The CLEF 2003 cross language image retrieval task

Paul Clough and Mark Sanderson
University of Sheffield
p.d.clough|m.sanderson@sheffield.ac.uk

Abstract

In this paper, we describe a pilot experiment run at CLEF 2003 for cross language image retrieval, called ImageCLEF. The task is this: given a user need expressed in a language different from the document collection, find as many relevant images as possible. To facilitate retrieval, textual captions are associated with each image, thereby enabling (but not limiting) retrieval using text-based retrieval methods. This paper describes our experiences of building a test collection for the ImageCLEF task, discusses the results from this campaign, and outlines our ideas for further ImageCLEF experiments.

1 Introduction

Retrieval from an image collection offers distinct characteristics from one in which the document to be retrieved is natural language text. For example, the way in which a query is formulated, the method used for retrieval (e.g. based on low-level features derived from an image, or based on an associated caption), the types of query, how relevance is assessed, the involvement of the user during the search process, and fundamental cognitive differences between the interpretation of visual versus textual media.

Within CLEF, the problem is further complicated by specifying user queries in a language different to that of the document collection. This requires that a bridge is built between the language barrier by translating the collection, the queries, or both into the same language. As multimedia collections grow and many organisations manage large image repositories, the instigation of ImageCLEF addresses an important problem that is not dealt with by existing CLEF and iCLEF tasks. Furthering research in cross language image retrieval is appealing both academically and commercially as organisations would be able to offer the same collections to a wider and more diverse range of users with differing language backgrounds.

As a retrieval task, cross language image retrieval encompasses two main research areas: (1) image retrieval, and (2) cross language information retrieval (CLIR). Providing a suitable test collection for such different retrieval tasks is a tall order, therefore the primary aim of this test collection is to evaluate systems built to accept user requests formulated in a language different from the image captions to find relevant images. However, unlike searching text collections, previous research (see, e.g. [9]) has shown that image retrieval does have its own characteristics, e.g. that users tend to browse as well as perform specific searches, that users will often consult associated textual metadata to decide whether an image is relevant or not, that users search for both abstract and concrete concepts, that search requests tend to be more specific than textual searches, and that users request specific instances of objects rather than a general category, e.g. "London bridge" rather than just "bridges" (see also [7], and [10]). Of course, like text retrieval, search requests may vary: depending on the domain and the users searching ability.

Rather than make assumptions on which retrieval methods will be used on this test collection, e.g. combining both text-based and content-based retrieval methods, or the utilisation of relevance

feedback, we aim to provide a test bed that can be used to evaluate a range of retrieval methods and for analysing the behavior of users during the search process, e.g. query formulation in both cross language and visual environments, iterative searching, and query reformulation. Goodrum [9] calls for a TREC-style test collection for image retrieval, which will provide a benchmark set of queries, relevance assessments and evaluation measures. We partially fulfill this call through ImageCLEF by creating a test collection of images and captions, and this year offered participants an ad hoc retrieval task for cross language image retrieval.

2 The cross language image retrieval task

Two retrieval tasks were proposed for ImageCLEF: (1) ad hoc retrieval, and (2) an interactive search task. Because of a lack of interest in the latter task, in this paper we focus on the former ad hoc retrieval task. The exploratory nature of ImageCLEF also meant we concentrated our efforts on building a suitable test collection, rather than identifying and designing different retrieval tasks. The aim this year was to determine whether CLEF participants would be interested in this kind of retrieval task, rather than expend our efforts on identifying the characteristics of cross language image retrieval.

The aim of the ImageCLEF ad hoc retrieval task is this: given a multilingual statement describing a user need, find as many relevant images as possible using automatic or manual retrieval methods. This task is similar to the classic TREC ad hoc retrieval task in that we simulate the situation in which a system knows the set of documents to be searched, but cannot anticipate the user requests (i.e. queries are not known in advance). Participants are free to use whatever methods they wish to retrieve relevant images, including content or text-based retrieval methods, relevance feedback and any translation method. This retrieval task simulates when a user is able to express their need in a natural language expression, but require a visual document to fulfil their search. Typical evaluation of content-based retrieval assumes the user performs "query-by-example" search using an exemplar of what it is they require, in ImageCLEF we provide search requests that consist of both a visual exemplar and textual description of the user need.

