The CLEF Cross-Language Image Retrieval Track (ImageCLEF) 2004

Paul Clough†, Henning Müller‡and Mark Sanderson†

†University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
‡University and University Hospitals of Geneva, Service of Medical Informatics,
rue Micheli–du–Crest 24,1211 Geneva 14, Switzerland.
p.d.clough@sheffield.ac.uk, henning.mueller@sim.hcuge.ch,
m.sanderson@sheffield.ac.uk

Abstract. The purpose of this paper is to outline efforts from the 2004 CLEF cross–language image retrieval campaign (ImageCLEF). The aim of this CLEF track is to explore the use of both text and content–based retrieval methods for cross–language image retrieval. Three tasks were offered in the ImageCLEF track: a TREC–style ad-hoc retrieval task, retrieval from a medical collection, and a user–centered (interactive) evaluation task. Eighteen research groups from a variety of backgrounds and nationalities participated in ImageCLEF. In this paper we describe the ImageCLEF tasks, submissions from participating groups and summarise the main findings.

1 Introduction

A great deal of research is currently underway in the field of Cross–Language Information Retrieval (CLIR) [1]. Campaigns such as CLEF and TREC have proven invaluable in providing standardised resources for comparative evaluation for a range of retrieval tasks. However, one area of CLIR which has received less attention is image retrieval. In many collections (e.g. historic or stock–photographic archives, medical databases and art/history collections), images are often accompanied by some kind of text (e.g. metadata or captions) semantically related to the image. Retrieval can then be performed using primitive features based on pixels which form an image's content (Content–Based Image Retrieval or CBIR [2]), using abstracted textual features assigned to the image, or a combination of both. The language used to express the associated texts or metadata should have a minimal effect on their usefulness for retrieval and be, as far as possible, language independent (e.g. an image with English captions should be searchable in languages other than English). Practically, this would enable organisations who manage image collections such as Corbis[1] or Getty Images[2] to be able to offer the same collection to a wider and more diverse range

[1] See http://www.corbis.com/

[2] See http://www.gettyimages.com/

of users with different language backgrounds. It is this area of CLIR which we address in ImageCLEF[3], the CLEF cross–language image retrieval campaign.

In 2003, we organised a pilot experiment with the following aim: given a multilingual statement describing a user need, find as many relevant images as possible [3]. A collection of historic photographs from St. Andrews University Library was used as the dataset and 50 representative search topics created to simulate the situation in which a user expresses their need in text in a language different from the collection and requires a visual document to fulfil their search request (e.g. searching an on–line art gallery or stock–photographic collection). Four groups from industry and academia participated using purely text–based retrieval methods and a variety of translation and query expansion methods.

To widen the scope of tasks offered by ImageCLEF and offer greater diversity to participants, in 2004 we offered both a medical retrieval and a user–centered evaluation task, along with a bilingual ad hoc retrieval task based on the St. Andrews photographic collection. To encourage participants to use content–based retrieval methods in combination with text–based methods, we did the following: (1) provided participants with access to a default CBIR system[4], and (2) created a medical retrieval task where initial retrieval is visual. These ideas payed off as many groups used visual retrieval, only [4–6], and the supplied visual system was also used several times [7, 8]. A number of groups combined visual and textual approaches [9, 10] Also, to promote ImageCLEF as the CLEF entry–level CLIR task, we offered topics in 12 languages rather than the 6 offered in 2003. In the following sections of this paper we describe the test collections, the search tasks, participating research groups, results from ImageCLEF 2004 and a summary of the main findings.

2 The ImageCLEF 2004 Tasks

Evaluation of a retrieval system is either system–focused (e.g. comparative performance between systems) or user–centered, e.g. a task–based user study. ImageCLEF offers the necessary resources and framework for comparative and user–centered evaluation. Two image collections were provided: (1) the St. Andrews collection of historic photographic images, and (2) the CasImage radiological medical database. In addition, example search topics and relevance assessments or ground truths (called *qrels*) based on submitted entries were also provided.

Two tasks were offered which used the St. Andrews collection: (1) a bilingual ad hoc retrieval task: given an initial topic find as many relevant images as possible, and (2) a known–item interactive task: given a target image, users

[3] See `http://ir.shef.ac.uk/imageclef2004/` for further information about the ImageCLEF 2004 campaign.

[4] We offered access to the VIPER system (`http://viper.unige.ch/`) through: (1) PHP, (2) a list of the top N images from a visual search using given exemplar images, and (3) via local download and installation of GIFT `http://www.gnu.org/software/gift/`.

Table 1. Participating Groups in ImageCLEF 2004.

