

# Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task

Thomas Arni<sup>1</sup>, Paul Clough<sup>1</sup>, Mark Sanderson<sup>1</sup> and Michael Grubinger<sup>2</sup>

<sup>1</sup>Sheffield University, Sheffield, UK

<sup>2</sup>Victoria University, Melbourne, Australia

## Abstract

ImageCLEFphoto 2008 is an ad-hoc photo retrieval task and part of the ImageCLEF evaluation campaign. This task provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval systems. In 2008, the evaluation task concentrated on promoting diversity within the top 20 results from a multilingual image collection. This new challenge attracted a record number of submissions: a total of 24 participating groups submitting 1,042 system runs. Some of the findings include that the choice of annotation language is almost negligible and the best runs are by combining concept and content-based retrieval methods.

## 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

Performance Evaluation, IAPR TC-12 Benchmark, Image Retrieval, Diversity, Clustering

## 1 Introduction

The evaluation of multilingual image retrieval systems (i.e. where associated texts are in languages different from written queries) has been the focus of ImageCLEF since its inception in 2003. The track has evolved over the years to address different domains (e.g. cultural heritage, medical imaging and Wikipedia), and different kinds of tasks (e.g. ad-hoc retrieval, automatic annotation and clustering). The focus of the ImageCLEFphoto task in 2008 has been to promote diversity in the top  $n$  results (see section 1.2). The resources provided enable system-centred evaluation for multilingual and diversity-based visual information retrieval based on a collection of “general” photographs (see section 2.1).

### 1.1 Evaluation Scenario

The evaluation scenario is similar to the classic TREC<sup>1</sup> 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. the search topics are not known to the system in advance) [6]. The goal of the simulation is: given an alphanumeric statement (and/or sample images) describing a user’s 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). For 2008, the scenario is slightly different in that systems must return relevant images from as many different sub-topics as possible (i.e. promote diversity) in the top  $n$  results.

### 1.2 Evaluation Objective for 2008

The main objective of ImageCLEFphoto for 2008 comprised the evaluation of ad-hoc multilingual visual information retrieval systems from a general collection of annotated photographs (i.e. image with accompanying semi-structured captions such as the title, location, description, date or additional notes). However, this year

---

<sup>1</sup> <http://trec.nist.gov/>

focused on a particular aspect of retrieval: diversity of the results set (see section 1.3). More recently, research in image search has concentrated on ensuring that duplicate or near-duplicate documents retrieved in response to a query are hidden from the user. This should ideally lead to a ranked list where images are both relevant *and* diverse. In 2007, the task considered maximising the number of relevant documents in the resulting ranked list. In 2008, the task is to promote diversity in the top  $n$  results, which has been shown to better satisfy a user’s information need [8, 9] (people often type in the same query but prefer to see results which represent different aspects of the results set). Hence, providing a diverse results list is especially important when a user types in a query that is either poorly specified or ambiguous.

This new challenge allows for the investigation of a number of research questions, including the following:

- Is it possible to promote diversity within the top  $n$  results?
- Which approaches work best at promoting diversity?
- Does promoting diversity reduce the number of relevant images in the top  $n$  results?
- Can “standard” text retrieval methods be used to promote diversity?
- How does the retrieval performance compare between bilingual and multilingual annotations?

One major goal of ImageCLEFphoto 2008 was to attract participants from various backgrounds and with different research interests. The collection developed for the 2008 task, in our view, provides a resource that can be used to evaluate both concept and content-based approaches for image retrieval.

### 1.3 An Example of Diversity

To illustrate what a diverse results set looks like, consider the following example. Given the search topic “images of typical Australian animals”, using traditional ranking methods (commonly based on the Probability Ranking Principle) will produce a result calculated on the similarity between query and documents. This often leads to a set of results that contains groups of similar documents. Figure 1 shows a typical example of the kind of results one could expect in the top 10 using traditional ranking methods. However, going down the ranked list one finds other types of animals such as koala bears.











