The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004

Paul Clough†, Mark Sanderson†and Henning Müller‡
†Department of Information Studies, University of Sheffield, Sheffield, UK.
{*p.d.clough, m.sanderson*}*@sheffield.ac.uk*
‡University and University Hospitals of Geneva, Service of Medical Informatics,
rue Micheli–du–Crest 24,1211 Geneva 14, Switzerland.
henning.mueller@sim.hcuge.ch

Abstract

The purpose of this paper is to outline research efforts for the 2004 CLEF cross–language image retrieval campaign (ImageCLEF). The aim of this CLEF track is to explore the use of 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 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 to 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 to be able to offer the same collection to a wider and more diverse range of users with different language backgrounds. It is this area of CLIR which we address in ImageCLEF 2004[2] , the CLEF cross–language image retrieval campaign.

In 2003, we organised a pilot experiment at CLEF with the 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 a language different from the collection and requiries a visual document to fulfil their search request (e.g. searching an on–line art gallery or stock–photographic collection). Four groups from

[1] See `http://www.corbis.com/`
[2] See `http://ir.shef.ac.uk/imageclef2004/` for further information about ImageCLEF campaign.

Table 1: Participating Groups in ImageCLEF.

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

industry and academia participated in ImageCLEF 2003 using purely text–based image retrieval 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 (see [4]). 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[3], and (2) created a medical retrieval task where initial retrieval is visual. 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 are provided by us: (1) the St. Andrews collection of historic photographic images, and (2) the CasImage radiological medical database. Further to the image collections, we also provided the search topics and performed relevance assessments based on submitted entries. Based on the St. Andrews collection, we offered two tasks: (1) a bilingual ad hoc retrieval task: given an initial topic find as many relevant images as possible, and (2) an known–item interactive task: given an image from the St. Andrews collection, find it again. For the CasImage collection we offered a query–by–example search task: given an initial medical image find as many relevant images as possible.

2.1 Participating groups

In total 18 groups participated in ImageCLEF 2004 (Table 1): 11 in the medical task, 12 in the ad hoc task and 2 in the interactive task. Six groups participated in more than 1 task. ImageCLEF attracted participants from a variety of backgrounds including textual, medical and visual. We received entries from research groups located globally in 3 continents. For the medical task we received 43 submissions (runs), 190 for the ad hoc task and 2 for the interactive task.

[3]We offered access to the VIPER system `http://viper.unige.ch/` through PHP, a list of the top N images for each topic retrieved using exemplar images and via download of GIFT `http://www.gnu.org/software/gift/`.

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.

Figure 1: An example image and caption from the ImageCLEF collection.

2.2 Bilingual Retrieval from St. Andrews

Similar to the TREC ad hoc retrieval task, this task simulates the situation in which a system knows the set of documents to be searched but topics are not known to the system in advance. The goal of the ad hoc task is to retrieve as many relevant images as possible from the St. Andrews image collection (described further in [3]) given multilingual topics. The collection consists of 28,133 images, all of which have associated textual captions (see, e.g. Figure 1). The captions consist of 8 fields written in British English by librarians working at St. Andrews Library. Caption fields include a title, the photographer, location, date and one or more pre–defined categories.

For ImageCLEF 2004, a new set of 25 topics was generated by the authors. We created these by first deciding on general topic areas in the St. Andrews collection, and then refining them to create a representative topic set to test the capabilities of both a translation and image retrieval system. Broad categories were obtained from log file analysis, a discussion with librarians at St. Andrews University Library and the categorisation scheme suggested by Armitage and Enser [5] for picture archives. Topics were refined by attributes such as photographer, date and location (see Table 2). Topics consist of a short title (3–4 words), a longer narrative describing in further detail the user need, and an exemplar image (see Figure 5). Topic titles were translated into French, German, Spanish[4], Dutch, Italian, Chinese, Japanese, Finnish, Swedish, Danish, Russian and Arabic by native speakers.

