User Experiments with the Eurovision Cross-Language Image Retrieval System

Paul Clough and Mark Sanderson

University of Sheffield

Sheffield, UK.

Abstract

In this paper we present Eurovision, a text-based system for cross-language (CL) image retrieval. The system is evaluated by multilingual users for two search tasks with the system configured in English and five other languages. To our knowledge this is the first published set of user experiments for CL image retrieval. We show that: (1) it is possible to create a usable multilingual search engine using little knowledge of any language other than English, (2) categorizing images assists the user's search, and (3) there are differences in the way users search between the proposed search tasks. Based on the two search tasks and user feedback, we describe important aspects of any CL image retrieval system.

1. User Experiments with the Eurovision Cross-Language Image Retrieval System

A great deal of research is currently being conducted in the field of Cross Language Information Retrieval (CLIR), where documents written in one language (referred to as the *target* language), are retrieved by a query written in another language (referred to as the *source* language). Most work up to now has concentrated on locating or creating translation resources and methods that automatically transform a user's query into the language of the documents. Additional methods to match queries and documents include translating the document collection into the source language and converting both the queries and documents into a common language (Oard, 1997). Query translation is the dominating approach because can be made to work successfully with simple translation methods, e.g. dictionary-lookup (Ballesteros & Croft, 1998), and does not require the overhead of translating collection documents which is often computationally expensive (e.g. Oard (1998) spent ten machine-months translating the TREC SDA/NZZ German collection – 251,840 newswire articles).

With the right approach, CLIR systems are able to achieve retrieval effectiveness that is only marginally degraded from the effectiveness achieved had the query been manually translated (referred to as monolingual retrieval). For example, Ballesteros & Croft (1998) achieved CLIR effectiveness at 90% of monolingual retrieval by using query expansion before and after query translation. Other CLIR work has concentrated on means of presenting the retrieved documents in some surrogate form such that users can judge relevance without having access to a full translation (Resnik, 1997; Gonzalo & Oard, 2003).

One area of CLIR research that has received less attention is retrieval from collections where text is only used to describe the collection objects, and the object's relevance to a query are hopefully clear to anyone regardless of their foreign language skills. One such collection is a picture archive where each image is accompanied by some kind of text, e.g. metadata or captions, semantically related to the image. 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 historic or stock-photographic archives, medical case notes and art and history collections. Here CLIR offers the opportunity of broadly expanding the range of potential searchers to an image archive through multilingual access. As Oard (1997) comments: “an image search engine based on cross-language free text retrieval would be an excellent early application for cross-language retrieval.”

Retrieval from an image collection offers distinct characteristics from one in which the document to be retrieved is natural language text (Armitage & Enser, 1997; Goodrum, 2000). For example, the way in which a query is formulated, methods used for retrieval (e.g. based on low-level features derived from an image, or based on associated texts), 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[1] (e.g. colour, shape, spatial layout and texture), or by text/metadata associated with the image (see, e.g. (Smeulders et al., 2000) and (Goodrum, 2000)).

Retrieval from such an archive presents a number of challenges and opportunities. The challenges come from matching queries to the relatively short descriptions associated with each image. Opportunities come from the unusual situation for CLIR systems of users being able to easily judge images for relevance. For those organisations managing image repositories in which text is associated with images (e.g. on-line art galleries), one way to increase user numbers is by enabling multilingual access to them. This paper is divided as follows: in section 2 we present previous work, section 3 the Eurovision system, section 4 our experimental methodology, section

5 our results and discussion, and section 6 our conclusions and avenues for future work in this area.

2. Previous Work

To date there appears to be little work in cross language image retrieval, so in this section we first review the two main component research areas separately: image retrieval by associated text and cross language IR; followed by a review of the proposals and occasional attempts at performing image CLIR.

2. 1 Image retrieval via associated text

Retrieval of images by text queries matched against associated text has been long researched, and approaches tend to reflect the nature of the application being addressed and the structure of the image collection. For example, collections such as the Web are networks of hypertext links which can be exploited to improve retrieval from hypermedia collections. As part of his work on multimedia retrieval, Dunlop (1993) examined the use of hyperlinked networks for image retrieval, using links to calculate representations for non-textual nodes. The ideas in this work were later extended to a study of image retrieval from art gallery Web sites by Harmandas et al. (1997), who showed that associated text was well-suited for retrieval over a range of query types.

More recently, Chen et al. (1999) presented a method of multi-modal caption retrieval involving both content-based and textual-based modalities, enabling images in Web collections to be browsed. Images are clustered according to similarity based on textual features extracted from the URL, from the image ALT tag, and the image link, together with a simple colour histogram content-based approach.

Other approaches have focused on extracting grammatical relations from the captions in order to describe image content. Although in the past it has been shown that Natural Language Processing (NLP) is detrimental to retrieval from documents (Smeaton, 1997), there are some

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

applications which involve short texts, e.g. image captions, where the situation becomes different and NLP can help improve retrieval. For example, Flank (1998) shows how searching on heads and head-modifier combinations obtains high precision and recall for image retrieval using captions. Elworthy et al. (2001) discuss an information retrieval system for searching a photographic database called ANVIL. The system deals with queries such as “camera with a lens” by creating a dependency structure where words and their relationships are captured. The grammatical relations of both query and caption are compared and a similarity score assigned based on matching dependency structures. The use of grammatical relations enable the query “yellow car” to match against variations of the request such as “car which is yellow” with high similarity score, but result in a low score for the caption “car which is not yellow” which would have otherwise been given a high score in a simple IR approach.

Given the typically small caption lengths, attention has also been given to expanding the query using lexical resources to reduce the effects of mismatch between query and caption words. Smeaton and Quigley (1996) show that by pre-computing word-word similarities based on a WordNet-derived knowledge base, the effectiveness of retrieval can be improved for image retrieval where the captions are typically of length one sentence. Other systems involving natural language processing for image retrieval include the MARIE system (Guglielmo,1996) in which captions and queries are translated into a logical form based on the meaning of nominal compounds (sequences of consecutive nouns). MARIE has been used on a naval warfare image collection where photographs include pictures of weapons, aircraft, aerial views and test equipment. The captions describe events in the images, or unique characteristics and features of weapon systems. Queries to this collection reflect the captions by either describing an action, e.g. “aircraft hitting a test drone”, or specific object(s), e.g. “Sidewinder missile”.

Flank et al. (1993) describe the SEYMOUR system which provides WWW access to more than 300,000 images via textual captions using NLP analysis on both the query and caption. Flank (2000) extends this work to deal with cross-language requests using both a dictionary lookup and online machine translation approach to translation. Finally, Srihari (1995) and Houghton (1999) describe the interaction of textual and photographic information in the Piction and NamedFaces systems respectively. Both aim to extract information from image captions, in particular extract proper names and grammatical relations such that people identified using image analysis can be labeled in the accompanying photographs, and photographs can be retrieved based on proper names assigned to the images. The approach taken in the NamedFaces system is simpler in that photographs of people are identified from Web pages, and hypertext links and image captions used to extract proper names for labeling.