Rank 1  Relevant	Rank 2  Relevant	Rank 3  Relevant	Rank 4  Relevant	Rank 5  Relevant
Rank 6  Relevant	Rank 7  Relevant	Rank 8  Relevant	Rank 9  Relevant	Rank 10  Relevant

Figure 1: Example top 10 results set, primarily showing Kangaroos

The 2008 ImageCLEFphoto task is to promote diversity in the top  $n$  results by including at least one relevant document from each sub-topic within the first  $n$  results (i.e. pictures of *different* animals in the top  $n$ ). A more diverse (and arguably improved) results set is illustrated in Figure 2.

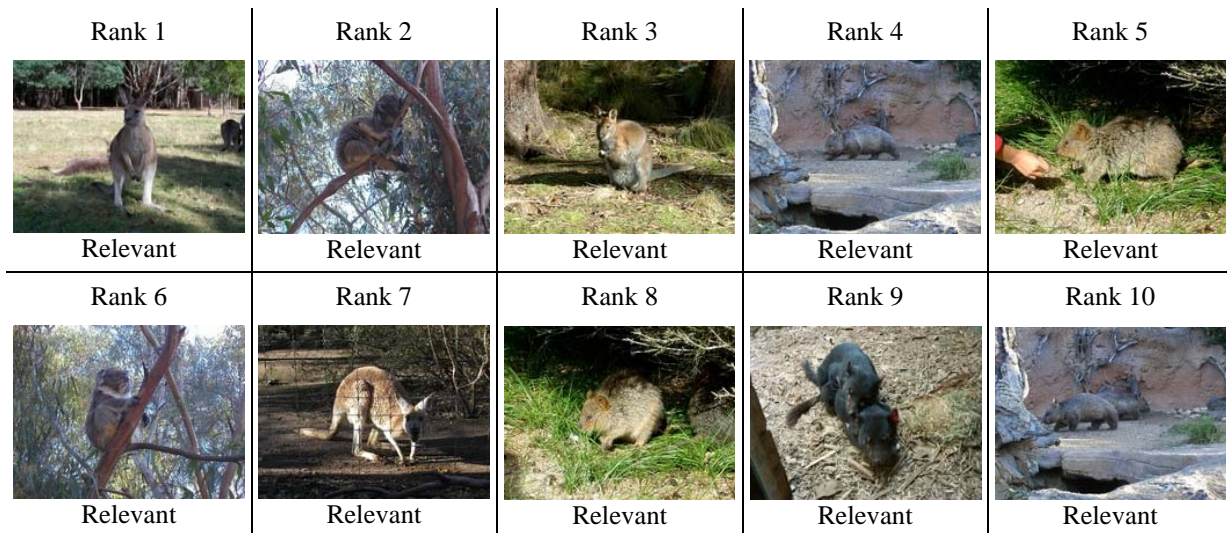


Figure 2: Example top 10 results set with a more diverse range of Australian animals

## 2 Evaluation Framework

Similar to the 2006 and 2007 ImageCLEFphoto tasks [3, 2], we generated a subset of the IAPR TC-12 Benchmark as an evaluation resource for 2008. This section provides more information on these individual components: the document collection, the query topics, relevance judgements, cluster relevance judgements and performance indicators. More information on the design and implementation of the IAPR TC-12 Benchmark itself, created under Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR<sup>2</sup>), can be found in [7].

### 2.1 Document Collection

The IAPR TC-12 Benchmark consists of 20,000 colour photographs taken from locations around the world and comprises a varying cross-section of still natural images. Figure 3 illustrates a number of sample images from a selection of categories.

