MEDIAMILL: SEARCHING MULTIMEDIA ARCHIVES BASED ON LEARNED SEMANTICS

C.G.M. Snoek, D.C. Koelma, J. van Rest, N. Schipper, F.J. Seinstra, A. Thean, and M. Worring

MediaMill
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
mediamill@science.uva.nl

Abstract

Video is about to conquer the Internet. Real-time delivery of video content is technically possible to any desktop and mobile device, even with modest connections. The main problem hampering massive (re)usage of video content today is the lack of effective content based tools that provide semantic access. In this contribution we discuss systems for both video analysis and video retrieval that facilitate semantic access to video sources. Both systems were evaluated in the 2004 TRECVID benchmark as top performers in their task.

1. INTRODUCTION

Query by keyword forms the foundation for machine-based interaction of humans and text repositories. Elaborating on the success of text-based search engines, the query by keyword paradigm is also gaining momentum in multimedia retrieval scenarios. For multimedia archives it is hard to achieve effective access, however, when based on keywords that appear in the text only. Video archives require semantic access where all modalities can contribute to the concept.

We have participated in the 2004 NIST TRECVID video retrieval benchmark to measure performance of our solution to semantic access. The benchmark aims to promote progress in video retrieval via open, metrics-based evaluation [1]. The video archive of the 2004 TRECVID benchmark is composed of 184 hours of ABC World News Tonight and CNN Headline News and is recorded in MPEG-1 format. Within this video archive we define a lexicon of 32 semantic concepts [3]. The lexicon contains both general concepts, like *people*, *car*, and *beach*, as well as specific concepts such as *airplane take off* and *news subject monologue*. In Section 2 we briefly discuss the system architecture that detects all 32 concepts. The search engines that were developed based on the detected concepts are discussed in Section 3.

This research is sponsored by the BSIK MultimediaN project.

2. VIDEO ANALYSIS

The core of semantic indexing is to reverse the authoring process [2]. We follow this path to arrive at a system architecture for semantic indexing in video. The proposed semantic value chain is composed of three links. The output of a link in the chain forms the input for the next one. The semantic value chain starts in the *content link*. In this link, we follow a data-driven approach of indexing semantics. The *style link* is the second link. Here we tackle the indexing problem by viewing a video from the perspective of production. Finally, to enhance the indexes, in the *semantic link*, we view semantics in context. The virtue of the semantic value chain is that concepts are incrementally adapted to the intention of the author. The links in the semantic value chain exploit a common architecture with a standardized input-output model to allow for semantic integration. We build this architecture on machine learning for robust detection of semantics. For an in depth discussion on the semantic value chain and its successful performance we refer to [3].

3. SEARCH ENGINES

We developed two prototype search engines based on the analysis results of the semantic value chain. The first one allows for interactive retrieval, while the second one is a personalized search engine for the Internet.

3.1. Interactive Retrieval

To shield the user from technical complexity we offer three basic query interfaces to a video archive: query by concept, query by keyword, and query by example. The set of concepts from the concept lexicon forms the basis for interactive filtering of the query results. Naturally, this aids for queries that contain concepts from this lexicon. In this case, users may rely on direct query by concept. Based on query by concept users can also make a first selection when

