Chapter 8
Photographic Image Retrieval

Monica Lestari Paramita and Michael Grubinger

Abstract CLEF[1] was the first benchmarking campaign that organized an evaluation event for image retrieval: the ImageCLEF photographic ad hoc retrieval task in 2003. Since then, this task has become one of the most popular tasks of ImageCLEF, providing both the resources and a framework necessary to carry out comparative laboratory–style evaluation of multi–lingual visual information retrieval from photographic collections. Running for seven years, several challenges have been given to participants, including: retrieval from a collection of historic photographs; retrieval from a more generic collection with multi–lingual annotations; and retrieval from a large news archive, promoting result diversity. This chapter summarizes each of these tasks, describes the individual test collections and evaluation scenarios, analyzes the retrieval results, and discusses potential findings for a number of research questions.

8.1 Introduction

At the turn of the millennium, several calls (Goodrum, 2000; Leung and Ip, 2000) were made to develop a standardized test collection for Visual Information Retrieval (VIR). In 2003, ImageCLEF[2] was the first evaluation event to answer these calls by providing a benchmark suite comprising an image collection, query topics, relevance assessments and performance measures for cross–language image retrieval, which encompasses two main domains of VIR: (1) image retrieval, and (2) Cross–Language Information Retrieval (CLIR).

Monica Lestari Paramita
University of Sheffield, United Kingdom e-mail: m.paramita@sheffield.ac.uk

Michael Grubinger
Carrera 83 Calle 33-93, Medellín, Colombia e-mail: michael.grubinger@gmx.at

[1] http://www.clef-campaign.org/
[2] http://www.imageclef.org/

H. Müller et al. (eds.), *ImageCLEF*, The Information Retrieval Series 32,
DOI 10.1007/978-3-642-15181-1_8,

Images by their very nature are language–independent; hence, the language used to express the associated texts or textual queries should not affect retrieval, i.e. an image with a caption written in English should be searchable in languages other than English. The main goals of ImageCLEF thereby include:

- to investigate the effectiveness of combining text and image for retrieval;
- to collect and provide resources for benchmarking image retrieval systems;
- to promote the exchange of ideas which may help improve retrieval performance;
- to evaluate VIR systems in a multi–lingual environment.

To achieve these goals, several tasks have been offered to participating groups between 2003 and 2009, including ad hoc retrieval (hereinafter, but also Chapter 13), object recognition and automatic classification tasks (see Chapters 11 and 12) as well as interactive evaluation of retrieval systems (see Chapter 7). ImageCLEF has provided these tasks within two main areas: retrieval of images from photographic collections and retrieval of images from medical collections.

One of the key tasks of ImageCLEF is concerned with evaluation of system performance for ad hoc image retrieval from photographic collections in a laboratory style setting. This kind of evaluation is system–centered and similar to the classic Text REtrieval Conference or TREC[3] ad hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but the query topics are not known to the system in advance. Evaluation thereby only concentrates on comparing algorithms and systems and not on aspects such as retrieval speed or user interaction, as such evaluation is carried out in other tasks (see Chapter 7).

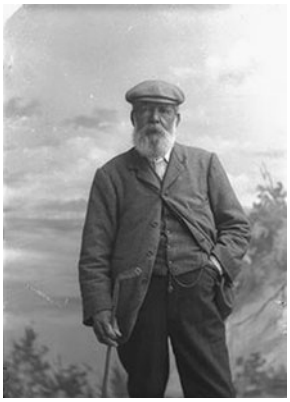
The specific goal of the photographic ad hoc retrieval task is: given a semantic statement (and/or sample images) describing a user information need, find as many relevant images as possible from a given photographic collection (with the query language either being identical to, or different from, that used to describe the images).

Three major phases can be identified in the history of photographic ad hoc retrieval evaluation at ImageCLEF: From 2003 to 2005, the evaluation was based on retrieval from a historic photographic collection (see Section 8.2). In 2006 and 2007, a generic photographic collection with multi–lingual annotations was used (see Section 8.3). Finally, in 2008 and 2009, the evaluation concentrated not only on retrieval precision, but also on retrieval diversity (see Section 8.4).

8.2 Ad hoc Retrieval of Historic Photographs: ImageCLEF 2003–2005

The ImageCLEF 2003 ad hoc retrieval task was the first evaluation event to finally fulfil the calls for a TREC–style evaluation framework for VIR. The research questions concentrated on the evaluation of retrieval from a collection of historic photographs within the first three years, including:

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



Title: Old Tom Morris, golfer, St Andrews.
Short title: Old Tom Morris, golfer.
Location: Fife, Scotland
Description: Portrait of bearded elderly man in tweed jacket, waistcoat with watch chain and flat cap, hand in pockets; painted backdrop.
Date: ca.1900
Photographer: John Fairweather
Categories: [golf - general], [identified male], [St. Andrews Portraits], [Collection - G M Cowie]
Notes: GMC-F202 pc/BIOG: Tom Morris (1821-1908) Golf ball and clubmaker before turning professional, later Custodian of the Links, St Andrews; golfer and four times winner of the Open Championship; father of Young Tom Morris (1851-1875). DETAIL: Studio portrait.

Fig. 8.1: Sample image and caption from the SAC.

- How can researchers be attracted to participate and submit their results?
- How can representative topics and objective relevance judgments be created?
- What methods can be applied to improve retrieval performance?
- How does monolingual retrieval compare to bilingual image retrieval?
- Is it possible to estimate retrieval difficulty in advance?

This section describes both the pilot task of 2003 as well as the follow–up tasks of 2004 and 2005. Further information can be found in the corresponding overview papers: (Clough and Sanderson, 2004; Clough et al, 2005, 2006).

8.2.1 Test Collection and Distribution

The St. Andrews Collection (SAC) of historic photographs is a subset of one of Scotland's most important archives of historic photography and was provided to ImageCLEF by St. Andrews University Library[4]. This collection of 28,133 photographs was the core component of the ImageCLEF ad hoc retrieval task from 2003 to 2005. Detailed information on the SAC can be found in Section 2.2.1.

Each image contains a semi–structured English annotation, describing the image content in detail (see Figure 8.1 for an example). Participants were provided with these annotations, a 368 x 234 large version and a 120 x 76 thumbnail of each image.