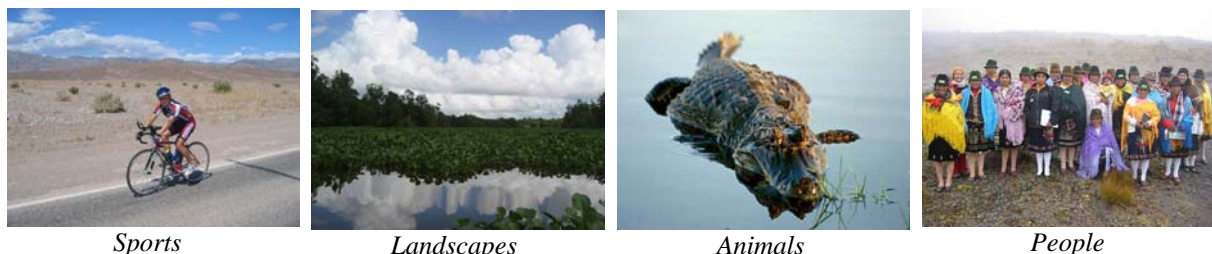


Figure 3: Sample images from the IAPR TC-12 collection

The majority of images have been provided by Viventura<sup>3</sup>, 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 trips made and general pictures of each location including accommodation, facilities and ongoing social projects. In addition to these photos, a number of photos from a personal archive have also been added to form the collection used in ImageCLEF. The collection is publicly available for research purposes and, unlike many existing photographic collections, can be used to evaluate image retrieval systems. The collection is general in content with 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.

Each image in the collection has a corresponding semi-structured caption consisting of the following six 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) where and (6) when the photo was taken. Figure 4 shows a sample image with its

<sup>2</sup> <http://www.iapr.org/>

<sup>3</sup> <http://www.viventura.de/>

corresponding textual annotation (in English).



```
<DOC>
<DOCNO>annotations/16/16392.eng</DOCNO>
<TITLE>Sunset in Salvador</TITLE>
<DESCRIPTION>a sandy beach at the sea with dark rocks behind
it; the setting sun in an orange sky in the background;
</DESCRIPTION>
<NOTES></NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>10 October 2004</DATE>
<IMAGE>images/16/16392.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16392.jpg</THUMBNAIL>
</DOC>
```

Figure 4: Sample image caption

By using a custom-built application for managing the images, various subsets of the collection can be generated with respect to a variety of particular parameters (e.g. using a selected subset of caption fields). For 2008, the following data was provided:

- **Annotation language:** two sets of annotations in (1) English and (2) Random. In the random set, the annotation language was randomly selected from for each of the images (i.e. annotations are either German or English image captions).
- **Caption fields:** all caption fields were provided for the 2008 task.
- **Annotation completeness:** each image caption exhibited the same level of annotation completeness - there were no images without annotations (as experimented with in 2006). The participants were granted access to the data set on 22<sup>nd</sup> April 2008 and had exactly one month to familiarise themselves with the new subset. Most participants had to modify their standard retrieval systems in order to generate diverse results in the top  $n$ .

## 2.2 Query Topics

From an existing set of 60 topics, 39 were selected and distributed to participants (Table 1) representing varying search requests (many of these are realistic and based on queries extracted during log file analysis – see [4] for more detailed information). We found that for the new retrieval challenge (promoting diversity), not all of the existing topics were suitable and therefore some were removed (see [1] for further details). Although 21 topics were removed, the remaining 39 topics are well-balanced, diverse and should present a retrieval challenge to participants wishing to use either text and/or low-level visual analysis techniques for creating clusters.

Similar to TREC, the query topics were provided as structured statements of user needs. The full description of a topic consists of (1) a topic titles (2) a topic narrative, (3) a newly added *cluster type* and (4) three example relevant images for that topic. An additional field was added called *cluster type*, which was augmented for easier assessment of the clusters as well as to facilitate the quantification of the result set diversity [1]. Below is an example augmented topic:

```
<top>
<num> Number: 48 </num>
<title> vehicle in South Korea </title>
<cluster> vehicle </cluster>
...
</top>
```

The cluster type in topic 48 is vehicle (in the <cluster> tag), which clearly defines how relevant images from this topic should be clustered. Different from previous years, topics were available in English only.