2.2 Cross-Language Information Retrieval (CLIR)

A considerable body of research has grown up around CLIR, and most research has concentrated on locating and exploiting translation resources. CLIR is basically a combination of machine translation and traditional monolingual IR and four approaches commonly used for translation include (Gollins, 2001): (1) a controlled vocabulary, (2) machine translation, (3) bilingual parallel corpora, and (4) bilingual dictionaries. Schäuble and Sheridan (1997) suggest that to cross the language boundary between source and target language, one can translate the query in the source language into the target language, translate each document in the collection into the same language as the query, or translate both queries and documents into an intermediate representation, i.e. using a controlled vocabulary. With reasonably accurate translation, effective cross-language retrieval is possible and this has been confirmed in large-scale evaluation forums such as TREC[2] and CLEF[3].

The effectiveness of retrieval is based on translation and some of the problems arise from (Flank, 2000): (1) the bilingual dictionary may not contain specialised vocabulary or proper names; (2) dictionary terms are ambiguous and can add extraneous terms to the query and (3) the effective translation of multi-word concepts such as phrases. Other problems include lexical ambiguity of words in both the target and source languages (e.g. polysemy) which becomes worse given queries of a few words, special terms, and cross-lingual spelling variants. Further information can be

[2] http://trec.nist.gov/ (site visited: 09/08/2004).
[3] http://www.clef-campaign.org/ (site visited: 09/08/2004).

found in (Grefenstette, 1998; Ballesteros & Croft, 1997).

2.3 Cross-Language Image Retrieval

Both Oard (1997) and Jones (2001) have discussed cross-language image retrieval, but neither reported any work in this area. However, at least five examples of cross-language image retrieval exist:

1. The IR Game system built at Tampere University (Sormunen, 1998) offers Finish/English cross-language image retrieval from an image archive. Images are ranked using a best match search, but no form of query or image caption expansion is used and little has been written about the system.

2. The European Visual Archive[4] (EVA) offers English/Dutch/German cross language searching of 17,000 historical photographs indexed by a standard set of 6,000 controlled vocabulary terms and searching is restricted to Boolean searching.

3. Flank (2000) presents a method for accessing 400,000 photographic images from a commercial image company called PictureQuest[5]. Associated with the images are English captions and cross-language image retrieval in Spanish, German and French is provided via a dictionary-lookup approach and the use of different on-line machine translation tools. Flank claims retrieval effectiveness ranging from 68% of human translator performance, to 100% for French for ten example queries.

4. In 2002, the ImageCLEF track of CLEF (Cross Language Evaluation Forum) was established with the release of one of the first publicly available test collections for cross

[4] http://www.eva-eu.org/ (site visited: 09/08/2004).
[5] http://www.picturequest.com/ (site visited: 09/08/2004).

language image retrieval: approximately 30,000 photographic images from a collection held at St. Andrews University Library in Scotland and fifty queries (Clough & Sanderson, 2003). The collection was used by four research groups working across a number of European languages and Chinese. Results confirmed Flank's conclusion that image CLIR can be made to work.

5. In preliminary experiments, we confirmed the feasibility of an image CLIR system using a test collection study for German and Portuguese (Sanderson et al., 2004). Together, the examinations showed that searching for images from a historic collection of Scottish photography where images are captioned in a language unknown to the searcher is feasible.

From these past works one might conclude that image CLIR is feasible; however, the conclusion would be based on test collection evaluation alone and limited usability testing.

3. The Eurovision System

The Eurovision system combines existing translation resources to provide Web-access to an image archive provided by St. Andrews University Library. Most image captions contain a number of textual fields which can be exploited during image retrieval. Although our retrieval system is built to search the St. Andrews collection, the information contained in captions for this collection can be found in most annotated pictures, e.g. a description of the image, its author and some kind of categorization. The following sections describe the architecture/interface, and translation.

3.1 The St. Andrews Image Collection

A collection of historic photographs from St. Andrews University Library (Reid, 1999) was used as the dataset for system development and evaluation. This dataset was used because: (1) it represents a real-world collection for which multilingual access can enhance its access, (2) it contains almost 30,000 images with captions of high quality generated by historians, (3) it is being used in a comparative evaluation competition called ImageCLEF (Clough & Sanderson, 2003), and (4) the varied quality and content of the collection makes CLIR access a challenge.

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

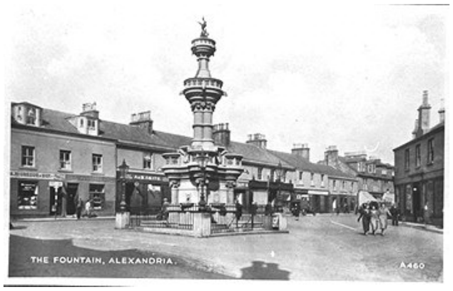


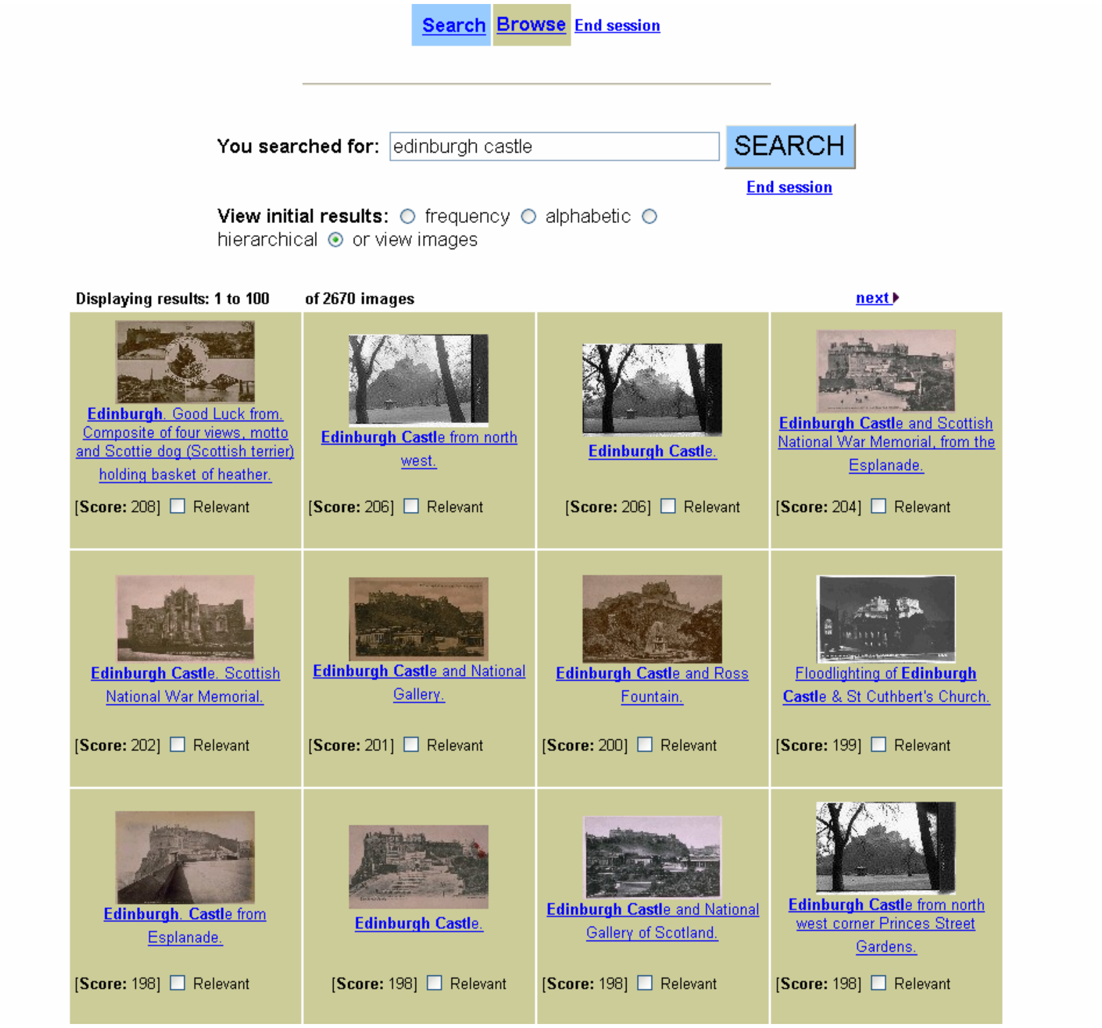
Figure 1 An example image and caption from the St. Andrews collection

