Multimedia Retrieval at INEX 2006

Thijs Westerveld
CWI
Amsterdam
The Netherlands
thijs@cwi.nl

Roelof van Zwol
Yahoo! Research
Spain
roelof@yahoo-inc.com

1 Introduction

Structured document retrieval allows for the retrieval of document fragments, i.e. XML elements, containing relevant information. The main INEX Ad Hoc track focuses on text-based XML element retrieval. Although text is dominantly present in most XML document collections, other types of media can also be found. Existing research on multimedia information retrieval has shown that it is far from trivial to determine the combined relevance of a document that contains several multimedia objects. The objective of the INEX multimedia track is to exploit the XML structure that provides a logical level at which multimedia objects are connected, to improve the retrieval performance of an XML-driven multimedia information retrieval system.

2 Collections Tasks and Resources

Since 2006, INEX uses a corpus based on the English part of Wikipedia, the free content encyclopedia (`http://en.wikipedia.org`). Wikitext pages have been converted to XML [1] making a structured collection of 659,388 XML pages. Statistics about the collection and its multimedia nature are listed in Table 1. This collection is also the basis for the multimedia track

Total number of XML documents	659,388
Total number of images	344,642
Number of unique images	246,730
Average number of images per document	0.52
Average depth of XML structure	6.72
Average number of XML nodes per document	161.35

Table 1: Wikipedia collection statistics

In addition, the multimedia track makes use of a collection of Wikipedia metadata documents. Each of the images on Wikipedia comes with such a document, usually containing a brief caption or description of the image, the user who uploaded the image, and the copyright information. For the images in the INEX Wikipedia collection, these metadata documents are downloaded and converted to XML. Some documents had to be removed because of copyright issues or parsing problems, leaving us with a collection of 171,900 images with metadata. Figure 1 shows an example of such a metadata document.

Figure 1: Example Wikipedia image and metadata document.

Both the main Wikipedia collection and the Wikipedia images and metadata collection are used in the INEX 2006 multimedia track, each for their own task. The main Wikipedia collection is used in the Ad Hoc XML retrieval task of finding relevant XML fragments given a multimedia information need, the images and metadata collection is used in a pure image retrieval task: Find images (with metadata) given an information need. The collection corresponding to a task is merely the target collection, both collections can be used for both tasks. In addition we provide a number of additional sources of information to help participants get to the relevant information in these collections.

Image classification scores: For each image the classification scores for 101 different concepts are provided by the University of Amsterdam [4]. The concepts are shown in Figure 2. The classifiers are trained on manually annotated TRECVID video data [6] and the concepts are picked for the broadcast news domain. It is unclear how well these concept classifiers will perform on the broad collection of Wikipedia images, but we believe it may be a useful source of information.

Image features: A set of 22D feature vectors, one for each image, is available that has been used to derive the image classification scores [2]. Participants can use these feature vectors to build a custom content-based image retrieval (CBIR) system, without having to pre-process the image collection.

CBIR system: An on-line content-based image retrieval system to get a ranked list of similar images given a query image (from the collection) is provided by RMIT [3].

3 Topics

The topics used in the INEX MM track are descriptions of (structured) multimedia information needs. Structural and multimedia hints are expressed in the NEXI query language [5], a variant of XPath developed for structural queries. For example to find articles about *Anne Frank*, one could type

```
//article[about(.,Anne Frank)].
```

For multimedia search, NEXI has been extended to incorporate visual hints. If a user wants to express that results should have images similar to a given example image, this can be indicated in an about clause with the keyword *src:*. For example to find images of cityscapes similar to the image with identifier 789744, one could type

```
//image[about(.,cityscape) and about(.,src:789744)]
```

To keep things manageable, only example images from within the MM collection are allowed. This has repercussions on the evaluation as it will give an unfair advantage to systems that use the example images. Therefore, topics containing references to example images are interpreted as \Find *other* images like this one", and the example images are removed from submissions as well as relevance judgements.

We introduced a second type of visual hints, directly related to the concept classifications that are provided as an additional source of information. If a user thinks the results should be of a given concept, this can be indicated with an about clause with the keyword *concept:*. For example, to search for cityscapes one could decide to use the concept building:

```
//image[about(.,cityscape) and about(.,concept:building)]
```

Terms following the keyword *concept:* are obviously restricted to the 101 concepts for which classification results are provided (cf. Figure 2).