Group	ID	Country	Medical (#Runs)	Ad-hoc (#Runs)	Interactive
National Taiwan University	ntu	Taiwan		⋆ (5)	
I–Shou University	KIDS	Taiwan	⋆ (3)	⋆ (4)	⋆
University of Sheffield	sheffield	UK		⋆ (5)	
Imperial College	imperial	UK	⋆ (1)		
Dublin City University	dcu	Ireland		⋆ (79)	
University of Montreal	montreal	Canada		⋆ (11)	
Oregon Health and Science U.	OSHU	USA	⋆ (1)		
State University of New York	Buffalo	USA	⋆ (3)		
Michigan State University	msu	USA		⋆ (4)	⋆
University of Alicante	alicante	Spain		⋆ (27)	
Daedalus	daedalus	Spain	⋆ (4)	⋆ (40)	
UNED	uned	Spain		⋆ (5)	
University Hospitals Geneva	geneva	Switzerland	⋆ (14)	⋆ (2)	
Dept. Medical Informatics, Aachen	aachen–inf	Germany	⋆ (2)		
Dept. Computer Science, Aachen	aachen–med	Germany	⋆ (8)	⋆ (4)	
University of Tilburg	tilburg	Netherlands	⋆ (1)		
CWI	cwi	Netherlands	⋆ (4)		
Commissariat Energie Automique	cea	France	⋆ (2)	⋆ (4)	
			11 (43)	12 (190)	2

must find it again. For the CasImage collection, a query–by–example search task was offered: given an initial medical image find as many relevant images as possible. It is, of course, difficult to create evaluation resources which test all kinds of retrieval system, but the tasks offered do pose different challenges and will appeal to researchers from a variety of backgrounds.

2.1 Participating Groups

In total 18 groups participated in ImageCLEF 2004 (Table 1): 11 in the medical task, 12 in the bilingual ad hoc task and 2 in the interactive task. This evaluation attracted research groups from 10 countries with various retrieval backgrounds including text, visual and medical. In total 43 submissions (*runs*) were submitted to the medical task, 190 to the ad-hoc task and 2 to the interactive task.

Short title: Rev William Swan.
Long title: Rev William Swan.
Location: Fife, Scotland
Description: Seated, 3/ 4 face studio portrait of a man.
Date: ca.1850
Photographer: Thomas Rodger
Categories: `[ ministers ][ identified male ][ dress - clerical ]`
Notes: ALB6-85-2 jf/ pcBIOG: Rev William Swan ( ) ADD: Former owners of album: A Govan then J J? Lowson. Individuals and other subjects indicative of St Andrews provenance. By T. R. as identified by Karen A. Johnstone " Thomas Rodger 1832-1883. A biography and catalogue of selected works".

Fig. 1. An example image and caption from the St. Andrews collection.

2.2 Ad Hoc Retrieval from the St. Andrews Collection

Similar to the TREC ad hoc retrieval task, we test retrieval when a system is expected to match a user's one-time query against a more or less static collection (i.e. the set of documents to be searched is known prior to retrieval, but the search requests are not). Multilingual text queries are used to retrieve as many relevant images as possible from the St. Andrews image collection. Queries for images based on abstract concepts rather than visual features are predominant in this task. This limits the effectiveness of using visual retrieval methods alone as either these concepts cannot be extracted using visual features and requires extra external semantic knowledge (e.g. the name of the photographer), or images with different visual properties may be relevant to a search request (e.g. different views of Rome).

The St. Andrews collection consists of 28,133 images, all of which have associated textual captions written in British English (the target language). The captions consist of 8 fields including title, photographer, location, date and one or more pre–defined categories (all manually assigned by domain experts). Examples can be found in [11] and St. Andrews University Library[5].

A new set of 25 topics was generated by the authors familiar with the St. Andrews collection. We first decided on general topic areas and then refined them to create representative search requests to test the capabilities of both cross-language and image retrieval systems. General categories were obtained from an analysis of log files from on–line access to the St Andrews' collection, a discussion with staff from St. Andrews University Library - the proprietors of the collection and categories identified by Armitage and Enser [12] for users of picture archives. The type of information that people typically search for in the St. Andrews collection include the following:

- Social history, e.g. old towns and villages, children at play and work.
- Environmental concerns, e.g. landscapes and wild plants.
- History of photography, e.g. particular photographers.
- Architecture, e.g. specific or general places or buildings.
- Golf, e.g. individual golfers or tournaments.
- Events, e.g. historic, war related.
- Transport, e.g. general or specific roads, bridges etc.
- Ships and shipping, e.g. particular vessels or fishermen.