All images in the St. Andrews collection are accompanied by a caption consisting of eight distinct fields which can be used individually or collectively to facilitate image retrieval (see Fig. 1). The captions consist of 44,085 terms and 1,348,474 word occurrences; the average caption length is 48 words, with a maximum of 316. All captions are written in British English; they 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 and the majority of images (82%) are black and white.

Like many image collections, pictures in the St. Andrews collection have been annotated manually by domain experts (historians). Part of the process is to assign images to one or more pre-defined categories for image storage and management. There are 971 categories in the St. Andrews collection, some more general (e.g. “flowers”, “landscapes”) than others (e.g. names of geographic regions and photographers). Most images are assigned to 3-4 categories.

3.2 Architecture and Interface

The interface is Web-based and generated dynamically using Perl/CGI scripts and JavaScript (see Figure 2). A simple modular architecture enables new functionality to be added with relative ease, e.g. changing the translation resource. Users log into the system and can begin by entering search requests in English, French, German, Italian, Spanish, Simplified Chinese or Japanese. Queries are passed our own in-house probabilistic retrieval system, based on the "best match" BM25 weighting operator (Robertson et al., 1998). Images are indexed by a number of caption fields, e.g. title, description and set of manually assigned categories (see Figure 1). The default settings of case normalization, removal of stopwords and word stemming are used and a document ranking scheme used where captions containing all query terms are ranked highest by their BM25 score, and then all other captions containing at least one query term ranked below.



You searched for: edinburgh castle SEARCH

End session

View initial results: ○ frequency ○ alphabetic ○ hierarchical ● or view images

Displaying results: 1 to 100 of 2670 images next

Edinburgh. Good Luck from. Composite of four views, motto and Scottie dog (Scottish terrier) holding basket of heather. [Score: 208] □ Relevant	Edinburgh Castle from north west. [Score: 206] □ Relevant	Edinburgh Castle. [Score: 206] □ Relevant	Edinburgh Castle and Scottish National War Memorial, from the Esplanade. [Score: 204] □ Relevant
Edinburgh Castle. Scottish National War Memorial. [Score: 202] □ Relevant	Edinburgh Castle and National Gallery. [Score: 201] □ Relevant	Edinburgh Castle and Ross Fountain. [Score: 200] □ Relevant	Floodlighting of Edinburgh Castle & St Cuthbert's Church. [Score: 199] □ Relevant
Edinburgh. Castle from Esplanade. [Score: 198] □ Relevant	Edinburgh Castle. [Score: 198] □ Relevant	Edinburgh Castle and National Gallery of Scotland. [Score: 198] □ Relevant	Edinburgh Castle from north west corner Princes Street Gardens. [Score: 198] □ Relevant

Figure 2 Part of the main Eurovision search screen

Results are presented as a 5 x 4 grid of thumbnails with titles from the captions. Users can browse the results set a page at a time, or re-iterate their query. Clicking on a thumbnail image shows the caption and a larger version of the image. The pre-assigned categories are used to help users navigate the results set and find similar relevant images.

An alternative to the grid display of images is to view returned images by their categories (see Figure 3). This provides a summary of the results indicating their contents as described by image categories and can be a more efficient method of searching than viewing images one page at a time. Three methods for viewing these categories are offered to the user: ranked in ascending order by the number of images assigned to each category, alphabetically, and hierarchically. Categories are organised automatically into a hierarchical structure using a co-occurrence relation called *subsumption*, proposed by Sanderson and Croft (1998) for IR. Up to four levels are generated with more frequent/dominant categories ordered at the top of the hierarchy leaving more specific categories nearer the bottom (e.g. "horses and ponies" > "farm implements" > "farming – ploughing"). The user can browse the search results by either navigating through pages of image thumbnails, or viewing the lists of categories.

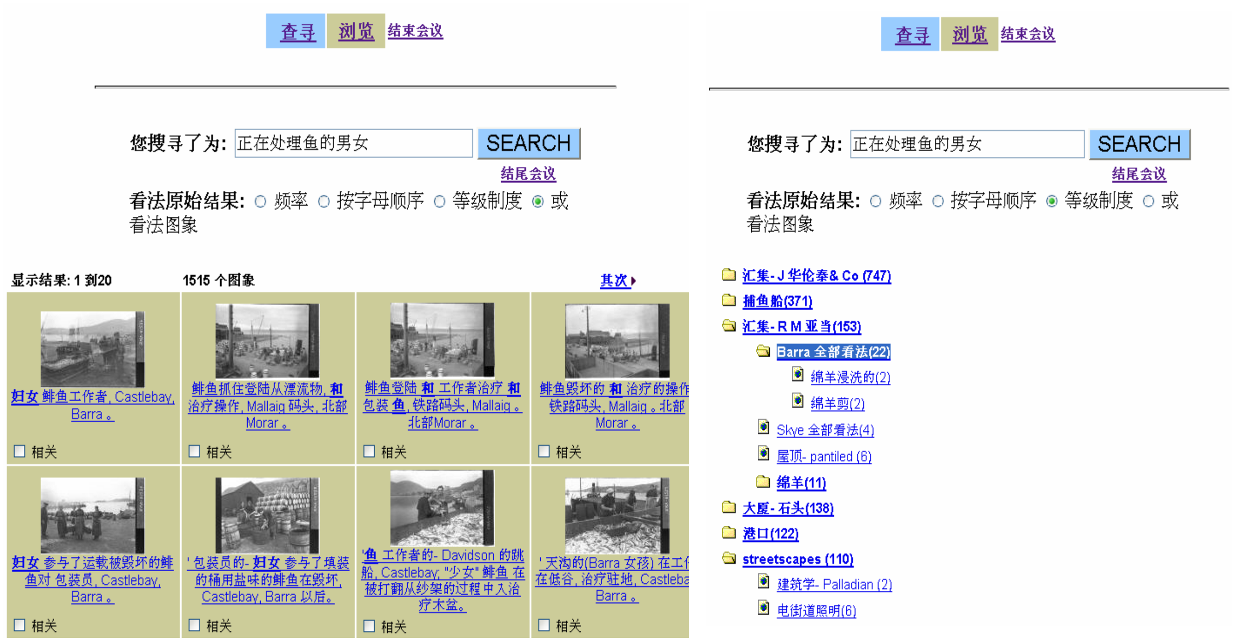
Similar views of the image categories can also be obtained when browsing, this time generated from the entire collection rather than the results of a search. By listing the categories, the user is able to obtain an overview of the contents of the collection. For example, displaying the categories by the frequency of images assigned to that category shows the most dominant.