The SAC was chosen as the basis for ImageCLEF because the collection represents a realistic archive of images with high quality captions, and because permission was granted by St. Andrews Library to download and distribute the collection for use in the ad hoc retrieval task.

[4] http://www-library.st-andrews.ac.uk/

Table 8.1: Ad hoc query topics at ImageCLEF 2003.

ID	Topic Title	ID	Topic Title
1	Men and women processing fish	26	Portraits of Robert Burns
2	A baby in a pram	27	Children playing on beaches
3	Picture postcard views of St. Andrews	28	Pictures of golfers in the nineteenth century
4	Seating inside a church	29	Wartime aviation
5	Woodland scenes	30	Glasgow before 1920
6	Scottish marching bands	31	Exterior views of Indian temples
7	Home guard on parade during World War II	32	Scottish fishing vessels by the photographer Thompson
8	Tea rooms by the seaside	33	Male portraits
9	Fishermen by the photographer Adamson	34	Dogs rounding-up sheep
10	Ships on the river Clyde	35	The mountain Ben Nevis
11	Portraits of Mary Queen of Scots	36	Churches with tall spires
12	North Street St. Andrews	37	Men holding tennis racquets
13	War memorials in the shape of a cross	38	People using spinning machines
14	Boats on Loch Lomond	39	Men cutting peat
15	Tay bridge rail disaster	40	Welsh national dress
16	City chambers in Dundee or Glasgow	41	A coat of arms
17	Great Yarmouth beach	42	University buildings
18	Metal railway bridges	43	British Windmills
19	Culross abbey	44	Waterfalls in Wales
20	Road bridges	45	Harvesting
21	Animals by the photographer Lady Henrietta Gilmour	46	Postcards by the Valentine photographic company
22	Ruined castles in England	47	People dancing
23	London bridge	48	Museum exhibits
24	Damage due to war	49	Musician and their instruments
25	Golf course bunkers	50	Mountain scenery

8.2.2 Query Topics

In the first year, the topic creation process was based on two different approaches. First, the task organizers browsed the SAC to familiarize themselves with the subjects, which are available throughout the collection. Second, an analysis of the log files taken from the St. Andrews Library Web server that hosted the SAC for several years was carried out to identify popular queries.

Based on the log file analysis, which found that queries are commonly short and specific, modifications were made on some of the original queries to make them more suitable for visual retrieval. For example, the query 'church' was changed to 'churches with tall spires'. A total of 50 English query topics (see Table 8.1) were created to test various aspects of query translation and image retrieval, e.g. pictures of specific objects vs. pictures containing actions, broad vs. narrow concepts, topics containing proper names, compound words, abbreviations, morphological variants and idiomatic expressions.

Each topic consisted of a title (i.e. a short phrase describing the search request), a narrative (i.e. a description of what constitutes a relevant or non–relevant image for

<top>
<num> Number: 1 </num>
<EN-title n="1"> Men and women processing fish </EN-title>
<EN-narr> A relevant image will show men and/or women processing fish after catching them. Processing may include gutting or curing and the picture must show the fish processors at work; not just mention fish processing, e.g. that fish processing takes place at this port. An example relevant document is [stand03_2093/stand03_2382]. </EN-narr>
</top>

<top>
<num> Number: 1 </num>
<IT-title n="1"> Uomini e donne che puliscono il pesce </IT-title>
<IT-title n="2"> Pulizia del pesce al porto </IT-title>
<IT-title n="3"> uomini e donne che lavorano il pesce </IT-title>
</top>

Fig. 8.2: ImageCLEF 2003 sample topic.

Table 8.2: Ad hoc query topics at ImageCLEF 2004.

ID	Topic Title	ID	Topic Title
1	Portrait pictures of church ministers by Thomas Rodger	14	Elizabeth the Queen Mother visiting Crail Camp, 1954
2	Photos of Rome taken in April 1908	15	Bomb damage due to World War II
3	St. Andrews cathedral by John Fairweather	16	Pictures of York Minster
4	Men in military uniform, George Middlemass Cowie	17	Pictures of Edinburgh Castle taken before 1900
5	Fishing vessels in Northern Ireland	18	All views of North Street, St. Andrews
6	Views of scenery in British Columbia	19	People marching or parading
7	Exterior views of temples in Egypt	20	River with a viaduct in background
8	College or university buildings, Cambridge	21	Photos showing traditional Scottish dancers
9	Pictures of English lighthouses	22	War memorials in the shape of a cross
10	Busy street scenes in London	23	Photos of swans on a lake
11	Composite postcard views of Bute, Scotland	24	Golfers swinging their clubs
12	Tay Bridge rail disaster, 1879	25	Boats on a canal
13	The Open Championship golf tournament, St. Andrews 1939		

that search request), and an example relevant image to facilitate Content–Based Image Retrieval (CBIR) as well. Moreover, both topic titles and narratives were translated into Italian, German, Dutch, French, Spanish and Chinese to encourage participants to research Cross–Language Information Retrieval (CLIR) methods, too. Each translation was carried out by native speakers, who were also asked to specify alternative translations if appropriate. Figure 8.2 shows one sample topic and its Italian translation.

In 2004, 25 new topics (see Table 8.2) were created using a similar approach. Further, several categories (e.g. queries modified by date/location/photographer) were defined and the topics were modified to be distributed evenly within these categories.

Participants at ImageCLEF 2004 had suggested the creation of more visually–based query topics to allow for a more meaningful application of CBIR methods.

Table 8.3: Ad hoc query topics at ImageCLEF 2005.

ID	Topic Title	ID	Topic Title
1	Aircraft on the ground	15	Golfer putting on green
2	People gathered at bandstand	16	Waves breaking on beach
3	Dog in sitting position	17	Man or woman reading
4	Steam ship docked	18	Woman in white dress
5	Animal statue	19	Composite postcards of Northern Ireland
6	Small sailing boat	20	Royal visit to Scotland (not Fife)
7	Fishermen in boat	21	Monument to poet Robert Burns
8	Building covered in snow	22	Building with waving flag
9	Horse pulling cart or carriage	23	Tomb inside church or cathedral
10	Sun pictures & Scotland	24	Close-up picture of bird
11	Swiss mountain scenery	25	Arched gateway
12	Postcards from Iona, Scotland	26	Portrait pictures of mixed sex group
13	Stone viaduct with several arches	27	Woman or girl carrying basket
14	People at the marketplace	28	Colour pictures of woodland scenes around St Andrews

Table 8.4: Languages researched at ImageCLEF 2003–2005.

