Lexicon-based Browsers for Searching in News Video Archives

M. Worring [*], C.G.M. Snoek, D.C. Koelma, G.P. Nguyen, O. de Rooij
Intelligent Systems Lab Amsterdam, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
worring@science.uva.nl, www.mediamill.nl

Abstract

In this paper we present the methods and visualizations used in the 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. The search system is evaluated within the TRECVID benchmark. 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. 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 much poorer. Indexing videos with semantic visual concepts is more appropriate.

In an ideal system there is a lexicon containing thousands of semantic visual concepts accurately detected in the video collection. The semantic gap between the concepts and the data, however, dictates that this is not realistic. Small lexicons with some of the concepts having high accuracy, is the best one can currently hope for. When concepts are not in the lexicon, or when the accuracy is limited the burden of finding relevant video fragments is still with the user. The user should interactively find her way through the collection.

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 [2]. In [6], clustering is used to improve the presentation of results to the user. Both [2] and [6] use simple grid based visualizations. More advanced visualization tools are employed in [1] and [4] based on collages of keyframes and dynamically updated graphs respectively, but no semantic lexicon is used there.

In this paper we present the MediaMill 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. To demonstrate the effectiveness of our interactive search engine, it is evaluated in the 2005 NIST TRECVID video retrieval benchmark, the standard for this field.

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

2.1 Visual similarity space

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 [3] (See also [8]). Wiccest features combine color invariance with natural image statistics. Color invariance aims to remove accidental lighting conditions, while natural image statistics efficiently represent image data.

The next step in the indexing is to compute a visual similarity function S_v allowing comparison of different shots in A. For this the function described in [3, 8] 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.

2.2 Semantic similarity space

Computing visual features and similarity is common practice in all interactive content based video retrieval systems. We now move on to the more specific topic of adding semantic indexing

[*] This work is partly sponsored by the BSIK MultimediaN project

to the data, which is the process of associating every shot s_j in the database with a measure of presence P_i^j for concept i.

The central assumption in our semantic indexing architecture is that any broadcast video is the result of an authoring process. When we want to extract semantics from a digital broadcast video this authoring process needs to be reversed. For authoring-driven analysis we proposed the semantic pathfinder [9], composed of three analysis steps. It follows the reverse authoring process. Each analysis step in the path detects semantic concepts. In addition, one can exploit the output of an analysis step in the path as the input for the next one. The semantic pathfinder starts in the *content analysis step*. In this analysis step, we follow a data-driven approach, using both visual and textual information, of indexing semantics. The *style analysis step* is the second analysis step. Here we tackle the indexing problem by viewing a video from the perspective of production. Finally, to enhance the indexes further, in the *context analysis step*, we view semantics in context. One would expect that some concepts, like *vegetation*, have their emphasis on content where the style (of the camera work that is) and context (of concepts like *graphics*) do not add much. In contrast, more complex events, like *people walking*, profit from incremental adaptation of the analysis to the intention of the author. The virtue of the semantic pathfinder is its ability to find the best path of analysis steps on a per-concept basis. The generic indexing structure is used to create a lexicon of 101 concepts so that every shot s_i is described by the vector $P^i = \{P_1^i, P_2^i, ..., P_{101}^i\}$. 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 semantic similarity function S_C to compare shots, now on the basis of their semantics. This yields the *semantic similarity space*.

$$S_C(s_j, s_k) = \sum_{i=1}^{101} min(P_i^j, P_i^k)$$

2.3 Semantic thread space

The semantic similarity space induced by S_C is complex as shots can be related to several concepts. Therefore, we propose to 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 question is how to select the different elements which constitute the path and the ordering of those elements.

When the whole collection is considered, the first obvious ordering is time. So our first thread is the time thread T^t. A complete set of threads $T^l = \{T_1^l, ..., T_{101}^l\}$ on the whole collection is defined by the concepts in the lexicon. The ranking based on P provides the ordering.

Now how to proceed if we want to compute semantic threads based on the semantic similarity space, but which are not in one to one correspondence with an element in the lexicon? This requires to consider the whole space and to find shots that share similar semantics. To find such groups we perform