A question we might ask ourselves is whether this CLEF task is any different from the other CLEF tasks? At this stage in ImageCLEF, the answer has to be yes and no. In one respect the ad hoc retrieval task is the same as any other cross language task if retrieval is based on the captions only, however because the document to be judged is an image, approaches which exploit other features such as low-level content-based cues can also be used to enhance retrieval. Also, from our experiences of manually judging the relevance of images with respect to the ad hoc search requests, we also know that the image itself plays an important part in the judgment process. For example, a caption containing query terms may not be judged as relevant because the image is too small, too dark or not taken from the desired angle or view. On the other hand, an image may be relevant to the user but contain only a few or none of the query terms in the caption, perhaps because the caption is of poor quality, the language used by the annotator does not match the search request, or the caption is very short in length.

ImageCLEF aims to provide the necessary collection and framework in which to analyse the link between the image and text, and promote the discovery of alternative methods of retrieval for cross language image retrieval. We imagine that ImageCLEF will appeal to researchers from a variety of communities including image retrieval, CLIR, and user interaction.

3 The test collection

The classic means of measuring the performance of an information retrieval (IR) system is based on a test collection. This provides the necessary resources and framework in which to assess performance and is typically used to compare different retrieval methods or systems. The design of a standardised resource for IR evaluation was first proposed by Cleverdon [8] and has since been used in major IR conferences such as TREC [3], CLEF [4] and NTCIR [6].

Much work has taken place in addressing particular methods of collection construction including the kinds of documents to include, the types of requests users are likely to make on the collection, and how relevant documents should be defined based on such requests. Vorhees [5] discusses the main assumptions behind TREC's method of building an IR test collection (based on Cleverdon's ideas) and although critics have argued against this approach, over the years the creation of a standard test environment has proven invaluable for design and evaluation of practical retrieval systems. So building upon previous research in test collection construction, the three main constituents of the ImageCLEF collection are:

1. **Document collection:** a set of documents (e.g. texts and images).
2. **Topics:** a set of user information needs.
3. **Relevance assessments:** a set of relevance judgments associated with the topics.

Voorhees and Harman [3] suggest that the set of texts used as the document collection is a sample which is likely to be encountered within an operational setting. Of course, it is not always possible to capture the entire population for a text collection (e.g. a constantly growing collection of news stories) therefore some kind of sampling is required to create a "representative" sample. Some test collections are built around specific domains (e.g. MEDLARS, CACM), or tasks (e.g. journalists finding images for an illustration task [10]), whereas other collections are more general (e.g. TREC). In ImageCLEF, we wanted to create a collection which represented a realistic domain, but on which a particular task was not imposed to create a more general collection suitable for further retrieval tasks and experiments in the future.

3.1 The document collection

Selecting a suitable collection for ImageCLEF proved to be a non-trivial task. Not only did we require a "large" collection of images, but also a collection with captions of high quality to facilitate text-based retrieval methods. Additional issues involving copyright were also encountered as typically photographs and images have a potentially high marketable value and therefore not permitted to be distributed to ImageCLEF participants. Our search was eased by our links with the library at St. Andrews University[1]. They hold one of the largest and most important collections of historic photographs in Scotland, exceeding over 300,000 photographs from a number of well-known Scottish photographers [11]. A cross-section of approximately 30,000 images from the main collection has been part of a large-scale digitisation project to enable public access to the collection via a web interface.

This collection was used as the basis for ImageCLEF because the collection represents a realistic image archive, high quality captions are associated with the images, and permission was granted by St. Andrews Library to downloaded and distribute the collection for use in ImageCLEF. Over a two day period, we automatically crawled the photographic collection, filtered out duplicate images and converted the captions into a TREC-style format. The collection consists of 28,133 images (a 368x234 large version and 120x76 thumbnail) and captions. The majority (82%) of images are in black and white (because of the historic nature of the collection) ranging from 1832 to 1992 (with mean of 1920). Images and captions of varying styles, presentation and quality exist in the collection from a diverse range of topics making it challenging for both image and text retrieval. Figure 1 shows an example image and caption from the collection.