ID	Topic title	ID	Topic title
2	church with more than two towers	3	religious statue in the foreground
5	animal swimming	6	straight road in the USA
10	destinations in Venezuela	11	black and white photos of Russia
12	people observing football match	13	exterior view of school building
15	night shots of cathedrals	16	people in San Francisco
17	lighthouse at the sea	18	sport stadium outside Australia
19	exterior view of sport stadium	20	close-up photograph of an animal
21	accommodation provided by host families	23	sport photos from California
24	snowcapped building in Europe	28	cathedral in Ecuador
29	views of Sydney's world-famous landmarks	31	volcanoes around Quito
34	group picture on a beach	35	bird flying
37	sights along the Inka-Trail	39	people in bad weather
40	tourist destinations in bad weather	41	winter landscape in South America
43	sunset over water	44	mountains on mainland Australia
48	vehicle in South Korea	49	images of typical Australian animals
50	indoor photos of a church or cathedral	52	sports people with prizes
53	views of walls with unsymmetric stones	54	famous television (and telecommunication) towers
55	drawings in Peruvian deserts	56	photos of oxidised vehicles
58	seals near water	59	creative group pictures in Uyuni
60	salt heaps in salt pan		

Table 1: Topics for the ImageCLEFphoto 2008 task.

### 2.3 Relevance Assessments

The relevance assessments, with the exception of removing any additional images considered as non-relevant, are exactly the same as in year 2007. No pooling of the images was carried out in 2008. Information about relevance assessments from previous years can be found in [2]. To enable diversity to be quantified, it was necessary to classify images relevant to a given topic to one or more sub-topics or clusters. This was performed by two assessors. In case of inconsistent judgements, a third assessor was used to resolve the inconsistencies. The resulting cluster assessment judgements are then used in combination with the normal relevance assessment to determine the retrieval effectiveness of each submitted system run (for further details see [1]).

### 2.4 Generating the Results

Once the relevance judgements and the cluster relevance assessments were completed, the performance of individual systems and approaches can be evaluated. The results for submitted runs were computed using the latest version of trec eval<sup>4</sup>, as well as a custom-built tool to calculate diversity of the results set. Submissions were evaluated using two metrics: (1) precision at rank 20 (P20) and (2) cluster recall at rank 20 (CR20). Rank 20 was selected as the cut-off point to measure precision and cluster recall because most online image retrieval engines (e.g. Google, Yahoo! and AltaVista) display 18 to 20 images by default. Further measures considered included uninterpolated (arithmetic) Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP) to test system robustness and binary preference (bpref), which is a good indicator of how complete relevance judgements are. To enable an absolute comparison between individual runs, a single metric is required: the F1-measure was used to combine scores from P20 and CR20 (representing the harmonic mean of P20 and CR20).

## 3 Overview of Participation and Submissions

In 2008, 43 groups registered for ImageCLEFphoto (32 in 2007, 36 in 2006), with 24 groups eventually submitting a total of 1,042 runs (all of which were evaluated by the organisers). This is an increase in the number of runs from previous years (20 groups submitting 616 runs in 2007, 12 groups submitting 157 runs in 2006, and 11 groups 349 runs in 2005 respectively). Table 2 provides an overview of the participating groups, the corresponding number of submitted runs and whether they are new or returning participants. The 24 participating

<sup>4</sup> [http://trec.nist.gov/trec\\_eval/trec\\_eval.8.1.tar.gz](http://trec.nist.gov/trec_eval/trec_eval.8.1.tar.gz)

groups are affiliated to 21 different institutions in 11 countries. New participants submitting in 2008 include joint work from four French labs (AVEIR), University of Waseda (GITS), Laboratory of Informatics of Grenoble (LIG), System and Information Science Lab (LSIS), Meiji University (Meiji), University of Ottawa (Ottawa), Telecom ParisTech (PTECH), University of Sheffield (Shef), University of Alicante (TEXTMESS) and Piere & Marie Curie University (UPMC). In total, 65% of the participants in 2007 returned and participated in 2008.

