The ImageCLEF benchmark on multimodal multilingual image retrieval

Henning Müller, Paul Clough

Introduction

Image retrieval has been an extremely active area of research in the fields of computer vision and pattern recognition for almost 20 years [1]. Many prototypes and techniques have been developed and explored, but still there is no general breakthrough in visual analysis and indexing techniques to bridge the semantic gap, although a few commercial companies such as LookThatUp technologies are very successful. Early systems used either purely visual features or textual metadata associated with images and involved little user interaction. However, modern systems increasingly use multimodal features (a combination of images, text, speech and structured data) and extensive user interaction to help improve the success of accessing visual information. Success is also based on analyzing user's information needs and searching behavior to dictate the design and functionality provided by multimedia retrieval systems (e.g. studies have shown that certain classes of users prefer to access images using text queries than visual features because they enable semantically-orientated searches [2]). However, accessing images using non-visual features relies on there being such information available in the first place. Even if available, issues such as quality, quantity and consistency will affect its usefulness. Recently interest in techniques such as automatic (and semi-automatic) image annotation can be seen as one way of propagating semantic information between visually similar images that have little or no other metadata [3].

With large multimedia retrieval projects such as Quaero[1] and search engine giants such as Google and Yahoo! investing massively in the multimedia retrieval market (e.g. Yahoo! buying the image exchange platform FlickR[2]), it is clear that multimedia retrieval is more than just a research domain, it is also an important strategic market.

To really advance the multimedia retrieval field, it has been increasingly accepted that systematic evaluation is needed. The evaluation of most early systems was limited, with semi-realistic queries (or examples) based on privately-held databases of images. Early initiatives such as the Benchathlon[3] stimulated the discussion of benchmarking issues but without a concrete evaluation event in which systems could be compared. Many papers discussed running an evaluation event similar to TREC[4] (Text REtrieval Conference), an initiative run by the US National Institute of Standards and Technology (NIST[5]) for Information Retrieval [4]. TREC has an annual circle of events for different search tasks: data release, topic release, results submission, evaluation, and finally a workshop in which to discuss and interchange ideas. A highly successful task has been TRECVid[6] which provides an evaluation framework for video retrieval. This started as part of TREC in 2001, but has since turned into an independent entity with an increasing number of participants. Another less-known benchmark is ImageEval[7], a French initiative that has successfully run a first test benchmark and will have its first official evaluation in 2006. CLEF[8] (Cross Language Evaluation Forum) is also a spin-off from TREC that focuses on multilingual information retrieval, held independently since 2000. In 2003 ImageCLEF[9] [5] began as part of CLEF, focused on text-based image retrieval from historic photographs. Since 2004 the focus has shifted towards combining visual and multilingual textual features for multimodal multilingual retrieval of images from medical and more general photographic collections. ImageCLEF has continued to address the barriers between research interests and real-world needs by offering application-driven evaluation tasks.

ImageCLEF 2005

ImageCLEF 2005 offered four separate evaluations: retrieval from historic photographs, medical image retrieval, medical image annotation (or image classification) and interactive image retrieval. In addition a one-day workshop on visual information retrieval evaluation was held the day before the CLEF workshop. The proceedings of this workshop are available on the web[10].

A total of 36 participants registered for ImageCLEF 2005 and 24 research groups from 14 countries submitted results.

Retrieval from a collection of historic photographs

The collection of this retrieval task contained 28,133 historical images from St. Andrews University Library. All images have a structured annotation in British English. Example queries were from analyzing typical user needs and selected to

[1] http://www.quaero.org/

[2] http://www.flickr.com/

[3] http://www.benchathlon.net/

[4] http://www.trec.nist.gov/

[5] http://www.nist.gov/

[6] http://www-nlpir.nist.gov/projects/trecvid/

[7] http://www.imageval.org/

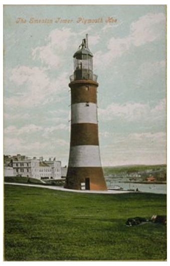
[8] http://www.clef-campaign.org/

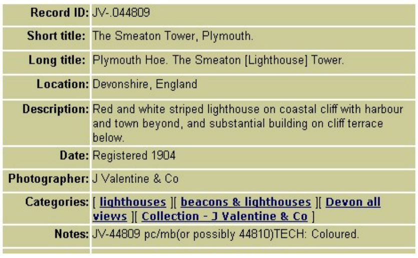
[9] http://ir.shef.ac.uk/imageclef/

[10] http://muscle.prip.tuwien.ac.at/ws_proceedings_2005.php

test different aspects of visual and textual search. 28 queries were given to participants consisting of a written statement (and translated into various languages) plus two example images.