Given the topics, participants were free to construct queries in any manner they desired (e.g. using iterative relevance feedback) and for any language. We encouraged participants to experiment with: different methods of translation (e.g. dictionary–based vs. MT), query expansion (e.g. based on thesaurus lookup or relevance feedback), indexing and retrieval on only part of the image caption, different models of retrieval, and combining text and content–based retrieval methods. 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 and without query expansion (QE), and use of title vs. title and narrative[5]. Table 3 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. Table 4 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.

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

[5]Some groups tried using the narrative for English monolingual experiments.

Table 2: Ad hoc topic titles.

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

Table 3: 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*
Total	-	190

Table 4: Ad hoc experiments listed by query dimension.

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

Table 5: Medical experiments listed by query dimension.

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

2.3 Medical Retrieval from CasImage

The use of content–based image retrieval (CBIR) systems is becoming an important factor in medical imaging research. The main goal of the medical task is to compare CBIR systems and in particular determine how associated cross–language text can be used in combination with CBIR to improve retrieval and ranking in this domain. Participants were not expected to require a deep clinical knowledge to perform well in this task. The goal of the medical task is to find images that are similar with respect to modality (e.g. CT, radiograph, MRI, ...), the shown anatomic region (e.g. lung, liver, head, ...) and sometimes with respect to the radiologic protocol (such as a contrast agent, T1/T2 for MRI), when 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 the ImageCLEF medical 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 may be tuned to features such as contrast and modality. The dataset for the medical retrieval task is called CasImage[6] and consists of 8,725 anonymised medical images, e.g. scans, and X–rays from the University Hospitals of Geneva (see Figure 3 for example images). The majority of images are associated with case notes, a written description of a previous diagnosis for an illness the image identifies (see, e.g. Figure 4). Case notes consist of several fields including: a diagnosis, a free textual description, clinical presentation, keywords and title. The task is multilingual because case notes are mixed language written in either English or French (approx. 20%). 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. query expansion). Around 1500 of the 8,725 images in the collection are not attched 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.

For the selection of the query tasks, a radiologist familiar with the database was asked to chose a number of topics (images only) that represent the database well. They corresponded to different

[6]See [6] and `http://www.casimage.com/` for more information about the CasImage collection.

modalities, different anatomic regions and several radiologic 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 Figure 3). 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 5 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 (not including the caption) 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 should be 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 and 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 a similar 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.

For this task, we had 2 submissions from I–Shou University (KIDS) and Michigan State University (MSU) and at the time of writing we are still processing the results. KIDS tested 2 retrieval systems: a baseline system using only text features and querying in Chinese, and an alternative

Table 6: Pool and qrels sizes for the ad hoc ($N = 50$) and medical ($N = 60$) tasks.

	Ad hoc retrieval task			**Medical retrieval task**		
Topic #	Pool Size	% Max Size	# Relevant	Pool Size	% Max Size	# Relevant
1	1035	10.7	30	641	28.9	235
2	1389	14.4	21	778	35.0	320
3	1178	12.2	18	994	44.8	72
4	964	10.0	110	877	39.5	43
5	1497	15.5	28	680	30.6	84
6	1190	12.3	14	537	24.2	252
7	1290	13.4	31	348	15.7	48
8	913	9.5	36	944	42.5	117
9	1441	14.9	29	671	30.2	43
10	1544	16.0	34	595	26.8	79
11	1754	18.2	17	861	38.8	9
12	833	8.6	12	617	27.8	179
13	736	7.6	61	789	35.5	95
14	592	6.1	10	654	29.5	11
15	1333	13.8	14	602	27.1	252
16	1071	11.1	23	669	30.1	141
17	1112	11.5	31	749	33.7	31
18	1940	20.1	14	643	29.0	78
19	1740	18.0	31	348	17.3	114
20	1569	16.3	45	761	34.3	27
21	1621	16.8	13	676	30.5	90
22	1855	19.2	8	767	34.5	171
23	1576	16.3	46	909	40.9	74
24	1756	18.2	109	753	33.9	409
25	1101	11.4	44	224	10.1	64
26				468	21.1	53

system which uses both text and visual features to retrieve images. Both systems also allow the user to choose relevant images for relevance feedback. Currently we have no further information on the MSU submission.