0-7803-9332-5/05/$20.00 ©2005 IEEE

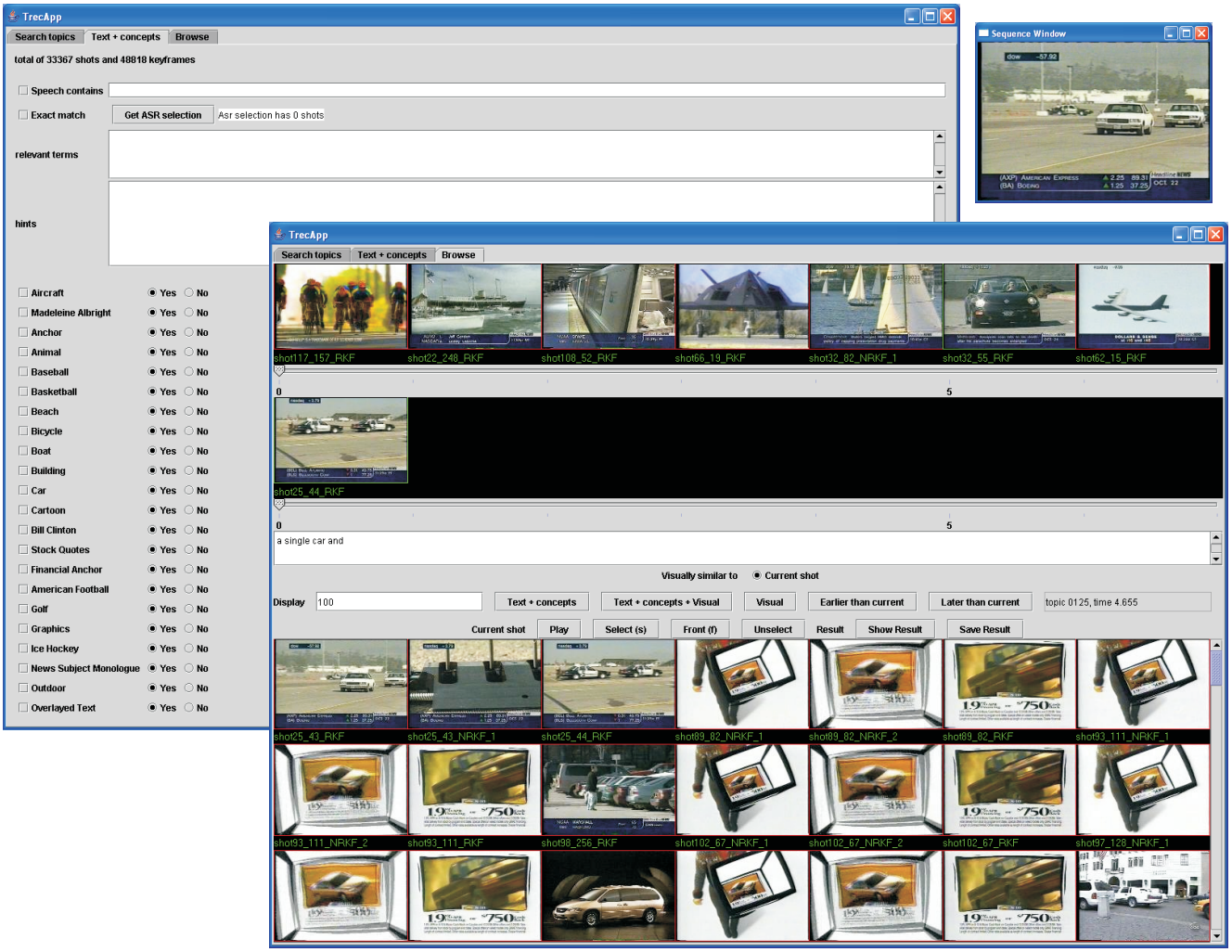

Figure 1: Interfaces of the interactive video search engine. The top row of the right panel shows selected results for *vehicle*, the bottom shows results for *car*.

a query includes a concept that is a super-class or a sub-class of a concept in the lexicon, e.g. when searching for *vehicles* one can use the concept *car*. For search topics not covered by on of the concepts in the lexicon, users have to rely on a combination of query by keyword and query by example. Applying query by keyword in isolation allows users to find very specific topics, e.g. *Senator Henry Hyde's Face*. Based on query by example shots that exhibit a similar color distribution or similar salient point distributions can augment results further. To select query interfaces and combine retrieval results we rely on interaction by a user. Ultimately, this retrieval approach results in semantic access to video archives. The interface of the search engine is depicted in Fig. 1. For performance statistics of interactive retrieval within TRECVID we refer to [3].

3.2. Personalized Retrieval

Apart from an interactive retrieval engine, we also developed the VIdeo PERsonalizer (VIPER). With this search engine we aim to provide a person with the most relevant information given the expected use and user preferences. Personas are used to model different users. VIPER obtains the user profiles explicitly from user preferences and implicitly from learned user interaction. The set of concepts from the lexicon are used for retrieval. In addition, an ontology based on WordNet was developed that provides a structure on the concepts and increases search possibilities. To decide whether semantic concepts are related to a query, a pruner module computes query-dependent thresholds. To present the results, VIPER provides five visualization modes, ranging from grid to spiral, see Fig. 2.

Figure 2: Interfaces of the VIPER search engine. The left panel shows the search interface, the top panel shows a spiral visualization of a search on *graphics* and the bottom panel shows a grid visualization of *ice hockey*.

4. FUTURE EXTENSIONS

For future work we aim to extend the query interface with query by region example and relevance feedback. Furthermore, to enhance display possibilities we need other visualizations, for example using 3D. However, the greatest challenge ahead is to extend the lexicon of semantic concepts to a set that is competitive with human knowledge. This will have a dazzling impact on multimedia repository usage scenarios.

5. ACKNOWLEDGMENTS

The authors are grateful to the students of the Personalized Information Delivery class of 2004 for developing VIPER.

6. REFERENCES

[1] NIST. TRECVID Video Retrieval Evaluation, 2004. `http://www-nlpir.nist.gov/projects/trecvid/`.

[2] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[3] C. Snoek, M. Worring, J. Geusebroek, D. Koelma, and F. Seinstra. The MediaMill TRECVID 2004 semantic video search engine. In *Proceedings of the TRECVID Workshop*, NIST Special Publication, Gaithersburg, USA, 2004.