Given these general categories (and others), topics were created by refinement based on attributes such as name of photographer, date and location. A list of topic titles can be found in [13]. These are typical of retrieval requests from picture archives where semantic knowledge is required in addition to the image itself to perform retrieval. Topics consist of title (a short sentence or phrase describing the search request in a few words) and a narrative (a description of what constitutes a relevant or non-relevant image for that search request). We also provided an example relevant image which we envisaged could be used for relevance feedback (both manual and automatic) and query–by–example searches. Topic

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

titles were translated into French, German, Spanish[6], Dutch, Italian, Chinese, Japanese, Finnish, Swedish, Danish, Russian and Arabic by native speakers. An example topic is shown in Figure 2.

```
<top>
<num> Number: 1 </num>

<title> Portrait pictures of church ministers by
Thomas Rodger </title>

<narr> Relevant images are portrait photographs
of ministers or church leaders by the photographer
Thomas Rodger. Images from any era are relevant,
but must show one person only taken within a studio,
i.e. posing for the picture. Pictures of groups are
not relevant. </narr>

</top>
```

Fig. 2. An example ad hoc topic in English.

Participants were asked to classify their runs according to four main query dimensions: query language, manual vs. automatic (automatic runs involve no user interaction; whereby manual runs are those in which a human has been involved in query construction), with or without query expansion[7] (QE), and use of title vs. title and narrative (narratives were translated by participants for French topics). As training data, 5 topics from 2003 were provided together with relevance assessments (197 relevant images). The main challenges of this task include: (1) captions and queries which are typically short in length (limited context), (2) images of varying content and quality (mostly black and white which limits the effectiveness of using colour as a visual feature), (3) captions containing text not directly associated with the visual content of an image (e.g. expressing something in the background). (4) use of colloquial and domain-specific language in the caption, and (5) filtering out images which contain query terms but are not judged relevant (e.g. the image is too dark or the subject of the query is not clearly visible).

Table 2 shows the 190 submitted experiments/runs for the ad hoc task listed by the query/topic language where predominant languages are Spanish and French. All groups were asked to submit an English monolingual run for comparison with cross–language retrieval (although not all groups did). Table 3 shows the proportion of submitted runs based on the query dimension. Almost all runs were automatic (99%) and pleasing to us were the large proportion of text+visual submissions (41%).

[6] UNED found errors in the original Spanish queries and released a revised topic set which was used by participants for the Spanish submission.

[7] Query expansion refers to adding further terms to a text query (e.g. through PRF or thesaurus lookup) or more images to a visual query.

Table 2. Ad hoc experiments listed by query/topic language.

Language	#Participants	#Runs
Spanish	6	41
English (mono)	9	29
French	6	23
German	5	20
Italian	5	20
Dutch	3	20
Chinese	5	18
Japanese	2	4
Russian	2	4
Swedish	2	2
Finnish	2	2
Danish	1	1
Visual only	2	*6*

Table 3. Ad hoc experiments listed by query dimension.

Query Dimension	#Runs
Manual	1 (1%)
Automatic	189 (99%)
With QE	135 (71%)
Visual only	6 (3%)
Text Only	106 (56%)
Text +Visual	78 (41%)
Title + Narrative	5 (3%)

2.3 Medical Retrieval from CasImage

The use of Content–Based Image Retrieval (CBIR) systems is becoming an important factor in medical imaging research making this a suitable domain for a second ImageCLEF task. The goal being to find similar images with respect to the following features: modality (e.g. CT, radiograph or MRI), anatomic region (e.g. lung, liver or head) and radiological protocol (e.g. contrast agent or T1/T2 weighting for MRI) where applicable. Identifying images referring to similar medical conditions is non–trivial and may require the use of visual content and additional semantic information not obtainable from the image itself. However, the first query step has to be visual and it is this which we test in ImageCLEF 2004. Participants were not expected to require a deep clinical knowledge to perform well in this task. Given the query image the simplest submission is to find visually similar images (e.g. based on texture and colour). However, more advanced retrieval methods can be tuned to features such as contrast and modality.

The dataset for the medical retrieval task is called CasImage[8] and consists of 8,725 anonymised medical images, e.g. scans, and X–rays from the University Hospitals of Geneva. The majority of images are associated with *case notes*, a written description of a previous diagnosis for an illness the image identifies. Case notes are written in XML and consist of several fields including: a diagnosis, free-text description, clinical presentation, keywords and title. The task is multilingual because case notes are mixed language written in either English or French (approx. 80%). An example case notes field for description and corresponding images is shown in Figure 3. Not all case notes have entries for each field and the text itself reflects real clinical data in that it contains mixed–case text, spelling errors, erroneous French accents and un–grammatical sentences as well as some entirely empty case notes. In the dataset there are 2,078 cases to be exploited during retrieval (e.g. for query expansion). Around 1,500 of the 8,725 images in the collection are not attached to case notes and 207 case notes are empty. The case notes may be used to refine images which are visually similar to ensure they match modality and anatomic region.

Table 4. Medical experiments listed by query dimension.