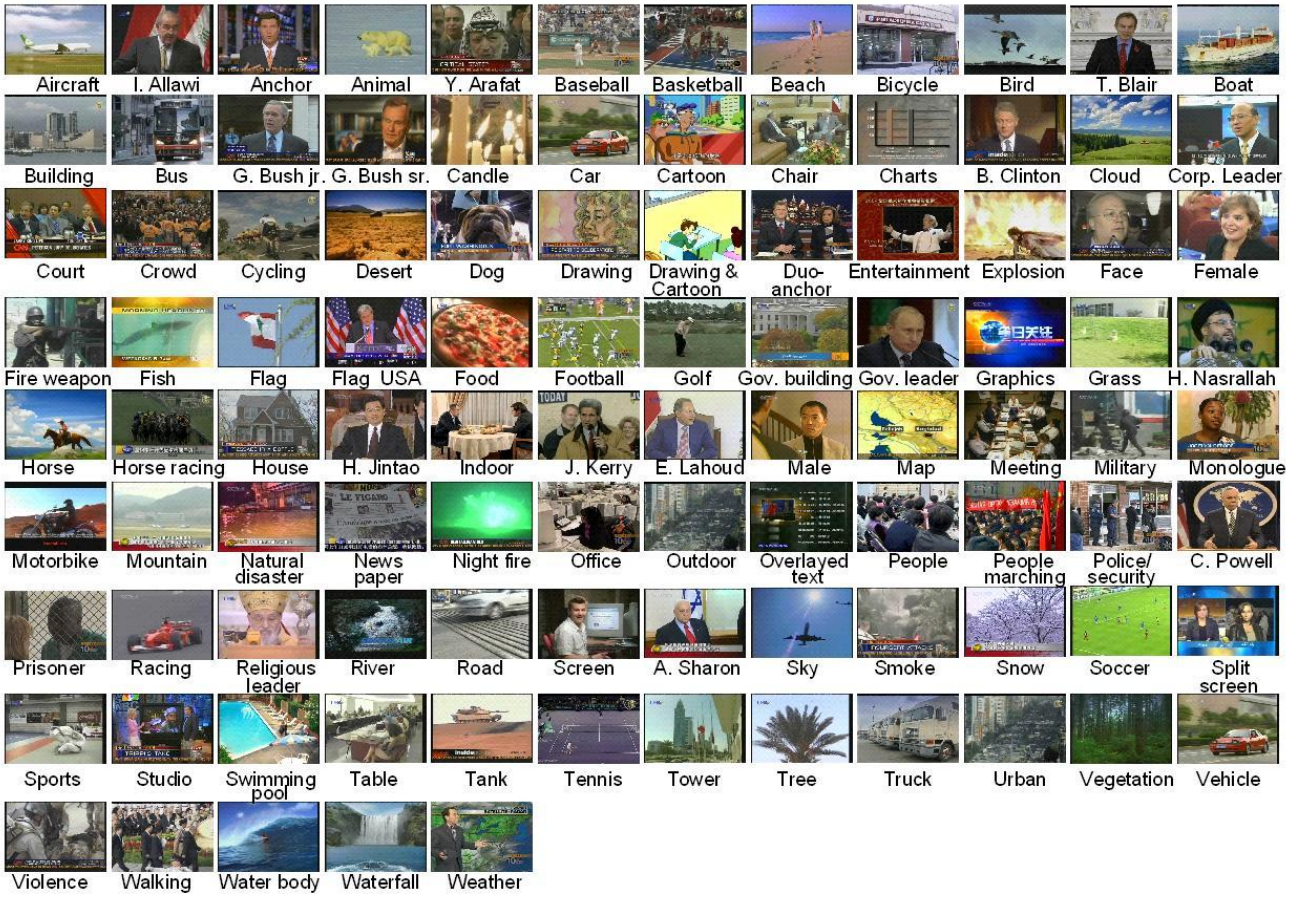

Figure 2: The 101 concepts for which classification scores are available. This image is taken from [4]

It is important to realise that all structural, textual and visual filters in the query should be interpreted loosely. It is up to the retrieval systems how to use, combine or ignore this information. The

relevance of a fragment does not directly depend on these elements, but is decided by manual assessments.

Topics for both the multimedia fragment retrieval task and the image and metadata retrieval task are contributed by participants and consist of the usual title, description and narrative fields. The title field comes in two flavours, the first variant a plain web-search-style set of keywords, the second a more structured NEXI query, potentially with visual hints as described above.

4 Assessments and Metrics

The images and metadata task is a document retrieval tasks. A document, i.e., an image with its metadata, is either relevant or not. For this task we adopted TREC style document pooling of the documents and binary assessments at the document level. We report the standard measures mean average precision and recall precision graphs.

