The MediaMill Large-lexicon Concept Suggestion Engine

Marcel Worring, Cees G.M. Snoek, Bouke Huurnink,
Jan C. van Gemert, Dennis C. Koelma, Ork de Rooij
ISLA, Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ
Amsterdam, The Netherlands
mediamill@science.uva.nl
www.mediamill.nl

ABSTRACT

In this technical demonstration we show 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 lexicon of 436 automatically detected semantic concepts. To handle such a large lexicon in retrieval, an engine is developed which automatically selects a set of relevant concepts based on the textual query and example images. The result set can be browsed easily to obtain the final result for the query.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Algorithms, Experimentation, Human Factors, Performance

Keywords

Semantic indexing, video retrieval, information visualization

1. INTRODUCTION

Most commercial video search engines such as Google and Blinkx 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, 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* system using a 101 concept lexicon [6] evaluated in the TRECVID benchmark [4]. For our current system we made a jump to a lexicon of 436 concepts. For a user, selecting the right topic from the large lexicon 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 for quering the dataset (see fig 1 and 2).

2. THE MEDIAMILL 2006 SYSTEM

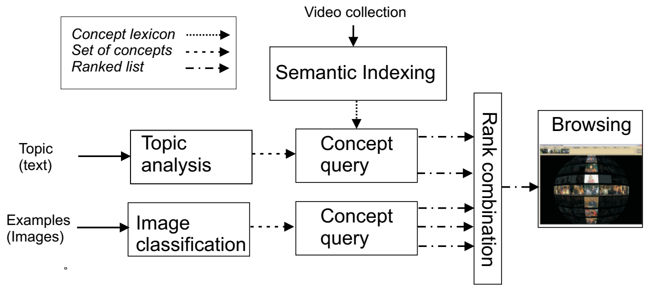


Figure 1: Overview of the different processing steps in the MediaMill system.

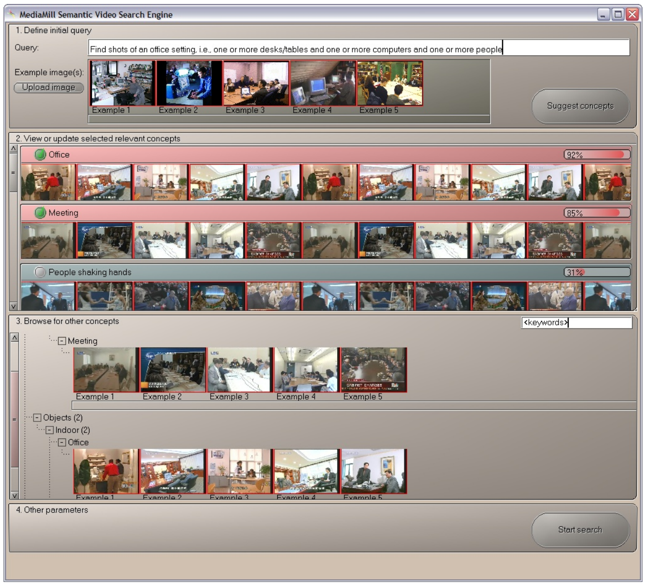
2.1 Semantic Indexing

For semantic indexing we proposed the semantic pathfinder, for details see [5]. First, it extracts features from the visual [7], 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 lexicon of 436 semantic concepts, covering *setting*, *object*, and *people*, is learned based on the LSCOM annotations [1] and the 101 concepts used in our 2005 engine [6].

2.2 Topic analysis

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 [2]. When a match has been found those words are eliminated from further lookups. Then we look up any

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.



Query: Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Figure 2: Example of a query for shots of an office setting, using both text and image examples, yielding office and meeting as most relevant concepts.

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 concepts are related to WordNet, we can compute the semantic distance between the textual concepts and the multimedia concepts. We use Resnik's algorithm [3] 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 concepts in order of expected utility.

2.3 Image classification

Concept suggestion based on query image analysis first extracts visual features [7]. 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.

2.4 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. In addition, they may specify that some concepts are more important than others by adding weights to individual concepts.

2.5 Browsing the result

The result of concept suggestion, the subsequent concept queries and their combination yields a ranked list of shots. To explore 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. 3).

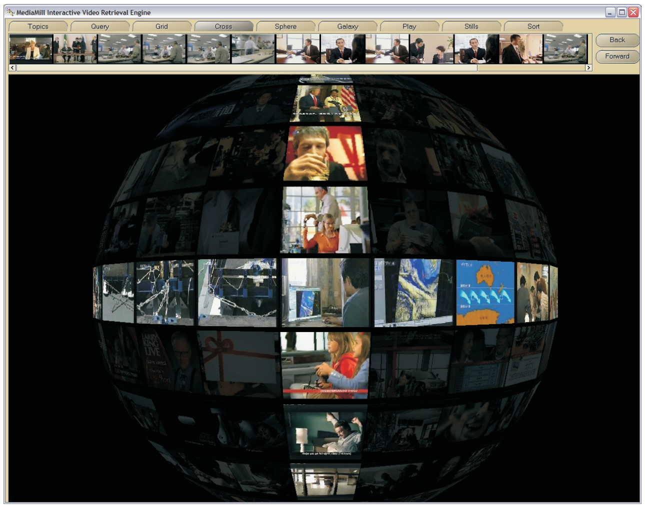


Figure 3: Result of the query in fig. 2 visualized in the *CrossBrowser*.

3. ACKNOWLEDGMENTS

This research is sponsored by the BSIK MultimediaN project.

4. REFERENCES

[1] L. Kennedy, A. Hauptmann, M. Naphade, J. Smith, and S.-F. Chang. Lscom lexicon definitions and annotations version 1.0. TR 217-2006-3, Columbia University, 2006.

[2] G. A. Miller. Wordnet: A lexical database for english. *Comm. of the ACM*, 38:39–41, 1995.

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

[4] A. Smeaton. Large scale evaluations of multimedia information retrieval: The TRECVid experience. In *CIVR*, volume 3569 of *LNCS*, 2005.

[5] 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 Trans. Pattern Analysis Machine Intell.*, 28(10), 2006.

[6] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, D.C. Koelma, G.P. Nguyen, O. de Rooij, and F.J. Seinstra. Mediamill: Exploring news video archives based on learned semantics. In *ACM Multimedia*, Singapore, 2005.

[7] J.C. van Gemert, J.-M. Geusebroek, C. Veenman, C.G.M. Snoek, and A.W.M. Smeulders. Robust scene categorization by learning image statistics in context. In *Int'l Workshop on Semantic Learning Applications in Multimedia*, 2006.