3 Evaluating Submissions

3.1 Methodology

In this section we describe the evaluation methodology for the ad hoc and medical retrieval tasks (similar to ImageCLEF 2003 [3]). Submissions were assessed in the following way: (1) the top N (for ad hoc $N = 50$; for medical $N = 60$) runs were extracted from each submission (190 submissions used for ad hoc; 43 for medical), (2) a document pool was created for each topic by computing the union overlap of submissions (Interactive Search and Judge also used to supplement the pools using the Eurovision image retrieval system [7]), (3) three sets of assessments created for each topic pool manually (images judged as relevant and partially relevant), (4) sets of relevant images created for each topic (qrels sets), (5) each system run compared against the qrels, and (6) uninterpolated mean average precision (MAP) computed across all topics using trec_eval. For evaluation of the medical task, we also added a set of previously identified ground truths to the pools to ensure maximum coverage.

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 subjective assessments, we obtained three sets of relevance judgements per topic and task. Relevance assessment is primarily based on the image, but for certain topics (especially the ad hoc task) the caption is also required to make a decision (e.g. "pictures of North Street St Andrews"). What constitutes a relevant image is a subjective decision, but typically a relevant image will have the subject of the topic in the foreground, the image will not be too dark in contrast, and maybe the caption confirms the judge's decision. 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).

Several assessors judged the image pools generated from pooling the submissions. To reduce the subjectivity of relevance assessments, we created several sets of qrels (6 for the ad hoc task; 9 for the medical task) based on the overlap of relevant images between assessors, and whether partially relevant images were included in the qrels set. For the ad hoc task, relevance assessments were performed by students and staff at the University of Sheffield; for the medical task we used three scientists from the University Hospitals Geneva (one radiologist, a medical doctor and a medical computer scientist). To compute the overlap between judgements, we used a voting scheme to rate each image. 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.

For the ad hoc task we created 6 relevance sets and used the *pisec–total* set for evaluation:

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
4. **pisec–total:** images judged as either relevant or partially relevant by the topic creator and 1 other assessor
5. **union–rel:** images judged as relevant by at least 1 assessor
6. **union–total:** images judged as either relevant or partially relevant by at least 1 assessor

For the medical task we created 9 relevance sets and used the *pisec–total* set for evaluation:

1. **isec–rel:** images judged as relevant by all three assessors
2. **isec–partial:** images judged as partially relevant by all three assessors
3. **isec–total:** images judged as either relevant or partially relevant by all three assessors
4. **pisec–rel:** images judged as relevant by at least 2 assessors
5. **pisec–partial:** images judged as partially relevant by at least 2 assessors
6. **pisec–total:** images judged as either relevant or partially relevant by at least 2 assessors
7. **union–rel:** images judged as relevant by all at least 1 assessor
8. **union–partial:** images judged as partially relevant by at least 1 assessor
9. **union–total:** images judged as either relevant or partially relevant by at least 1 assessor

Table 6 shows the size of pools which assessors for the ad hoc and medical retrieval tasks. The maximum pool size is computed as nm where n is the number of images in the pool and m is the size of top N documents used to create the pool. The number of relevant show the proportion of images judged relevant based on the *pisec-total* relevance set and included in the final qrels.

