A Proposal for the CLEF Cross Language Image Retrieval Track 2004

Paul Clough[1], Mark Sanderson[1] and Henning Müller[2]

[1]Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK.
{p.d.clough,m.sanderson}@sheffield.ac.uk

[2]University Hospitals of Geneva, Division of Medical Informatics, 21 rue Micheli-du-Crest, CH-1211 Geneva 4, Switzerland.
{henning.mueller}@dim.hcuge.ch

Abstract. In this paper we describe our proposal for a cross language image retrieval task called ImageCLEF[1] being run as part of the Cross Language Evaluation Forum (CLEF[2]). A pilot experiment was organised for 2003 in which participants performed an ad hoc bilingual image retrieval task on a collection of historic photographs to simulate the situation in which users express their search request in natural language but require visual documents in return. For 2004 we plan to extend the tasks to include a medical image retrieval task and user-centred evaluation.

1 Introduction

A great deal of research is currently underway in the field of Cross Language Information Retrieval (CLIR) where documents written in one language are retrieved by a query written in another (see, e.g. [11] and [16]). One can consider CLIR as basically a combination of machine translation (MT) and traditional monolingual information retrieval (IR). Most CLIR research has focused on locating and exploiting translation resources with which the user's search requests or target documents (or both) are translated into the same language. Campaigns such as the Cross Language Evaluation Forum (CLEF) [16] and the Text REtrieval Conference (TREC) [20] cross language task have helped encourage and promote international research for multilingual retrieval and create standardised resources for evaluation.

However, one area of CLIR research which has received less attention is image retrieval. In many collections images are accompanied by some kind of text, e.g. metadata or captions, semantically related to the image [2][12]. Images can then be retrieved using standard IR methods based on textual queries. Many image collections exist where textual captions accompany individual or groups of images such as his-

[1] ImageCLEF: http://ir.shef.ac.uk/imageclef2004/
[2] CLEF: http://www.clef-campaign.org

toric or stock-photographic archives, medical case notes and art and history collections.

Retrieval from an image collection offers distinct characteristics from one in which the document to be retrieved is natural language text [1][10]. 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 associated text), 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. Methods of image retrieval are typically based on visual content[3] (e.g. colour, shape, spatial layout and texture), or by text/metadata associated with the image (see, e.g. Smeulders et al. [18] and Goodrum [10]).

For those organisations managing image repositories in which text is associated with images (e.g. on-line art galleries), one way to exploit these is by enabling multilingual access to them. To promote research in this area we instigated a cross language image retrieval task called ImageCLEF [5] as part of the Cross Language Evaluation Forum. We felt this contribution to the CLEF campaign would address an important and timely problem not dealt with by existing cross language evaluations.

We envisage ImageCLEF will appeal to both commercial and academic research communities including: cross language information retrieval, image retrieval, and user interaction. The main aims of the ImageCLEF campaign are:

- To promote and initiate international research for CL image retrieval.
- To further our understanding of the relationships between CL texts and images for IR.
- To create a set of useful standardised resources for CL image retrieval to scientific communities in the whole.

The paper divides into the following: in section 2 we describe the ImageCLEF 2003 test collection for an ad hoc retrieval task, in section 3 we describe the proposed tasks for ImageCLEF 2004, finally in section 4 we summarise the contents of this paper and provide some ideas for future work in cross language image retrieval.

2 Building a Test Collection for Multilingual Image Retrieval

Evaluation of retrieval systems is either system-focused, e.g. comparative performance between systems or user-centered, e.g. a task-based user study. For many years IR evaluation has been dominated by comparative evaluation of systems in a competitive environment. The design of a standardised resource for IR evaluation was first proposed over 30 years ago by Cleverdon [4] and has since been used in major IR conferences such as TREC [20], CLEF [16] and NTCIR [3]. Over the years the creation of a standard test environment has proven invaluable for the design and evaluation of practical retrieval systems both within and outside a competitive environment. The main components of a TREC-style test collection are: (1) document collection, (2) topics, and (3) relevance assessments.

[3] These are called Content-Based Information Retrieval (CBIR) systems.

In TREC, NTCIR and CLEF, participants are given test collection data and topics and asked to submit their entries. A subset is chosen by the organisers and used create document pools, one for each topic. Domain experts (assessors) are then asked to judge which documents in the pool are relevant or not. These document pools are created because for large collections it is infeasible to judge every single document for relevance. These assessments are then used to assess the performance of submitted systems.