Group ID	Institution	Country	Runs	Status
AVEIR	Joint project of the four French labs : LIG,LIP6, LSIS, PTECH	France	4	Returning / New
Budapest-ACAD	Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, Hungary	Hungary	8	Returning
CLAC	Computational Linguistics at Concordia (Clac) Lab, Concordia University, Montreal, Canada	Canada	6	Returning
CUT	Chemnitz University of Technology, Chemnitz, Germany	Germany	4	Returning
DCU	School of Computing, Dublin City University, Dublin Ireland	Ireland	733	Returning
GITS	KAMEYAMA Lab,GITS, Waseda University, Japan	Japan	4	New
INAOE	National Institute of Astrophysics, Optics and Electronics, Computer Science Department, Puebla, Mexico	Mexico	16	Returning
IPAL	Image Perception, Access & Language (IPAL), Singapore & National Center for Scientific Research, France & Institute for Infocomm Research, Singapore & University of Joseph Fourier, Grenoble, France	Singapore / France	10	Returning
LIG	Laboratory of Informatics of Grenoble (LIG), Grenoble, France	France	4	New
LSIS	System and Information Sciences Lab, France	France	15	New
Meiji	Department of Computer Science, Meiji University, Japan	Japan	8	New
MirFI	Computer Science Faculty, Daedalus, Madrid, Spain	Spain	41	Returning
MirGSI	Intelligent System Group, Daedalus, Madrid, Spain	Spain	14	Returning
MMIS	Imperial College London & Open University, UK	UK	9	Returning
NII	National Institute of Informatics, Tokyo, Japan	Japan	10	Returning
NTU	National Taiwan University, Taipei, Taiwan	Taiwan	7	Returning
Ottawa	School of Information Technology and Engineering, University of Ottawa, Canada	Canada	13	New
PTECH	Institut TELECOM, TELECOM ParisTech, Paris, France	France	15	New
Shef	Department of Information Studies, University of Sheffield, Sheffield, UK	UK	37	New
SINAI	Sinai group of the University of Jaén, Jaén, Spain	Spain	6	Returning
TEXTMESS	Department of Software and Computing Systems, University of Alicante, Spain & University of Jaén, Jaén, Spain	Spain	17	New
UA GPLSI	Department of Software and Computing Systems, University of Alicante, Spain	Spain	18	Returning
UPMC	Pierre & Marie Curie University, Paris, France	France	15	New
XRCE	Xerox Research Centre Europe	France	28	Returning

Table 2: Participating groups

Increased participation might be an indicator for (1) the growing need for evaluation of visual information retrieval from more general photographic collections, (2) the growing need for comparative evaluation of diversity and/or (3) an interest by researchers world-wide to participate in evaluation events such as ImageCLEFphoto. Although the total number of runs has risen, the geometric mean of runs per participating group is slightly lower than in 2007 (12.4 in 2008 / 13.8 in 2007). The reason for the increasing number of total runs is mainly due to the larger number of submissions from Dublin City University (DCU), who submitted a total of 733 runs.

### 3.1 Overview of Submissions

Overall, 1042 runs were submitted and categorised with respect to the following dimensions: (1) annotation language, (2) modality (text only, image only or combined) and (3) run type (automatic or manual). Table 3 provides an overview of all submitted runs according to these dimensions. Most submissions (96.8%) used the provided image annotations, with 22 groups submitting a total of 404 purely concept-based (textual) runs and 19 groups a total of 605 runs using a combination of content-based (visual) and concept-based features. A total of 11

groups submitted 33 purely content-based runs. Of all retrieval approaches, 61.2% involved the use of image retrieval (53.4% in 2007 and 31% in 2006), 79% of all groups used content-based (i.e. visual) information in their runs (60% in 2007 and 58% in 2006). Almost all of the runs (99.7%) were automatic (i.e. involving no human intervention); only 3 submitted runs were manual. Only one participating group made use of additional data, which was available from the Visual Concept Detection Task<sup>5</sup>.

Dimensions	Type	2008		2007		2006	
		Runs	Groups	Runs	Groups	Runs	Groups
Annotation language	EN	514	24	271	17	137	2
	RND	495	2	32	2		
Modality	Text Only	404	22	167	15	121	2
	Mixed (text and image)	605	19	255	13	21	1
	Image Only	33	11	52	12		
Run type	Manual	3	1	19	3		
	Automatic	1039	25	455	19	142	2

Table 3: Submission overview by dimensions.

