffic in Asia
18	sport stadium outside Australia	48	vehicle in South Korea
19	exterior view of sport stadia	49	images of typical Australian animals
20	close-up photograph of an animal	50	indoor photos of churches or cathedrals
21	accommodation provided by host families	51	photos of goddaughters from Brazil
22	tennis player during rally	52	sports people with prizes
23	sport photos from California	53	views of walls with asymmetric stones
24	snowcapped buildings in Europe	54	famous television (and telecommunication) towers
25	people with a flag	55	drawings in Peruvian deserts
26	godson with baseball cap	56	photos of oxidised vehicles
27	motorcyclists racing at the Australian Motorcycle Grand Prix	57	photos of radio telescopes
28	cathedrals in Ecuador	58	seals near water
29	views of Sydney's world-famous landmarks	59	creative group pictures in Uyuni
30	room with more than two beds	60	salt heaps in salt pan

The participants were also given access to the results of a visual baseline run for each topic, provided by the FIRE system. The run thereby used colour histograms (compared with JSD, weight 3), Tamura texture histograms (compared with JSD, weight 2), and 32x32 thumbnails (compared with Euclidean distance, weight 1). More information on FIRE can be found in [5].

2.3 Relevance Assessments

Relevance assessments were carried out by the two topic creators using a custom-built online tool. The top 40 results from all submitted runs were used to create image pools giving an average of 2,299 images (max: 3237; min: 1513) 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 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 sets of relevant images (qrels).

Finally, these qrels were complemented with the relevant images found at *ImageCLEFphoto 2006* in order to avoid missing out on relevant images not found this year due to the reduced captions.

2.4 Result Generation

Once the relevance judgments were completed, we were able to evaluate the performance of the individual systems and approaches. The results for submitted runs were computed using the latest version of trec_eval[3] (Version 8.1).

The 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 system robustness, and the *binary preference* (bpref) measure which is an indicator for the completeness of relevance judgments.

3 Participation and Submission Overview

ImageCLEFphoto 2007 saw the registration of 32 groups (4 less than in 2006), with 20 of them eventually submitting 616 runs (all of which were evaluated). This is an increase in comparison to previous years (12 groups submitting 157 runs in 2006, and 11 groups submitting 349 runs in 2005 respectively).

Table 2. Participating groups

Group ID	Institution	Runs
Alicante	University of Alicante, Spain	6
Berkeley	University of California, Berkeley, USA	19
Budapest	Hungarian Academy of Sciences, Budapest, Hungary	11
CINDI	Concordia University, Montreal, Canada	5
CLAC	Concordia University, Montreal, Canada	6
CUT	Technical University Chemnitz, Germany	11
DCU-UTA	Dublin City University, Dublin, Ireland & University of Tampere, Finland	138
GE	University and Hospitals of Geneva, Switzerland	2
HongKong	Nanyang Technological University, Hong Kong	62
ImpColl	Imperial College, London, UK	5
INAOE	INAOE, Puebla, Mexico	115
IPAL	IPAL, Singapore	27
Miracle	Daedalus University, Madrid, Spain	153
NII	National Institute of Informatics, Tokyo, Japan	3
RUG	University of Groningen, The Netherlands	4
RWTH	RWTH Aachen University, Germany	10
SIG	Universite Paul Sabatier, Toulouse, France	9
SINAI	University of Jaén, Jaén, Spain	15
Taiwan	National Taiwan University, Taipei, Taiwan	27
XRCE	Cross-Content Analytics, Meylan, France	8

Table 2 provides an overview of the participating groups and the corresponding number of submitted runs. The 20 groups are from 16 countries, with one

[3] http://trec.nist.gov/trec_eval/

institution (Concordia University) sending two separate groups (CINDI, CLAC), while DCU and UTA joined forces and submitted as one participating group. New participants submitting in 2007 include Budapest, CLAC, UTA, NTU (Hongkong), ImpColl, INAOE, RUG, SIG and XRCE. The number of runs per participating group has risen as well, with participants submitting an average of 30.8 runs in 2007 (13.1 runs in 2006). However, this may be attributed to the fact that four sets of annotations were offered (compared to two in 2007) and that the participants were allowed to submit as many runs as they desired.