User-centred evaluation is important to assess the overall success of a retrieval system which takes into account other factors other than just system performance, e.g. the design of the user interface and system speed (Dunlop argues this in [7]). A number of researchers have highlighted the advantages of user-centred evaluation, particularly in image retrieval systems (see, e.g. [10], [14] and [7]). One of the main aims of ImageCLEF is to provide both the CLIR and image retrieval communities a number of useful resources (datasets and relevance assessments) to facilitate and promote further research in multilingual image retrieval.

Calls for a TREC-style evaluation for image retrieval systems have been suggested [10][15][19], although Forsyth [9] argues that the evaluation of CBIR systems at the moment is useless because systems are too bad (hence the interest in combining both textual and visual features). We are unaware of existing test collections for CL image retrieval, although evaluation resources do exist to evaluate specific image retrieval tasks, e.g. journalism [13] and more generally Benchathalon to evaluate CBIR systems, TRECvid [] for video retrieval systems and CLEF and TREC for speech retrieval.

One of the largest obstacles in creating a test collection for public use is securing a suitable collection of images for which copyright permission is agreed. This has been a major factor influencing the datasets used in the ImageCLEF campaigns. The ImageCLEF test collection provides a unique contribution to publicly available test collections and complements existing evaluation resources.

2.1 The Existing ImageCLEF Test Collection

Because cross language image retrieval encompasses at least two research areas: (1) image retrieval and (2) cross language information retrieval, building a suitable test collection is a tall order. Therefore, in 2003 we organised a pilot experiment at CLEF with the following aim: given a multilingual statement describing a user need, find as many relevant images as possible. More formally the task was a bilingual ad hoc retrieval task in which a static collection was searched using previously unseen topics.

The retrieval task was designed to simulate the situation in which a user expresses their need in a language different from the collection and requires a visual document to fulfil their search request. Participants were not restricted in their method of retrieval enabling either text or content-based searches or a combination of both. As a retrieval task there are several challenges as well as translation including: (1) captions which are typically short in length, (2) images that vary widely in their content and quality, (3) matching the language used in CL queries with captions annotated by historians, and (4) short user search requests which provide little context for translation.

Record ID:	JV-A.000460
Short title:	The Fountain, Alexandria.
Long title:	Alexandria. The Fountain.
Location:	Dunbartonshire, Scotland
Description:	Street junction with large ornate fountain with columns, surrounded by rails and lamp posts at corners; houses and shops.
Date:	Registered 17 July 1934
Photographer:	J Valentine & Co
Categories:	[columns unclassified][street lamps - ornate][electric street lighting][shepherds & shepherdesses][streetscapes][shops]
Notes:	JV-A460 jf/mb

Fig. 1. An example image and caption (see: http://www-library.st-andrews.ac.uk).

The dataset used consisted 28,133 historic photographs from the library at St Andrews University [17]. All images are accompanied by a caption consisting of 8 distinct fields which can be used individually or collectively to facilitate image retrieval (see Fig. 1). The 28,133 captions consist of 44,085 terms and 1,348,474 word occurrences; the maximum caption length is 316 words, but on average 48 words in length. All captions are written in British English and contain colloquial expressions and historical terms. Approximately 81% of captions contain text in all fields, the rest generally without the description field. In most cases the image description is a grammatical sentence of around 15 words. The majority of images (82%) are black and white, although colour images are also present.

We generated fifty representative search requests in English (called *topics*) and translated them into 6 different languages: Dutch, Spanish, German, French, Italian and Chinese (provided by the National Taiwan University or NTU). In TREC, CLEF and NTCIR final topics are chosen from a pool of suggestions generated by searchers familiar with the domain of the document collection. Frequently searched subject areas in the St Andrews were identified by analysing log files generated from accesses to a web search engine used by the library.

Based on these subject areas we created queries 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 idioms. Each topic consisted of a short title, a longer narrative describing the search request and an exemplar relevant image. For ImageCLEF 2003 only topic titles were translated due to limited resources available to us.

2.2 Relevance Assessments and Evaluation

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. 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 document, and searching experience), therefore to minimise this two assessors judged each topic.

Ideally every document in the collection would be judged for relevance for each topic, however this is infeasible. We therefore adopted the pooling method as used in TREC, CLEF and NTCIR where a set of candidate documents is created (called 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. Ideally, the ranked lists should come from a diverse range of systems to ensure maximal coverage.