## 4 Results

This section provides an overview of results with respect to the various submission dimensions (1) annotation language, (2) retrieval modality and (3) run type. The task for the participants was to maximise the number of relevant images in the *top 20* results. At the same time the relevant images in the top 20 results should be from as many different sub-topics as possible. Simply getting lots of relevant images from one sub-topic or filling the ranking with diverse, but non-relevant images, results in a poor overall effectiveness score. Measures such as MAP are not suitable since it does not take into account diversity. To determine the diversity of a result set, S-Recall (sub-topic recall) proposed by Zhai et al [5] was used. S-recall at rank  $K$  is defined as the percentage of sub-topics covered by the first  $K$  documents in the list:

$$S\text{-recall at } K \equiv \frac{\left| \cup_{i=1}^K \text{subtopics}(d_i) \right|}{n_A}$$

where  $d_i$  represents the  $i^{\text{th}}$  document,  $\text{subtopics}(d_i)$  the number of sub-topics  $d_i$  belongs to, and  $n_A$  the total number of sub-topics in a particular topic. Thus the evaluation is based on two measures: precision at 20 and cluster recall at rank 20 (S-recall). As previously mentioned, it was important to maximise both measures in order to get a high overall ranking. To provide a single measure of effectiveness, we used the F1-measure (harmonic mean) to combine P20 and CR20:

$$\text{F1-measure} = \frac{2 \times (P20 \times CR20)}{(P20 + CR20)}$$

The order of the diverse and relevant documents within the first top 20 result is not considered for the calculation of the cluster recall. This means that relevant documents from different sub-topics can be in a random order, without affecting the cluster recall score.

### 4.1 Results by annotation language

Tables 4 and 5 show the runs which achieved highest F1-measure scores for the two annotation languages: ENG and RND. Taking into account that only two groups submitted 495 runs with a random annotation language, the result shows the same trend as in previous years: the highest monolingual run still outperforms the highest bilingual run, which consists of a random annotation language. However, as in previous years, the margin of difference is low and can be attributed to significant progress of the translation and retrieval methods using these languages. The best performing runs using random annotations performed with an F1-measure score at 97.4% of the highest monolingual run. Hence, the language barrier is no longer a critical factor in achieving good retrieval results.

<sup>5</sup> <http://www.imageclef.org/2008/iaprconcepts>

Query language	Annotation language	Group	Run-ID	Run type	Modality	P20	CR20	F1-Measure
English	English	PTECH	PTECH-EN-EN-MAN-TXTIMG-MMBQL.run	MAN	TXTIMG	0.6885	0.6801	0.6843
English	English	PTECH	PTECH-EN-EN-MAN-TXTIMG-MMBML.run	MAN	TXTIMG	0.6962	0.6719	0.6838
English	English	PTECH	PTECH-EN-EN-MAN-TXT-MTBTN.run	MAN	TXT	0.5756	0.5814	0.5785
English	English	XRCE	xrce_tilo_nbdiv_15	AUTO	TXTIMG	0.5115	0.4262	0.4650
English	English	DCU	DCU-EN-EN-AUTO-TXTIMG-ge.txt	AUTO	TXTIMG	0.4756	0.4542	0.4647
English	English	XRCE	xrce_tilo_nbdiv_10	AUTO	TXTIMG	0.5282	0.4146	0.4646
English	English	XRCE	xrce_cm_nbdiv_10	AUTO	TXTIMG	0.5269	0.4111	0.4619
English	English	DCU	DCU-EN-EN-AUTO-TXTIMG.txt	AUTO	TXTIMG	0.4628	0.4546	0.4587
English	English	XRCE	xrce_cm_mmr_07	AUTO	TXTIMG	0.5282	0.4015	0.4562
English	English	XRCE	xrce_tfidf_nbdiv_10	AUTO	TXTIMG	0.5115	0.4081	0.4540