Language	2003	2004	2005	Language	2003	2004	2005
Arabic			✓	Hungarian			✓
Bulgarian			✓	Indonesian			✓
Chinese	✓	✓	✓	Italian	✓	✓	✓
Croatian			✓	Japanese		✓	✓
Czech			✓	Norwegian			✓
Danish		✓		Polish			✓
Dutch	✓	✓	✓	Portuguese			✓
Finnish		✓	✓	Romanian			✓
French	✓	✓	✓	Russian		✓	✓
English	✓	✓	✓	Spanish	✓	✓	✓
Filipino			✓	Swedish		✓	✓
German	✓	✓	✓	Turkish			✓
Greek			✓	Visual	✓	✓	✓

Hence, in 2005 the task organizers not only based the topic creation process on the log file analysis and Text–Based Image Retrieval (TBIR) challenges, but also on CBIR baseline runs and provided two sample images (compared to only one in the first two years). These query topics are depicted in Table 8.3.

The number of topic languages increased every year thanks to the help of many participants who contributed translations for the query topics in their native languages. Each translation was double–checked by another native speaker of the same language. By 2005, the topic titles had been translated into up to 31 different languages; yet, not all of them were used by participating groups. The actual use of languages in the retrieval experiments from 2003 to 2005 is summarized in Table 8.4.

Table 8.5: Highest MAP for each query language at ImageCLEF 2003.

Language	Group	MAP	Language	Group	MAP
English	Daedalus	0.5718 (monolingual)			
French	Sheffield	0.4380	Italian	Sheffield	0.4047
Spanish	Daedalus	0.4323	Dutch	Sheffield	0.3904
German	Sheffield	0.4285	Chinese	NTU	0.2888

8.2.3 Relevance Judgments and Performance Measures

The creation of relevance judgment was based on a pooling method and Interactive Search and Judge (ISJ). Both approaches are explained in Chapter 4.

In 2003, the top 100 results from all submitted runs were used to create image pools to be assessed for each topic (and in 2004 and 2005, the top 50 results respectively). To reduce judging subjectivity, each image in the topic pools was assessed by the topic creator and at least two other assessors using a ternary classification scheme: (1) relevant, (2) partially relevant, or (3) not relevant.

Based on these judgments, various combinations could be used to create the final set of relevant images (qrels). In all three years, the qrels were based on the pisac–total set: all images judged as relevant or partially relevant by the topic creator and at least one other assessor. ISJ was also used to supplement the image pools with further relevant images that had not been retrieved by any of the participants.

To evaluate the runs, the retrieval results were computed using the newest version of trec_eval[5]. In 2003 and 2004, only the (arithmetic) mean average precision (MAP) was used, while in 2005 methods were also compared using Precision at 10 and 100 images, P10 and P100 respectively, and the number of relevant images retrieved (RelRet). These and other performance measures are defined in Chapter 5.

8.2.4 Results and Analysis

Four groups participated at ImageCLEF 2003 and experimented with different translation methods, such as dictionary and on–line translation tools, and used Query Expansion (QE) to improve TBIR performance. Monolingual runs thereby consistently achieved higher performance than bilingual runs. Table 8.5 provides an overview of the highest MAP for each topic language.

All runs submitted in 2003 retrieved images based on their captions only. To encourage the use of visual methods, a CBIR system[6] was made available for the participants and query topics were also modified to be more visual in 2004 and 2005. As shown in Table 8.6, these measures taken by the ImageCLEF organizers

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

[6] GIFT system (http://www.gnu.org/software/gift/)

Table 8.6: Number and percentage of runs with respect to query dimensions.

Query Dimension	2003	2004	2005
Text only	45 (100%)	106 (56%)	318 (91%)
Combined	-	78 (41%)	27 (8%)
Visual only	-	6 (3%)	4 (1%)
TOTAL	45	190	349

Table 8.7: Highest MAP values for the top six languages at ImageCLEF 2004.

Language	Group	Run ID	MAP	QE	Text	Visual	Title	Narr
English	daedalus	mirobaseen	0.5865		✓		✓	
German	dcu	delsmgimg	0.5327	✓	✓	✓	✓	
Spanish	UNED	unedesent	0.5171	✓	✓		✓	
French	montreal	UMfrTFBTI	0.5125	✓	✓	✓	✓	
Italian	dcu	itlsstimg	0.4379	✓	✓		✓	
Dutch	dcu	nllsstimg	0.4321	✓	✓		✓	
Visual	geneva	GE_andrew4	0.0919	✓		✓		

Table 8.8: Top six languages with highest MAP at ImageCLEF 2005.

Language	Group	Run ID	MAP	QE	Text	Visual	Title	Narr	Image
English	CUHK	ad-eng-tv-kl-jm2	0.4135	✓	✓	✓	✓		✓
Chinese	NTU	CE-TN-WEprf-Ponly	0.3993	✓	✓	✓	✓	✓	
Spanish	Alicante, Jaen	R2D2vot2SpL	0.3447	✓	✓		✓		
Dutch	Alicante, Jaen	R2D2vot2Du	0.3435	✓	✓		✓		
Visual	NTU	NTU-adhoc05-EX-prf	0.3425	✓		✓			✓
German	Alicante, Jaen	R2D2vot2Ge	0.3375	✓	✓		✓		

subsequently proved to be effective as more participating groups submitted runs exploring the use of CBIR, or the combination of CBIR and TBIR, respectively.

Table 8.7 provides an overview of the highest MAP values for the languages in 2004. Popular translation methods included machine translation (73%), bilingual dictionaries and parallel corpora. A number of groups also improved their retrieval results by performing structured and constrained searches in order to identify named entities such as the photographer, date and location.

In most combined approaches, CBIR and TBIR were first performed separately, and then the ranked lists from both searches were merged. However, the combination of visual and textual approaches only managed to improve the performance of some topics. Also, purely visual searches performed poorly. This was (1) due to the fact that the query topics in 2004 did not involve enough visually–related topics, and (2) due to the nature of the images in the SAC which made CBIR difficult.

In 2005, the task organizers created query topics exhibiting more visual features. As a result, the results of visual approaches improved significantly. Table 8.8 provides the MAP scores for the top six highest performing languages. Most of these runs used QE and/or Relevance Feedback (RF). Twenty seven runs combined CBIR

Table 8.9: Average MAP by different modalities for ImageCLEF 2004 and 2005.

