Automatic video search using semantic concepts

Mats Sjöberg
mats.sjoberg@tkk.fi

Ville Viitaniemi
ville.viitaniemi@tkk.fi

Markus Koskela
markus.koskela@tkk.fi

Jorma Laaksonen
jorma.laaksonen@tkk.fi

Adaptive Informatics Research Centre
Aalto University School of Science and Technology
P.O. Box 15400, FI-00076 Aalto, Finland

ABSTRACT

This paper describes automatic video search using semantic concept detection, and how it is implemented in our PicSOM system. The demand for such methods is growing because of the exploding growth of digital video data produced today, ranging from professionally produced TV programming to individuals sharing videos online. Content-based methods address the problem of finding relevant information in large amounts of visual data, which by nature are very hard to search and index automatically. The semantic gap between the high-level concepts understood by humans and the low-level information of computer systems is partially bridged by modelling mid-level semantic concepts. The performance of our system is judged in the video retrieval setting of the international TRECVID 2008 and 2009 evaluations, comparing favourably with the competing state-of-the-art systems.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation, Performance

Keywords

video search, concept detection, support vector machine

1. INTRODUCTION

Digital video has become commonplace, both in professional use and in various consumer products such as camcorders, webcams, digital TV, mobile phones, video sharing websites, CCTV surveillance, and virtual and augmented reality applications. While the capturing, storing, and transmitting of digital video has steadily become easier and more cost-effective, the current methods for the automatic analysis and semantic representation of the video content itself are considerably less mature.

Content-based video retrieval addresses the problem of finding visual data relevant to the users' information needs from video repositories [4]. This is generally done with querying by examples and measuring the similarity of the database objects (keyframes, video clips) with *low-level features* automatically extracted from the objects. Generic low-level features are often, however, insufficient to discriminate content well on a conceptual level. This "semantic gap" is the fundamental problem in content-based multimedia retrieval.

The modelling of *mid-level semantic concepts* (events, objects, locations, people, etc.) attempts to fill, or at least reduce, the semantic gap. Indeed, in recent studies it has been observed that, despite the fact that the accuracy of the concept detectors is far from perfect, they can be useful in supporting *high-level indexing and querying* on multimedia data [1]. This is mainly because such semantic concept detectors can be trained off-line with computationally more demanding supervised learning algorithms and with considerably more positive and negative training examples than what are typically available at query time.

In this paper, we first describe briefly the relevant parts of our PicSOM multimedia analysis and retrieval framework [2], including video corpus preparation, concept detection and concept-based automatic search. We then use the large-scale experimental setup of the TRECVID evaluation campaign to relate the performance of our system to other systems in the video retrieval community for automatic search (i.e. without human input in the loop).

2. PARTS OF A VIDEO SEARCH SYSTEM

This section gives an overview of the components of a video retrieval system in general and the implementation in our PicSOM system in particular. For brevity, only a brief summary of the system is provided in this paper. More detailed descriptions can be found in our annual TRECVID workshop papers (e.g. [3]). The general architecture of the PicSOM video search system is shown in Figure 1.

The operation of a video retrieval system generally consists of two phases: an offline *preparation phase*, where a database is processed and indexed, and the online *search phase*, in which responses to queries must be generated in a reasonable time for human interaction.

In the preparing phase the video corpus is usually first segmented into shots and a number of low-level visual, audio and textual feature descriptors are extracted from each shot and content-based indices are prepared based on the features. Shot-wise detectors are trained on annotated data, either from the same corpus or they might be trained previously on similar data. The detectors apply supervised learning techniques to learn the mapping between low-level shot features and the annotation concepts. The preparation is allowed to be time-consuming as it is intended to be performed off-line prior to the on-line use of the retrieval system.

