MediaMill: Video Search using a Thesaurus of 500 Machine Learned Concepts

Cees G.M. Snoek, Marcel Worring, Bouke Huurnink, Jan. C. van Gemert,
Koen E.A. van de Sande, Dennis C. Koelma, and Ork de Rooij

***Abstract*— In this technical demonstration we showcase the current version of the MediaMill system, a search engine that facilitates access to news video archives at a semantic level. The core of the system is a thesaurus of 500 automatically detected semantic concepts. To handle such a large thesaurus in retrieval, an engine is developed which automatically selects a set of relevant concepts based on the textual query and user-specified example images. The result set can be browsed easily to obtain the final result for the query.**

***Index Terms*— Semantic indexing, video retrieval, information visualization.**

I. Introduction

Most commercial video search engines such as Google, Blinkx, and YouTube provide access to their repositories based on text as this is still the easiest way for a user to describe an information need. The indices of these search engines are based on the filename, surrounding text, social tagging, or a transcript. This results in disappointing performance when the visual content is not reflected in the associated text. In addition, when the videos originate from non-English speaking countries, such as China or the Netherlands, querying the content becomes even harder as automatic speech recognition results are so much poorer. Additional visual analysis yields more robustness. Thus, in video retrieval a recent trend is to learn a lexicon of semantic concepts from multimedia examples and to employ these as entry points in querying the collection.

Last year we presented the *MediaMill 2005* video search engine [1] using a 101 concept lexicon [2] evaluated in the TRECVID benchmark [3]. For our current system we made a jump to a thesaurus of 500 concepts. The items vary from pure format like a detected *split screen*, or a style like an *interview*, or an object like a *horse*, or an event like an *airplane take off*. Any one of those brings an understanding of the current content. The elements in such a thesaurus offer users a semantic entry to video by allowing them to query on presence or absence of content elements. For a user, however, selecting the right topic from the large thesaurus is difficult. We therefore developed a suggestion engine that analyzes the textual topic, and possible image examples given by the user, to automatically derive the most relevant concept detectors for querying the video archive (see Fig. 1 and Fig. 2).

This research is sponsored by the BSIK MultimediaN project. The authors are with the Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands (e-mail: info@mediamill.nl, http://www.mediamill.nl).

II. The MediaMill 2006 System

The data flow of the MediaMill 2006 system is depicted in Fig. 1. We will now highlight its components in more detail.

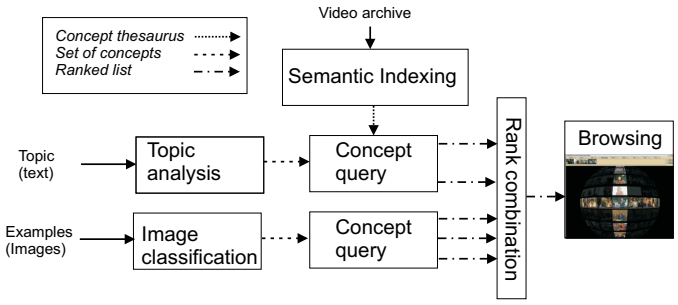

Fig. 1. Overview of the different processing steps in the MediaMill semantic video search engine.

A. Semantic Indexing

For semantic indexing we proposed the semantic pathfinder, for details see [4]. First, it extracts features from the visual [5], textual, and auditory modality. The architecture exploits supervised machine learning to automatically label segments with semantic concepts. In the first step learning is on the content features only. In the second step, the video is analyzed based on its style properties. Finally, semantic concepts are analyzed in context, with the potential to boost index results further. The resulting thesaurus of 500 semantic concepts, covering *setting*, *objects*, and *people*, is learned based on the LSCOM annotations [6] and the 101 concepts used in our 2005 engine [2].

B. Topic Analysis

We map the richness and subjectivity of semantics in user queries to concept detectors available in our thesaurus. To derive the most relevant concepts for a given user topic, we first assign syntactic categories to groups of words in the input text using a chunking algorithm. We then assign a grammatical classification to each word by using a part-of-speech tagger. From there, looking up each noun chunk in WordNet [7]. When a match has been found those words are eliminated from further lookups. Then we look up any remaining nouns in WordNet. The result is a number of WordNet words related to the input text. Now that both the concepts in the text and the multimedia concept detectors are related to WordNet, we can