4 Results and Main Findings

4.1 Bilingual Ad Hoc Retrieval Task

Table 7 shows the top 5 runs for each query language (ordered by MAP). The %monolingual is computed as a proportion of the highest monolingual submission (0.5865) and parameters for each

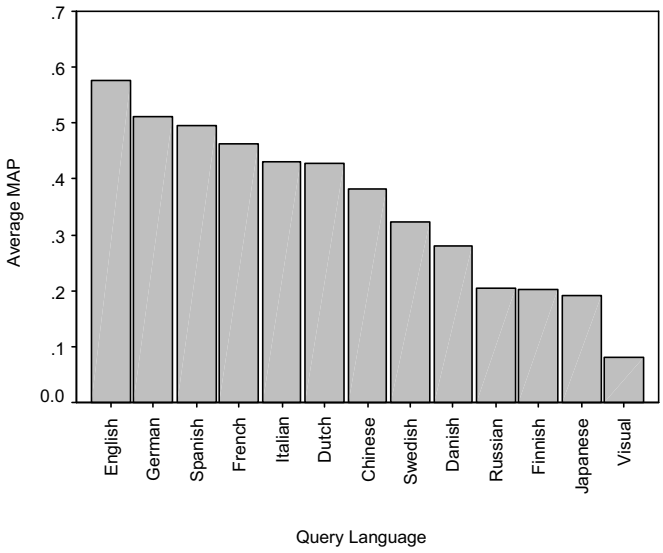


Figure 2: Average MAP for the top 5 systems shown in Table for each language.

run are also shown. Figure 2 shows the average MAP score for each language (scores derived from Table 7). On average, German retrieval performs highest based on MAP, closely followed by Spanish and French. On average, for runs with query expansion, average $MAP = 0.4155$. Without query expansion average $MAP = 0.2805$ ($t = 3.255$ $p = 0.002$) indicating that query expansion is beneficial to the ad hoc retrieval task.

For runs using text 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 $p < 0.05$). Two groups submitted runs using a purely visual search which performed poorly. 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. Further analysis of results will take place after the workshop.

4.2 Medical Retrieval Task

Table 9 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 $MAP = 0.2586$ and with text+visual average $MAP = 0.2652$, although these differences are not statistically significant ($t = -0.195$, $p = 0.847$). The *kids_run3* run is low MAP due to a misconfiguration in their submission. Table 9 shows the top 10 results for medical task using automatic runs only.

State University of New York achieved the highest result using both text and visual features; although Imperial came close using visual features only (difference is not statistically significant). On average, we find that for runs using relevance feedback, average $MAP = 0.2444$; without relevance feedback, average $MAP = 0.2678$ ($t = 0.859$, $p = 0.397$). 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 expansion and mix of textual and visual features delivered significant improvements in retrieval quality. This will need further discussion after the workshop when all systems and techniques can be compared.

Table 7: Top 5 results for each language for the ad hoc retrieval task.