Table 4: Systems with the highest F1-Measure for English annotations

Query language	Annotation language	Group	Run-ID	Run type	Modality	P20	CR20	F1-Measure
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr.txt	AUTO	TXTIMG	0.4397	0.4673	0.4531
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-ge.txt	AUTO	TXTIMG	0.4423	0.4529	0.4475
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tf-all.txt	AUTO	TXTIMG	0.4038	0.4967	0.4455
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tfidf-all.txt	AUTO	TXTIMG	0.3974	0.4948	0.4408
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tfidf-all.txt	AUTO	TXTIMG	0.3897	0.5049	0.4399
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tf-ge-all.txt	AUTO	TXTIMG	0.4013	0.4806	0.4374
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tf-all.txt	AUTO	TXTIMG	0.3910	0.4936	0.4363
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tfidf-ge-all.txt	AUTO	TXTIMG	0.4013	0.4766	0.4357
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-kx-tfidf-ge-all.txt	AUTO	TXTIMG	0.3897	0.4768	0.4289
English	RND	DCU	DCU-EN-RND-AUTO-TXTIMG-tr-d50-k40-tf-ge-all.txt	AUTO	TXTIMG	0.3897	0.4678	0.4252

Table 5: Systems with the highest F1-Measure for Random annotations (German / English)

## 4.2 Results by Retrieval Modality

In 2006 and 2007, the results showed that by combining visual features from the image and semantic knowledge derived from the captions offered optimum performance for retrieval from a general photographic collection with fully annotated images [2, 3]. As indicated in Table 6, the results of ImageCLEFphoto 2008 show that this also applies for our modified task, which promotes diversity in the results set. However, contrary to 2007 (24% MAP improvement over averages for combining techniques over solely text-based approaches), the improvement is not as clearly visible when combining visual features from the image and semantic information. The difference between “Mixed” and “Text Only” runs is across the averages from all runs, and differs only marginally. However, looking at the best runs in each modality, the “Mixed” runs (F1-Measure = 0.4650) outperform the “Text Only” runs by 16% (F1-measure = 0.4008). Purely content-based approaches still lag behind, although with a smaller gap than in previous years. The best “Image Only” runs (F1-Measure = 0.3396) is higher than both averages for the “Mixed” and “Text only” runs.

Modality	Precision at 20		Cluster Recall at 20		F1-measure (P20/CR20)	
	Mean	SD	Mean	SD	Mean	SD
Mixed	0.2538	0.1023	0.3998	0.0977	0.3034	0.0932
Text Only	0.2431	0.0590	0.3915	0.0819	0.2957	0.0576
Image Only	0.1625	0.1138	0.2127	0.1244	0.1784	0.1170

Table 6: Results by retrieval modality



### 4.3 Results by Run Type

Table 7 shows the average scores and the standard deviations across all systems runs with respect to the run type. Unsurprisingly, F1-Measure results of manual approaches are significantly higher than purely automatic runs. All submitted manual runs are done with English annotation, whereas the average of the automatic runs is both from English as well as Random annotation. However, as previously shown the translation does not have a big impact and can therefore be neglected. In case of the automatic runs the F1-measure is practically identical for the English (ENG) annotations and those with the language randomly selected (RND).

Technique	Precision at 20		Cluster Recall at 20		F1-measure (P20/CR20)	
	Mean	SD	Mean	SD	Mean	SD
Manual	0.6534	0.0675	0.6445	0.0548	0.6489	0.0610
Automatic	0.2456	0.0873	0.3899	0.0975	0.2955	0.0829
Automatic RND Only	0.2353	0.0651	0.4191	0.0731	0.2992	0.0679
Automatic ENG Only	0.2609	0.0990	0.3731	0.1002	0.2994	0.0879
Automatic IMG Only	0.1625	0.1138	0.2127	0.1244	0.1784	0.1170