Query Dimension	# Runs
Manual	9 (21%)
Automatic	34 (79%)
With RF	13 (30%)
Visual only	29 (67%)
Text +Visual	14 (33%)

[8] See [14] and `http://www.casimage.com/` for more information about the CasImage collection.

```
<?xml version='1.0' encoding='iso-8859-1' ?>
<CASIMAGE_CASE>
<ID>
2526
</ID>
<Description>
Bassin du 28.02.1985 :

Status avant et aprËs rÉduction. Avant rÉduction, luxation
compl`Ete du fÉmur, avec fracture avec fragments du cotyle.
Apr`Es rÉduction, interposition de l'un de ces fragments entre
la tte fÉmorale et le toit du cotyle.

</Description>

<Diagnosis>
Luxation postÉrieure du fÉmur gauche associÉe ? une fracture
multifragmentaire d
</Diagnosis>
........
```

Fig. 3. An example medical case note (in French) and associated images.

For the selection of topics, a radiologist familiar with CasImage was asked to chose a number of topics (images only) that represented the database well. They corresponded to different modalities, different anatomic regions and several radiological protocols such as contrast agents or weightings for the MRI. This resulted in 30–35 images being chosen. One of the authors then used these images for query–by–example searches to find further images in the database resembling the query using feedback and the case notes and selected 26 of these for the final topic set (see [14],[13]). Similar to the ad hoc task, participants were free to use any method for retrieval, but were asked to identify their runs against three main query dimensions: with and without relevance feedback, visual vs. visual+text, and manual vs. automatic. Table 4 shows submissions to the medical task categorised according to these query dimensions.

2.4 User–Centered Search Task

The user–centered search task aims to allow participants to explore variations of their retrieval system within a given scenario, rather than compare systems in a competitive environment. There are at least four aspects of a cross–language image retrieval system to investigate including: (1) how the CLIR system supports user query formulation for images with English captions, particularly for users in their native language which may be non–English; (2) whether the CLIR system supports query re–formulation, e.g. the support of positive and negative feedback to improve the user's search experience; (3) browsing the image collection; and (4) how well the CLIR system presents the retrieval results to the user to enable selection of relevant images. The interactive task is based on the St. Andrews collection with a known–item search.

Given an image from the St Andrews collection, the goal for the searcher is to find the same image again using a cross–language image retrieval system. This aims to allow researchers to study how users describe images and their methods of searching the collection for particular images, e.g. browsing or by conducting specific searches. The scenario models the situation in which a user searches with a specific image in mind (perhaps they have seen it before) but without knowing key information thereby requiring them to describe the image instead, e.g. searches for a familiar painting whose title and painter are unknown. This task can be used to determine whether the retrieval system is being used in the manner intended by the system designers and determine how the interface helps users reformulate and refine their search requests.

Participants compared two interactive cross–language image retrieval systems (one intended as a baseline) that differ in the facilities provided for interactive query refinement. For example, the user is searching for a picture of an arched bridge and starts with the query "bridge". Through query modification (e.g. query expansion based on the captions), or perhaps browsing for similar images and using feedback based on visual features, the user refines the query until relevant images are found. As a cross–language image retrieval task, the initial query is in a language different from the collection (i.e. not English) and translated into English for retrieval. The simplest approach is to translate the query and display only images to the user (assuming relevance can be based on the image only, i.e. that images are language independent), maybe using relevance feedback on visual features only, enabling browsing, or categorising the images in some way and allowing the user to narrow their search through selecting these categories. Any text displayed to the user must be translated into the user's source language. This might include captions, summaries, pre–defined image categories etc.

A minimum of 8 users (who can search with non–English queries) and 16 example images (topics) are required for this task (we supply the topics). The interactive ImageCLEF task is run similar to iCLEF 2003 using the same experimental procedure. However, because of the type of evaluation (i.e. whether known items are found or not), the experimental procedure for iCLEF 2004 (Q&A) is also very relevant and we make use of both iCLEF procedures. Given the 16 topics, participants get the 8 users to test each system with 8 topics. Users are given a maximum of 5 minutes only to find each image. Topics and systems are presented to the user in combinations following a latin–square design to ensure user/topic and system/topic interactions are minimised.

3 Evaluating Submissions

3.1 Methodology

In this section we describe the evaluation methodology for the ad hoc and medical retrieval tasks (which is similar to ImageCLEF 2003 [3]). Submissions were assessed in the following way: (1) the top N runs (for ad-hoc $N = 50$; for medical $N = 60$) were extracted from each submission (190 submissions for ad hoc; 43

for medical), (2) a document pool was created for each topic by computing the union overlap of submissions, (3) obtained three sets of assessments for documents in each topic pool (images judged as relevant, partially relevant and not relevant), (4) computed different sets of relevant images for each topic (called *qrels*), (5) compared each system run against one of the sets of qrels and (6) computed uninterpolated mean average precision[9] (MAP). To ensure maximum pool coverage, we used Interactive Search and Judging [15] for the ad hoc task and added a set of previously identified ground truths to the medical pools.