The captions, a vital part of the test collection, consist of data in a semi-structured format added manually by domain experts at St Andrews. The caption contains 8 fields, all or a combination of which can be used for text-based retrieval. Information in the caption ranges from specific date, location and photographer to a more general description of the image. Approximately 81% of captions have text in all fields, the rest generally without a description. In most cases the description is a grammatical sentence of around 15 words enabling possible use of NLP technology to

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

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

extract linguistic information, e.g. subject-object relations. The categories have been added manually by St. Andrews annotators and could be used for future image categorisation experiments. The captions exist only in English.

3.2 Selecting suitable topics

System effectiveness for the ad hoc retrieval task is evaluated against a set of user needs called topics. Deciding on which topics to include in the test collection is crucial as if they are not representative of the collection, or they differ from real user requests, effectiveness measured with the test collection will not be realistic of that one might expect to obtain in a practical setting. In TREC, NTCIR and CLEF, final topics are chosen from a pool of suggestions generated by searchers familiar with the domain of the document collection. This initial set is narrowed-down based on several conditions, including the estimated number of relevant documents for each topic, the variation of task parameters (e.g. for a multilingual task the topics are chosen to test different translation problems), the difficulty of the topic, and its scope (e.g. broad or narrow, general or specific). The goal is to "achieve a natural, balanced topic set accurately reflecting real world user statements of information needs" [2, pg. 1069].

The first author created a set of fifty topics that would test the capabilities of both a translation and image retrieval system, e.g. pictures of specific objects versus pictures containing actions, broad versus narrow concepts, topics containing proper names, compound words, abbreviations, morphological variants and idiomatic expressions. To determine potential subject areas for ImageCLEF topics, we first became familiar with the St. Andrews collection by browsing its contents, and analysing log files taken from the St. Andrews Library web server hosting the photographic collection over a two year period.

The result of a simple analysis of the log files (together with the subject categories used to group images) gave a set of 2796 distinct queries (of which 67% occur only once). Like previous results from examining image searches from Web search engines [1], we find on average that individual queries against the St. Andrews collection tend to be short and specific, i.e. requests for specific objects or locations. However, a more extensive click-through analysis would be necessary to verify

our initial findings. By ranking queries by frequency, we were able to find the most common search requests and topics (e.g. from the top 15 most common queries, 60% of these were proper names).

After deciding the initial subject areas, we used combinations of keywords related to that area to explore possible topics by adding or removing words to make searches more specific or general. For example, the subject area "fishermen" might become "fishermen by the photographer Adamson", "churches" become "churches with tall spires", or "tay bridge railway disaster" might become "metal railway bridges". This results in finding more or less possibly relevant documents. The aim of this process was to create a set of topics that would test various issues involved with both query translation and image retrieval (and relevance assessment of visual information). The set of keywords used in this initial stage were used as the titles in the topic description, a few keywords describing the required topic. Table 1 lists the selected 50 topics and section 3.5 discusses some of the comments from assessors regarding those topics chosen for ImageCLEF.

3.3 Creating the topic statement

Given the fifty selected subject areas, we expressed the user's need using both a natural language statement and an example image. Because we used several judges for the relevance assessment stage, the topic must convey to both the judge assessing relevance and the participants using the topics what is expected from the topic by defining what are relevant and non-relevant images. ImageCLEF topics consist of 2-3 keywords, the title, a short description of relevance, a narrative, and an example relevant image. Because the task involves image retrieval, we decided to supply an example image as part of the topic statement which would not typically accompany a topic definition for texts. Participants were free to use this image, maybe as part of a content-based search, or part of a relevance feedback cycle.

Given the multilingual nature of ImageCLEF, topic titles were translated into Spanish, Dutch, German, French, and Italian by native speakers of each language[2]. We did not translate the narratives because of limitations in time and resources available to us. Translators were also given an example of a relevant image and also asked to specify alternative translations where appropriate (e.g. a colloquial version of the translation). These translation variations were supplied as part of the ImageCLEF topic statement.

