Mediamill: advanced browsing in news video archives

Marcel Worring*, Cees Snoek, Ork de Rooij Giang Nguyen, Giang Nguyen, Richard van Balen, and Dennis Koelma

Intelligent Systems Lab Amsterdam, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
worring@science.uva.nl
http://www.mediamill.nl

Abstract. *In this paper we present our Mediamill video search engine. The basis for the engine is a semantic indexing process which derives a lexicon of 101 concepts. To support the user in navigating the collection, the system defines a visual similarity space, a semantic similarity space, a semantic thread space, and browsers to explore them. It extends upon [1] with improved browsing tools. The search system is evaluated within the TRECVID benchmark [2]. We obtain a top-3 result for 19 out of 24 search topics. In addition, we obtain the highest mean average precision of all search participants.*

1 Introduction

Despite the emergence of commercial video search engines, such as Google and Blinkx, video retrieval is by no means a solved problem. Present day video search engines rely mainly on text in the form of closed captions or transcribed speech. Indexing videos with semantic visual concepts is more appropriate.

In literature different methods have been proposed to support the user beyond text search. Some of the most related work is described here. Informedia uses a limited set of high-level concepts to filter the results of text queries [3]. In [4], clustering is used to improve the presentation of results to the user. Both [3] and [4] use simple grid based visualizations. More advanced visualization tools are employed in [5] and [6] based on collages of keyframes and dynamically updated graphs respectively, but no semantic lexicon is used there.

In this paper we present our semantic search engine. This system computes a large lexicon of 101 concepts, clusters and threads to support interaction. Advanced visualization methods are used to give users quick access to the data.

2 Structuring the video collection

The aim of our interactive retrieval is to retrieve from a multimedia archive A, which is composed of n unique shots $\{s_1, s_2, \ldots, s_n\}$, the best possible answer set in response to a user information need. Examples of such needs are ”find me shots of dunks in

* This research is sponsored by the BSIK MultimediaN project.

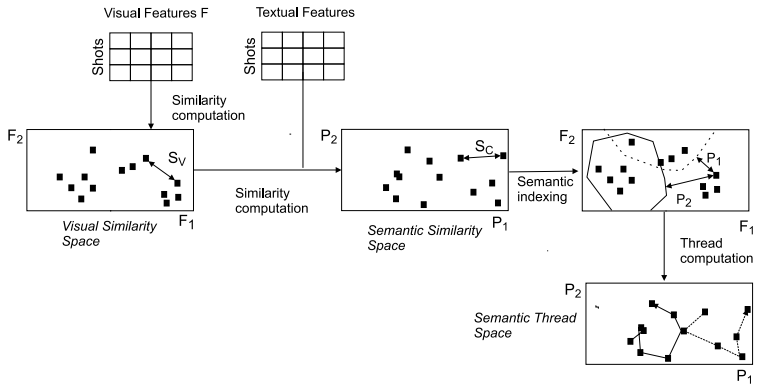

Fig. 1. A simplifi ed overview of the computation steps required to support the user in interactive access to a video collection. Note, that for both the vectors F and P only two dimensions are shown.

a basketball game" or "find me shots of Bush with an American flag". To make the interaction most effective we add different indices and structure to the data.

The visual indexing starts with computing a high-dimensional feature vector F for each shot s. In our system we use the Wiccest features as introduced in [7].

The next step in the indexing is to compute a similarity function S_v allowing comparison of different shots in A. For this the function described in [7] to compare two Weibull distributions is used. The result of this step is the *visual similarity space*. This space forms the basis for visual exploration of the dataset.

We employ our generic semantic pathfinder archictecture [8], to create a lexicon of 101 concepts so that every shot s_i is described by a vector $\{P_1, P_2, ..., P_{101}\}$. Elements in the lexicon range from specific persons to generic classes of people, generic settings, specific and generic objects etc. See [8] for a complete list.

Given two probability vectors, we use similarity function S_C to compare shots, now on the basis of their semantics. This yields the *semantic similarity space*.

The semantic similarity space induced by S_C is complex as shots can be related to several concepts. Therefore, we add additional navigation structure composed of a collection of linear paths, called *threads* through the data. Such a linear path is easy to navigate by simply moving back and forth. The first obvious thread is the time thread T^t. A complete set of threads $T^l = \{T^l_1, ..., T^l_{101}\}$ on the whole collection is defined by the concepts in the lexicon. The ranking based on P provides the ordering. Finally, groups are identified by clustering. Each cluster is then linearly ordered using a shortest path algorithm yielding the threads $T^s = \{T^s_1, ..., T^s_k\}$. The *Semantic thread space* is composed of T^t, T^l and T^s.

