Workshop on the Evaluation of Multimedia Retrieval

Thijs Westerveld[†], **Arjen P. de Vries**[†] and **Franciska M.G. de Jong**[‡]

[†]CWI
The Netherlands
thijs,arjen@cwi.nl

[‡]University of Twente
The Netherlands
fdejong@cs.utwente.nl

1 Introduction

The evaluation of multimedia retrieval is a subject that has gained momentum in the last couple of years. CWI, the National Research Institute for Mathematics and Computer Science in the Netherlands, organised a workshop organised on the subject on 24 November 2004. The main aim of the workshop was to bring together researchers and practitioners in the field of multimedia retrieval to discuss the area of multimedia in general and methodology for evaluation within this area in particular. The workshop, organised by Franciska de Jong (Utwente/TNO, NL), Arjen de Vries (CWI, NL) and Thijs Westerveld (CWI, NL) was an informal half-day meeting without papers or proceedings. Because the workshop was co-located with Thijs Westerveld's PhD defence, we were able to invite Alex Hauptmann (CMU, PA, USA) to give a talk. In total six speakers were invited to present their work related to the evaluation of multimedia retrieval. The workshop started with a talk presenting multimedia retrieval in practise. Then, three talks discussed experiments in the laboratory context of the TRECVID video retrieval benchmark. The afternoon ended with a presentation of a study of interactive experiments and a talk discussing metrics for measuring multimedia retrieval effectiveness.

2 Presentations

Annemieke de Jong (Netherlands Institute for Sound and Vision, NL) presented the archive's point of view on multimedia retrieval, discussing both the way the institute catalogues multimedia material and the types of queries it has to deal with. Sound and vision manually indexes a variety of Dutch television broadcasts, but also a wide range of stock shots. They catalogue the information content (what is the programme about?), the audiovisual content (what is seen and heard?) and metadata (e.g. where is the data stored?). In cataloguing the audiovisual content, no detail is indexed; thus, for example the landscape is annotated, but not the trees. Of course, exceptions are made for outstanding elements. The archive receives a wide variety of queries, mostly from program makers, but also from the general public. The queries range from known item queries (a particular broadcast) to subject queries (queries for shots relating to a subject) and generic and specific queries for shots or quotes (e.g., ethnic minorities in a quiz show, or a man dressed in green in a suburban environment). Most users want recent shots, and, typically, a certain aesthetic value is wanted.

Wessel Kraaij (TNO, NL) presented an overview of the TRECVID workshop, see also [3] and [4]. Wessel presented the four TRECVID tasks: shot boundary detection, story segmentation, feature extraction and search. Shot boundary detection for cuts seems to be more or less solved, while for gradual shot transitions there is still some room for improvement. It is unclear how well the results transfer to video material other than news broadcasts. Some techniques applied in the scene segmentation, and feature extraction tasks though, are generic and independent of the CNN/ABC format. Scene segmentation exploiting visual

information proofs better than segmentation based on ASR only, i.e., the TRECVID segmentation results are better than the TDT baseline. The topics for the search task are created with the sound of the video turned off, but still textual information from the ASR transcripts is an important source of information. In general, scores on the search task are low compared to the numbers known from text retrieval, but results for interactive runs are close.

Cees Snoek (UvA, NL) discussed *weak retrieval*, the MediaMill approach to video retrieval [5]. They propose a semantic retrieval solution that combines indexing of a limited set of generic concepts with interaction. Since this requires some user effort it is considered to be weak retrieval. They developed detectors for a lexicon of 32 semantic concepts that allow for query by semantic concept. A generic approach called the semantic value chain is used to detect these concepts. The approach is based on the idea that produced video is the result of an authoring process, where the author starts from a semantic intention. This semantic message is conveyed using stylistic elements and thus a multimedia document is produced. Multimedia analysis is seen as the reverse of this process: starting from basic features, the semantic value chain gradually adds more and more semantics. Cees explained the details of the approach and showed their evaluation results on the TRECVID 2004 search and feature detection tasks. Top ranking performance in both tasks indicate the potential of the approach.

Alex Hauptmann (CMU, PA, USA) discussed two main subjects. First, CMU's participation in the TRECVID workshop series. Second, the statistical analysis of TRECVID 2003 and 2004 results. CMU has participated in TRECVID from its start in 2001. Alex quickly reviewed some of the techniques used by CMU including exploiting relationships between different concepts (e.g. cars are typically outdoors) and *co-retrieval*. Co-retrieval uses the results found using text retrieval on ASR transcripts in a blind relevance feedback loop to find the optimal weights for mixing in the other modalities and system components. This approach improves a bit over text only results, but, as Alex remarked, "It makes everything look similar to text, so the approach will not find you new stuff."

Alex also presented an analysis of the results and showed many differences between runs are actually meaningless. Using Newman-Keuls' test of pairwise significance, Alex found that actually large groups of submissions did not differ significantly. For example, the difference between all CMU runs submitted in 2003 were insignificant and the top 15 automatic runs in 2004's search task proved indistinguishable.

Laura Hollink (VU, NL) presented the results of a study in which she analysed search behaviour of people querying an interactive news video retrieval system [2]. The results of the study show that topics concerning 'specific' people or objects were better retrieved than topics concerning 'generic' objects and scenes. Some users therefore used specific queries to solve general information needs (e.g., querying for Michael Jordan when searching for basketball shots). Users were able to estimate the overall quality of a search but did not know when the optimal result was reached within the search process. Analysis of the results at various stages in the retrieval process suggests that retrieval based on transcriptions of the speech in video data adds more to the average precision of the result than content-based retrieval. The latter is particularly useful in providing the user with an overview of the dataset and thus an indication of the success of a search.

Arjen de Vries (CWI, NL) discussed evaluation metrics for search tasks without a pre-defined retrieval unit, like video and XML retrieval (in both cases document fragments without pre-defined borders are wanted). The use of traditional recall and precision metrics is problematic in these settings due to issues caused by overlap between result and reference items. Arjen proposed evaluation metrics derived from a user-effort oriented view of information retrieval to address these problems [1]. It builds on the Expected Search Length metric of Cooper, revived by Dunlop for the Expected Search Duration metric. His work extends these

previous works by demonstrating how to handle systematically the overlap problems introduced when the assumption of a fixed, predefined retrieval unit is removed from the benchmark setting.

3 Conclusion

Multimedia retrieval is a lively area. Although experimentation in this field is perhaps not as important yet as it has been since long in text retrieval, it is gaining attention. This workshop served as a forum to discuss evaluation methodology and lessons learned from experimentation. With around 30 participants and lots of discussion after each talk, we feel the workshop was a great success. Abstracts and slides for the talks are available from `http://www.cwi.nl/projects/trecvid/MRE/`.

References

[1] Arjen P. de Vries, Gabriella Kazai, and Mounia Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference Proceedings*, pages 463–473, 2004.

[2] L. Hollink, G.P. Nguyen, D.C. Koelma, A.T. Schreiber, and M. Worring. User strategies in video retrieval: A case study. In *Proceedings of The International Conference on Image and Video Retrieval (CIVR2004)*, Dublin, Ireland, 2004.

[3] W. Kraaij, A.F. Smeaton, P.Over, and J. Arlandis. Trecvid 2004 – an introduction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[4] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Trecvid: evaluating the effectiveness of information retrieval tasks on digital video. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 652–655. ACM Press, 2004.

[5] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma, and F.J. Seinstra. The mediamill trecvid 2004 semantic viedo search engine. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.