Pictures of English lighthouses
Fotos de faros ingleses
Kuvia englantilaisista majakoista
Bilder von englischen Leuchttürmen
صور لمنارات انجليزيه
Изображения английских маяков
イングランドにある灯台の写真





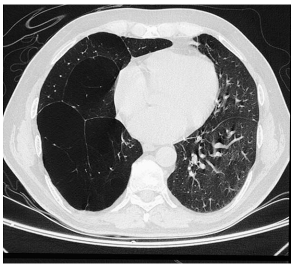
Record ID:	JV-.044809
Short title:	The Smeaton Tower, Plymouth.
Long title:	Plymouth Hoe. The Smeaton [Lighthouse] Tower.
Location:	Devonshire, England
Description:	Red and white striped lighthouse on coastal cliff with harbour and town beyond, and substantial building on cliff terrace below.
Date:	Registered 1904
Photographer:	J Valentine & Co
Categories:	[**lighthouses**][**beacons & lighthouses**][**Devon all views**][**Collection - J Valentine & Co**]
Notes:	JV-44809 pc/mb(or possibly 44810)TECH: Coloured.

Figure 1: An example of a query for the photographic retrieval task (left) and example image and caption (right).

Figure 1 shows an example query (left) plus a relevant image (middle) and its structured annotation (right). Challenges included: the use of British English and colloquial language for the annotations, short captions presenting problems of vocabulary mismatch between query-captions, and the majority of images being grey-scale and varying quality (making visual analysis hard). Best systems reached a Mean Average Precision (MAP) of 0.4135 for English-English (monolingual) retrieval and of 0.3993 for X-English (bilingual) retrieval.

Medical image retrieval

The medical retrieval task combined four datasets giving a total of 50,000 images with varying annotations partially in English, French and German. 25 written queries based on a survey among medical professionals were made available in the three collection languages, plus 1-3 query images (with one query also containing a negative feedback image).



Show me chest CT images with emphysema

Zeige mir Lungen CTs mit Emphysem

Montre moi des CTs pulmonaires avec un emphysème

Figure 2: An example query for the medical image retrieval task.

Figure 2 shows an example written query and example image. Challenges in this task included: varying quality and quantity of annotation of the images, domain-specific knowledge including unusual abbreviation and spelling errors. Best systems reached a MAP of 0.2821 for multimodal retrieval, 0.2084 for textual retrieval and 0.1455 for purely visual retrieval.

Medical image annotation

The automatic annotation task was a purely visual task. 9,000 images were given to participants as training data, each one labeled with one out of 57 classes. Participants then had to automatically assign class labels to 1,000 previously unseen images. The distribution of images among classes was very heterogeneous: the largest class containing 2,563 images; the smallest class containing 9 training images. Inter-class differences between the images were sometimes small. Best systems reached an error rate of 12.6%.

Interactive image retrieval

The goal of this track was to evaluate interaction strategies for cross-language image retrieval. The same database was used as the historic photographic task and participants had to design a system to offer multilingual access to it. An evaluation framework was provided that included 16 example images from the collection which users were required to find (a target search task). Participation in this task was low due to a higher demand on resources needed. However, user-centered evaluation is an extremely important topic in determining the success of a visual retrieval system

ImageCLEF 2006

For 2006 another pre-CLEF workshop on visual information retrieval evaluation is foreseen and submissions of papers are invited[11]. As the deadlines for participation have not yet passed, we encourage people to contact us for more information. Until the submission deadline it is possible to register and submit results for ImageCLEF, later it is still possible to register for getting access to the data but no official submission is possible. New information will be made available regularly on the ImageCLEF web pages[10]. The coming description only gives a short introduction.