Group	Run ID	MAP	%Mono	Rank	QE	Text	Visual	Title	Narr
English									
daedalus	mirobaseen	0.5865	-	1		X		X	
daedalus	enenrunexp1	0.5838	-	2	X	X	X	X	
sheffield	en_en_fb	0.5829	-	3	X	X		X	
daedalus	mirosbaseen	0.5623	-	4	X	X		X	
montreal	UMenTNFBTI	0.5620	-	5	X	X	X	X	X
Chinese									
ntu	NTU-adhoc-CE-T-WE	0.4171	71.12	53		X	X	X	
ntu	NTU-adhoc-CE-T-WEI	0.4124	70.32	54	X	X	X	X	
ntu	NTU-adhoc-CE-T-W	0.3977	67.81	68		X		X	
ntu	NTU-adhoc-CE-T-WI	0.3969	67.67	70	X	X	X	X	
msu	msustat2	0.2935	50.04	102		X	X	X	
Danish									
daedalus	mirobaseda	0.2799	47.72	107		X		X	
Dutch									
dcu	nllsstimg	0.4321	73.67	39	X	X		X	
dcu	nlstimgfbk3	0.4319	73.64	40	X	X		X	
dcu	nlstimgal	0.4273	72.86	46		X		X	
dcu	nlmgimgal	0.4219	71.94	48	X	X	X	X	
dcu	nlmgimgfbk3	0.4207	71.73	50	X	X		X	
Finnish									
montreal	UMfiTFBTI	0.2347	40.02	125	X	X	X	X	
daedalus	mirobasefi	0.1700	28.99	141		X		X	
French									
montreal	UMfrTFBTI	0.5125	87.40	15	X	X	X	X	
dcu	frintimgfbk1	0.4662	79.50	27	X	X		X	
dcu	frlsintimg	0.4656	79.40	28	X	X	X	X	
sheffield	fr_fr_fb	0.4365	74.44	36	X	X		X	
dcu	frstimgfbk1	0.4310	73.50	41	X	X		X	
German									
dcu	delsmgimg	0.5327	90.84	10	X	X	X	X	
dcu	demgimgfbk3	0.5318	90.69	11	X	X		X	
dcu	demgimgal	0.5312	90.59	12	X	X	X	X	
dcu	delssdlimg	0.5017	85.56	16	X	X	X	X	
dcu	desdlimgfbk3	0.5005	85.35	17	X	X		X	
Italian									
dcu	itlsstimg	0.4379	74.68	35	X	X		X	
sheffield	it_it_fb	0.4355	74.27	37	X	X		X	
dcu	itstimgal	0.4341	74.03	38		X		X	
dcu	itbasest	0.402	68.55	61		X		X	
dcu	itlssdlimg	0.3708	63.23	78	X	X	X	X	
Japanese									
daedalus	mirobaseja	0.2358	40.21	124		X		X	
alicante	ALCim04jp0	0.2256	38.47	126		X		X	
alicante	ALCim04jp1	0.1555	26.52	144		X		X	
alicante	ALCim04jp2	0.1427	24.33	151	X	X		X	
Russian									
daedalus	mirobaseru	0.3866	65.93	73		X		X	
alicante	ALCim04ru0	0.1472	25.10	147		X		X	
alicante	ALCim04ru2	0.1441	24.57	149	X	X		X	
alicante	ALCim04ru1	0.1360	23.19	155		X		X	
Spanish									
sheffield	es_es_fb	0.5211	88.86	13	X	X		X	
uned	UNEDESENT	0.5171	88.18	14	X	X		X	
montreal	UMesTFBTI	0.4890	83.39	18	X	X	X	X	
uned	UNEDES	0.4827	82.32	19		X		X	
dcu	reessdlimg	0.4732	80.70	22	X	X	X	X	
Swedish									
montreal	UMsvTFBTI	0.3400	57.98	85	X	X	X	X	
daedalus	mirobasesw	0.3043	51.89	99		X		X	
Visual									
geneva	GE_andrew4	0.0919	15.67	179	X		X		
aachen–inf	i6-010101	0.0859	14.65	180			X		
aachen–inf	i6-111111	0.0859	14.65	181			X		
aachen–inf	i6-rfb1	0.0839	14.31	182	X		X		
aachen–inf	i6-010012	0.0773	13.18	185			X		

Table 8: All results for the medical manual experiment.

Group	Run ID	MAP	Rank	With RF	Visual	Text
geneva	GE_rfvistex20	0.4214	1		X	X
geneva	GE_rfvistex10	0.4189	2		X	X
geneva	GE_rfvistex1	0.3824	3		X	X
geneva	GE_4d_4g_rf	0.3791	4		X	
KIDS	kids_run2	0.3457	6		X	
aachen–inf	i6-rfb1	0.3437	8	X	X	
geneva	GE_8d_16g_rf	0.3380	10		X	
geneva	GE_4d_16g_rf	0.3259	14		X	
KIDS	kids_run3	0.0784	43		X	

Table 9: Top 10 results for the medical automatic experiment.