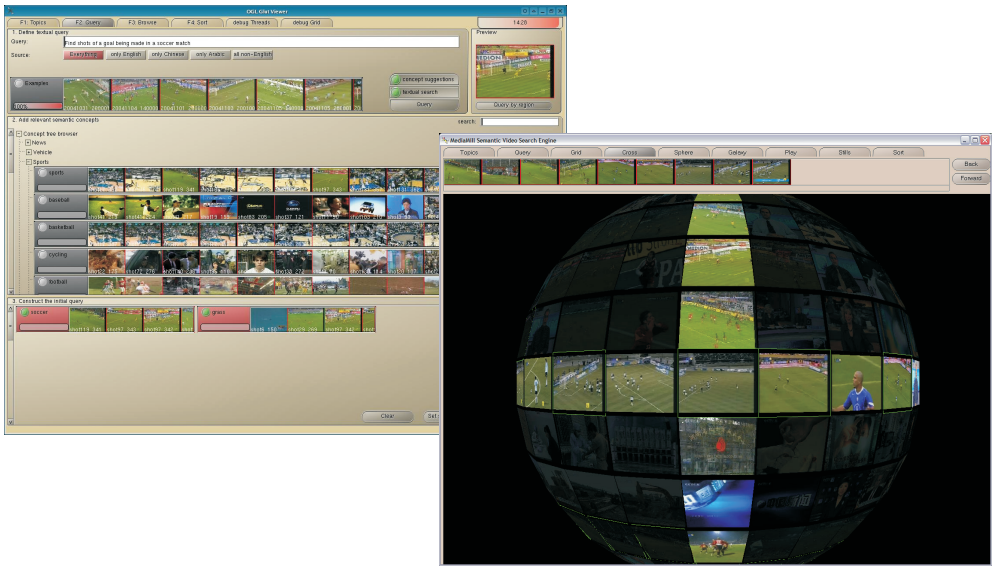

Fig. 2. On the left an example of a query for shots of a goal being made in a soccer match, using both text and image examples, yielding *soccer* and *grass* as most relevant concepts. Result of the query are visualized in the *CrossBrowser* on the right.

compute the semantic distance between the textual concepts and the multimedia concepts. We use Resnik's algorithm [8] which calculates the similarity of a concept to each of the WordNet nouns from the query text. Based on the combined scores we rank each multimedia concept detector in order of expected utility.

C. Image Classification

Concept suggestion based on query image analysis first extracts visual features [5]. Based on the features we predict for each image a concept using pre-learned visual-only models. Rather than selecting the concept with maximal score –which are often the most robust but also least informative ones, e.g. *people*, *face*, *outdoor* – we select the model that maximizes the probability of observing this image given the concept. To compute, Bayes' theorem is applied using training set statistics. Hence, we prioritize less frequent, but discriminative, concepts with reasonable probability scores over frequent, but less discriminative, concepts with high probability scores.

D. Rank Combination

We offer users several possibilities to combine the various ranked lists. They can employ standard combination methods such as min, max, sum, and product [9]. In addition, they may specify that some concepts are more important than others by adding weights to individual concepts.

E. Browsing the Result

The result of concept suggestion, the subsequent concept queries and their combination yields a ranked list of shots. To aid human interpretation in exploring this result the *CrossBrowser* visualizes the ranked list (vertical axis) versus the time (horizontal axis) of the program containing the shot. The two dimensions are projected onto a sphere to allow easy navigation. It also enhances focus of attention on the most important elements. Remaining elements are still visible, but much darker (see Fig. 2).

III. Demonstration

We demonstrate semantic exploration of news video archives with the MediaMill system. We will show how a thesaurus of 500 concepts can be exploited for effective access to video at a semantic level. In addition, we will exhibit novel browsers that present retrieval results using advanced visualizations. Taken together, the search engine provides users with semantic access to news video archives.

References

[1] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.-M. Geusebroek, D.Koelma, G.P. Nguyen, O. de Rooij, and F. Seinstra, "MediaMill: Exploring news video archives based on learned semantics," Singapore, November 2005, pp. 225–226.

[2] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.-M. Geusebroek, and A.W.M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proceedings of the ACM International Conference on Multimedia*, Santa Barbara, USA, October 2006, pp. 421–430.

[3] A. Smeaton, "Large scale evaluations of multimedia information retrieval: The TRECVid experience," in *CIVR*, ser. LNCS, vol. 3569. Springer-Verlag, 2005, pp. 19–27.

[4] C.G.M. Snoek, M. Worring, J.-M. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W.M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1678–1689, October 2006.

[5] J.C. van Gemert, J.-M. Geusebroek, C.J. Veenman, C.G.M. Snoek, and A.W.M. Smeulders, "Robust scene categorization by learning image statistics in context," in *International Workshop on Semantic Learning Applications in Multimedia, in conjunction with CVPR'06*, New York, USA, June 2006.

[6] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[7] G.A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, pp. 39–41, 1995.

[8] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *IJCAI*, 1995.

[9] J. Lee, "Analysis of multiple evidence combination," in *Proceedings of ACM SIGIR*, 1997, pp. 267–276.