Modality	2004	2005
Text only	0.3787	0.2121
Combined text & image	0.4508	0.3086

and TBIR results, including the best monolingual run. On average, combined modality runs outperformed text–only runs in 2004 and 2005, as shown in Table 8.9.

Even though more visual queries were used in 2005, the number of runs using CBIR decreased compared to 2004. CBIR approaches did not seem to benefit from the visual features that could be extracted from the SAC. The evaluation using the SAC had reached a plateau due to several limitations with the collection: mainly black–and–white and grey–scale images (limiting the use of colour, as visual feature playing a vital role in CBIR), domain–specific annotation vocabulary in only one language (English), and restricted retrieval scenarios (i.e. search for historic photographs).

8.3 Ad hoc Retrieval of Generic Photographs: ImageCLEFphoto 2006-2007

At ImageCLEF 2005, participants had called for a test collection with richer visual features and multi–lingual annotations. Hence, in 2006 the SAC was replaced by a more generic photographic collection, the IAPR TC–12 database, created under Technical Committee 12 (TC–12) of the International Association of Pattern Recognition[7] (IAPR). Furthermore, the general photographic ad hoc retrieval task was given a new name (ImageCLEFphoto) in order to avoid confusion with the medical ad hoc retrieval task (ImageCLEFmed). Evaluation objectives and questions included:

- Are evaluation results obtained from the SAC also applicable to generic photos?
- Can combining CBIR and TBIR methods as well as using RF and/or QE improve retrieval performance also with generic photos?
- How does retrieval using short captions compare to using extensive captions?
- Are traditional TBIR methods still applicable for short captions?
- How significant is the choice of the retrieval and/or annotation language?

This section summarizes ImageCLEFphoto 2006 and 2007. More information can be found in the related overview papers: (Clough et al, 2007; Grubinger et al, 2008).

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

Title: Flamingo Beach
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;
Notes: Original name in Portuguese: 'Praia do Flamengo'; Flamingo Beach is considered as one of the most beautiful beaches of Brazil
Location: Salvador, Brazil
Dates: 2 October 2004

Fig. 8.3: Sample image and caption of the IAPR TC–12 database.

8.3.1 Test Collection and Distribution

The photographic collection of the IAPR TC–12 database contains 20,000 colour photos taken from locations around the world and comprises a varying cross–section of still natural images. This test collection, which was specifically built to support the evaluation needs of ImageCLEF, was the core component of ImageCLEFphoto 2006 and 2007. Detailed information on the creation and contents of the IAPR TC–12 database can be found in Section 2.2.2 of Chapter 2 and (Grubinger et al, 2006).

Figure 8.3 illustrates a sample image from the IAPR TC–12 database. Each image in the collection comprises corresponding semi–structured annotations in three different languages: English, German and Spanish. The annotation structure was thereby very similar to that used in the SAC (compare Table 8.1) to provide a smooth transition for returning participants. Only the 'categories' field was missing as it had hardly been used in retrieval from the SAC.

The ImageCLEF organizers used the parametric nature of the test collection and created a different subset of the test collection each year. Consequently, the participants of ImageCLEFphoto 2006 were provided with 20,000 images and the corresponding English and German captions exhibiting a varying degree of annotation 'completeness':

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

One year later, ImageCLEFphoto 2007 focused on whether TBIR methods would still be suitable to find images with short captions. Thus, the description field was removed from the annotations and participants were provided with annotations only containing title, notes, location and date. The lack of textual information should encourage participants to use CBIR techniques. Four sets of annotations were provided: (1) English, (2) German, (3) Spanish, and (4) one set whereby the annotation language was randomly selected for each of the images.

Table 8.10: Query topics in the IAPR TC–12 database.

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 Inka-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	destination 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 unsymmetric 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

8.3.2 Query Topics

The participants were given 60 query topics (see Table 8.10) representing typical search requests for the photographic collection of the IAPR TC–12 database.

The creation of these topics had been based on several factors, including: the analysis of a log file from on–line access to the image collection; knowledge of the collection content; various types of linguistic and pictorial attributes; the use of geographic constraints; and the estimated difficulty of the topic.

In particular, 40 topics were directly taken from the log files with slight syntactic modification (e.g. 'lighthouse sea' was changed to 'lighthouse at the sea'). Another ten were derived from the logs (e.g. 'straight roads in Argentina' was changed to

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

Fig. 8.4: Sample query topic at ImageCLEFphoto 2006.

'straight roads in the USA'). The rest of the queries was not taken from the logs, but created to assess specific aspects of CBIR and TBIR (e.g. 'black and white photos of Russia'). There were 24 queries which contained geographical constraints (e.g. 'tourist accommodation near Lake Titicaca') since these queries were quite common in the log. Half of the topics were classified as 'semantic', one third as 'neutral' and the rest as 'visual'. CBIR approaches were not expected to improve retrieval results in semantic topics, while the visual topics would benefit from the use of visual approaches. More information can be found in (Grubinger, 2007).

The format of the topics (see Figure 8.4) was identical with the one used in previous years, again to provide a smooth transition for returning participants: each topic contained a title, a narrative description and three sample images. The same queries were used in the two year period to allow for a comparison of retrieval from collections of fully annotated (2006) and lightly annotated (2007) photographs.

In both years, the topic titles were provided in 16 languages, including: English, German, Spanish, Italian, French, Portuguese, Chinese, Japanese, Russian, Polish, Swedish, Finnish, Norwegian, Danish and Dutch. All translations were provided by native speakers and verified by at least one other native speaker. Since the annotations were provided in two languages in 2006 (and four sets in 2007), this created 32 potential bilingual retrieval pairs (and even 64 in 2007, respectively).

8.3.3 Relevance Judgments and Performance Measures

Similar to the first three years, the relevance assessments at ImageCLEFphoto 2006 and 2007 were based on the pooling method and ISJ. Image pools were created by taking the top 40 results from all participants' runs, yielding an average 1,045 images to be judged per query topic in 2006 (and 2,299 images in 2007, respectively). ISJ was also being deployed to find more relevant images that were not returned by any methods within the top 40 results, and the resulting pools in 2007 were

Table 8.11: Top results at ImageCLEF 2006.