3.3 Translation using SYSTRAN

Query, interface and document translation is provided by the free on-line version of SYSTRAN[6], one of the oldest and most widely used Machine Translation (MT) systems (Hutchins & Somers, 1986; Systran, 2002). Cross-language queries are translated into English and passed to the retrieval system in which the English captions have been indexed. Results are displayed as English Web pages and translated dynamically as users interact with the system by calls to SYSTRAN. This method translates the flat and hierarchical category lists, the image captions, and the whole

[6] http://www.systransoft.com/ (site visited: 09/08/2004). We used this version to assess what could be done using freely available on-line translation resources requiring minimal cross-language tools and knowledge.

interface in the user's source language. Figure 3 shows an example multilingual version of the Eurovision interface (in Chinese) where non-translatable terms are maintained by SYSTRAN. Most un-translated terms are proper names which are either not found in the SYSTRAN dictionaries, or would not commonly be translated from English even manually. The quality of SYSTRAN varies across language due to a range of translation errors (see, e.g. Qu et al. (2000)). For short queries of 2-3 words, SYSTRAN is essentially used for dictionary-lookup as they carry little grammatical structure to help the MT algorithm.



您搜寻了为: 正在处理鱼的男女

看法原始结果: ○ 频率 ○ 按字母顺序 ○ 等级制度 ● 或 看法图象

显示结果: 1 到20 1515 个图象

妇女 鲱鱼工作者, Castlebay, Barra 。 □ 相关

鲱鱼抓住登陆从漂流物, 和 治疗操作, Mallaig 码头, 北部 Morar 。 □ 相关

鲱鱼登陆 和 工作者治疗 和 包装 鱼, 铁路码头, Mallaig 。北部Morar 。 □ 相关

鲱鱼毁坏的 和 治疗的操作 铁路码头, Mallaig 。北部 Morar 。 □ 相关

妇女 参与了运载被毁坏的鲱鱼对 包装员, Castlebay, Barra 。 □ 相关

' 包装员的- 妇女 参与了填装的桶用盐味的鲱鱼在毁坏, Castlebay, Barra 以后。 □ 相关

'鱼 工作者的- Davidson 的跳船, Castlebay; "少女" 鲱鱼 在被打翻从纱架的过程中入治疗木盆。 □ 相关

' 天洵的(Barra 女孩) 在工作在低谷, 治疗驻地, Castleba Barra 。 □ 相关

您搜寻了为: 正在处理鱼的男女

看法原始结果: ○ 频率 ○ 按字母顺序 ● 等级制度 ○ 或 看法图象

- 汇集- J 华伦泰& Co (747)
- 捕鱼船(371)
- 汇集- R M 亚当(153)
  - Barra 全部看法(22)
    - 绵羊浸洗的(2)
    - 绵羊剪(2)
  - Skye 全部看法(4)
  - 屋顶- pantiled (6)
  - 绵羊(11)
- 大厦- 石头(138)
- 港口(122)
- streetscapes (110)
  - 建筑学- Palladian (2)
  - 电街道照明(6)

Figure 3 Example interface in Chinese

4. User Experiments

Cox et al. (1996) 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 the user to perform and has clear measures of effectiveness. Both a known-item and category search is used for the evaluation of Eurovision.

4.1 Participants

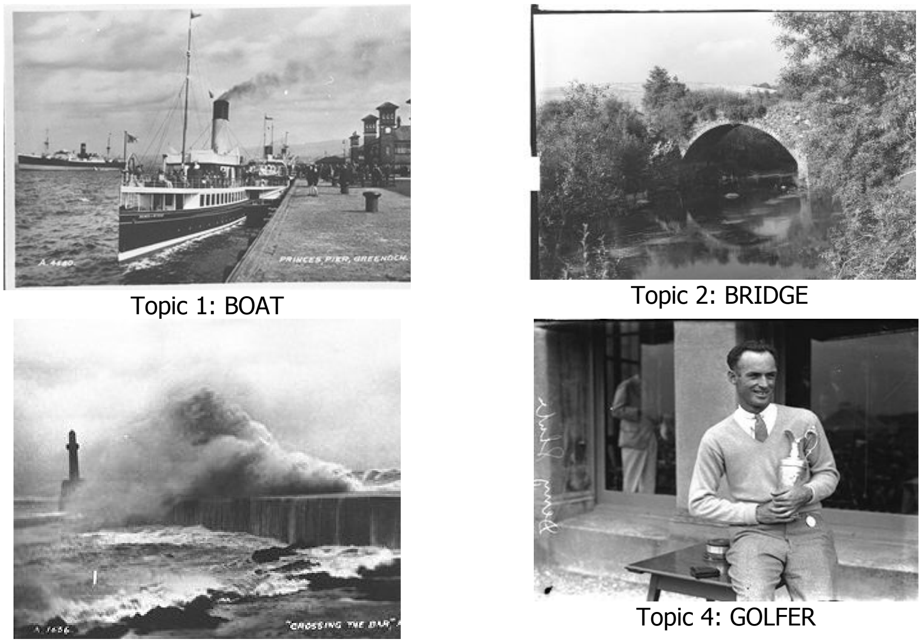
As the collection being searched was captioned in English, the queries were written in (and searchers had to be fluent in) a different language. With such a collection, it would be preferable for users not to know English. Locating such people within the UK, however, proved to be too hard a task. However, it was possible to locate a large number of bilingual people instead. Eight undergraduate/postgraduate students currently studying at the University of Sheffield in the Computer Science and Information Studies Departments were recruited.