3.2 Relevance Assessments

Judging whether an image is relevant or not is highly subjective (e.g. due to knowledge of the topics or domain, different interpretations of the same image and searching experience). Therefore to minimise subjectivity we obtained three sets of relevance judgements per topic and task. For the ad hoc task, relevance assessments were performed by students and staff at the University of Sheffield (each assessor given 5 topics to judge); for the medical task three scientists familiar with the CasImage collection from the University Hospitals Geneva (one radiologist, a medical doctor and a medical computer scientist) each judged all 26 topics.

An on-line system built specifically for ImageCLEF was used by assessors to judge the relevance of documents in the topic pools. No time limit was specified for carrying out assessments and judges could alter their assessments before submitting final results. A ternary relevance scheme was used by assessors consisting of relevant, partially relevant and not relevant. The partially relevant judgement was used to pick up images where the judge thought it was in some way relevant, but could not be entirely confident (e.g. the required subject is in the background of the image in the case of ad hoc retrieval).

Given three sets of assessments per topic, we used a "voting" scheme to generate sets of relevant images (qrels) based on the overlap of relevant images between assessors, and whether partially relevant images were included. For each topic the assessments were used to vote for each image in the document pool. For the medical task, all assessors were given an equal vote of 1; in the ad hoc task the topic creator was given a count of 2 and other assessors a vote of 1. We created 6 basic relevance sets based on the voting score obtained for each image:

1. **isec–rel:** images judged as relevant by all three assessors.
2. **isec–total:** images judged as either relevant or partially relevant by all three assessors.
3. **pisec–rel:** images judged as relevant by the topic creator and 1 other assessor (ad hoc) or at least two assessors (medical).
4. **pisec–total:** images judged as either relevant or partially relevant by the topic creator and 1 other assessor (ad hoc) or at least two assessors (medical).
5. **union–rel:** images judged as relevant by at least 1 assessor.

[9] A version of `trec_eval` from U. Massachusetts and `ireval.pl` from the Lemur IR toolkit distribution - `http://www-2.cs.cmu.edu/~lemur/` were used for evaluation.

6. **union–total:** images judged as either relevant or partially relevant by at least 1 assessor.

Any of these qrels sets can be used for evaluation, ranging from the strictest set of judgments (isec–rel) to the most relaxed (union-total). In ImageCLEF 2004 we used *pisec–total* as a compromise between the two extremes.

4 Results and Main Findings

4.1 Bilingual Ad Hoc Retrieval Task

Table 5 shows the top run for each query language (ordered by MAP) and parameters used. The %monolingual score is computed as a proportion of the highest English submission (0.5865). Excluding the English and visual results, 45% of the best runs used CBIR to complement text retrieval, and 64% used some kind of query expansion (either text-based or by adding "relevant" images to a visual query). In Table 5, 73% of runs used MT systems for translation, although statistical models trained on parallel corpora [10] and bilingual dictionaries were also used [16, 17]. Finnish is a particularly difficult language to process and results in the lowest MAP score. This was also observed in results from other CLEF tracks in 2004. Query translation proved to the predominant translation approach, although Clough [18] combined query and document translation and found a combination of both approaches gave highest retrieval effectiveness.

Taking the top 5 runs for each language, the average MAP score for runs with QE is 0.4155. Without QE, average MAP=0.2805 ($t = 3.255$ $p = 0.002$) indicating that some kind of text or visual QE based on PRF is beneficial. For runs using text-based methods only, average MAP=0.3787; for text+visual runs average MAP=0.4508 ($t = -2.007$, $p = 0.052$). On average it appears that combining text and visual features for ad-hoc multilingual retrieval improves effectiveness, although the results are not significant (at $p < 0.05$). However, some groups did observe improvements for individual topics [17, 10] where visual features can distinguish relevant images.

Two groups submitted runs using a purely visual search which performed poorly [5, 4]. We would expect this because for topics for the ad-hoc task, pure visual similarity plays a marginal role; whereas semantics and background knowledge are extremely important. A number of groups used methods to identify named entities such as photographer, date and location to try and improve retrieval by performing structured or constrained searches [8, 19, 16, 10]. Retrieval was performed by using the text or image (the exemplar image supplied by ImageCLEF) as initial query and then combining results. More often than not iterative searches would then include both text and visual retrieval methods. One of the main problems tackled by groups was how best to combine ranked lists from separate text and visual searches. Two groups experimented with using "bi-media" dictionaries where text is mapped to visual representatives showing promising further areas for research [10, 17].

Table 5. Systems with highest MAP for each language in the ad-hoc retrieval task.