In the search phase, the system is set to respond to video

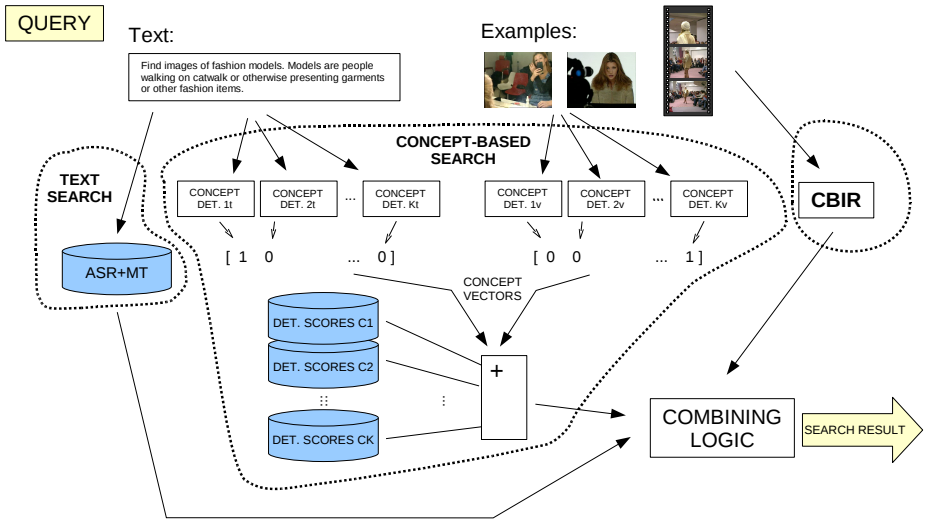

Figure 1: Overview of video search with PicSOM.

queries. The queries may constitute of textual phrases, image and video examples, or of any combination of them (see an example in the top part of Figure 1). The retrieval system can then employ some or all of the following search modalities: *text search*, *concept-based retrieval* and *content-based retrieval*. The retrieval result is a list of video shots, ranked in the order of decreasing predicted likelihood to match the query. The system operation in the search phase is intended to be sufficiently fast for interactive use.

2.1 Shot segmentation and keyframe selection

The first task of the preparing phase for a comprehensive video retrieval system is to segment the video corpus temporally into sequential basic units. Depending on the video material, such a segmentation can be performed on various levels, such as stories, events, scenes, groups, sequences, and shots. For scripted content, the basic semantic unit is the shot, as shots are intrinsically linked to the production of the video. As shots can usually be identified by automatic methods with a reasonable accuracy and they provide a suitable representation level for the higher-level video analysis tasks, the contemporary video retrieval systems customarily treat the shot as the basic unit of retrieval. In the TRECVID evaluations, a master definition of shots is provided by the organisers. In cases where the shot boundaries have not been available, we have applied one of two shot segmentation algorithms available in the PicSOM system [3].

To facilitate the usage of image-based retrieval, we also extract keyframe images from each video shot. If a video shot is short and contains only one visually homogeneous scene, a single well-chosen keyframe can compactly express the most central visual characteristics of that shot. The keyframes are also useful as still visual representations the shots when presenting a collage of several results to the user. In the PicSOM system the keyframe is selected by a heuristic method which tries to pick a "typical frame" (i.e. similar to the average of all frames) which is also close to the temporal centre of the shot, and that does not have rapid movement.The last requirement is to avoid motion blurring effects.

2.2 Low-level features

Using the video or image data directly in retrieval is typically not feasible due to the high dimensionality of the data. Extracted low-level features should thus ideally be of reasonable dimensionality and at the same time be discriminative of semantic differences in the data, i.e. the feature extractors should be sensitive to those characteristics of the raw data that are relevant to the human perception of the media contents. For video analysis in particular there is an opportunity to combine several data modalities, such as keyframe images, video motion, audio and text from speech recognition or subtitling. From these modalities diverse feature representations can be extracted to represent different relevant and complementary aspects of the underlying data.

Image features. The PicSOM system uses a wide range of image features that have been added over the years of its development. In addition to standardised MPEG-7 descriptors and some non-standard image features developed in-house the PicSOM system employs state-of-the-art bag-of-visual-words (BoV) features, in which images are represented by histograms of local image descriptors.The latter category includes in particular SIFT-based descriptors together with the Harris-Laplace interest point detector or dense sampling. These features are described further in [3].

Video features. In many cases the static visual properties of a single video frame are not enough to describe the salient features of the full scene. In some situations the motion of objects or the camera might be semantically significant, for example in distinguishing between a ball that is rolling from one that is still. Also, the dynamic properties may in some cases make the computational learning problem easier. For example it may be easy for a human to recognise a running person even from a still keyframe image, but such videos are surely easier to distinguish based on the temporal properties of the person moving across the scene. The set of video features used in PicSOM include temporal extensions to some of the still-image features, and specialized motion features [3].

Audio features. Most video shots include a sound track, containing for example music or different environment sounds.

A certain distinctive musical tune may perhaps indicate the beginning of a news broadcast, or indicate for example the occurrence of an action scene in a movie. A crowd cheering in a football game is a strong cue of an important event such as a goal being scored. In PicSOM we have used implementations of the popular mel-scaled cepstral coefficients feature (MFCC) as an audio feature.

2.3 Textual search

Often, the video material includes textual data or meta-data that can facilitate text-based indexing and retrieval. Textual data for video shots may originate e.g. from speech recognition, closed captions, subtitles, or video OCR. The textual search module can easily be implemented as a separate component whose output is fused afterwards with the other modalities. The module can then utilise all common text processing methods, such as stemming and part-of-speech tagging, and any of the existing tools for text-based indexing. In the experiments of this paper, we use textual search for indexing text obtained with automatic speech recognition and machine translation.