Table 7: Results by run type

### 4.4 Approaches Used by Participants

Some of the participating groups started by using a baseline run, carried out using different weighting methods (e.g. BM25, DFR, LM), with or without query expansion (e.g. using Local Content Analysis, Pseudo Relevance Feedback, thesaurus-based query expansion, Conceptual Fuzzy Sets, using a location hierarchy, and using Wordnet), and using content- and/or concept-based retrieval methods. The aim of this initial step was obtaining the best possible ranking (i.e. maximising the number of relevant documents returned in the top  $n$ ). The most common following step was to re-rank the initial baseline run in order to promote diversity. One approach of re-ranking is to cluster the top  $n$  documents into sub-topics or clusters and then select the highest ranked document in each cluster and promote higher in the ranked list (i.e. to the top  $n$ ). Clustering was mostly based on the associated textual information using various clustering algorithms (e.g. k-means, k-medoids, knn-density, and latent dirichlet allocation) and different weighting parameters. Some groups also tried to re-rank results using Maximal Marginal Relevance. Other approaches included merging different kind of runs (e.g. calculating image ranking with average/min/mean) or combining scores (novelty/ranking score) to get a diverse and relevant results list. Overall, a majority of approaches applied post-processing methods in one way or another.

## 5 Conclusions

This paper has reported on the 2008 ImageCLEFphoto task, a general photographic ad-hoc retrieval task. The focus this year is different from this year and based on promoting diversity in the top  $n$  results. The challenge for participants was to maximise both the number of relevant images, as well as the number of sub-topics represented within the top 20 results. The 2008 task attracted a record number of submissions: 24 participating groups submitting a total of 1,042 system runs. The participants were provided with a subset of the IAPR TC-12 Benchmark: 20,000 colour photographs and two sets of semi-structured annotations in (1) English and (2) one set whereby the annotation language was randomly selected from English and German for each of the images. To measure the diversity of a ranked list, the existing collection was augmented with cluster assessments. Cluster assessments describe to which sub-topic a relevant image belongs to. Participants experimented with both content- and concept-based retrieval techniques. The main findings of this year include:

- Bilingual retrieval performs nearly as well as monolingual retrieval;
- Combining concept and content-based retrieval methods improves retrieval performance;
- A large number of participants used visual retrieval techniques (similar to previous years).

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 number of participants who are using content-based retrieval techniques at ImageCLEFphoto is still increasing.

## Acknowledgements

We would like to thank Michael Grubinger for providing the data collection and queries which formed the basis of the ImageCLEFPhoto task for 2008. Work undertaken in this paper is supported by the EU-funded TrebleCLEF project (Grant agreement: 215231) and by the project Multimatch (contract IST-2005-2.5.10).

## References

- [1] Arni, T., Tang, J., Sanderson, M. and Clough, P. (2008) Creating a test collection to evaluate diversity in image retrieval, In Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments, held at SIGIR2008.
- [2] Grubinger, M., Clough, P., Hanbury, A. and Müller, H. (2007) Overview of the ImageCLEFPhoto 2007 photographic retrieval task. In Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, 19-21 September 2007.
- [3] Clough, P., Grubinger, M., Deselaers, T., Hanbury, A. and Müller, H. (2006) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In Evaluation of Multilingual and Multi-modal Information Retrieval: Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006), Lecture Notes in Computer Science (LNCS), Alicante, Spain, September 19–21 2006. Springer. (in press).
- [4] Grubinger, M. and Clough, P. (2007) On the Creation of Query Topics for ImageCLEFPhoto, In Proceedings of the third MUSCLE / ImageCLEF workshop on image and video retrieval evaluation, Budapest, Hungary, 19-21 September 2007.
- [5] Zhai, C., Cohen, W. W. and Lafferty, J. (2003) Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In Proceedings of ACM SIGIR 2003, pp. 10–17.
- [6] Voorhees, E. M. and Harman, D. (1998) Overview of the Seventh Text REtrieval Conference (TREC-7). In The Seventh Text Retrieval Conference, pp. 1–23, Gaithersburg, MD, USA, November 1998.
- [7] Grubinger, M., Clough, P., Müller, H. and Deselears, T. (2006) 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, pp. 13–23, Genoa, Italy, 22<sup>nd</sup> May 2006.
- [8] Tian, S. K., Gao, Y. and Huang, T. (2006) Diversifying the image retrieval results. In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006). MULTIMEDIA '06. ACM, New York, NY, pp. 707-710.
- [9] Chen, H. and Karger, D. R. (2006) Less is more: probabilistic models for retrieving fewer relevant documents. In Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, USA, August 06 - 11, 2006). SIGIR '06. ACM, New York, NY, pp. 429-436.