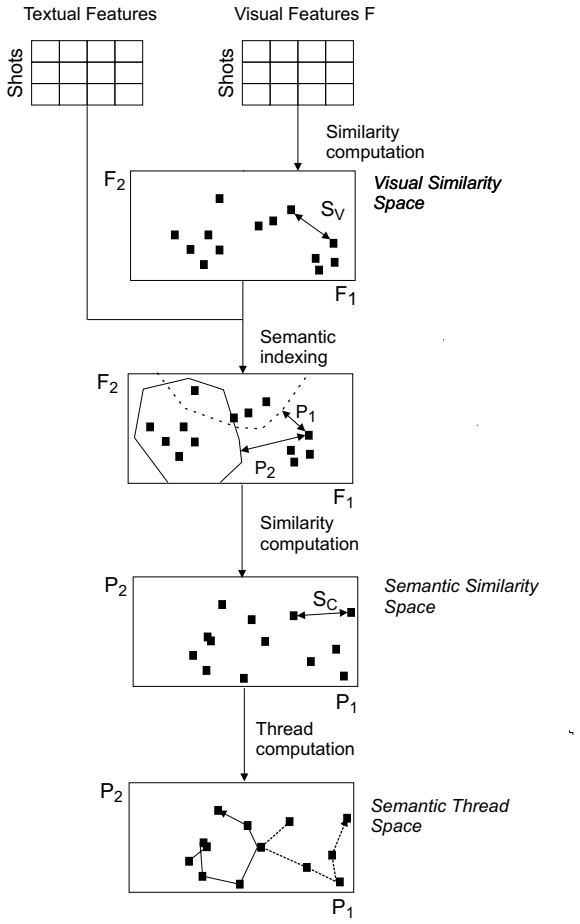

Figure 1: A simplified overview of the computation steps required to support the user in interactive access to a video collection. Note, that for both the vectors F and P only two dimensions are shown.

k-means clustering in the semantic similarity space. The elements of each group define the elements of the set of threads $T^s = \{T_1^s, ..., T_k^s\}$. Ordering of these elements is done by applying a shortest path algorithm inside the cluster. So, shots with similar semantic content are near each other in the thread.

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.

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 three different browsers, which one to use depends on the information need. 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 sim-

ilarity space and this is supported in the GalaxyBrowser. The different browsers are visualized in Fig. 2.

3.1 The CrossBrowser

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.

3.2 The SphereBrowser

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 containing this particular element. 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 through time or by navigating along a semantic thread.

3.3 The GalaxyBrowser

Browsing visual similarity space is the most difficult task as there are no obvious dimensions on which to base the display. We have developed a complete system for this purpose [5][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. When applied to images it leads to visualizations where images are overlapping one another. We therefore extended the methods to assure that images show very limited overlap. The result is a two-dimensional space where images are well visible and images next to each other have similar visual characteristics. A great advantage is that similar images are typically all relevant to the information need and can thus be selected by one user interaction.

4 Experiments

4.1 Experimental setup

We performed our experiments within the interactive search task of the 2005 NIST TRECVID benchmark. The video archive used is composed of 169 hours of US, Arabic, and Chinese broadcast news sources, recorded in MPEG-1 during November 2004. The test data contains about 85 hours. Together with the video archive came automatic speech recognition results and machine translations donated by a US government contractor. The Fraunhofer Institute provided a camera

Figure 2: Browsers in the our semantic video search engine. On the left the CrossBrowser showing results for tennis. On top the SphereBrowser, displaying several semantic threads. Bottom right: active learning using a semantic cluster-based visualization in the GalaxyBrowser.

shot segmentation. The camera shots serve as the unit for retrieval. The semantic pathfinder detects the 101 concepts with varying performance [8], based on partial annotation of the training set.

The goal of the interactive search task, as defined by TRECVID, is to satisfy an information need. Given such a need, in the form of a search topic, a user is engaged in an interactive session with a video search engine. To limit the amount of user interaction and to measure search system efficiency, all individual search topics are bounded by a 15-minute time limit. The interactive search task contains 24 search topics in total. They became known only a few days before the deadline of submission. Hence, they were unknown at the time we developed our 101 semantic concept detectors. In line with the TRECVID submission procedure, a user was allowed to submit, for assessment by NIST, up to a maximum of 1,000 ranked results for the 24 search topics.

Following the standard in TRECVID evaluations [7], we use *average precision* to determine the retrieval accuracy on individual search topics. The average precision is a single-valued measure that corresponds to the area under a recall-precision curve. As an indicator for overall search system quality TRECVID reports the mean average precision averaged over all search topics from one run by a single user.

4.2 Results

Four users participated in the search experiment. Three of them focussed on using only one browser. The fourth user mixed all browsers, but had limited experience. Results in Fig. 3 indicate that for most search topics, users of the proposed interactive retrieval system score well above average. They obtain a top-3