Language	Group	Run ID	MAP (%mono)	QE	Text	Visual	Title	Narr
English	daedalus	mirobaseen	0.5865		⋆		⋆	
German	dcu	delsmgimg	0.5327 (90.8)	⋆	⋆	⋆	⋆	
Spanish	UNED	unedesent	0.5171 (88.2)	⋆	⋆		⋆	
French	montreal	UMfrTFBTI	0.5125 (87.4)	⋆	⋆	⋆	⋆	
Italian	dcu	itlsstimg	0.4379 (74.7)	⋆	⋆		⋆	
Dutch	dcu	nllsstimg	0.4321 (73.7)	⋆	⋆		⋆	
Chinese	ntu	NTU-adhoc-CE-T-WE	0.4171 (71.1)		⋆	⋆	⋆	
Russian	daedalus	mirobaseru	0.3866 (65.9)		⋆		⋆	
Swedish	montreal	UMsvTFBTI	0.3400 (58.0)	⋆	⋆	⋆	⋆	
Danish	daedalus	mirobaseda	0.2799 (47.7)		⋆		⋆	
Japanese	daedalus	mirobaseja	0.2358 (40.2)		⋆		⋆	
Finnish	montreal	UMfiTFBTI	0.2347 (40.0)	⋆	⋆	⋆	⋆	
Visual	geneva	GE_andrew4	0.0919 (15.7)	⋆		⋆		

4.2 User–Centered Retrieval Task

For the interactive task, we had 2 submissions: one from I–Shou University (KIDS) and another from Michigan State University (MSU). No formal evaluation was undertaken this year. KIDS [20] tested 2 retrieval systems: a baseline system allowing users to search and refine queries with text only (`T_ICLEF`), and an alternative system enabling users to refine queries using both text and based upon the colour of the target image (`VCT_ICLEF`). Both systems provided text retrieval in Chinese and they found that allowing users to refine queries using a colour palate did improve retrieval effectiveness (89% of searchers found the target image if permitted to select colours compared to 56.25% without; on average a 63% reduction in time spent looking for the target image and 82% reduction in the number of retrieval iterations).

MSU [21] focused on methods of term selection for query expansion. They compared two systems in their user study: a baseline system where users were able to search for images in Chinese, refining and modifying queries using their own terms (`Standard Interface`) and an alternative system where 10 additional terms were suggested automatically to the user allowing them to add to and remove from existing query terms (`URF`). Results showed that the `Standard Interface` performed significantly better than `URF`. The main cause was found to be due to the suggestion of terms by the system which were unfamiliar with the user and hence not useful, or suggested terms not useful in identifying the target image. The results for MSU highlight some of the issues involved in interactive cross-language image retrieval when the collection is specialised like the St. Andrews collection of photographs and unfamiliar to multilingual users.

4.3 Medical Retrieval Task

Table 7 shows the results for the medical task using manual runs only (the rank position is the rank position within all runs ordered by descending MAP score). The highest MAP score is obtained for systems using both visual and text features. Based on all submissions (manual and automatic) average MAP=0.2882.

For visual only submissions, average MAP=0.2863; visual+text submissions average MAP=0.2922, although these differences are not statistically significant ($t = 0.140$, $p = 0.084$). The *kids_run3* run is low MAP due to a misconfiguration in their submission. Table 7 shows the top 10 results for medical task using automatic runs only.

State University of New York [22] achieved the highest result using both text and visual features; although University of Aachen [5] and Imperial [6] came close using visual features only (difference is not statistically significant). On average, we find that for runs using relevance feedback, average MAP=0.2675; without relevance feedback, average MAP=0.2972 ($t = 0.805$, $p = 0.337$). It would appear that some kind of relevance feedback helps (but the average difference is not statistically significant). Still, for single systems and techniques such as manual relevance feedback, automatic query expansions and mix of textual and visual features delivered significant improvements in retrieval quality. Best overall results were obtained combining visual and textual features in manual relevance feedback queries [23].

Table 6. All results for the medical manual experiment.

Group	Run ID	MAP	Rank	With RF	Visual	Text
geneva	GE_rfvistex20	0.4764	1		⋆	⋆
geneva	GE_rfvistex10	0.4757	2		⋆	⋆
geneva	GE_rfvistex1	0.4330	3		⋆	⋆
geneva	GE_4d_4g_rf	0.4303	4		⋆	
aachen–inf	i6-rfb1	0.3938	5	⋆	⋆	
KIDS	kids_run2	0.3799	8		⋆	
geneva	GE_8d_16g_rf	0.3718	12		⋆	
geneva	GE_4d_16g_rf	0.3584	14		⋆	
KIDS	kids_run3	0.0843	43		⋆	

When analysing the manual submissions, we find that the three best runs combine both visual and textual features, whereas the third and fourth use only visual searching. Low level visual features such as Gabor filters and simple grey level distributions seem to perform best. It would appear that combined systems result in better performance when including text than without, but the contribution of text retrieval should be weighted fairly low (10

When comparing several features [6], individually, the Gabor filters perform best, which are used in four out of the five best automatic systems. Still, a mixture of several features performs better as the performance of features for the various topics varies strongly. Having a topic-dependent feature selection could help improve results. Two of the top five automatic systems are based on the same visual methods but different text search strategies. This implies that even with the same visual starting point, significant differences are possible depending upon the text-retrieval strategy chosen.