An overview of all the steps performed in the structuring of the video collection is given in Fig. 1.

Fig. 2. Browsers in the our semantic video search engine. Top left ConceptBrowser showing several topics. Top right the CrossBrowser showing results for soccer. Bottom left the SphereBrowser, displaying several semantic threads. Bottom right: active learning using a semantic cluster-based visualization in the GalaxyBrowser.

3 Interactive Search

The visual similarity space and the thread space define the basis for interaction with the user. Both of them require different visualization methods to provide optimal support. We developed four different browsers, which one to use depends on the information need. The different browsers are visualized in Fig. 2.

For many search tasks the initial query is formed by selecting one of the concepts from the lexicon of 101. To aid the user in this selection the *ConceptBrowser* presents the concepts in a hierarchy. Whenever the user comes to a leaf containing the concept j, the single thread T_j^l is shown as a filmstrip of keyframes corresponding to shots. By looking at those keyframes she gets a clear understanding of the meaning of the concept and whether it is indeed relevant to the search topic.

The *CrossBrowser* visualizes a single thread T_j^l based on a selected concept j from the lexicon versus the time thread T^t [8]. They are organized in a cross, with T_j^l along the vertical axis and T^t along the horizontal axis. Except for threads based on the lexicon, this browser can also be used if the user performs a textual query on the speech recognition result associated with the data, as this also leads to a linear ranking. The two dimensions are projected onto a sphere to allow easy navigation. It also enhances focus of attention on the most important element, the remaining elements are still visible, but much darker.

In the *SphereBrowser* the time thread T_j^l is also presented along the horizontal axis [8]. For each element in the time thread, the vertical axis is used to visualize the semantic thread T_j^s this particular element is part of. Users start the search by selecting a current point in the semantic similarity space by taking the top ranked element in a

textual query, or a lexicon based query. The user can also select any element in one of the other browsers and take that as a starting point. They then browse the thread space by navigating time or by navigating along a semantic thread.

Browsing visual similarity space is the most difficult task as there are no obvious dimensions on which to base the display. We have developed the *GalaxyBrowser* for this purpose [9] [8]. A short overview is given here. The core of the method is formed by a projection of the high-dimensional similarity space induced by S_v to the two dimensions on the screen. This projection is based on ISOMAP and Stochastic Neighbor Embedding. However, in these methods an element is represented as a point. In our method great care is taken to assure image visibility by reducing overlap. Two other techniques are used to support the user. Clustering is employed to give users overview of the data and active learning is used to speed up the interaction process based on relevance feedback from the user.

4 Conclusion

We have presented the Mediamill video search engine and its four browsers. The ConceptBrowser allows intuitive concept based queries. The CrossBrowser is defined for those cases where there is a direct relation between the information need and one of the concepts in the lexicon. If a more complex relation between the need and the lexicon is present, the SphereBrowser is most appropriate. Finally, when there is no semantic relation, we have to interact directly with visual similarity space and this is supported in the GalaxyBrowser.

References

1. Snoek, C., Worrind, M., van Gemert, J., Geusebroek, J., Koelma, D., Nguyen, G., de Rooij, O., Seinstra, F.: Mediamill: Exploring news video archives based on learned semantics. In: ACM Multimedia, Singapore (2005)
2. Smeaton, A.: Large scale evaluations of multimedia information retrieval: The TRECVid experience. In: CIVR. Volume 3569 of LNCS. (2005)
3. Christel, M., Hauptmann, A.: The use and utility of high-level semantic features. In: CIVR, LNCS. Volume 3568. (2005)
4. Rautiainen, M., Ojala, T., Seppnen, T.: Clustertemporal browsing of large news video databases. In: IEEE International Conference on Multimedia and Expo. (2004)
5. Adcock, J., Cooper, M., Girgensohn, A., Wilcox, L.: Interactive video search using multilevel indexing. In: Conference on Image and Video Retrieval, LNCS. Volume 3568. (2005)
6. Heesch, D., Ruger, S.: Three interfaces for content-based access to image collections. In: Conference on Image and Video Retrieval, LNCS. Volume 3115. (2004)
7. Geusebroek, J.: Distinctive and compact color featuress for object recognition (2005) Submitted for publication.
8. Snoek, C., et al.: The MediaMill TRECVID 2005 semantic video search engine. In: Proc. TRECVID Workshop. NIST (2005)
9. Nguyen, G., Worring, M.: Similarity based visualization of image collections. In: Proceedings of 7th International Workshop on Audio-Visual Content and Information Visualization in Digital Libraries. (2005)