The runs submitted were categorised with respect to the following dimensions: query and annotation language, run type (automatic or manual), use of relevance feedback or automatic query expansion, and modality (text only, image only or combined). Most submissions (91.6%) used the image annotations, with 8 groups submitting a total of 312 bilingual runs and 18 groups a total of 251 monolingual runs; 15 groups experimented with purely concept-based (textual) approaches (288 runs), 13 groups investigated the combination of content-based (visual) and concept-based features (276 runs), while a total of 12 groups submitted 52 purely content-based runs, an increase in comparison with previous events (in 2006, only 3 groups had submitted a total of 12 visual runs). Furthermore, 53.4% of all retrieval approaches involved the use of image retrieval (31% in 2006).

Based on all submitted runs, 50.6% were bilingual (59% in 2006), 54.7% of runs used query expansion and pseudo-relevance feedback techniques (or both) to further improve retrieval results (46% in 2006), and most runs were automatic (*i.e.* involving no human intervention); only 3.1% of the runs submitted were manual. Two participating groups made use of additional data (*i.e.* the description field and the qrels) from *ImageCLEFphoto 2006*. Although all these runs were evaluated (indicated by "Data 2006"), they were not considered for the system performance analysis and retrieval evaluation described in Section 4.

Table 3 displays the number of runs (and participating groups in parenthesis) with respect to query and annotation languages. The majority of runs (66.2%) was concerned with retrieval from English annotations, with exactly half of them (33.1%) being monolingual experiments and all groups (except for GE and RUG) submitting at least one monolingual English run. Participants also showed increased interest in retrieval from German annotations; a total of eight groups submitted 88 runs (14.5% of total runs), 20.5% of them monolingual (compared with four groups submitting 18 runs in 2006). Seven groups made use of the new Spanish annotations (5.4% of total runs, 48.5% of them monolingual), while only two participants experimented with the annotations with a randomly selected language for each image (5.3%).

The expanded multilingual character of the evaluation environment also yielded an increased number of bilingual retrieval experiments: while only four query languages (French, Italian, Japanese, Chinese) had been used in 10 or more bilingual runs in 2006, a total of 13 languages were used to start retrieval approaches in 10 or more runs in 2007. The most popular languages this year were German (43 runs), French (43 runs) and English (35 runs). Surprisingly, 26.5% of the bilingual experiments used a Scandinavian language to start the

Table 3. Submission overview by query and annotation languages

Query / Annotation	English	German	Spanish	Random	None	Total
English	204(18)	18 (5)	6 (3)	11 (2)		239(18)
German	31 (6)	31 (7)	1 (1)	11 (2)		74 (9)
Visual	1 (1)				52 (12)	53(12)
French	32 (7)	1 (1)	10 (2)			43 (7)
Spanish	20 (5)		16 (7)	2 (1)		38 (9)
Swedish	20 (3)	12 (1)				32 (3)
Simplified Chinese	24 (4)	1 (1)				25 (4)
Portuguese	19 (5)			2 (1)		21 (5)
Russian	17 (4)	1 (1)		2 (1)		20 (4)
Norwegian	6 (1)	12 (1)				18 (1)
Japanese	16 (3)					16 (3)
Italian	10 (4)			2 (1)		12 (4)
Danish		12 (1)				12 (1)
Dutch	4 (1)			2 (1)		6 (1)
Traditional Chinese	4 (1)					4 (1)
Total	408 (18)	88 (8)	33 (7)	32 (2)	52 (12)	616(20)

retrieval approach: Swedish (32 runs), Norwegian (18 runs) and Danish (12 runs) – none of these languages had been used in 2006. It is also interesting to note that Asian languages (18.6% of bilingual runs) were almost exclusively used for retrieval from English annotations (only one run experimented with the German annotations), which might indicate a lack of translation resources from Asian to European languages other than English.

4 Results

This section provides an overview of the system results with respect to query and annotation languages as well as other submission dimensions such as query mode, retrieval modality and the involvement of relevance feedback or query expansion techniques. Although the description fields were not provided with the image annotations, the absolute retrieval results achieved by the systems were not much lower compared to those in 2006 when the entire annotation was used. We attribute this to the fact that more than 50% of the groups had participated at ImageCLEF before, improved retrieval algorithms (not only of returning participants), and the increased use of content-based retrieval approaches.