Topic	Caption	Group	Run ID	MAP	P20	GMAP	bpref
EN	EN	CINDI	Cindi_Exp_RF	0.385	0.530	0.282	0.874
PT	EN	NTU	PT-EN-AUTO-FB-TXTIMG-T-WEprf	0.285	0.403	0.177	0.755
ZH	EN	NTU	ZHS-EN-AUTO-FB-TXTIMG-TOnt-WEprf	0.279	0.464	0.154	0.669
RU	EN	NTU	RU-EN-AUTO-FB-TXTIMG-T-WEprf	0.279	0.408	0.153	0.755
SP	EN	NTU	SP-EN-AUTO-FB-TXTIMG-T-WEprf	0.278	0.407	0.175	0.757
DE	DE	NTU	DE-DE-AUTO-FB-TXTIMG-T-WEprf	0.311	0.335	0.132	0.974
EN	DE	DCU	combTextVisual_ENDEEN	0.122	0.175	0.036	0.524
FR	DE	DCU	combTextVisual_FRDEEN	0.104	0.147	0.002	0.245
Vis.	-	RWTH	RWTHi6-IFHTAM	0.063	0.182	0.022	0.366

complemented with further relevant images found in 2006 to avoid missing out on relevant images not found in 2007 due to the reduced captions. The assessments were, again, based on a ternary classification system, whereby this time, only those images judged relevant by both assessors were considered for the qrels.

The runs were evaluated using MAP and P20. The latter was chosen because most on–line image retrieval search engines display 20 images by default. Other measures used include the GMAP, which tests system robustness, and the binary preference (bpref) to indicate the bias due to incompleteness of relevance judgments.

8.3.4 Results and Analysis

There was an increasing number of participating groups: 12 groups submitted in 2006, and 20 groups in 2007. This was the highest number of participants at ImageCLEF thus far, which was an indication that the need for evaluation of VIR had increased over the years, and that ImageCLEFphoto was considered as a suitable track to explore this field of research. As a consequence, many novel retrieval methods and ideas were investigated. Tables 8.11 and 8.12 show the results for the best performing language pairs (MAP) in both years, but also indicate that the choice of the performance measure does affect system ranking. An overview of all retrieval methods and complete results are available in the ImageCLEF overview papers (Clough et al, 2007; Grubinger et al, 2008).

Comparing the results from both years, it is interesting to see how monolingual results were more affected by the annotation reduction than bilingual results. While monolingual retrieval produced better results than bilingual retrieval in 2006, the results at ImageCLEFphoto 2007 suggested that, on average, bilingual results were as competitive as monolingual results. This might be due to the short image captions provided in 2007, but could also be credited to improved translation resources. Moreover, the choice of the query language was almost negligible in 2007, most likely because many of the short captions contained proper nouns.

Table 8.12: Top results at ImageCLEF 2007.

Topic	Caption	Group	Run ID	MAP	P20	GMAP	bpref
EN	EN	CUT	cut-EN2EN-F50	0.318	0.459	0.298	0.162
PT	EN	NTU	PT-EN-AUTO-FBQE-TXTIMG	0.282	0.388	0.266	0.127
ZH	EN	NTU	ZHT-EN-AUTO-FBQE-TXTIMG	0.257	0.360	0.240	0.089
RU	EN	NTU	RU-EN-AUTO-FBQE-TXTIMG	0.273	0.383	0.256	0.115
ES	EN	NTU	ES-EN-AUTO-FBQE-TXTIMG	0.279	0.383	0.259	0.128
DE	DE	NTU	DE-DE-AUTO-FBQE-TXTIMG	0.245	0.379	0.239	0.108
EN	DE	DCU	combTextVisual_ENDEEN	0.278	0.362	0.250	0.112
FR	DE	DCU	combTextVisual_FRDEEN	0.164	0.237	0.144	0.004
Vis.	-	XRCE	AUTO-NOFB-IMG_COMBFK	0.189	0.352	0.201	0.102

Table 8.13: Results by retrieval modalities at ImageCLEFphoto 2006 and 2007.

Year	2006				2007			
Modality	MAP	P20	bpref	GMAP	MAP	P20	bpref	GMAP
Image	0.041	0.134	0.296	0.014	0.068	0.157	0.080	0.022
Text	0.129	0.173	0.465	0.027	0.120	0.152	0.141	0.018
Combined	0.199	0.281	0.650	0.095	0.149	0.225	0.203	0.050

Table 8.13 shows that combined text and image retrieval outperformed text–only and visual–only retrieval approaches. This trend had already been indicated for retrieval from historic photographic collections and has now continued for retrieval from generic photographic collections as well. The same is true for the use of RF and/or QE, which were also shown to improve retrieval performance in 2006 and 2007.

8.3.5 Visual Sub–task

To attract more visually–orientated groups, a visual sub–task was run in 2006 to investigate CBIR–only techniques: all image captions were deleted, and retrieval had to rely on CBIR techniques only. Thirty queries were selected from the original 60 query topics, with some modifications being made to remove non–visual constraints such as location. For example, the query 'black and white photos from Russia' was changed to 'black and white photos'. Some examples of the visual topics are shown in Table 8.14.

Only two out of 36 registered participants eventually submitted runs to this subtask. The highest performing run, submitted by RWTH University Aachen, Germany, used invariant and Tamura texture feature histograms. The evaluation showed promising results for P20, which was 0.285. However, MAP was very low (0.101 for the best run). This was due to the fact that relevant images found in P20 were quite similar to sample images given in the query (Clough et al, 2007).

Table 8.14: Example of topics in the visual sub–task of ImageCLEFphoto 2006.

ID	Topic Title	Level
82	sunset over water	Easy
78	bird flying	Easy
67	scenes of footballers in action	Medium
84	indoor photos of churches or cathedrals	Medium
83	images of typical Australian animals	Difficult
61	church with more than two towers	Difficult

8.4 Ad hoc Retrieval and Result Diversity: ImageCLEFphoto 2008–2009

The ImageCLEF ad hoc retrieval tasks had followed the evaluation scenario similar to the classical TREC ad hoc retrieval task during the first five years (see Section 8.1). However, in 2008 this scenario was slightly changed: systems were not only expected to return relevant images for a given search request, but also to return these relevant images from as many different sub-topics as possible (to promote retrieval diversity) in the top n results. This novel challenge allowed for the investigation of a number of new research questions, including:

- Is it possible to promote diversity within the top n results?
- Which retrieval approaches work best at promoting diversity?
- Does promoting diversity sacrifice relevance (i.e. precision)?
- How do results compare between bilingual and multi–lingual annotations?
- Do mixed approaches still outperform text or image only methods?
- How much does a priori knowledge about query clusters help to increase diversity?