Table 7. Top 10 results for the medical automatic experiment.

Group	Run ID	MAP	Rank	With RF	Visual	Text
Buffalo	UBMedImTxt01	0.3904	6		⋆	⋆
aachen-inf	i6-025501	0.3858	7		⋆	
imperial	ic_cl04_base	0.3784	9		⋆	
aachen-inf	i6-qe0255010	0.3741	10	⋆	⋆	
Buffalo	UBMedImTxt03	0.3722	11		⋆	⋆
Buffalo	UBMedImTxt02	0.3696	13		⋆	⋆
aachen-inf	i6qe02100010	0.3535	15	⋆	⋆	
geneva	GE_4g_4d_qe1	0.3500	16		⋆	
geneva	GE_4d_4g_vis	0.3499	17	⋆	⋆	
KIDS	kids_run1	0.3273	18		⋆	

5 Conclusions

5.1 Comments from participants

For ImageCLEF 2005 we will take into account comments received from participants at the 2004 workshop. In general, ImageCLEF was seen as a valuable effort: it is currently the only image retrieval evaluation event and the accessibility of datasets for image retrieval evaluation including ground truths was regarded as very important.

A negative comment was the lack of training data. This can be remedied in 2005 by the provision of topics and ground truths used in 2004. Another comment was with respect to the time from the release of the topics to the time that the results had to be sent in. Several groups remarked that a shorter time frame would be better to not allow research groups to optimize their system too much for perfect results. Participants also commented on the topics and data used in the ad hoc task. The St. Andrews collection, although realistic, proved very hard to use for CBIR and topics did not involve enough use of visual features.

5.2 ImageCLEF 2005

The bilingual ad hoc task will use more general topics to provide more suitable searches for CBIR systems. We will also provide more exemplar images to enable more effective use of CBIR systems (one image is not enough for effective retrieval). The task, however, will remain predominantly text-based involving multilingual topics (where the entire topic statement is translated).

The medical image retrieval task will be performed with a larger set of images and a new set of queries. The goal will be to obtain at least one or two radiology teaching files that can be added to the current casImage database. The retrieval task will again be single images, although tests will be run using using several images as a query for case-based retrieval or by adding short multi-lingual texts to an image that describes visual content.

A new automatic annotation task is planned for ImageCLEF 2005. This task will be similar to the medical image retrieval task based on a visual analysis of the images. It will be undertaken with help from the IRMA group[10] (Image Retrieval in Medical Applications) of the Technical University of Aachen. It will use a database of 10,000 images that are classified according to a four-code axis - the IRMA code. This code allows image annotation in several languages. Half of the database will be given out as training data, and then the other half given to participants for classification based on visual features in the images only. We hope this task will attract interest from the machine learning community.

5.3 Summary

In this paper we have described the ImageCLEF 2004 campaign for evaluating cross–language image retrieval. We were successful at attracting a range of groups from a variety of research backgrounds for two retrieval tasks in different domains. The ImageCLEF task was very successful this year and by encouraging the use of a CBIR system, we are able to compare systems based on a large–scale evaluation.

Participants applied a variety of methods to bridge the language and media barriers and the fact that many of the best performing systems all used a combination of visual and textual methods shows that there is a potential for improving retrieval effectiveness over any single method. Some tasks, such as the ad hoc retrieval task, are better suited to text-based image retrieval (assuming that metadata is associated with the images to be retrieved), but other tasks, such as the medical retrieval task, are naturally better suited to visual retrieval (although requiring extra information provided by associated texts to enable more advanced retrieval). Although several systems in ImageCLEF used visual and textual features together, we assume that there is still much potential for further research. Better results for one can help the other through automatic query expansion, for example. If the best visual and textual techniques are combined, we can expect optimal results.

The high participation at ImageCLEF 2004 has shown that there is a need for such an evaluation event, especially given the multilingual and multimedia environment in which current retrieval systems must operate. To create more dynamic research in the field of multi–modal visual/textual retrieval we need to attract visual and multilingual information retrieval groups for the future and promote combined submissions of different research groups.

The rather visual medical task and the rather textual ad hoc task should be complemented with tasks that are somewhere in between. This could be realised by using collections that are closer to existing CBIR evaluation collections containing colour images with a limited number of objects and themes, having more search requests which include an element of both textual and visual search, having more exemplar images and maybe also negative examples. For the medical collection we can well imagine having a short description of the image written

[10] `http://www.irma-project.org/`

by a medical doctor that can be used in addition to the image. Simple semantic retrieval tasks may also help attract further visual retrieval research groups. These could be based on the visual content of images, such as finding all images than contain sunsets or at least three faces. Another community to attract for the medical task would be the image analysis and classification community. This could be achieved through a simple classification task.