Group	Run ID	MAP	Rank	With RF	Visual	Text
Buffalo	UBMedImTxt01	0.3488	5		X	X
imperial	ic_cl04_base	0.3450	7		X	
aachen–inf	i6-025501	0.3407	9		X	
aachen–inf	i6-qe0255010	0.3323	11	X	X	
Buffalo	UBMedImTxt02	0.3309	12		X	X
Buffalo	UBMedImTxt03	0.3291	13		X	X
geneva	GE_4g_4d_vis	0.3157	15		X	
aachen–inf	i6qe02100010	0.3115	16	X	X	
geneva	GE_4d_4g_qe1	0.3100	17	X	X	
KIDS	kids_run1	0.2960	18		X	

5 Conclusions

In this paper we have described the ImageCLEF 2004 campaign for evaluating cross–language image retrieval. In general, for two retrieval tasks across different domains we have found that a combination of visual and textual features provides retrieval effectiveness which is higher than retrieval based on text or visual features alone. We have also shown that query expansion and relevance feedback improves retrieval performance. 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.

Although in the medical retrieval task there were no significant improvements shown for neither manual relevance feedback nor automatic query expansion, nor of adding textual to the visual features, single system had significant improvement for these techniques. This implies that more work is needed and that several of the fairly low results are due to inexperience with the task and the unavailability of test data to try out the performance in advance. The fact that the three best systems all use visual/textual combinations show potential and that more research is needed on how to combine the two.

The high participation at imageCLEF 2004 has shown that there is a need for such an evaluation event and that visual analysis of images can support multi–lingual retrieval. Although several systems use visual and textual features together, we assume that there much potential for combining the two. 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. To create more varied research in the field of multi–modal visual/textual retrieval we need to attract visual and multi–lingual information retrieval groups for the future and promote combined submissions of different research groups.

We believe that 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 through a collection that is closer to a typical visual retrieval collection (e.g. the Corel database) containing colour images with a limited number of objects and themes, query topics that include text and several images, and maybe also negative examples. For the medical collection we can well imagine having a short description of the user need expressed by a medical doctor that can be used in addition to the image. Simpler semantic retrieval tasks could also be imagined to attract further visual retrieval research groups. These tasks could be based on the visual content of the 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 who may be attracted by a simple classification task.

All these ideas will require some help for the organisation of the tasks, relevance assessments as well as the possibility to use image collections that are large enough and usable in these contexts.

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, Tristan Zand and Antoine Rosset. 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. IEEE Transactions on Pattern Analysis and Machine Intelligence **22 No 12** (2000) 1349–1380

[3] Clough, P., Sanderson, M.: The CLEF cross language image retrieval track. In: Working Notes of the CLEF 2003 Workshop. (2003)

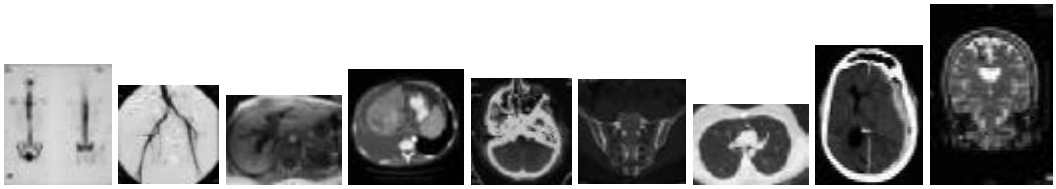
[4] Clough, P., Sanderson, M., Müller, H.: A proposal for the clef cross-language image retrieval track 2004. In: Poster at the Third International Conference for Image and Video Retrieval (CVIR 2004). (2004) 243–251

[5] Armitage, L., Enser, P.: Analysis of user need in image archives. Journal of Information Science (1997) 287–299