This section summarizes the ImageCLEFphoto 2008 and 2009 tasks. More information can be found in the respective overview papers: Arni et al (2009); Paramita et al (2010).

8.4.1 Test Collection and Distribution

As in previous years, the IAPR TC–12 database provided the resources for ImageCLEFphoto 2008. Since the evaluation concentrated on diversity within the top retrieval results, a different collection subset to that used in 2006 and 2007 was generated: participants were given two sets of complete annotations (i.e. all caption fields were provided) in (1) English and (2) 'Random', whereby the language for each caption was randomly selected from either English or German.

Reusing the same image collection as in previous years allowed for the investigation of whether precision is affected when diversity is implemented. However, ImageCLEF participants felt in 2008 that the time had come to move on to a bigger

837661 MOS06-20020212-MOSCOW, RUSSIAN FEDERATION: Russian Foreign Minister Igor Ivanov (L) shakes hands with Afghanistan's interim Defence Minister General Mohammad Qasim Fahim (R) to start their talks at the Foreign Ministry in Moscow on Tuesday 12 February 2002. Russia will give the technical and logistical assistance to Afghanistan's army but will not train Afghan military specialists, it was announced. EPA PHOTO EPA-SERGEI CHIRIKOV-vk-fob

Fig. 8.5: Sample image and caption from Belga.

Table 8.15: Examples of different clusters at ImageCLEFphoto 2008.

ID	Topic Title	Cluster
2	church with more than two towers	city
3	religious statue in the foreground	statue
5	animal swimming	animal
12	people observing football match	venue
23	sport photos from Australia	sport
50	indoor photos of a church or cathedral	country

image archive for evaluation. Hence, in 2009 a new challenge was offered by replacing the IAPR TC–12 database with a database that was nearly 25 times larger: the photographic collection of *Belga*[8], a Belgian news agency (see also Section 2.2.3 in Chapter 2).

This data set comprised 498,920 photos with unstructured, English–only annotations (see Figure 8.5). This offered new challenges to the participants in comparison to the SAC and IAPR TC–12 collections. For example, the unstructured nature of the image captions required the automatic extraction of information about, for example the location, date or photographic source of the image as a part of the indexing and retrieval process. In addition, it contained many cases where pictures had not been orientated correctly, thereby making CBIR more difficult (Paramita et al, 2009).

8.4.2 Query Topics

ImageCLEFphoto 2008 used a subset of the previous year's queries: 39 topics were identified that would also be useful for the evaluation of retrieval diversity. The annotation structure was thereby identical to that used in 2006 and 2007, apart from an additional cluster field that was included to represent the diversity need.

For example, the query 'vehicle in South Korea' would benefit from retrieval diversity with respect to 'vehicle types' (see Figure 8.6). A selection of query examples together with their corresponding clusters is illustrated in Table 8.15.

[8] http://www.belga.be/

<top>
<num> Number: 48 </num>
<title> vehicle in South Korea </title>
<**cluster**> **vehicle type** </**cluster**>
<narr> Relevant images will show vehicles in South Korea, including cars, trains, buses, forklifts, boats, and so on. Images with vehicles outside of South Korea are not relevant. Images from South Korea without a single vehicle are not relevant either. </narr>
<image> SampleImages/48/35645.jpg </image>
<image> SampleImages/48/35705.jpg </image>
<image> SampleImages/48/35982.jpg </image>
</top>

Fig. 8.6: Sample query topic at ImageCLEFphoto 2008.

<top>
<num> 12 </num>
<title> clinton </title>
<**clusterTitle**> **hillary clinton** </**clusterTitle**>
<clusterDesc> Relevant images show photographs of Hillary Clinton. Images of Hillary with other people are relevant if she is shown in the foreground. Images of her in the background are not relevant. </clusterDesc>
<image> belga26/05859430.jpg </image>
<**clusterTitle**> **obama clinton** </**clusterTitle**>
<clusterDesc> Relevant images show ... </clusterDesc>
<image> belga... </image>
<**clusterTitle**> **bill clinton** </**clusterTitle**>
<clusterDesc> Relevant images show ... </clusterDesc>
<image> belga... </image>

Fig. 8.7: Example of Query Part 1 at ImageCLEFphoto 2009.

The topic creation process for ImageCLEFphoto 2009 was based on search query logs from Belga. In contrast to 2008, where the cluster fields had been estimated based on the query topics, the information on query variations could also be extracted from the log file. For example, 'Victoria Beckham' and 'David Beckham' were variations (and at the same time clusters) for a query looking for 'Beckham'. Eventually, 50 topics (with an average number of four clusters each) were generated, divided in two sets of 25 topics each and released in two different formats: 'Query Part 1' and 'Query Part 2'.

Figure 8.7 provides an example for Query Part 1, which includes a topic title, cluster title, cluster description and an example image. All potential retrieval clusters were provided as a part of the query topic, simulating the situation in which search engines have access to query logs telling the system what variations to expect.

However, in real–life scenarios, often little or no query log information is available to indicate potential clusters. Thus, in the second set of query topics, Query Part 2, little evidence was given for what kind of diversity was expected: the `clusterTitle` and `clusterDesc` fields were hidden, and only the topic title and three example images were provided for CBIR approaches (which, in many

```
<top>
<title> obama </title>
<num> 26 </num>
<image> belga30/06098170.jpg </image>
<image> belga28/06019914.jpg </image>
<image> belga30/06017499.jpg </image>
```

Fig. 8.8: Example of Query Part 2 at ImageCLEFphoto 2009.

cases, would not cover all clusters). Figure 8.8 provides an example for Query Part 2. Further information regarding query and cluster development at ImageCLEFphoto 2009 is available in Paramita et al (2009).

8.4.3 Relevance Judgments and Performance Measures

The relevance assessments from 2007 were reused for ImageCLEFphoto 2008. In addition, the images were assigned one (or more) predefined clusters to enable the quantification of retrieval diversity. Two assessors carried out the classification process, while a third assessor was used to resolve any inconsistent judgments.