We also used the method adopted by NTCIR: that of supplementing the pooling method with manual interactive searches (also known as *interactive search and judge* or ISJ) to ensure good quality pools. We found assessors were able to judge the relevance of images very quickly (especially eliminating non-relevant ones) enabling *all* ImageCLEF submissions to be used in creation of the pools. One of the authors familiar with the collection assessed all fifty topics to provide a "gold" set of judgments; in addition, ten assessors from the University of Sheffield judged five topics each to provide a second judgment for each topic using a custom-built assessment tool.

Images were judged relevant if *any* part of the image was deemed relevant. Primary judgment was made on the image, but assessors were able to also consult the image caption. Assessors were asked to judge the relevance of images using a ternary scheme: relevant, partially relevant and not relevant 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).

Unlike other test collections we provided four sets of relevance assessments (called *qrels*) - strict/relaxed union/intersection - with which to assess system performance based on the overlap of relevant images between assessors and whether the relevance sets include images judged as partially relevant or not. These are further described in [5]. The strict relevance set can be contrasted with a high-precision task; the relaxed set providing an assessment that promotes higher recall.

3 The Proposed ImageCLEF 2004 track

3.1 The Ad Hoc Image Retrieval Task

As a cross language retrieval task, Flank [8] and results from ImageCLEF 2003 have shown that multilingual image retrieval based on query translation can achieve high performance compared with a monolingual baseline. In particular results from ImageCLEF showed: for Chinese retrieval transliteration of proper names was beneficial, and for other languages thesaurus-based query expansion improved performance. The ad hoc task as it stands makes a suitable entry-level task to CLEF and CLIR because it is possible to participate without using content-based retrieval approaches.

A similar task will be run in 2004 to enable further experiments on the St Andrews dataset. For example to compare: (1) different methods of query translation (e.g. dictionary-lookup versus MT), (2) query expansion (e.g. global versus local methods), (3) the use of text-based and CBIR methods used either separately or combined, (4)

different retrieval models, (5) different indexing methods (e.g. indexing all or some fields) and (6) manual vs. automatic relevance feedback.

A new set of 25 topics will be produced this year in the same manner as before, however as well using query logs we will also use subject areas supplied by library staff from St Andrews. In this way topics will be based on real search requests which library staff have been asked to perform for external clients. As before we will alter these subject areas to create the final topic set so they test a range of different CL and image search parameters. Topics will be translated into languages used before, plus Japanese, Danish, Russian and Swedish.

One non-intentional but interesting "feature" of translated topics in ImageCLEF 2003 was the introduction of translation errors, e.g. spelling mistakes and erroneous diacritics, resulting in low retrieval performance for some topics. These problems are not addressed by existing CLEF tasks. We will provide two sets of topics: one set will contain spelling errors; the other will be checked and free of such errors.

3.2 The Medical Image Retrieval Task

To offer participants a different domain/scenario and encourage the use of CBIR system we have introduced a task based on medical retrieval. In the ad hoc task it is the query which is multilingual; in the medical retrieval task the document collection is multilingual presenting different retrieval challenges.

In general, medical practitioners are not satisfied with retrieving images by text and the implicit knowledge stored in the images plus attached text is rarely used. As a diagnostic aid being able to search a database of images with a new example would enable them to get more evidence on it. The goal of this task is to investigate the use of CBIR and text-based retrieval systems for this kind of medical task. This task is being run by University Hospitals of Geneva who are supplying the medical data, topics and relevance judgments.

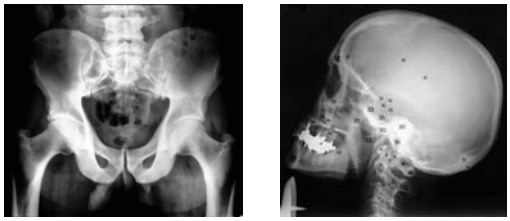

Fig. 2. Example images from the CasImage dataset: http://www.casimage.com/.

The aim of the task is this: given an example medical image, find similar images which will be helpful in confirming the initial diagnosis. Because the initial retrieval has to be visual, we expect the case notes to be useful in finding additional similar images complementary to CBIR. We also aim to evaluate whether relevance feedback can improve performance, compare relevance feedback using either image/text or both, and whether images alone can be used for pseudo relevance feedback.

The dataset (CasImage) consists of 8,751 anonymised medical images, e.g. scans, and x-rays (see Fig. 2). The majority of images are associated with *case notes*, a writ-

ten description of a previous diagnosis for an illness the image identifies. Case notes consist of several fields including: a diagnosis, a description, clinical presentation, keywords and title. The task is multilingual because case notes are mixed language written in either English or French. 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. In the dataset there are 2,078 cases to be exploited during retrieval (e.g. query expansion).