The pre-test questionnaire established that 63% of participants searched in a language other than their native language daily, 75% searched for images occasionally and most agreed they were strong searchers. By their own admission 63% regarded their command of English as fluent. The native language of participants reflects a variety of cross-language users; although with a dominance towards Chinese.

4.2 Methodology

The user experiments were undertaken in the following manner. At arrival time participants filled in a pre-test questionnaire to establish their search and language abilities. Participants were then briefed on the goals of the experiments, and shown how to use the Eurovision system. Then followed 10 minutes when participants were able to perform their own searches and become familiar with the Eurovision system. Two experiments or tasks were performed by the users: a known-item and category search, comprising of 4 topics each. The condition varied in each task was the search language: either English or their native language. Users performed half the topics in English and half in their native language, the system providing the same functionality in each case. The presentation order of topic and search language was rotated according to a Latin-Square arrangement as used in iCLEF (Gonzalo & Oard, 2002). This reduces bias from users performing the same tasks with the same system in the same order (counter-balancing). Users were allowed a maximum of 10 minutes to complete each topic with no restriction on their search methods.

After completing the first task, users were asked to complete a post-experiment questionnaire regarding their search experiences and given a 10 minute break before completing the next task. During each experiment, users were able to ask the investigators any questions regarding the system. After completing the second experiment, users were asked to complete a final questionnaire to capture user feedback and comments about the Eurovision system.



Topic 1: BOAT

Topic 2: BRIDGE

Topic 3: STORM

Topic 4: GOLFER

Figure 4 Images used in the known-item search task

4.3 Experiment 1: Known-Item Search

The first experiment was a series of known-item searches in which users were shown 4 images from the collection and asked to find them again. Unlike being given a textual description of the task, 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.

The 4 images selected for this task are shown in Figure 4. These images were selected to give a range of difficulty of locating a known item. For example, the bridge image is expected to be relatively easy to find, whereas the boat and golfer are harder without knowing, for example, the name of the ship or man. Measures of success for this task are: the number of images successfully found, the time taken to find each image, and total number of images viewed. However, because the cross-language version of the interface is slower than in English (approximately 3 seconds for an English search and 7 seconds for a cross-language one), the time taken to find a relevant image and total images viewed do not accurately measure success for this search task. Cross-language retrieval is slower because of the costs involved in translating the query, captions and interface.

4.4 Experiment 2: Category Search

The second experiment was a category search (i.e. a TREC ad hoc search task). Users are provided with a statement of four different information needs in their native language describing a search topic. They must then find as many images as possible they consider relevant to that topic. This differs from the first experiment because users are given a textual statement of the information need rather than having to interpret the image to generate a textual search request. The four topics given to the users were taken from the 2003 ImageCLEF ad hoc task (Clough & Sanderson, 2003) and included: (1) pictures of the beach in Great Yarmouth, (2) pictures of damage due to war, (3) fishermen by the photographer Adamson, and (4) pictures showing a coat of arms.

The four topics in the second experiment, again, represent search tasks of varying complexity, designed to present various challenges to a translation system, e.g. the use of proper names in topics 1 and 3. Evaluation in this task is based on the number of relevant images found and calculated as a proportion of the overall number of relevant images found (a recall measure). This was computed by pooling together all relevant images identified by the users to give 13 relevant for topic 1, 18 for topic 2, 8 for topic 3 and 44 for topic 4. These are similar to images defined in the ImageCLEF relevance assessments for the relaxed union set (13, 17, 7 and 46).

5. Results and Discussion

5.1 Overall Performance

Table 1 shows the results broken down by experiment and task for cross-language (CL) and monolingual (English) retrieval. The measure of success for the known-item task is the proportion of times the known-item was found (e.g. 0.50 indicates that on average the desired image was found 50% of the time). In the category task, the score indicates the proportion of relevant images found from the document pool. On average across both experiments, Eurovision performs at 86% of monolingual performance (differences are not statistically significant using paired t-test, $p<0.01$). The results indicate that although the absolute performance of the CL system is quite poor (i.e. 50% of relevant images found for the known-item search and 35% for the category search), the results for English retrieval are also poor. The relative measure of performance as a proportion of monolingual performance is therefore much higher.

Table 1. A summary of overall performance for tasks 1 and 2

Known item		**1**	**2**	**3**	**4**	**Avg**
	Mono	0.25	0.50	0.75	0.75	*0.56*
	CL	0.50	0.25	0.75	0.50	*0.50*
	%Mono	200	50	100	67	*89*
Category		**1**	**2**	**3**	**4**	
	Mono	0.58	0.56	0.25	0.29	*0.42*
	CL	0.34	0.43	0.38	0.24	*0.35*
	%Mono	59	77	152	83	*83*

The most noticeable result is the overall relative performance of the CL system compared to the monolingual version. Remembering that the entire system has been created with little or no knowledge of any language other than English, a result of 86% of monolingual on average is pleasing, especially 89% of monolingual for known-item searching. Although we previously found SYSTRAN to be poor for query translation (Clough & Sanderson, 2003), we find that users are still able to use the multilingual interface to search for and judge the relevance of images. In general users said they preferred known-item retrieval because it had a clearer goal than the category search because users had no idea how many relevant images might be in the collection.

5.2 Known-Item Search Results

Known-item performs highest, on average 89% of monolingual; although performance varies dramatically across topics (differences are not statistically significant). The known-items found least successfully were topics 1 and 2; the most successful topic 3. The challenge of the known-item search is generating suitable query term from visually interpreting the image and overcoming vocabulary mismatch between the terms selected by the image annotator (which are British English) and those used by the searcher. For example, in topic 1 the word "ship" is used in a query many times. However, although this term matches the caption (41 distinct terms), it returns 324 images with the relevant image not appearing until page 12 which is far beyond the limit of most users. Terms such as "ferry" (used only by one user) are more useful as the known-item is found on page 7 (this term is used in cross-language searches which explain the higher CL score).

Vocabulary mismatch occurs for both the monolingual and CL searches, but there are specific cross-language problems which result from errors in the translation resource and cultural differences. For example, in topic 1 the English equivalent of "harbour" is translated as "harbor" in cross-language searches because SYSTRAN makes use of bilingual dictionaries which use American spelling variants. We also observe that spelling errors are also present in some queries (e.g. "brigde" rather than "bridge" in topic 2) which causes vocabulary mismatch. Topic 4 also proves harder for cross-language searchers because important keywords like "trophy" are not used by the searcher, or not translated correctly. For example in Chinese, users typically searched with the word "cup" which would not find the correct image.

Further difficulties of this task are unfamiliarity with the St. Andrews collection by users, and the frequent use of colloquial language used in the collection which introduces vocabulary mismatch. This is worsened by query translation where errors or differences in the colloquial language of the translation resource cause further query-caption mismatches.