In 2009, the relevance assessments were performed using Distributed Information Retrieval Evaluation Campaign Tool[9] (DIRECT) and were carried out in two phases: (i) the relevant images for each query were identified; and (ii) these relevant images were assigned to the clusters. Due to the large collection, the pool sizes rose drastically compared to previous years; thus, each image was only evaluated by one assessor. An average of 700 images were found to be relevant for each query, and around 200 images were relevant for each cluster.

To evaluate the search results, standard IR measures were used: MAP, GMAP and bpref. Retrieval diversity was evaluated using cluster recall CR(n), which represents the percentage of clusters retrieved in the top n documents (Zhai et al, 2003). Moreover, F_1 was used to combine P20 and CR20 in 2008, and P10 and CR10 in 2009 respectively, because the number of clusters had an upper bound of 10 in that year. For a definition of these performance measures, see Chapter 5.

8.4.4 Results and Analysis

ImageCLEFphoto managed to attract more than 40 groups, which registered in both years of the task; 24 submitted results in 2008, and 19 in 2009 respectively. Most participants employed post–processing methods to achieve result diversity. They started the retrieval process by using TBIR baseline runs enhanced by RF/QE to

[9] http://direct.dei.unipd.it/

Table 8.16: Systems with highest F_1 across all 39 topics at ImageCLEFphoto 2008.

Group	Run-ID	Run Type	Modality	P20	CR20	F_1
PTECH	EN-EN-MAN-TXTIMG	MAN	TXT-IMG	0.6885	0.6801	0.6843
PTECH	EN-EN-MAN-TXTIMG-MMBMI	MAN	TXT-IMG	0.6962	0.6718	0.6838
PTECH	EN-EN-MAN-TXT-MTBTN	MAN	TXT	0.5756	0.5814	0.5785
XRCE	xrce_tilo_nbdiv_15	AUTO	TXT-IMG	0.5115	0.4262	0.4650
DCU	EN-EN-AUTO-TXTIMG-QE	AUTO	TXT-IMG	0.4756	0.4542	0.4647
XRCE	xrce_tilo_nbdiv_10	AUTO	TXT-IMG	0.5282	0.4146	0.4646

Table 8.17: Systems with highest F_1 across all 50 topics at ImageCLEFphoto 2009.

Group	Run Name	Topic Fields*	Modality	P10	CR10	F_1
XEROX-SAS	XRCEXKNND	T-CT-I	TXT-IMG	0.794	0.8239	0.8087
XEROX-SAS	XRCECLUST	T-CT-I	TXT-IMG	0.772	0.8177	0.7942
XEROX-SAS	KNND	T-CT-I	TXT-IMG	0.800	0.7273	0.7619
INRIA	LEAR5_TI_TXTIMG	T-I	TXT-IMG	0.798	0.7289	0.7619
INRIA	LEAR1_TI_TXTIMG	T-I	TXT-IMG	0.776	0.7409	0.7580
InfoComm	LRI2R_TI_TXT	T-I	TXT	0.848	0.6710	0.7492

* T = Title, CT = Cluster Title, I = Image

Table 8.18: Performance measures for different query formats.

Queries	Runs	P10	CR10	F_1
Queries part 1 with CT	52	0.6845	0.5939	0.6249
Queries part 1 without CT	32	0.6641	0.5006	0.5581
Queries part 2	84	0.6315	0.5415	0.5693

maximize the number of relevant images in the top n results. Diversity was then promoted by re–ranking the initial run, clustering the top n documents, and selecting the highest ranked document in each cluster to create diverse results.

The top six results across all query topics of ImageCLEFphoto 2008 and 2009 are shown in Tables 8.16 and 8.17. In 2008, the top ten results were all monolingual (English), with the highest bilingual run exhibiting P20 of 0.4397, CR20 of 0.4673 and F_1 of 0.4531. On average, however, the margin between monolingual and bilingual runs was low, continuing the trend of previous years. In 2009, only monolingual runs were evaluated since English was the only language for both annotations and topics. Retrieval results were much higher than in 2008, which was due to less semantic and hence easier topics compared to those used the year before.

Table 8.18 provides the results of the analysis on whether the different query formats influence retrieval effectiveness. Since participants could choose which query fields to use for retrieval, the scores for Query Part 1 were divided into runs which used the cluster title (CT) and runs which did not. The scores between Query Parts 1 and 2 were found to be significantly different.

Table 8.19 shows average scores of the top 20 results across all runs with respect to their retrieval modalities for 2008 and 2009. Mixed CBIR and TBIR methods

Table 8.19: Results by retrieval modalities at ImageCLEFphoto 2008 and 2009.

Year	2008			2009		
Modality	P20	CR20	F_1	P20	CR20	F_1
Image only	0.1625	0.2127	0.1784	0.0787	0.2986	0.1244
Text only	0.2431	0.3915	0.2957	0.6915	0.622	0.6454
Combined	0.2538	0.3998	0.3034	0.6994	0.6883	0.6913

Table 8.20: Participation overview for ImageCLEFphoto 2003-2009.

Queries	2003	2004	2005	2006	2007	2008	2009
Registered groups			19	36	32	43	44
Participating groups	4	12	11	12	20	24	19
Submitted runs	45	190	349	157	616	1042	84

provided the best results also in evaluation scenarios promoting retrieval diversity, although the difference to TBIR–only methods was, on average, only marginal. Yet, looking at the best runs (see Tables 8.16 and 8.17), mixed approaches still outperform TBIR–only approaches. CBIR methods have slightly caught up, but still lag behind.

8.5 Conclusion and Future Prospects

After the image retrieval community had been calling for resources similar to those used by TREC in its ad hoc retrieval tasks for the text retrieval domain, ImageCLEF began in 2003 to also provide similar resources within the context of VIR to facilitate standardised laboratory–style testing of cross–language image retrieval systems. While these resources have predominately been used by systems applying a TBIR approach, there has also been an increasing number of groups using CBIR approaches over the years. Benchmark resources created for ad hoc retrieval from photographic collections include the following:

- historic photographs with extensive semi–structured annotations;
- generic photographs with multi–lingual semi–structured annotations;
- a large press collection containing photos with unstructured annotations.