3.4 The relevance assessments

What turns a set of documents and queries into a test collection are the relevance judgments, manual assessments of which documents are relevant or not for each topic. There are two areas of concern with relevance assessments: (1) the quality or subjectivity of the judgment, and (2) their coverage or completeness. The first concern is based on disagreement between assessors about what constitutes a relevant document caused by subjectivity (e.g. knowledge of the topics or domain, different interpretations of the same document, and their experience of searching). This is not a trivial problem to solve, but we did two things to reduce this: (1) have two assessors judge each topic, and (2) capture information from the assessors about their judgments.

The second issue deals with the coverage of relevance assessments. Ideally every document in the collection would be judged for relevance for each topic, but with large collections this becomes infeasible as it requires too much manual effort (even though from ImageCLEF we find it is much quicker to judge the relevance of an image versus text). To make assessment feasible, pooling has been used by TREC, CLEF and NTCIR. In this technique, a set of candidate documents is created (the pool) by merging together the results of the top n documents from the ranked lists provided by participants. This assumes that highly ranked documents from each entry will contain relevant documents; questions left to deal with include what size of n is chosen, how many systems are used in the pooling process, and which systems are used to create the pool. Ideally, the ranked lists should come from a diverse range of systems to ensure maximal coverage, however because we had only 4 participants in ImageCLEF (resulting in 45 runs for each topic), the method proposed by

[2] One of the participating groups, NTU, manually translated the topics into Chinese (traditional and simplified) and submitted Chinese to English runs.

Kuriyama et al. [6] and used in NTCIR, that of supplementing the pooling method with manual interactive searches, was applied (also known as *interactive search and judge* or *ISJ*). This was found to enhance recall and improve the coverage of relevant documents (particularly with queries requiring a more general image) - see Table 1.

To assess the topics, the topic creator assessed all fifty topics to provide a "gold" set of judgments (this involved assessing around 50,000 images); in addition, ten assessors from the University of Sheffield judged five topics each to provide a second judgment for each topic (enabling the agreement between assessors to be evaluated). Judging was made easier by creating a custom Web-based assessment tool which enabled the judgment of images in the pools (ranked by the proportion of systems which included the image in the top 100 documents), as well as providing an interactive search environment to supplement the pools as necessary. This tool enabled assessors to make judgments from any location; combined with term highlighting features in the `Google` toolbar, judgments were made quickly and easily. Assessors were asked to judge the relevance of *all* images from the topic pool based on the relevance description and using a ternary scheme: relevant, partially relevant, and not relevant. To convey the topic statement to the assessors, they were provided with the topic as given to the participants and an example image. They were free to contact the first author with any questions. Primary judgment was made on the image, but assessors were able to also consult the image caption. Images were to be judged relevant if *any* part of the image was deemed relevant.

The ternary scheme was adopted to deal with potential uncertainty in the assessor's judgment (i.e. it is possible to determine that the image is relevant, but less certain whether it fulfils the need described by the topic exactly), and to enable evaluation based on a strict set of relevance judgments (i.e. those documents marked as relevant only) and a more relaxed set (i.e. those marked as relevant and partially relevant). We believe this is particularly important in practical image retrieval evaluation to deal with situations where a binary judgment might be too restricting, e.g. only part of the image is relevant, the required object is obscured (e.g. perhaps in the background), the image is too small, or the image appears relevant, but the caption is unable to confirm its contents.

As assessment of images for relevance was found to be much faster than assessment of text documents, *all* 45 submitted runs were assessed. By using all runs for both monolingual and cross lingual entries (automatic and manual) to create the pools and then using interactive searches to supplement the pools, we hoped to find as many relevant images as possible from the collection and maximise coverage. Rather than create a single set of relevant images (qrels) for each topic, we created four sets based on the overlap of relevant images between the assessors. We created the following four relevance sets from which we evaluated participant's entries:

1. **Union-strict:** the union of images judged as relevant by the two assessors and evaluated using only those marked as relevant.
2. **Union-relaxed:** the union of images judged as relevant by the two assessors and evaluated using those marked as relevant or partially relevant.
3. **Intersection-strict:** the intersection of images judged as relevant by the two assessors and evaluated using only those marked as relevant.
4. **Intersection-relaxed:** the intersection of images judged as relevant by the two assessors and evaluated using those marked as relevant or partially relevant.