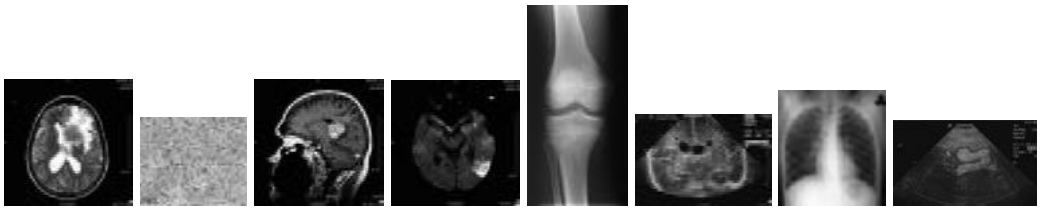
[6] Müller, H., Rosset, A., Geissbuhler, A., Terrier, F.: A reference data set for the evaluation of medical image retrieval systems. CMIG (2004 (to appear))

[7] Clough, P., Sanderson, M.: User experiments with the eurovision cross-langauge image retrieval system. In: JASIST, in submission. (2004)

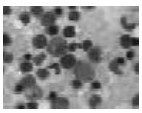
(a) Hip X–Ray (b) Thorax CT (c) MRI legs (d) MRI Lombaire (e) Arteriography (f) Mediastin CT (g) Photo (h) X–Ray



(i) Szintigraphie (j) Arteriography (k) MRI abdominal (l) Abdominal CT (m) Head CT (n) Bone CT (o) Thorax CT (p) Head CT (q) Head MRI T2 Cor.



(r) Head MRI T2 axial (s) Cell cut (t) Head MRI sagittal (u) Head MRI Diffusion (v) X–Ray (w) Head Echography (x) Thorax X–Ray (y) Echography



(z) Cell cut

Figure 3: The 26 topics for the medical retrieval task.

An example case and images (selected)

<Description>

X ray: Mass effect within the soft tissues of the proximal part of the left calf, difficult to outline, seen only as it displaces the fat planes. There are no calcifications. The adjacent bone is normal.

MRI: Oval mass within the medial **gastrocnemius** muscle, very well delineated, slightly **lobulated** (axial cuts). Its structure is heterogeneous. On T1, it is slightly **hyperintense** compared to the adjacent muscle (red arrow). It is **isointense** to fat on proton density images (DP), very **hyperintense** on T2 and IR with some **hypointense** areas in its centre. After injection of contrast medium, there is marked enhancement except for the central area, which remains **hypointense.**

Arteriography: there is **hypervascularity** of the soft tissues outside the medial tibial plateau by vessels **arrising** mostly from the genicular arteries.

.....

</Description>

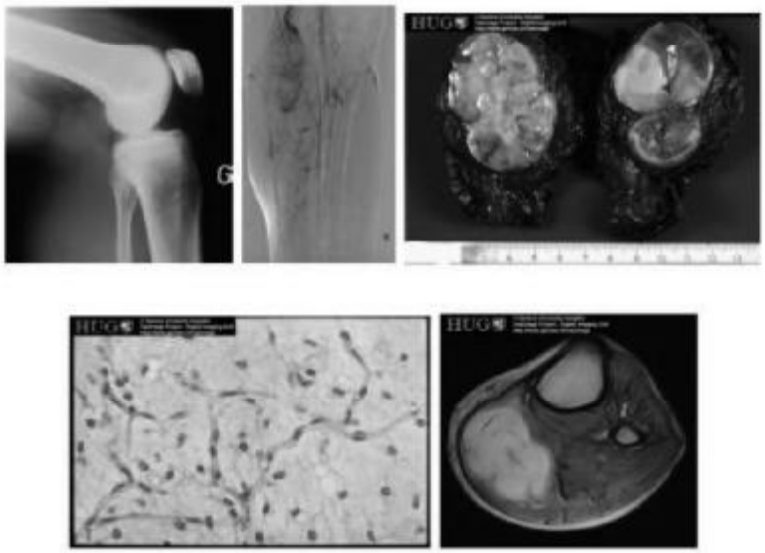


Figure 4: An example case note from the medical task (all images link to this diagnosis).