4.1 Results by Language

Table 4 shows the runs which achieved the highest MAP for each language pair (ranked by descending order of MAP scores).

Of these runs, 90.6% use query expansion or relevance feedback, and 78.1% use both visual and textual features for retrieval. It is noticeable that submissions from CUT, DCU, NTU (Taiwan) and INAOE dominate the results. As

Table 4. Systems with highest MAP for each language

Query (Caption)	Group/Run ID	MAP	P(20)	GMAP	bpref
English (English)	CUT/cut-EN2EN-F50	0.318	0.459	0.298	0.162
German (English)	XRCE/DE-EN-AUTO-FB-TXTIMG_MPRF	0.290	0.388	0.268	0.156
Portuguese (English)	Taiwan/NTU-PT-EN-AUTO-FBQE-TXTIMG	0.282	0.388	0.266	0.127
Spanish (English)	Taiwan/NTU-ES-EN-AUTO-FBQE-TXTIMG	0.279	0.383	0.259	0.128
Russian (English)	Taiwan/NTU-RU-EN-AUTO-FBQE-TXTIMG	0.273	0.383	0.256	0.115
Italian (English)	Taiwan/NTU-IT-EN-AUTO-FBQE-TXTIMG	0.271	0.384	0.257	0.114
S. Chinese (English)	CUT/cut-ZHS2EN-F20	0.269	0.404	0.244	0.098
French (English)	Taiwan/NTU-FR-EN-AUTO-FBQE-TXTIMG	0.267	0.374	0.248	0.115
T. Chinese (English)	Taiwan/NTU-ZHT-EN-AUTO-FBQE-TXTIMG	0.257	0.360	0.240	0.089
Japanese (English)	Taiwan/NTU-JA-EN-AUTO-FBQE-TXTIMG	0.255	0.368	0.241	0.094
Dutch (English)	INAOE/INAOE-NL-EN-NaiveWBQE-IMFB	0.199	0.292	0.191	0.038
Swedish (English)	INAOE/INAOE-SV-EN-NaiveWBQE-IMFB	0.199	0.292	0.191	0.038
Visual (English)	INAOE/INAOE-VISUAL-EN-AN_EXP_3	0.193	0.294	0.192	0.039
Norwegian (English)	DCU/NO-EN-Mix-sgramRF-dyn-equal-fire	0.165	0.275	0.174	0.057
German (German)	Taiwan/NTU-DE-DE-AUTO-FBQE-TXTIMG	0.245	0.379	0.239	0.108
English (German)	XRCE/EN-DE-AUTO-FB-TXTIMG_MPRF_FLR	0.278	0.362	0.250	0.112
Swedish (German)	DCU/SW-DE-Mix-dictRF-dyn-equal-fire	0.179	0.294	0.180	0.071
Danish (German)	DCU/DA-DE-Mix-dictRF-dyn-equal-fire	0.173	0.294	0.176	0.073
French (German)	CUT/cut-FR2DE-F20	0.164	0.237	0.144	0.004
Norwegian (German)	DCU/NO-DE-Mix-dictRF-dyn-equal-fire	0.167	0.270	0.165	0.070
Spanish (Spanish)	Taiwan/NTU-ES-ES-AUTO-FBQE-TXTIMG	0.279	0.397	0.269	0.113
English (Spanish)	CUT/cut-EN2ES-F20	0.277	0.377	0.247	0.105
German (Spanish)	Berkeley/Berk-DE-ES-AUTO-FB-TXT	0.091	0.122	0.072	0.008
English (Random)	DCU/EN-RND-Mix-sgramRF-dyn-equal-fire	0.168	0.285	0.175	0.068
German (Random)	DCU/DE-RND-Mix-sgram-dyn-equal-fire	0.157	0.282	0.167	0.064
French (Random)	DCU/FR-RND-Mix-sgram-dyn-equal-fire	0.141	0.264	0.148	0.059
Spanish (Random)	INAOE/INAOE-ES-RND-NaiveQE-IMFB	0.124	0.228	0.136	0.027
Dutch (Random)	INAOE/INAOE-NL-RND-NaiveQE	0.083	0.156	0.094	0.011
Italian (Random)	INAOE/INAOE-IT-RND-NaiveQE	0.080	0.144	0.086	0.018
Russian (Random)	INAOE/INAOE-RU-RND-NaiveQE	0.076	0.136	0.085	0.017
Portuguese (Random)	INAOE/INAOE-PT-RND-NaiveQE	0.030	0.043	0.032	0.001
Visual	XRCE/AUTO-NOFB-IMG_COMBFK	0.189	0.352	0.201	0.102

in previous years, the highest English monolingual run outperforms the highest German and Spanish monolingual runs (MAPs are 22.9% and 12.1% lower).