The strict relevance set can be contrasted with a high-precision task, and the relaxed set providing an assessment that promotes higher recall. The most stringent category of relevance assessment is strict intersection as this produces the smallest number of relevant documents, and most relaxed category is relaxed union (see Table 1. Included in the pools are documents which were marked as relevant or partially relevant using the interactive search. In summary, the following procedure was used to assess relevance and evaluate participant's entries:

Topic	Title	Pool size	Added by ISJ	Strict ∩	Strict ∪	Relaxed ∩	Strict ∪
1	Men and women processing fish	504	0	9	14	15	25
2	A baby in a pram	908	3	5	16	25	64
3	Picture postcard views of St. Andrews	1312	118	23	100	25	136
4	Seating inside a church	1071	0	91	131	115	138
5	Woodland scenes	1239	0	137	300	238	502
6	Scottish marching bands	1145	5	5	8	7	16
7	Home guard on parade during World War II	772	2	5	8	7	16
8	Tea rooms by the seaside	1262	4	4	13	10	31
9	Fishermen by the photographer Adamson	601	0	4	6	4	7
10	Ships on the river Clyde	795	6	10	24	19	26
11	Portraits of Mary Queen of Scots	885	1	2	4	5	7
12	North Street St. Andrews	785	1	29	35	31	35
13	War memorials in the shape of a cross	803	6	12	27	14	31
14	Boats on Loch Lomond	1012	0	33	42	38	48
15	Tay bridge rail disaster	648	12	11	14	11	14
16	City chambers in Dundee or Glasgow	653	0	17	112	17	118
17	Great Yarmouth beach	937	1	9	11	10	13
18	Metal railway bridges	647	3	94	125	106	139
19	Culross abbey	643	0	3	3	3	3
20	Road bridges	1269	0	31	183	48	191
21	Animals by the photographer Lady Henrietta Gilmour	641	91	48	145	49	145
22	Ruined castles in England	698	41	42	85	53	114
23	London bridge	465	0	2	2	2	2
24	Damage due to war	695	4	12	14	12	17
25	Golf course bunkers	1383	3	12	18	22	37
26	Portraits of Robert Burns	831	4	6	6	6	6
27	Children playing on beaches	578	1	26	68	43	98
28	Pictures of golfers in the nineteenth century	1155	5	11	31	14	40
29	Wartime aviation	661	2	11	72	34	100
30	Glasgow before 1920	389	1	21	46	33	46
31	Exterior views of Indian temples	776	0	35	53	41	59
32	Male portraits	993	100	280	422	316	436
33	People using spinning machines	1192	1	7	10	8	10
34	Dogs rounding-up sheep	642	2	12	2	12	17
35	The mountain Ben Nevis	739	5	56	65	62	72
36	Churches with tall spires	1065	45	35	94	57	130
37	Men holding tennis racquets	736	0	2	3	3	4
38	Scottish fishing vessels by the photographer Thompson	929	1	14	16	14	19
39	Men cutting peat	881	0	5	5	5	7
40	Picture postcards by the Valentine photographic company	1218	266	104	726	194	899
41	A coat of arms	1006	6	10	25	13	46
42	University buildings	579	9	112	157	120	159
43	British Windmills	579	0	13	14	13	14
44	Waterfalls in Wales	731	0	24	27	24	28
45	Harvesting	595	2	14	22	18	24
46	Welsh national dress	1122	0	4	26	7	69
47	People dancing	882	5	10	19	12	19
48	Museum exhibits	467	7	17	37	25	49
49	Musicians and their instruments	440	3	7	17	7	21
50	Mountain scenery	1320	0	225	514	284	617
	Average	846	15	35	78	46	95

Table 1: Topics used in ImageCLEF and their assessment characteristics

1. Extracted the top 100 runs from each submission (45 submissions in total) for each topic.
2. Computed the union of documents from these runs to create the document pool.
3. Two assessors manually judged each image in the document pool using the ternary scheme.
4. An interactive search was then performed to supplement the document pools with relevant images not found in the pool.
5. Four relevant document sets (qrels) for each topic were created from the pools.
6. Documents from each system run were compared against each set of qrels.
7. Measures of effectiveness were computed using trec_eval for each topic and across all 50.