[1011] http://muscle.prip.tuwien.ac.at/ws_overview_2006.php

[12] http://ir.shef.ac.uk/imageclef/

Retrieval from a personal photographic collection (ImageCLEFphoto)

A new database of photographs taken from an independent travel company will replace the St. Andrews collection. This database will help to improve the use of visual retrieval methods, particularly for multimodal retrieval, as images are high-quality color photographs. 20,000 images have been annotated in English and German and realistic queries will be made available in a variety of languages enabling both monolingual and multilingual search tasks. Visual queries in the form of query images will also be made available. The task fits well with the growing interest in information access to personal photographic collections.

Medical image retrieval

For medical image retrieval the databases will stay the same but the tasks will be based on two user studies, plus the analysis of search terms for a medical web search engine (HONmedia search). Three groups of tasks will be created: visual query tasks mainly aimed at purely visual retrieval, mixed queries tasks where visual and textual information seem important and semantic tasks, where visual information does not seem important.

Medical image annotation

The medical automatic annotation task will have 10,000 images as training data set for this year and will offer a larger number of classes: 122 in place of the 57 classes from 2005. This is expected to create a harder task and also pave the way for multi-hierarchical classification in a future task (maybe 2007).

Non-medical image annotation

The non-medical automatic annotation task will take place for the first time in 2006. Thanks to LTU technologies[11] (LookThatUp), a large database of common objects gathered from the web is available. New objects need to be classified into one of the available classes. 20 object classes will be used in 2006. The goal is to evaluate the quality of algorithms attaching automatic text labels to images which are more general than in the medical annotation task.

Interactive image retrieval

In 2006, a collection of FlickR images is foreseen for the interactive task. This should increase the interest in the user-centered search task as FlickR is definitely one of the most popular image sharing places on the Internet and the number as well as the quality of the images is extremely high.

Conclusions

ImageCLEF is aiming to create a publicly-available benchmark for the evaluation of multilingual multimodal information retrieval systems. The goal is to create collections and realistic topics (based on real user needs) to help researchers evaluate and compare their algorithms. This is only possible if the research community is giving feedback to us and helping to create new tasks following active research area. Participation in ImageCLEF enables access to freely accessible databases for image retrieval system evaluation. Registration for ImageCLEF is free of charge and there are no hidden obligations to submit results. We appreciate active participation and hope to see many of you at the CLEF workshop where techniques from the participating systems are compared. ImageCLEF is not a “competition” to be won but a place to share experiences and problems with other researchers to improve future retrieval systems.

Important dates:

April 2006:	Topic release to participants
June 2006:	Submission of results
July 2006:	Release of ground truth
August 2006:	Submission of working notes papers of all participants
September 19, 2006:	MUSCLE/ImageCLEF workshop
September 20-22, 2006:	CLEF

Acknowledgements

ImageCLEF would be impossible without the help of several people, We would particularly like to thank Michael Grubinger for supplying and annotating the collection for ImageCLEFphoto, Bill Hersh and Jeffery Jensen for their help on the medical retrieval task, Thomas Deselaers and Thomas Lehmann for organizing the medical annotation task, Allan Hanbury for organizing the non-medical annotation task and Carol Peters for organizing CLEF.

References

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

[2] John P. Eakins, Pamela Briggs, Bryan Burford: Image Retrieval Interfaces: A User Perspective. *CIVR 2004,* pp. 628-637

[3] J Jeon, V Lavrenko, R Manmatha, Automatic image annotation and retrieval using cross-media relevance models, *ACM SIGIR*, Toronto, Canada, 119-126, 2003.

[4] H Müller, W Müller, DM Squire, S Marchand-Maillet, T Pun, Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals, *Pattern Recognition Letters*, 22(**5**): 593-601, 2001.

[11] http://www.ltutech.com/

[5] P Clough, H Müller, T Deselaers, M Grubinger, T Lehmann, J Jensen, W Hersh, The CLEF 2005 Cross-Language Image Retrieval Track, *Working Notes of the Cross Language Evaluation Forum*, Vienna, Austria, 2005.