6 Acknowledgements

We thank everyone who participated in ImageCLEF 2004 to make this such an interesting and successful evaluation. In particular we thank St. Andrews University Library (esp. Norman Reid) for use of the St. Andrews collection and University Hospitals Geneva for use of CasImage images. What makes this evaluation possible are the relevance assessments and we want to thank Hideo Joho, Simon Tucker, Steve Whittaker, Wim Peters, Diego Uribe, Horacio Saggion, Paul Fabry, and Tristan Zand. Our thanks also go out to those people involved in translating the captions including Jian–Yun Nie, Jesper Kallehauge, Assad Alberair, Hiedi Christensen, Xiao Mang Shou, Michael Bonn, Maarten de Rijke, Diego Uribe, Jussi Karlgren, Carol Peters, Eija Airio, Natalia Loukachevitch and Hideo Joho.

References

1. Grefenstette, G.: Cross Language Information Retrieval. Kluwer Academic Publishers, Norwell, MA, USA (1998)
2. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–Based Image Retrieval at the end of the early years. T-PAMI **22 No 12** (2000) 1349–1380
3. Clough, P., Sanderson, M.: The CLEF Cross Language image retrieval track. In: Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway. http://clef.isti.cnr.it/2003/WN_web/45.pdf. (2003)
4. Müller, H., Geissbuhler, A., Ruch, P.: Report on the ImageCLEF Experiment: How to Visually Retrieve Images from the St. Andrews Collection using GIFT. In: these proceedings. (2005)
5. Deselaers, T., Keysers, D., Ney, H.: FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In: these proceedings. (2005)
6. Howarth, P., Yavlinsky, A., Heesch, D., Rüger, S.: Visual Features for Content-based Medical Image Retrieval. In: these proceedings. (2005)
7. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection. In: these proceedings. (2005)
8. Martinez-Fernandez, J.L., Garcia Serrano, A., Villena, J., Mendez Saenz, V.: MIRACLE approach to ImageCLEF 2004: merging textual and content-based image retrieval. In: these proceedings. (2005)
9. van Zaanen, M., de Croon, G.: FINT: Find Images aNd Text. In: these proceedings. (2005)
10. Alvarez, C., Oumohmed, A.I., Mignotte, M., Nie, J.Y.: Toward Cross-Language and Cross-Media Image Retrieval. In: these proceedings. (2005)

11. Clough, P., Sanderson, M., Müller, H.: A proposal for the CLEF Cross-Language Image Retrieval Track 2004. In: Poster at the Third International Conference for Image and Video Retrieval (CVIR 2004). (2004) 243–251
12. Armitage, L., Enser, P.: Analysis of User Need in Image Archives. Journal of Information Science (1997) 287–299
13. Clough, P., Müller, H., Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. http://clef.isti.cnr.it/2004/working_notes/CLEF2004WN-Contents.html. (2004)
14. Müller, H., Rosset, A., Geissbuhler, A., Terrier, F.: A reference data set for the evaluation of medical image retrieval systems. CMIG (2004 (to appear))
15. Chen, K., Chen, H., Kando, N., Kuriyama, K., Lee, S., Myaeng, S.: Overview of clir task. In: Third NTCIR Workshop, Japan. (2002)
16. Peinado, V., Artiles, J., Lopez-Ostenero, F., Gonzalo, J., Verdejo, F.: UNED@ImageCLEF 2004: Using Image Captions Structure and Noun Phrase Based Query Expansion for Cross-Language Image Caption Retrieval. In: these proceedings. (2005)
17. Lin, W.C., Chang, Y.C., Chen, H.H.: From Text to Image: Generating Visual Query for Image Retrieval. In: these proceedings. (2005)
18. Clough, P.: Caption and Query Translation for Cross-Language Image Retrieval. In: these proceedings. (2005)
19. Saiz-Noeda, M., Vicedo, J.L., Izquierdo, R.: Pattern-based Image Retrieval with Constraints and Preferences on ImageCLEF 2004. In: these proceedings. (2005)
20. Cheng, P.C., Yeh, J.Y., Chien, B.C., Ke, H.R., Yang, W.P.: NCTU-ISU's Evaluation for the User-Centered Search Task at ImageCLEF 2004. In: these proceeedings. (2005)
21. Bansal, V., Zhang, C., Joyce, C.Y., Jin, R.: MSU at ImageCLEF: Cross Language and Interactive Image Retrieval. In: these proceeedings. (2005)
22. Ruiz, M.E., Srikanth, M.: UB at CLEF2004: Part 2 Cross Language Medical Image Retrieval. In: these proceedings. (2005)
23. Müller, H., Geissbuhler, A., Ruch, P.: Report on the CLEF Experiment: Combining Image and Multilingual Search for Medical Image Retrieval. In: these proceedings. (2005)