3.5 Feedback from translators and assessors

The opinions of translators and assessors was sought for their views on the tasks they undertook. For the translators, some English queries were found to have no similar concepts in another language and therefore had to be expressed using different terms. Some of the ten relevance assessors spoke English as a second language (all albeit very well) and they struggled to understand some of the image captions due to colloquial or historic language usage, e.g. "perambulator" and "infant". Assessors noted that is was very easy to judge irrelevant images based on the image alone. The copies of the images in the collection were found to be sometimes not detailed enough to allow effective assessment to take place. Judging if an object was in the foreground or background was also not always straightforward. The partially relevant category was not well used by some of the assessors.

4 Evaluation

4.1 The participants

Four groups participated in ImageCLEF 2003: the University of Surrey, National Taiwan University (NTU), Daedalus (Spain), and the University of Sheffield, exhibiting variety in the retrieval systems used and their methods for translating topic titles. None of the groups made use of content-based image retrieval methods during the retrieval process; all groups used information derived from the captions only. Sheffield and Surrey provided one automatic run for each language (including monolingual), whereas Daedalus submitted 25 runs based on different system parameters for all languages except Dutch. NTU translated the topic titles into Chinese and submitted 4 manual runs and 4 automatic runs for Chinese to English only.

NTU used the `OKAPI` retrieval system and two methods for translating Chinese queries into English based on dictionary lookup using four resources: the LDC Chinese-English dictionary, Denisowski's CEDICT, the BDC Chinese-English dictionary v2.2, and an in-house resource. Two methods were used to select translations: (1) a co-occurrence model, and (2) the two most frequent translations. In both cases, a backward transliteration model was used to translate person and location names not contained in the bilingual dictionary.

Daedalus used a probabilistic IR system called Xapian and lingusitic resources to deal with tokenisation, word decompounding, morphological variants and removal of stopwords. For translation, a dictionary-based lookup strategy was used based on three resources: FreeTranslation.com, LangToLang.com, and ERGANE (word-by-word translation). For English monolingual runs, WordNet was also used to expand queries with their synonyms, and in all retrieval experiments queries were constructed using an OR-ing approach.

Sheffield University also used a probabilistic system based on the BM25 weighting function for retrieval, but compared to the other groups used minimal language processing made possible through the use of Systran (on-line version). All fields in the caption were used for retrieval, and documents ranked by their BM25 score.

Source	Surrey	NTU	Sheffield	Daedalus
MONO (MAP)	0.0624	-	0.5616	0.5718 (Qor)
(#failed topics)	27	-	1	1
IT	0.0539	-	0.4047	0.4043 (QTdoc)
	30	-	7	5
DE	0.0503	-	0.4285	0.4083 (QTdoc)
	29	-	8	5
NL	0.0289	-	0.3904	-
	32	-	7	-
FR	0.0529	-	0.4380	0.3710 (QTor1)
	30	-	3	5
ES	0.0320	-	0.4076	0.4323 (QTdoc)
	32	-	3	5
CN	-	0.2888 (NTUiaCoP)	0.2850	-
	-	12	12	-

Table 2: Highest MAP and number of topics with no relevant image in the top 100 for each language and group (strict intersection)

The University of Surrey used two free Internet translation resources and expanded queries using WordNet. Unfortunately, due to a misconfiguration problem with their system, the results submitted by them were not correct.

4.2 The results

Systems are evaluated by comparing the output of submitted runs with the relevance sets (qrels) to determine how many relevant documents appear in the retrieval, and their rank position. We use a form of `trec_eval`[3] to compute retrieval effectiveness. Table 2 shows average precision for each participant's best run on a particular part of the ImageCLEF task. In addition, the number of topics for which no relevant images were returned in the top 100 is also shown; here it was assumed that users would be willing to examine the top 100 images from a topic and would regard finding no relevant images in that top set as a failure of the retrieval system for that topic. Somewhat unusually for effectiveness statistics, the two measures, especially between Sheffield and Daedalus, sometimes contradict each other: with one system scoring higher in average precision, while the other scores lower in number of failed topics.

5 Acknowledgments

The ImageCLEF track would not have run had it not been for St. Andrews University Library - in particular the curator of the St. Andrews image collection, Norman Reid - allowing a portion of the collection to be released for the track. This work was carried out within the Eurovision project at the University of Sheffield, funded by the EPSRC (Eurovision: GR/R56778/01).