5.3 Category Search Results

The absolute and relative results for the category search are, on average, lower than for known-item (differences are not statistically significant). The least proportion of relevant images is found in topic 4 "coat of arms"; the highest in topic 2 "pictures of damage due to war". The lowest CL score as a proportion of monolingual is topic 1 “beach at Great Yarmouth”. This can be explained due to, in general, the incorrect translation of the proper noun “Great Yarmouth” for almost all languages. When asked by users whether they were satisfied with the number of relevant images found in allotted time for task 2, 25% fully agreed with, 25% partially agreed, 25% were neutral and 25% partially disagreed. This indicates that although the proportion of relevant found in task 2 was generally low (for both the CL and English systems), this does not necessarily reflect that the user is unsatisfied.

Search Characteristics

Tables 2 and 3 summarise user interaction as captured in the log files. Results are summarised across system, task and user and represent the average number of user operations, e.g. the number of times they select a category, during their search. Query length is the number of English terms used for a single search, images viewed the number of images which are enlarged to show the caption text, images displayed represents the number of images the user views during their entire search, category link click represents the number of times the user selects a category when viewing an image and category click represents the number of times a user selects a category from listing these either using the concept hierarchy (CS), alphabetically or by frequency. From the results, we make a number of observations and suggest implications for CLIR.

Firstly, we observe that users select to view a large version of the image more often in the category search task than known-item. This is often because users can identify a relevant known-item without need for the caption; in category searching the caption is may be required, e.g. in topic 3 users view the image to check the name of the photographer. The implication for CLIR is that for some search tasks, i.e. known-item, the user does not have to view the image caption thereby removing the need for time-consuming and erroneous caption translation.

Secondly, we observe that for both search tasks users appear willing to search many images to find the relevant ones by browsing through pages or using the categories. Assuming on average that 20 images per page are displayed this amounts to 16 pages in known-item and 8.5 in ad hoc searching. This is typically a much larger number than text retrieval and would allow relevant images to be found in lower rank positions.

Table 2. A summary of retrieval operations for the known-item search

		Queries	Query	Images	Next	Prev	Images	Category	Category	View categories by		
			length	viewed	page	page	displayed	click	link click	alpha	freq	CS
Language	Mono	3.2	2.4	3.0	10.3	0.3	325.3	3.2	1.2	0.6	0.6	1.2
	CL	4.8	2.0	2.2	8.8	0.3	318.8	4.0	2.6	0.4	0.6	1.8
Task	1	4.4	2.3	2.3	13.1	0.3	408.9	1.8	2.0	0.5	0.1	1.0
	2	3.8	2.0	4.3	11.6	0.6	396.3	3.3	1.9	0.8	1.3	1.0
	3	2.6	2.1	1.8	3.0	0.0	151.6	5.1	1.8	0.1	0.3	2.3
	4	5.3	2.2	2.1	10.3	0.1	331.3	4.3	1.9	0.5	0.6	1.8
User	1	4.0	2.1	4.5	11.8	0.0	428.5	5.3	5.3	0.3	1.3	3.8
	2	1.8	1.4	3.5	6.5	1.0	256.8	4.5	1.8	0.5	1.0	1.0
	3	4.8	1.9	1.0	9.0	0.5	249.5	0.5	0.0	1.3	0.0	0.5
	4	4.0	2.5	1.8	5.0	0.0	289.0	7.8	3.8	0.5	0.0	3.8
	5	4.0	1.0	5.8	14.3	0.3	404.5	4.3	3.3	0.3	1.8	0.0
	6	3.0	2.6	1.3	5.0	0.0	241.8	5.5	0.5	1.0	0.3	2.0
	7	4.0	2.9	2.3	10.5	0.3	300.3	0.0	0.5	0.0	0.0	0.0
	8	6.5	2.7	0.8	14.0	0.0	405.8	1.0	0.0	0.0	0.3	1.0
Average		**4.0**	**2.2**	**2.6**	**9.5**	**0.3**	**322**	**3.6**	**1.9**	**0.5**	**0.6**	**1.5**

Table 3. A summary of retrieval operations for the category search

		Queries	Query	Images	Next	Prev	Images	Category	Category	View categories by		
			length	viewed	page	page	displayed	click	link click	alpha	freq	CS
Language	Mono	2.6	1.8	11.9	5.9	0.3	201.1	1.4	0.7	0.6	0.6	0.9
	CL	4.0	1.8	10.6	2.4	0.1	141.8	3.9	1.3	0.7	0.4	1.1
Task	1	2.9	2.0	8.1	0.8	0.0	103.5	3.1	1.1	0.9	1.0	0.5
	2	2.0	1.8	7.8	5.0	0.0	169.8	2.1	0.4	0.8	0.1	1.3
	3	4.5	1.4	18.0	2.9	0.4	150.4	3.9	1.4	0.9	0.6	2.1
	4	3.6	1.9	11.1	8.0	0.3	262.1	1.5	1.0	0.1	0.1	0.1
User	1	5.0	2.0	11.0	3.5	0.0	168.0	5.0	1.3	0.0	0.0	3.0
	2	1.8	0.8	8.3	6.5	0.3	220.5	3.5	2.0	2.0	0.3	0.3
	3	2.0	2.3	10.8	6.8	0.8	195.0	0.3	0.0	0.3	0.0	0.5
	4	4.3	1.4	10.3	1.8	0.0	124.8	1.3	1.5	0.0	0.0	1.0
	5	3.3	1.3	12.3	8.3	0.0	235.8	4.5	0.5	1.0	0.3	0.5
	6	4.5	1.8	8.8	1.3	0.0	75.0	0.3	0.3	0.8	1.3	0.3
	7	2.8	2.1	9.0	3.5	0.3	159.0	2.0	1.0	0.5	0.8	1.5
	8	3.0	2.7	19.8	1.8	0.0	193.5	4.5	1.3	0.8	1.3	1.0
Average		**3.3**	**1.8**	**11.3**	**4.2**	**0.2**	**171.4**	**2.7**	**1.0**	**0.7**	**0.5**	**1.0**

These could then be used for relevance feedback which may help deal with poor query translation by expanding queries based on captions marked as relevant. For example, we show that selecting just one relevant image can dramatically improve retrieval performance for an ad hoc retrieval task (Clough & Sanderson, 2004).

Thirdly, on average users make more use of the categories for known-item than category searching. The results would seem to suggest that more use of categories is made in CL searching than in English. Query failure is the most likely cause where users resort to browsing categories as well as searching through pages of images. For CLIR this means that offering alternative methods of finding relevant images other than a text query is important, as well as providing accurate query translation. The use of categories varies depending on user preference as, for example, users 1 and 4 make more use of them than users 3 and 7.

Fourthly, when categories are selected the concept hierarchy is used the most frequently. Users typically do not find viewing a flat list of hundreds of categories particularly effective and prefer the hierarchical structure which limits the results. This might imply that providing a summary of the results set using hierarchically-organised categories is beneficial, particularly in CLIR where translating the interface slows down the speed at which users can browse. Again, some users prefer the hierarchy when viewing the categories, e.g. users 1 and 4.