Currently 25 example images (topics) have been chosen as representative from the dataset. A set of ground truths for each topic has already been identified by domain experts based on the CBIR system developed by the third author[4] and these will form part of the document pools created from participant's entries. Pools will be formed in a manner similar to the ad hoc task and medical practitioners will help judge the relevance of the pools after final submissions. In this task images will be judged using a binary relevant or not relevant judgement and the pools will be used to evaluate participant's entries.

This retrieval task offers a number of challenges including: (1) combining text and content-based methods of retrieval after an initial visual search, (2) dealing with domain-specific medical terminology, (3) case notes of varying quality in more than one language (i.e. a mixed language index), and (4) the high cost of returning non-relevant images (i.e. mis-diagnosis) which is always inevitable when using visual-only search methods.

3.3. The Interactive Image Retrieval Task

Campaigns such as iCLEF[6] have shown the value of user-centred evaluation for CLIR. Multilingual image retrieval offers a rich source for user-centred experiments. Past research has shown that the search activities of a user in an image retrieval system vary between searching for specific images and browsing the image collection (see, e.g. [10] and [6]). For a CL image retrieval system, the issue is how best the system can support the user's search in locating relevant images as quickly, easily and accurately as possible.

User-centred evaluation in a variety of contexts and domains will help us determine how CL image retrieval systems can best help users to: (1) formulate their queries (e.g. whether text or visual queries alone are best or can be used in combination), (2) refine the search request - query reformulation will depend on the outcome of the system and could involve refinements using textual and/or visual features, (3) browse the collection, and (4) identify relevant images (e.g. what additional information would help the user judge the relevance of an image and how best is this displayed).

Cox et al. [6] suggest three classes of image search: (1) target or known-item search (i.e. find a specific image), (2) category search (e.g. "find pictures of the Eiffel Tower") and (3) open-ended browsing (i.e. wandering through the collection). They argue that the target search encompasses the other categories of search; it is simple for

[4] See http://viper.unige.ch/ for a list of publications about the VIPER CBIR system.
[6] See http://terral.lsi.uned.es/iCLEF/ for information about iCLEF.

the user to perform and has clear measures of effectiveness. The goal for the user in such a task is given an image, to find it again from the collection. Unlike being given a textual description of a topic, the user must interpret the given image and generate suitable query terms in a given language (different from the document collection).

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 the user to describe the image instead, e.g. searches for a familiar painting whose title and painter are unknown. Assessing whether different users generate the same keywords will be an interesting aspect of this task.

This task will use the St Andrews dataset and our experimental setup will follow the guidelines for user-centred experiments as suggested by iCLEF. This task will be undertaken with collaboration from iCLEF organisers to ensure a consistency in CLEF methodologies. Participants are asked to follow the experimental setup but can perform whatever experiments they like.

A minimum of 8 users and 8 topics will be required for this task. Users will be given 10 minutes to find each image using non-English searches. Captions must also be translated into this language before being displayed (if at all) to the user. The aim of this experiment will be to observe users search habits for this task and to determine what kind of interface best supports *query refinement*. For example the user is shown a picture of an arched bridge but starts with the query "bridge". By finding similar images and maybe using keywords from their captions, the user refines the query until the relevant image is found. Query refinement based on CBIR is also a possibility. Topics and systems will be presented to the user in combinations following a *Latin-square* design to ensure user/topic and system/topic interactions are minimised.

Qualitative performance measures will be captured using questionnaires provided by us, and quantitative measures will include whether the given image is found or not, the time taken to find the image, the number of images viewed before finding the image and number of user interactions required. We will provide 8 example images and users will be given 10 minutes to find each one. Topics will be general enough so that people unfamiliar with the collection can still perform the searches.

3.4 Evaluation Measures

We will use a number of measures of retrieval effectiveness to assess system performance. Evaluation will be based on un-interpolated average precision across topics, the number of topics with no relevant images in the top 100 (called *failed topics)*, the proportion of the top 100 which contain relevant images (precision at 100), and the proportion of relevant images found in the top 100 (a normalized precision at 100 which is not affected by the size of the qrels set). Also provided will be the "standard" measures of retrieval effectiveness including single-valued summaries and precision-recall graphs. In the medical retrieval task we may penalize systems which return non-relevant images in the final run because in this domain the results are potentially life-threatening if irrelevant images are used in an incorrect diagnosis.