The highest bilingual to English run (German – English) performed with a MAP of 91.3% of the highest monolingual run MAP, with the highest bilingual run in most other query languages such as Portuguese, Spanish, Russian, Italian, Chinese, French and Japanese all exhibiting at least 80% of that highest monolingual English run. Hence, there is no longer much difference between monolingual and bilingual retrieval, indicating a significant progress of the translation and retrieval methods using these languages. Moreover, the highest bilingual to Spanish run (English – Spanish) had a MAP of 99.2% of the highest monolingual Spanish run, while the highest bilingual to German run (English – German) even outperformed the highest German monolingual run MAP by 13.3%.

4.2 Results by Query Mode

This trend is not only true for the highest runs per language pair, but also for all submissions and across several performance indicators. Table 5 illustrates the average scores across all system runs (and the standard deviations in parenthesis) with respect to monolingual, bilingual and purely visual retrieval.

Again, monolingual and bilingual retrieval are almost identical, and so are the average results for monolingual Spanish, English and German retrieval (see

Table 5. Results by query mode

Query Mode	MAP	P(20)	BPREF	GMAP
Monolingual	0.138 (0.070)	0.192 (0.102)	0.132 (0.066)	0.038 (0.036)
Bilingual	0.136 (0.056)	0.199 (0.088)	0.136 (0.054)	0.037 (0.027)
Visual	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

Table 6): Spanish shows the highest average MAP and BPREF values, while German exhibits the highest average for P(20) and English for GMAP.

Table 6. Monolingual results by annotation language

Annotation	MAP	P(20)	BPREF	GMAP
Spanish	0.145 (0.059)	0.195 (0.092)	0.134 (0.056)	0.036 (0.034)
English	0.139 (0.075)	0.190 (0.108)	0.132 (0.071)	0.038 (0.038)
German	0.133 (0.043)	0.200 (0.083)	0.132 (0.048)	0.034 (0.031)

Across all submissions, the average values for bilingual retrieval from English and German annotations are even slightly higher than those for monolingual retrieval (see Table 7), while bilingual retrieval from Spanish annotations and from annotations with a randomly selected language does not lag far behind.

Table 7. Bilingual results by annotation language

Annotation	MAP	P(20)	BPREF	GMAP
English	0.150 (0.055)	0.204 (0.089)	0.143 (0.054)	0.037 (0.029)
German	0.138 (0.040)	0.217 (0.075)	0.145 (0.040)	0.045 (0.021)
Spanish	0.117 (0.079)	0.176 (0.108)	0.108 (0.070)	0.027 (0.037)
Random	0.099 (0.048)	0.169 (0.084)	0.108 (0.052)	0.028 (0.021)
None	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

These results indicate that the query language does not play a major factor for visual information retrieval for lightly annotated images. We attribute this (1) to the high quality of the state-of-the-art translation techniques, (2) to the fact that such translations implicitly expand the query terms (similar to query expansion using a thesaurus) and (3) to the short image captions used (as many of them are proper nouns which are often not even translated).

4.3 Results by Retrieval Modality

In 2006, the system results had shown that combining visual features from the image and semantic knowledge derived from the captions offered optimum performance for retrieval from a generic photographic collection with fully annotated images.

Table 8. Results by retrieval modality

Modality	MAP	P(20)	BPREF	GMAP
Mixed	0.149 (0.066)	0.225 (0.097)	0.203 (0.081)	0.050 (0.031)
Text Only	0.120 (0.040)	0.152 (0.051)	0.141 (0.045)	0.018 (0.018)
Image Only	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

As indicated in Table 8, the results of *ImageCLEFphoto 2007* show that this also applies for retrieval from generic photographic collections with lightly annotated images: on average, combining visual features from the image and semantic information from the annotations gave a 24% improvement of the MAP over retrieval based solely on text.