ImageCLEF ran seven ad hoc cross–language image retrieval tasks for the domain of photographic collections from 2003 to 2009, thereby addressing two main fields of information retrieval research: (1) image retrieval and (2) CLIR. The tasks were modelled on scenarios found in multimedia use at the time and proved to be very popular among researchers as shown by an increasing number of participants over the years (see Table 8.20).

Moreover, each year a large number of participants also registered without eventually submitting results, only to get access to the valuable benchmark resources. In 2009, the much lower number of submitted runs was due to a limitation of five

runs for each participating group (before that, an unlimited number of runs could be submitted, all of which were evaluated). In the first four years (2003 to 2006), retrieval from collections with extensive captions suggested the following trends for both historic and generic photographs:

- Using QE and/or RF improves retrieval performance.
- Combining CBIR and TBIR methods improves retrieval performance.
- Monolingual runs outperform bilingual runs.
- Retrieval success does still depend on the annotation language.
- The retrieval difficulty of a topic can be pre–determined.
- The choice of qrels and performance measures can affect system ranking.

At ImageCLEFphoto 2007, most of these trends could be verified also for retrieval from image collections with light annotations, with the following exceptions that indicated that for short captions:

- Bilingual runs perform as well as monolingual runs.
- The choice of query or annotation language hardly affects retrieval success.

The challenge of ImageCLEFphoto in 2008 and 2009 was slightly different to that in previous years and was based on promoting diversity in the search results. Results from both years showed that:

- It is possible to present a diverse result without sacrificing precision.
- A priori information about the cluster title is essential for retrieval diversity.
- A combination of title, cluster title and image maximizes diversity and relevance.
- Mixed runs (CBIR and TBIR) outperform runs based on TBIR or CBIR alone.
- Bilingual retrieval performs nearly as well as monolingual retrieval.

The change of direction in the evaluation objective in 2008 showed that, as the field of VIR develops, test collections and evaluation events need to evolve and react to those changes as well. ImageCLEFphoto is not an exception and will, hence, continue to provide resources to the VIR community in the future to facilitate standardized laboratory–style testing of image retrieval systems.

Acknowledgements We would like to thank the providers of the data sets: St. Andrews University Library for ImageCLEF 2003–2005, viventura for ImageCLEFphoto 2006–2008; Belga Press Agency for ImageCLEFphoto 2009; Michael Grubinger for the IAPR TC–12 database; Theodora Tsikrika for the Belga query log analysis; and Giorgio di Nunzio and the DIRECT team for their support at ImageCLEFphoto 2009. ImageCLEFphoto was funded by the EU under the projects: TrebleCLEF (Grant agreement: 215231), Multimatch (contract IST–2005–2.5.10) and the MUSCLE NoE (FP6–507752).

References

Arni T, Clough PD, Sanderson M, Grubinger M (2009) Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In: Peters C, Deselaers T, Ferro N, Gonzalo J, Jones G, Kurimo

M, Mandl T, Peñas A, Petras V (eds) Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross–Language Evaluation Forum (CLEF 2008). Lecture Notes in Computer Science (LNCS), vol 5706. Springer, Aarhus, Denmark, pp 500–511

Clough PD, Sanderson M (2004) The CLEF 2003 Cross Language Image Retrieval Track. In: Peters C, Gonzalo J, Braschler M, Kluck M (eds) Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003). Lecture Notes in Computer Science (LNCS), vol 3237. Springer, Trondheim, Norway, pp 581–593

Clough PD, Müller H, Sanderson M (2005) The CLEF Cross Language Image Retrieval Track 2004. In: Peters C, Clough P, Gonzalo J, Jones G, Kluck M, Magnini B (eds) Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Lecture Notes in Computer Science (LNCS), vol 3491. Springer, Bath, United Kingdom, pp 597–613

Clough PD, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W (2006) The CLEF 2005 Cross–Language Image Retrieval Track. In: Peters C, Gey FC, Gonzalo J, Müller H, Jones GJF, Kluck M, Magnini B, de Rijke M, Giampiccolo D (eds) Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Lecture Notes in Computer Science (LNCS), vol 4022. Springer, Vienna, Austria, pp 535–557

Clough PD, Grubinger M, Deselaers T, Hanbury A, Müller H (2007) Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In: Peters C, Clough PD, Gey FC, Karlgren J, Magnini B, Oard DW, de Rijke M, Stempfhuber M (eds) Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science (LNCS), vol 4730. Springer, Alicante, Spain, pp 579–594

Goodrum A (2000) Image Information Retrieval: An Overview of Current Research. Informing Science. Special Issue on Information Science Research 3(2):63–66

Grubinger M (2007) Analysis and Evaluation of Visual Information Systems Performance. PhD thesis, School of Computer Science and Mathematics. Faculty of Health, Engineering and Science. Victoria University, Melbourne, Australia

Grubinger M, Clough PD, Müller H, Deselaers 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 2006, Genoa, Italy, pp 13–23

Grubinger M, Clough PD, Hanbury A, Müller H (2008) Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In: Peters C, Jijkoun V, Mandl T, Müller H, Oard DW, Peñas A, Petras V, Santos D (eds) Advances in Multilingual and Multimodal Information Retrieval. 8th Workshop of the Cross–Language Evaluation Forum (CLEF 2007). Lecture Notes in Computer Science (LNCS), vol 5152. Springer, Budapest, Hungary, pp 433–444

Leung CHC, Ip H (2000) Benchmarking for Content–Based Visual Information Search. In: Laurini R (ed) Fourth International Conference On Visual Information Systems (VISUAL'2000). Lecture Notes in Computer Science (LNCS), vol 1929. Springer, Lyon, France, pp 442–456

Paramita ML, Sanderson M, Clough PD (2009) Developing a Test Collection to Support Diversity Analysis. In: Proceedings of the ACM SIGIR 2009 Workshop: Redundancy, Diversity, and Interdependence Document Relevance, Boston, MA, USA, pp 39–45

Paramita ML, Sanderson M, Clough PD (2010) Diversity in Photo Retrieval: Overview of the ImageCLEFphoto Task 2009. In: Peters C, Tsikrika T, Müller H, Kalpathy-Cramer J, Jones JFG, Gonzalo J, Caputo B (eds) Multilingual Information Access Evaluation Vol. II Multimedia Experiments: Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum (CLEF 2009), Revised Selected Papers. Lecture Notes in Computer Science (LNCS). Springer

Zhai CX, Cohen WW, Lafferty J (2003) Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM press, Toronto, Canada, pp 10–17