4 Conclusions and Future Work

In this paper we have discussed our proposal for three cross language image retrieval tasks as part of the ImageCLEF campaign. The tasks vary across domain, scenario, where multilingual retrieval is used, whether content-based image retrieval is required and whether the task is system or user-centred. Our aim is to promote CL image retrieval and provide a standardised set of resources in the form of test collections (i.e. a collection, topics and relevance assessments) which can be used in further CL image retrieval experiments. In future work we plan to expand the collections and tasks offered in ImageCLEF. In particular we would like to offer an ad hoc retrieval task in which the document collection is non-English, provide an image retrieval task from the Web, and translate case notes for the medical retrieval task into other non-English languages other than French.

References

1. Armitage, L.H. and Enser, P.: Analysis of User Need in Image Archives. In Journal of Information Science **Vol. 23(4)** (1997) 287-299
2. Chen, F., Gargi, U., Niles, L. and Schütze, H.: Multi-Modal Browsing of Images in Web Documents. In Proceedings of SPIE Doc. Recognition and Retrieval VI (1999) 122-133
3. Chen, K., Chen, H., Kando, N., Kuriyama, K., Lee, S. and Myaeng, S.: Overview of CLIR Task, Third NTCIR Workshop, Japan (2002)
4. Cleverdon, C.W.: The Cranfield Tests on Index Language Devices. In: K. Spark-Jones and P. Willett (eds), Readings in Information Retrieval, Morgan Kaufmann (1997) 47-59
5. Clough, P. and Sanderson, M.: The CLEF 2003 Cross Language Image Retrieval Track. In Proceedings of the Cross Language Evaluation Forum (CLEF) Workshop, Norway (2003)
6. Cox, I.J., Miller, M.L., Omohundro, M. and Yianilos, P.N.: Target Testing and the PicHunter Bayesian Multimedia Retrieval System. In Proceedings of Advanced Digital Libraries (ADL'96) Forum, Washington D.C. (1996)
7. Dunlop, M.: Reflections on MIRA: Interactive Evaluation in Information Retrieval. In Journal of the American Society for Information Science Vol. 51(14) (2000) 126-1274
8. Flank, S.: Cross-Language Multimedia Information Retrieval. In Proceedings of Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (2000)
9. Forsyth, D.A.: Benchmarks for Storage and Retrieval in Multimedia Databases. In Proceedings of SPIE International Society for Optical Engineering **Vol. 4676** (2001) 240-247
10. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. In Informing Science **Vol. 3(2)** (2000) 63-66
11. Grefenstette, G.: Cross-Language Information Retrieval. Kluwer Academic Publishers, Norwell, MA, USA. (1998).
12. Harmandas, V., Sanderson, M. and Dunlop, M.D.: Image Retrieval by Hypertext Links. In Proceedings of the 20th ACM SIGIR conference (1997) 296-303
13. Markkula, M. and Sormunen, E.: Searching for Photos – Journalist's Practices in Pictorial IR. In Proceedings of Conference on Image Retrieval (CIR'98) (1998)
14. McDonald, S., Lai, T.S., Tait, J.: Evaluating a Content Based Image Retrieval System. In Proceedings of SIGIR'01 (2001) 232-240
15. Müller, H., Müller, W., McG. Squire, D., Marchand-Mailet, S. and Pun, T.: Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. In Pattern Recognition Letters **Vol. 22(5)** (2001) 593-601

16. Peters, C. and Braschler, M.: Cross-Language System Evaluation: The CLEF Campaigns. In Journal of the American Soc. for Inf. Sci. and Tech. **Vol. 52(12)** (2001) 1067-1072
17. Reid, N.: The Photographic Collections in St Andrews University Library. In Scottish Archives **Vol. 5** (1999) 83-90
18. Smeulders, A.W.M, Worring, M. Santini, S. Gupta, A. and Jain, R.: Content-Based Image Retrieval at the End of the Early Years. In IEEE Transactions on Pattern Analysis and Machine Intelligence **Vol. 22(12)** (2000) 1349-1380
19. Smith, J.R.: Image Retrieval Evaluation. In Proceedings of the IEEE Workshop of Content-Based Access to Image and Video Databases (2001) 112-113
20. Voorhees, E.M. and Harman, D.: Overview of TREC 2001, In NIST Special Publication 500-250: Proceedings of TREC2001, NIST. (2001)