The multimedia fragment retrieval task requires assessments at the sub-document level, a simple binary judgement at the document level is not sufficient. Still, for ease of assessment, retrieved fragments are grouped by document. Once all participants have submitted their runs, the top N fragments for each topic are pooled and grouped by document. To keep assessment loads within reasonable bounds, we used pools with a depth of 500 fragments. The documents are alphabetized so that the assessors do not know how many runs retrieved fragments from a certain document or at what rank(s) the fragments were found. Assessors than look at the documents in the pool and highlight the relevant parts of each document. The assessment system stores the relevance or non-relevance of the underlying XML elements. Systems are compared using effort-precision/gain-recall graphs, the cumulative gain based measures used in many INEX tasks. We also report the summary statistic of these, mean average effort precision.

For the images and metadata task, we were able to do some duplicate assessments as well as some assessing with a deeper pool. The studies based on this show a high inter assessor agreement of 95%. Furthermore, we learned our pool is of good quality: comparative results are hardly affected by shallower pools or by groups not contributing to the pool. Full details of the inter assessor and pool quality studies are available in the INEX multimedia track overview paper [7].

5 Approaches and Results

Only four participants submitted runs for the multimedia track: CWI together with the University of Twente (CWI/UT), IRIT (IRIT), RMIT University (RMIT) and Queensland University of Technology in Australia (QUTAU). These groups submitted a total of 15 multimedia fragments runs and 16 images and metadata runs. Table 2 shows for each of the tasks in how many submissions each of the resources was used. The Wikipedia collections are mainly used for the tasks for which they are the target collection, but CWI/UT experimented with using both the main Wikipedia collection and the images and metadata collection on the fragment retrieval task. The visual resources provided are used mostly in the images and metadata task, although RMIT used their GIFT tool also on some fragment retrieval runs. QUTAU is the only group that used the concepts and features provided by the University of Amsterdam; they used them for images and metadata runs.

For both the multimedia fragment retrieval tasks and the images and metadata retrieval task, the top ranked submissions were text and structure based runs that did not make use of any visual processing. Detailed results and short descriptions of the approaches taken by the participants are available from the INEX multimedia track overview paper [7].

resource	Number of multimedia fragments runs using resource	Number of image and metadata runs using resource
main Wikipedia	15	0
Wikipedia image and metadata	3	16
image features	0	3
image classification scores	0	1
CBIR system	2	10

Table 2: Resources used by the submitted runs

6 Conclusions

The INEX multimedia track has been running for two years now. While the first year functioned mainly as a pilot study, using limited data sets, in 2006 we worked with a realistic and sizable collection based on Wikipedia documents, the images they contain and the related metadata. In addition we provided extracted image features, concept classification scores for 101 concepts, and a content based information retrieval system. The total set of data provided makes a unique collection of related resources.

The number of participants in the multimedia track was disappointing with only four groups submitting runs. This makes it hard to draw general conclusions from the results. What we could see so far is that the top runs in both tasks did not make use of any visual resources. More detailed analyses of the results and the participants' system descriptions is needed to see if groups managed to improve over a textual baseline using visual indicators of relevance. Also, a topic by topic analysis could shine some light. Perhaps these techniques did contribute for only a limited number of topics and hurt for others.

For next year's multimedia track, we hope to draw more participants, from inside as well as outside INEX. The set of related collections and resources, makes this track an interesting playing ground, both for groups with a background in databases or information retrieval, and for groups with a deeper understanding of computer vision or image analysis.

References

[1] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[2] Jan C. van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, Cees G. M. Snoek, and Arnold W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 105, Washington, DC, USA, 2006. IEEE Computer Society.

[3] RMIT University School of Computer Science and Information Technology. Wikipedia CBIR system for the multimedia track. *http://www.cs.rmit.edu.au/.*

[4] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421-430, New York, NY, USA, 2006. ACM Press.

[5] Andrew Trotman and Börkur Sigurbjörnsson. Narrowed extended xpath I (NEXI). In Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Zoltan Szlavik, editors, *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, volume 3493. Springer-Verlag GmbH, may 2005. *http://www.springeronline.com/3-540- 26166-4.*

[6] Timo Volkmer, John R. Smith, and Apostol (Paul) Natsev. A web-based system for collaborative annotation of large image and video collections: an evaluation and user study. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 892-901, New York, NY, USA, 2005. ACM Press.

[7] Thijs Westerveld and Roelof van Zwol. The inex 2006 multimedia track. In Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors, *Advances in XML Information Retrieval: Fifth International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006*, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence (LNCS/LNAI). Springer- Verlag, 2007.