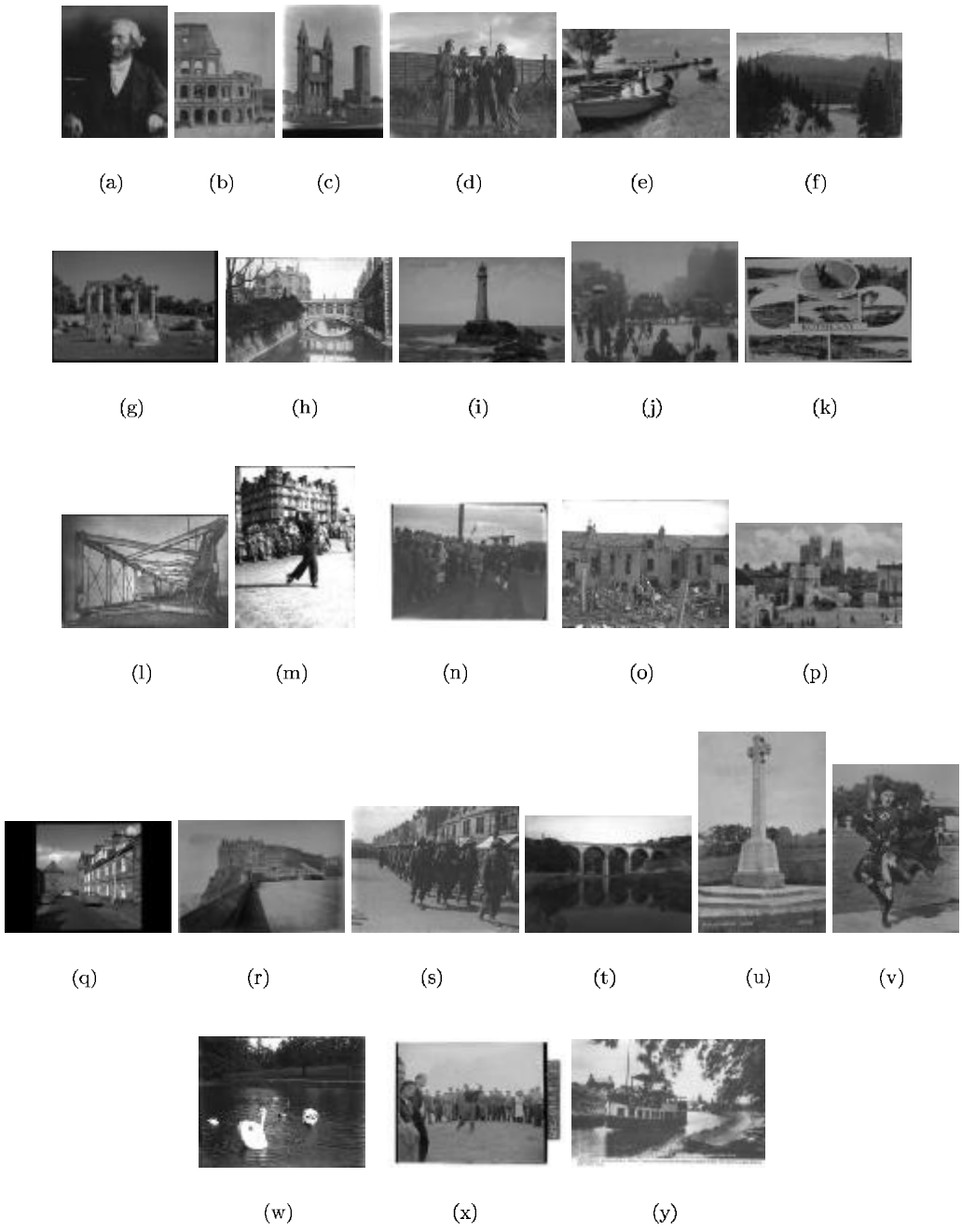


Figure 5: Exemplar images given to participants for the ad hoc retrieval task.