Upon observing users and analysing the log files, we find users develop quite different approaches to searching. However, the most typical search strategy for either task is that users begin with a query, if the initial page of results is not useful, they perhaps view the categories, or they reformulate the query. We find that users tend to view many pages of results more often when their search term is more general (e.g. for task 1 if the user begins with the search “bridge” in topic1, or “boat” in topic 2) and they are motivated to keep browsing because they find images which are similar to those being searched for. Some users will query and then browse pages of results, others will query, view one or two pages and then reformulate.

An interesting search tactic that users we found to develop, especially in the known-item search, was to view similar images to the one being searched, identify some key terms from the caption and then reformulate. This approach ensures that terms in the collection are used during retrieval and provides users with suggestions on what terms to search on. For example users found pictures of men with similar clothes to the man in task 4, they found out that most men dressed like this were golfers, and then proceeded searching in this direction.

User Feedback and Comments

From the post-test questionnaire, users provided their final comments on the Eurovision system. Overall, 86% of users rated the system as either good or very good at finding images, with all users indicating positively that they would use the system again. Of all post-questions asked regarding the user's search experiences, the majority of users rated the system highly. When asked to rate the Eurovision system according to a number of adjectives, users made the following remarks. Most users found the translation provided by SYSTRAN to be adequate for the query, caption and interface, although 5 users commented that query translation was poor. The majority of users found having images categorized was helpful (38% very helpful, 50% partially helpful and 12% neutral) and most users found the hierarchies useful for both search tasks. Users commented that categories were particularly useful for cross-language searches where failure of their queries to return images perceived as relevant often left browsing by categories the only option left to them. User comments during and after the experiment about Eurovision included the following:

- Bilingual searching is preferable where users can search in English and their native language and create mixed language searchers.
- Users would like to view the translated query in English and be able to modify it if wrong prior to searching.
- The option to search using a Boolean AND rather than the BM25 similarity score. In the situation when many images are returned, users want to restrict the search to captions containing all search terms only.
- Being able to search each field of the caption would enhance searching, e.g. restricting the search to the name of the photographer.

- Many of the categories were specific and unknown to the users. More general categories would help their searching.
- Users would like to sort images by their contents, e.g. shape and colour, and by orientation and size.

Conclusions and Future Work

In this paper we have discussed an area of CLIR research which to date has received little attention, that of CL image retrieval. We have presented Eurovision, a system for searching image collections by matching user's queries to associated captions. A multilingual search environment is created for the user without any knowledge of a language other than English by using the SYSTRAN MT system to translate user queries, image captions and the interface. We use a collection of historic photographs as a representative dataset and exploit pre-defined categories to enable users to browse images and view the results of a search by category. We also implement a version of concept hierarchies to re-organise the categories into a hierarchical structure to reduce the number of categories users have to view.

For two search tasks: known-item and category searching, we find that although the absolute success of CL retrieval is poor (43% success), relative to searching in English the system Eurovision operates at 86% monolingual (89% for the known item search and 83% for the ad hoc search). We believe that this high performance figure is a result of the nature of image retrieval: that users do not have to view the image caption to judge relevance and they are willing to view many pages of images during retrieval. As it is likely that caption translation would exhibit the most translation errors, by users being able to judge relevance in most cases without viewing the caption the impact of poor translation only affects user's queries. We also believe that providing browsing through the categories also leads to this degree of success because users have an alternative search method to use when their queries appear to fail. Given retrieval success it would appear that translation of the categories and interface is good enough to enable users to browse with success.

As a CLIR task, image retrieval is one application where even poor translation resources can still achieve good performance. Our experiments have confirmed previous work that image retrieval via associated text is possible, although both English and CL searching could be improved greatly. The two search tasks exhibit different search characteristics and in particular users appear to perform more browsing and searching in the known-item search than the ad hoc search. This confirms Cox et al. (1996) view that the known-item search is a more useful task for image retrieval as it encourages users to perform more varied methods of searching than for an ad hoc task. There are a number of avenues which we plan to investigate to improve retrieval performance of the Eurovision system. These include the following:

- We are organising the interactive task within the framework of the ImageCLEF campaign. We are planning a larger user evaluation (involving 16-32 users) to compare different user interfaces for cross-language image retrieval. In 2004 we are testing methods to aid query reformulation and refinement[7].
- Given that users can judge relevance without viewing captions in most cases, we believe that the focus for CLIR researchers should be improving query translation, not captions associated with the images. We plan to investigate query translation using alternative methods such as bilingual dictionaries and parallel corpora.
- Given that vocabulary mismatch arises from translation errors and differences between the collection and user’s vocabulary, we plan to investigate the use of query expansion through external resources such as a thesaurus, and based on relevance feedback from the user. Given that users are able to generally judge the relevance of images with ease; this would be an obvious source of information to improve the search results. Further enhancements such as spelling detection and correction could also help reduce vocabulary mismatch.
- Categorizing images would appear to help the searcher. We would like to

[7] ImageCLEF 2004: http://ir.shef.ac.uk/imageclef2004/ (site visited: 09/08/2004).

experiment with alternative methods of generating categories for the images automatically (e.g. clustering captions and extracting dominant concepts).

- The concept hierarchy would appear a promising method of re-organising a results set. Because users are uncertain of categories in the St Andrews collection and general categories are not represented which inexperienced users may find useful, we are currently investigating generating the concept hierarchy based on the entire image caption rather than just the categories. Initial results appear promising.
- Our current system is primarily text-based. However, we plan to experiment with combining CBIR and text-based methods to improve searching and browsing. These methods could be combined to improve document ranking, and provide an alternative “more like this” function to the user who would be able to specify which aspect of the image they want to find similar images to.

Acknowledgments

We would like to thank members and students of the Department of Information Studies and Computer Science for their time and effort in evaluating the Eurovision system. Thanks also to Hideo Joho for help and support with the GLASS system and concept hierarchies. This work was carried out within the Eurovision project at Sheffield University, funded by the EPSRC (Eurovision: GR/R56778/01).

References

Armitage, L. H., & Enser, P. (1997). Analysis of User Need in Image Archives. *Journal of Information Science, 23*(4), 287-299.

Ballesteros, L., & Croft, B. W. (1998). Resolving Ambiguity for Cross-Language Retrieval. *Proceedings of the 21st International Conference on Research and Devlelopment in Information Retrieval,* 64-71.

Chen, F., Gargi, U., Niles, L., & Schütze, H. (1999). Multi-Modal Browsing of Images in Web Documents. *Proceedings of SPIE Document Recognition and Retrieval VI,* 122-133.

Clough, P. D., & Sanderson, M. (2003). The CLEF 2003 Cross Language Image Retrieval Track. *Working Notes for the CLEF 2003 Workshop.*

Clough, P. D., & Sanderson, M. (2003). Assessing Translation Quality for Cross Language Image Retrieval. *Working Notes for the CLEF 2003 Workshop.*

Clough, P. D., & Sanderson, M. (2004). The Effects of Relevance Feedback in Cross Language Image Retrieval. *Proceedings of the 26th European Conference on IR Research (ECIR'04),* 238-252.