2.4 Concept detection

After having extracted low-level video features from each shot, supervised learning techniques are applied for learning the associations between the low-level features and the concepts in the annotations of the video corpus. Figure 2 illustrates the overall architecture of the concept detection module of the PicSOM video database search system. In this architecture, concepts are first detected independently from each video shot. Shot-wise concept detection is achieved by fusing together the outcomes of numerous elementary detectors, each of which is based a different visual feature of the shot. The feature-wise elementary detectors build upon the support vector machine (SVM) supervised learning algorithm. See [3] for more details.

Following the shot-wise detection phase, the detector outcomes are post-processed so that advantage is taken of the correlations that contents of temporally adjacent video shots often exhibit. Technically, this is achieved via concept-wise inter-shot N-gram modelling [6].

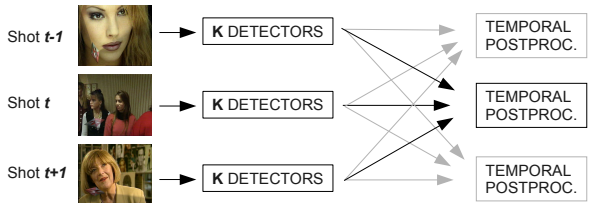

Figure 2: Overview of concept detection.

2.5 Video search

The goal of video retrieval is to find relevant video content for a specific information need of the user. The conventional approach has been to rely on textual descriptions, keywords, and other meta-data to achieve this functionality, but this requires manual annotation and does not usually scale well to large and dynamic video collections.

Content-based video retrieval, on the other hand, utilises techniques from related research fields, such as image and audio processing, computer vision, and machine learning, to automatically index the video material with low-level features. Content-based queries are typically based on provided examples (i.e. *query-by-example*). The video collection is ranked based on its similarity to the examples. See [2] for a description of the content-based retrieval methodology employed in the PicSOM system.

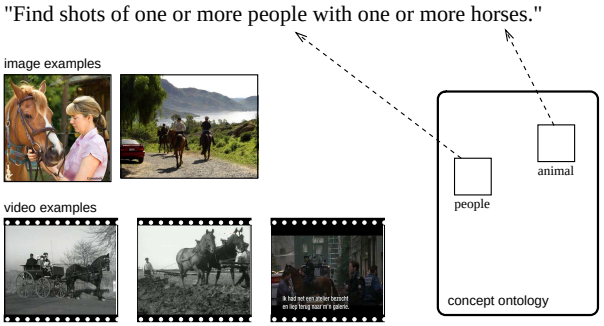

Figure 3: An example TRECVID search topic, with a possible concept mapping from a concept ontology.

In concept-based video retrieval, the fundamental problem is how to map the user's information need into the space of available concepts in the used concept ontology. The basic approach is to select a small number of concept detectors as active and weight them based either on the performance of the detectors or their estimated suitability for the current query. We have employed two methods for automatic concept selection: *text-based* and *visual-example-based*. The text-based approach uses lexical analysis and synonym lists to match the words of the textual query to the available concepts. In the visual-example-based method the PicSOM system measures the similarity of the given visual examples (images and/or videos) to the concept detectors, answering the question of how visually similar are the examples to the annotated training set. See [3] for details.

3. EXPERIMENTS IN TRECVID

The video material and the search topics used in these experiments are from the TRECVID evaluations [5] in 2008–2009. TRECVID is an annual workshop series organised by the National Institute of Standards and Technology (NIST) and arguably the leading venue for evaluating research on content-based video analysis and retrieval. It provides the participating organisations large test collections, uniform scoring procedures, and a forum for comparing the results. Each year the TRECVID evaluation contains a set of video analysis tasks, such as semantic concept detection, video search, video summarisation, and content-based copy detection. In the experiments of this paper, we focus on the settings of the automatic video search task.

In 2008–2009 the type of video material used in TRECVID consisted of documentaries, news reports, and educational programming from Dutch TV. The video data is always divided into separate development and test sets. The same development set of approximately 100 hours in length is used both in 2008 and 2009. The amount of test data was approximately 100 and 280 hours in 2008 and 2009, respectively. The training data for semantic concept detection is obtained with a collaborative annotation effort among the participants.

NIST also defines sets of standard search topics for the

video search tasks and then evaluates the results submitted by the participants. The search topics contain a textual description along with a small number of both image and video examples of an information need. Figure 3 shows an example of a search topic, including a possible mapping of concept detectors from a concept ontology based on the textual description. The number of topics evaluated for automatic search was 48 in the year 2008 and 24 in 2009. Due to the limited space, the search topics are not listed here, but are available in the TRECVID guidelines documents[1].