Purely content-based approaches still lag behind, but the average MAP for retrieval solely based on image features shows an improvement of 65.8% compared to the average MAP in 2006.

4.4 Results by Feedback and/or Query Expansion

Table 9 illustrates the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the use of query expansion or relevance feedback techniques.

Table 9. Results by feedback or query expansion

Technique	MAP	P(20)	BPREF	GMAP
None	0.109 (0.052)	0.178 (0.075)	0.110 (0.047)	0.027 (0.024)
Query Expansion	0.112 (0.040)	0.158 (0.053)	0.106 (0.036)	0.024 (0.019)
Relevance Feedback	0.131 (0.055)	0.185 (0.084)	0.132 (0.054)	0.038 (0.026)
Expansion & Feedback	0.218 (0.062)	0.324 (0.076)	0.209 (0.053)	0.073 (0.046)

While the use of query expansion (*i.e.* the use of thesauri or ontologies such as WordNet) does not necessarily seem to dramatically improve retrieval results for retrieval from lightly annotated images (average MAP only 2.1% higher), relevance feedback (typically in the form of query expansion based on pseudo relevance feedback) appeared to work well on short captions (average MAP 19.9% higher), with a combination of query expansion and relevance feedback techniques yielding results almost twice as good as without any of these techniques (average MAP 99.5% higher).

4.5 Results by Run Type

Table 10 shows the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the run type. Unsurprisingly, MAP results of manual approaches are, on average, 58.6% higher than purely automatic runs — this trend seems to be true for both fully annotated and lightly annotated images.

Table 10. Results by run type

Technique	MAP	P(20)	BPREF	GMAP
Manual	0.201 (0.081)	0.302 (0.116)	0.189 (0.074)	0.066 (0.051)
Automatic	0.127 (0.058)	0.187 (0.084)	0.126 (0.055)	0.034 (0.029)

5 Conclusion

This paper reported on *ImageCLEFphoto 2007*, the general photographic adhoc retrieval task of the *ImageCLEF 2007* evaluation campaign. Its evaluation objective concentrated on visual information retrieval from generic collections of lightly annotated images, a new challenge that attracted a large number of submissions: 20 participating groups submitted a total of 616 system runs.

The participants were provided with a subset of the *IAPR TC-12 Benchmark*: 20,000 colour photographs and four sets of semi-structured annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images. Unlike in 2006, the participants were not allowed to use the semantic description field in their retrieval approaches. The topics and relevance assessments from 2006 were reused (and updated) to facilitate the comparison of retrieval from fully and lightly annotated images.

The nature of the task also attracted a larger number of participants experimenting with content-based retrieval techniques, and hence the retrieval results were similar to those in 2006, despite the limited image annotations in 2007. Other findings for multilingual visual information retrieval from generic collections of lightly annotated photographs include:

- bilingual retrieval performs as well as monolingual retrieval;
- the choice of the query language is almost negligible as many of the short captions contain proper nouns;
- combining concept and content-based retrieval methods as well as using relevance feedback and/or query expansion techniques can significantly improve retrieval performance;

ImageCLEFphoto will continue to provide resources to the retrieval and computational vision communities to facilitate standardised laboratory-style testing of image retrieval systems. While these resources have predominately been used by systems applying a concept-based retrieval approach thus far, the rapid increase of participants using content-based retrieval techniques at *ImageCLEFphoto* calls for a more suitable evaluation environment for visual approaches (*e.g.* the preparation of training data). For *ImageCLEFphoto 2008*, we are planning to create new topics and will therefore be able to provide this year's topics and qrels as training data for next year.

Acknowledgements

We would like to thank *viventura* and the *IAPR TC-12* for providing their image databases for this year's task. This work was partially funded by the EU MultiMatch project (IST-033104) and the EU MUSCLE NoE (FP6-507752).

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.D., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview ofthe ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 579–594. Springer, Heidelberg (2007)
3. Grubinger, M., Clough, P.D., Müller, H., Deselears, T.: The IAPRTC12 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 2006, Genoa, Italy, pp. 13–23 (2006)
4. Grubinger, M.: On the Creation of Query Topics for ImageCLEFphoto. In: Third MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation, Budapest, Hungary, pp. 50–63 (2007)
5. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: An experimental comparison. Information Retrieval (in press, 2008)