Cox, I. J., Miller, M. L., Omohundro, M., & Yianilos, P. N. (1996). Target Testing and the PicHunter Bayesian Multimedia Retrieval System. *Proceedings of Advanced Digital Libraries Forum (ADL'96).*

Dunlop, M. D., & van Rijsbergen, C. J. (1993). Hypermedia and Free Text Retrieval. *Journal of Information Processing and Management, 29*(3), 287-298.

Elworthy, D., Rose, T., Clare, A., & Kotcheff, A. (2001). A Natural Language System for Retrieval of Captioned Images. *Journal of Natural Language Engineering, 7*(2), 117-142.

Flank, S. (1998). A Layered Approach to NLP-Based Information Retrieval. *Proceedings of 36th ACL and 17th COLING Conferences,* 397-403.

Flank, S. (2000). Cross-Language Multimedia Information Retrieval. *Proceedings of Applied Natural Language Processing and te North American Chapter of the Association for Computational Linguistics.*

Flank, S., Martin, P., Balogh, A., & Rothey, J. (1996). Photofile: A Digital Library for Image Retrieval. *Proceedings of IEEE International Conference on Multimedia and Computing Systems,* 292-295.

Gollins, T. (2000). Dictionary Based Transitive Cross-Language Information Retrieval using Lexical Triangulation. *Masters Dissertation.* Department of Information Studies, University of Sheffield.

Gonzalo, J., & Oard, D. (2003). The CLEF 2002 Interactive Track. In *Springer Lecture Notes in Computer Science (LNCS 2785): Vol. . CLEF 2002* (pp. 372-382). Berlin Heidelberg: Springer-Verlag.

Goodrum, A. A. (2000). Image Information Retrieval. *Informing Science, 3*(2), 63-66.

Grefenstette, G. (1998). *Cross-Language Information Retrieval.* Norwell, MA, USA: Kluwer Academic Publishers.

Guglielmo, E. J., & Rowe, E. J. (1996). Natural Language Retrieval of Images based on Descriptie Captions. *ACML Transactions on Information Systems, 14*(3), 237-267.

Harmandas, V., Sanderson, M., & Dunlop, M. D. (1997). Image Retrieval by Hypertext Links. *Proceedings of the 20th International Conference on Research and Development in Information Retrieval,* 296-303.

Houghton, R. (1999). Named Faces: Putting Names to Faces. *IEEE Intelligent Systems, 14*(5), 45-50.

Hutchins, W. J., & Somers, H. (1986). *An Introduction to Machine Translation.* London, England: Academic Press.

Jones, G. J. F. (2000). New Challenges for Cross-Language Information Retrieval: Multimedia Data and the User Experience. In *Working Notes for the CLEF 2000 Workshop* (pp. 71-81).

López-Ostenero, F., Gonzalo, J., Peñas, A., & Verdejo, F. (2003). Interactive Cross-Language Searching: Phrases are better than Terms for Query Reformulation and Refinement. In *Springer Lecture Notes in Computer Science (LNCS 2785)* (pp. 416-429). Berlin Heidelberg: Springer-Verlag.

Oard, D. (1997). Serving Users in Many Languages. *D-Lib.*

Oard, D. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas*, 472-483.

Peters, C., & Braschler, M. (2001). Cross-Language System Evaluation: The CLEF Campaigns. *Journal of the American Society for Information Science and Technology, 52*(12), 1067-1072.

Qu, Y., Eilerman, A. N., Jin, H., & Evans, D. (2000). The Effect of Pseudo Relevance Feedback on MT-Based CLIR. *Proceedings of RIAO 2000.*

Reid, N. (1999). The Photographic Collection in St. Andrews University Library. *Scottish Archives, 5,* 83-90.

Resnik, P. (1997). Evaluating Multilingual Gisting of Web Pages. *Proceedings of AAAI Symposium on Cross-Language Text and Speech Retrieval.*

Robertson, S., Walker, S., & Beaulieu, M. (1998). *Proceedings of TREC-7* (NIST Special Publiation 500-242, pp. 253-264).

Sanderson, M., & Croft, B. W. (1999). Deriving Concept Hierarchies from Text. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval,* 206-213.

Sanderson, M., Clough, P. D., Paterson, C., & Tung Lo, W. (2004). Measuring a Cross Language Image Retrieval System. *Proceedings of the 26th European Conference on IR Research (ECIR'04),* 353-363.

Schäuble, P., & Sheridan, P. (1997). *Proceedings of the 6th Text Retrieval Conference (TREC-6): NIST Special Publications 500-226. Cross Language Information Retrieval (CLIR) Track Overview.*

Smeaton, A. F. (1997). Information Retrieval: Still Butting Heads with Natural Language Processing? *International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology,* 115-138.

Smeaton, A. F., & Quigley, I. (1996). Experiments on Using Semantic Distances Between Words in Image Caption Retrieval. *Proceedings of the 19th International Conference on Research and Development in Information Retrieval,* 174-180.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380.

Sormunen, E., & Laaksonen, J. (1998). The IR Game - A Tool for Rapid Query Analysis in Cross-Language IR Experiments. *Proceedings of PRICAI'98 Workshop on Cross Language Issues in Artificial Intelligence,* 22-32.

Srihari, R. K. (1991). Piction: A System that uses Captions to Label Human Faces in Newspaper Photographs. *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91),* 80-85.

Systran Ltd. (2002). The Systran Linguistics Platform: A Software Solution to Manage Multilingual Corporate Knowledge. *Systran On-line Documentation* Retrieved July 29, 2004, from http://www.systransoft.com/Technology/SLP.pdf

Voorhees, E. M., & Harman, D. (2001). *Proceedings of the 10th Text Retrieval Conference (TREC2001): NIST Special Publication 500-250. Overview of TREC 2001.*

Appendix

[Insert Appendix Here]

Author Note

[Insert Author Note(s) Here]

Footnotes

[1] These are called Content-Based Information Retrieval (CBIR) systems.
[2] http://trec.nist.gov/ (site visited: 09/08/2004).
[3] http://www.clef-campaign.org/ (site visited: 09/08/2004).
[4] http://www.eva-eu.org/ (site visited: 09/08/2004).
[5] http://www.picturequest.com/ (site visited: 09/08/2004).
[6] http://www.systransoft.com/ (site visited: 09/08/2004). We used this version to assess what could be done using freely available on-line translation resources requiring minimal cross-language tools and knowledge.
[7] ImageCLEF 2004: http://ir.shef.ac.uk/imageclef2004/ (site visited: 09/08/2004).

Table 1

A summary of overall performance for tasks 1 and 2.

Table 2

A summary of retrieval operations for the known-item search.

Table 3

A summary of retrieval operations for the category search.

Figure Caption

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

Figure 2 Part of the main Eurovision search screen

Figure 3 Example interface in Chinese

Figure 4 Images used in the known-item search task.