The video material used in the search tasks is divided into shots in advance and these reference shots are used as the unit of retrieval. The shot segmentation step resulted in 36 000 shots for the training set and 97 000 shots for the full 2009 test set. The output from an automatic speech recognition (ASR) software is provided to all participants. In addition, the ASR result from all non-English material is translated into English by using automatic machine translation. It is therefore quite unsurprising that the quality of the textual data is remarkably poor and pure text queries can only obtain a very modest performance.

Search performance in TRECVID is measured using *mean average precision* (MAP) and *mean inferred average precision* (MIAP). MIAP is an approximation of MAP, but requires only a subset of the results to be evaluated manually.

In Figure 4 the automatic search performance of the current PicSOM system is compared with the top automatic search systems submitted to TRECVID 2008 and 2009. Our own submissions in those years are included as well (with the year indicated). We see that the PicSOM system compares very well with the other top systems.

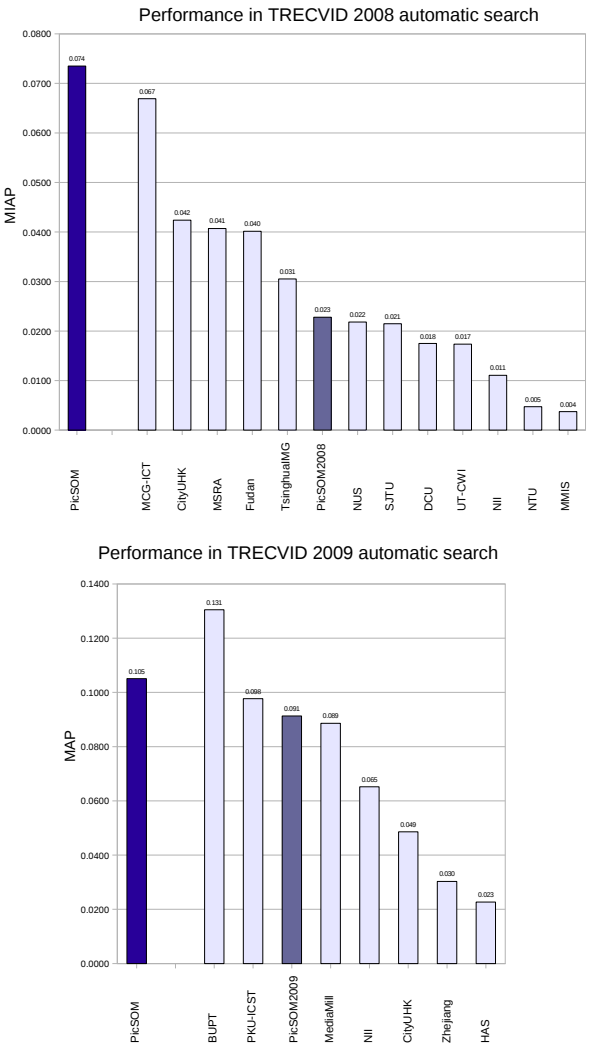

Figure 4: MIAP/MAP performance in TRECVID

4. CONCLUSIONS

This paper has given an overview of video search using the PicSOM system utilising concept-detection for helping to bridge the semantic gap between the low-level features and the high semantic level of a human query. By comparing with other state-of-the-art systems in the TRECVID 2008 and 2009 competitions we have demonstrated the competitiveness of the proposed retrieval method for automatic search.

Automatic video search systems can find practical application in web-wide search engines and in the public archives of a TV station or other content-producing institutions. For example, the data set of previous TRECVID evaluations have been TV broadcasts and the upcoming TRECVID 2010 data comes from the Internet Archive's collection of public domain videos. Providing efficient access to the growing amount of visual data created today is not only a matter of convenience, but also crucial for the preservation of our contemporary culture. Only information that can be found and retrieved with reasonable effort is useful in practice.

5. REFERENCES

[1] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[2] J. Laaksonen, M. Koskela, and E. Oja. PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks*, 13(4):841–853, July 2002.

[3] M. Sjöberg, V. Viitaniemi, M. Koskela, and J. Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[4] A. F. Smeaton. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, 32(4):545–559, 2007.

[5] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006.

[6] V. Viitaniemi, M. Sjöberg, M. Koskela, and J. Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of WIAMIS 2008*, pages 12–15, Klagenfurt, Austria, May 2008.

[1] *http://www-nlpir.nist.gov/projects/trecvid/*

ACKNOWLEDGEMENT

This work has been supported by TKK MIDE programme, project UI-ART.