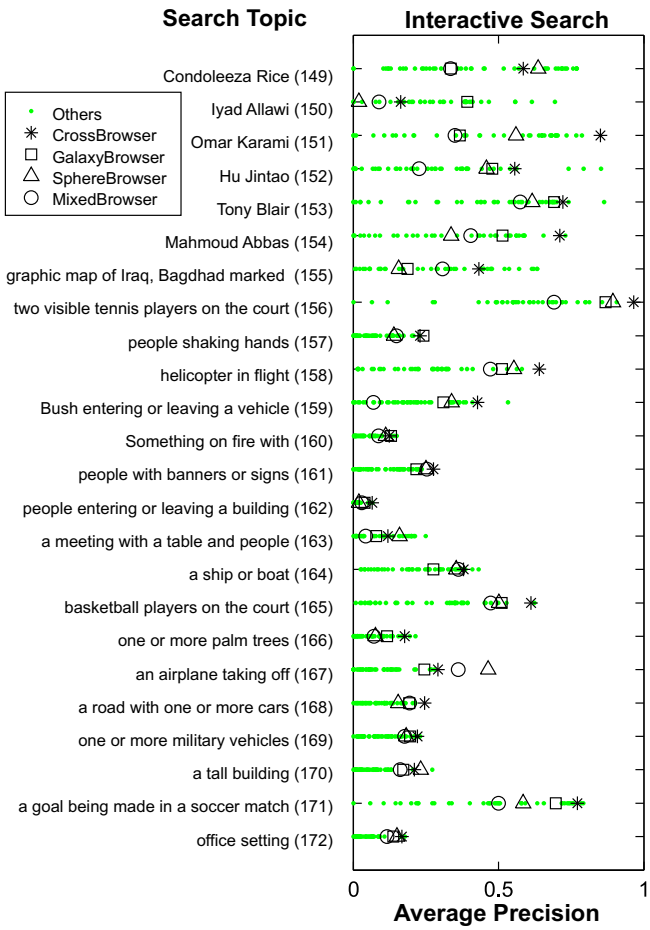

Figure 3: Interactive search results for 24 topics, results for the users of the different browsers are indicated with special markers.

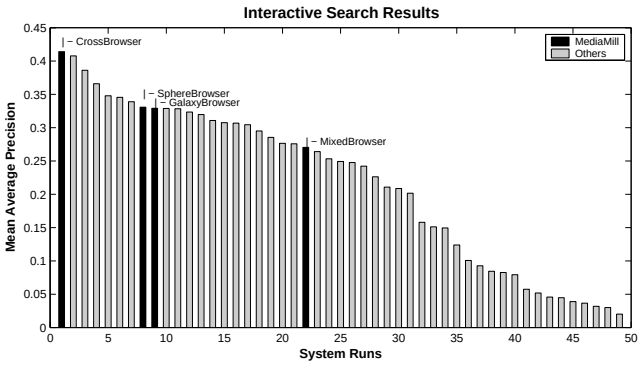

Figure 4: All results of interactive search in TRECVID 2005 ranked according to mean average precision.

average precision result for 19 out of 24 topics. Best performance is obtained for 7 topics.

To gain insight in the overall quality of our system, we compare the results of our users with all other users that participated in the retrieval tasks of the 2005 TRECVID benchmark. We visualized the results for all submitted search runs in Fig. 4.

5 Discussion and conclusion

The success of the Crossbrowser indicates that having a large lexicon of concepts, such that a direct match between information need and a lexicon concept is likely to exist, is the best method for effective search. It yields the best overall performance.

The SphereBrowser was successful when multiple semantic concepts are relevant such as *Tennis*, *People with banners or signs*, *Meeting* and *Tall building*. It was also successful for topics such as *Airplane takeoff* and *Office setting*. Here there were only a limited number of consecutive valid shots visible in each thread, but because of the combination of both time and semantic threads there was always another valid, but not yet selected, shot visible.

Finally, the GalaxyBrowser works well in case shots for an information need are visually similar e.g. topics related to *tennis*, *car* or *fire*. When topics have large variety in visual settings, for instance *person x* topics, visual features hardly yield additional information to aid the user in the interactive search process.

In conclusion we have developed a number of different browsing methods based on a lexicon of 101 concepts, where the optimal method follows from the information need and the availability of reliable concepts in the lexicon.

References

[1] J. Adcock, M. Cooper, A. Girgensohn, and L. Wilcox. Interactive video search using multilevel indexing. In *Conference on Image and Video Retrieval, LNCS*, volume 3568, 2005.

[2] M. Christel and A. Hauptmann. The use and utility of high-level semantic features. In *CIVR, LNCS*, volume 3568, 2005.

[3] J. Geusebroek and A. W. M. Smeulders. A six-stimulus theory for stochastic texture. *International Journal of Computer Vision*, 62(1/2):7–16, 2005.

[4] D. Heesch and S. Ruger. Three interfaces for content-based access to image collections. In *Conference on Image and Video Retrieval, LNCS*, volume 3115, 2004.

[5] G. Nguyen and M. Worring. Similarity based visualization of image collections. In *Proceedings of 7th International Workshop on Audio-Visual Content and Information Visualization in Digital Libraries*, 2005.

[6] M. Rautiainen, T. Ojala, and T. Seppnen. Clustertemporal browsing of large news video databases. In *IEEE International Conference on Multimedia and Expo*, 2004.

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

[8] C. Snoek et al. The MediaMill TRECVID 2005 semantic video search engine. In *Proc. TRECVID Workshop*, NIST, 2005.

[9] C. Snoek, M. Worring, J.-M. Geusebroek, D. Koelma, F. Seinstra, and A. Smeulders. The semantic pathfi nder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. in press.