6 Bibliography

References

[1] Visual information seeking: A study of image queries on the world wide web. In *In Proceedings of the 1999 Annual Meeting of the American Society for Information Science*, 1999.

[3] We make use of the UMASS version of `trec_eval`.

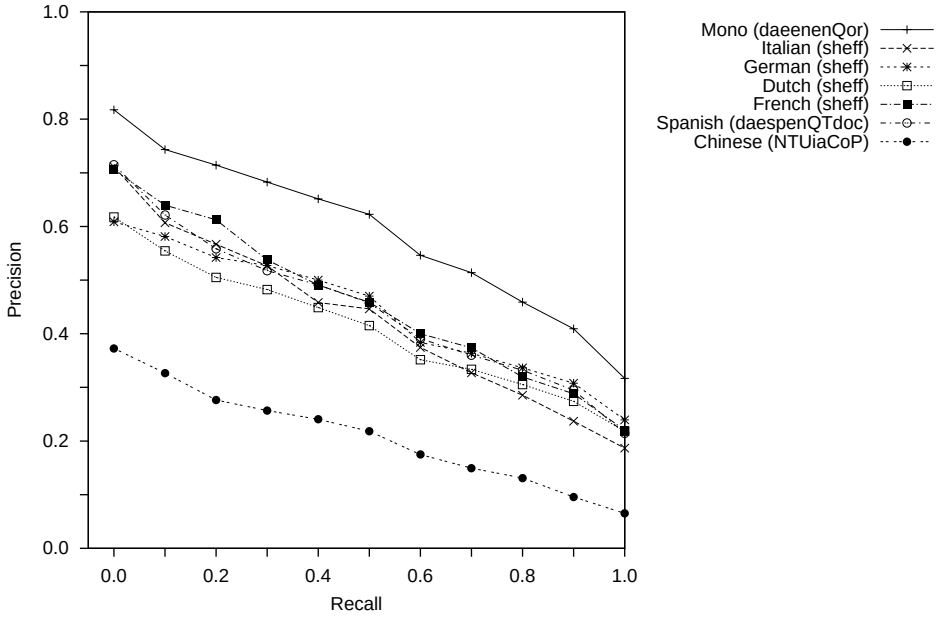

Figure 2: Precision-Recall for best performing systems for each language (using strict intersection)

[2] Cross-language system evaluation: The clef campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001.

[3] Overview of trec 2001. In *In Proceedings of TREC2001, NIST*, 2001.

[4] Clef methodology and metrics. In *In C. Peters (Ed.), Cross-language information retrieval and evaluation: Proceedings of the CLEF2001 Workshop, Lecture Notes in Computer Science 2406, Springer Verlag*, pages 394–404, 2002.

[5] The philosophy of information retrieval evaluation. In *In C. Peters (Ed.), Cross-language information retrieval and evaluation: Proceedings of the CLEF2001 Workshop, Lecture Notes in Computer Science 2406, Springer Verlag*, pages 355–370, 2002.

[6] Pooling for a large-scale test collection: An analysis of the search results from the first ntcir workshop. *Information Retrieval*, Vol. 5(1):41–59, 2002.

[7] L.H. Armitage and P. Enser. Analysis of user need in image archives. *Journal of Information Science*, Vol. 23(4):287–299, 1997.

[8] C.W. Cleverdon. *The Cranfield tests on index language devices*, pages 47–59. In: K, Spark-Jones and P. Willett (eds), Readings in Information Retrieval, Morgan Kaufmann, 1997, 1997.

[9] A.A Goodrum. Image information retrieval: An overview of current research. *Informing Science*, Vol. 3(2):63–66, 2000.

[10] M. Markkula, M. Tico, B. Sepponen, K. Nirkkonen, and E. Sormunen. A test collection for the evaluation of content-based image retrieval algorithms - a user and task-based approach. *Information Retrieval*, Vol. 4(3/4):275–294, 2001.

[11] N.H. Reid. The photographic collections in st andrews university library. *Scottish Archives*, Vol. 5:83–90, 1999.