Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Zhirong Yang, Ville Viitaniemi, Mats Sjöberg, He Zhang

7.1 Introduction

The Content-Based Information Retrieval Research Group studies and develops efficient methods for content-based information retrieval (CBIR) and analysis tasks and implements them in the PicSOM[1] CBIR system. During the years 2008 and 2009, the PicSOM search engine has been used in various old and new applications.

In the PicSOM CBIR system, parallel Self-Organizing Maps (SOMs) and Support Vector Machine (SVM) classifiers have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different classifiers and their underlying feature extraction schemes impose different similarity functions on the images, videos, texts and other media objects. In the PicSOM approach, the system is able to discover those of the parallel classifiers that provide the most valuable information for retrieving relevant objects in each particular query.

7.2 Semantic concept detection from images and videos

Extracting semantic concepts from multimedia data has been studied intensively in recent years. The aim of the research on the multimedia retrieval research community has been to facilitate semantic indexing and concept-based retrieval of unannotated multimedia content. The modeling of mid-level semantic concepts is often essential in supporting high-level indexing and querying on multimedia data as such concept models can be trained off-line with considerably more positive and negative examples than what are available at query time.

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for multimedia retrieval tasks. Detection of concepts from multimedia data—e.g. images and video shots—forms an important part of the architecture. In the PicSOM system, concept detection problem is formulated as a standard supervised learning problem. Our concept detection technology is fundamentally based on fusion of a large number of elementary detections, each based on a different low-level audiovisual feature extracted from the multimedia data. During the reporting period 2008–2009 we have continued to develop the technologies involved in the concept detection component of the PicSOM architecture. We have also successfully evaluated the concept detection performance of the PicSOM architecture by participating in international evaluation campaigns. These include PASCAL NoE Visual Object Classes (VOC) Challenge 2008 image analysis evaluation [1] and the annual TRECVID video analysis evaluations [2, 3]. Figure 7.1 illustrates the architecture for detecting concepts from video shots that was used in our TRECVID 2009 system.

We have recently enriched the set of audiovisual features that we use as the basis for for concept detection. As a part of that work, we have studied various aspects and extensions of the bag of visual words (BoV) model. In the BoV model images are represented with histograms of local image features. In our studies of BoV features we have addressed e.g. methods for quantisation of local image features [4, 5, 6], their distance measures [7] and spatial extensions of the BoV methodology [8]. In addition to feature extraction, we have continuosly developed the techniques for feature-wise elementary detection and detector fusion. We have also developed inter-shot temporal and cross-concept techniques for taking into account the dependecies that temporally adjacent video shots on one hand, and related concepts on the other hand typically exhibit [9].

[1] http://www.cis.hut.fi/picsom

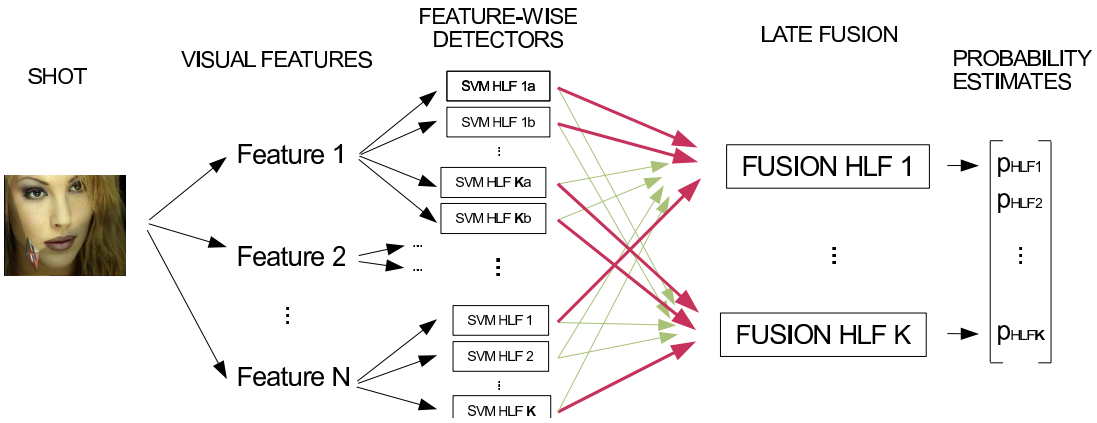

Figure 7.1: Fusion-based shot-wise concept detection module of the PicSOM system that was used for participating in the TRECVID 2009 evaluation.

7.3 Video search and retrieval

In recent studies it has been observed that, despite the fact that the accuracy of the concept detectors may be far from perfect, they can still be useful in supporting *high-level indexing and querying* on multimedia data [10]. We have found this to be true in particular for video search [11]. This is mainly because such semantic concept detectors can be trained off-line with computationally more demanding algorithms and considerably more training examples than what are typically available during interactive use.

Figure 7.2 gives an overview of the automatic video search process within PicSOM, with a detailed view of the concept-based submodule. In the top part of the figure a search query is presented, typically containing a *text query* and possibly also *visual examples*. The visual examples may consist of videos and/or images, demonstrating the visual properties of the desired retrieval response. Either or both of these two modalities of the search query are then used as input to the three parallel submodules of the search system: *text search*, *concept-based search* and *content-based search*. Based on its input, each module produces an estimate of the relevance of each database video to the given query. These scores are finally fused to produce the final search result which is a list of video shots ordered with decreasing estimated relevance to the query.

An important catalyst for research in video retrieval is provided by the annual TREC Video Retrieval Evaluation (TRECVID) workshop. The goal of the workshop series is to encourage research in multimedia retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested to compare their results. The search task in TRECVID models the task of an intelligence analyst who is looking for specific segments of video containing persons, objects, events, locations, etc. of current interest. The task is defined as follows: given a search test collection and a multimedia statement of information need, return a ranked list of shots which best satisfy the need.

We have successfully participated in TRECVID annually since 2005. In TRECVID 2008 we participated in the high-level feature extraction, automatic search, video summarization, and video copy detection tasks, using the PicSOM system framework [12]. In the high-level feature extraction experiments, we used SOM-based semantic concept modeling followed by a post-processing stage that utilizes the concepts' temporal and inter-concept co-occurrences. We also studied the effects of a more comprehensive feature selection scheme and the inclusion of audio features and face detection. The results show that more thorough feature selection can be useful, and that the temporal and inter-concept

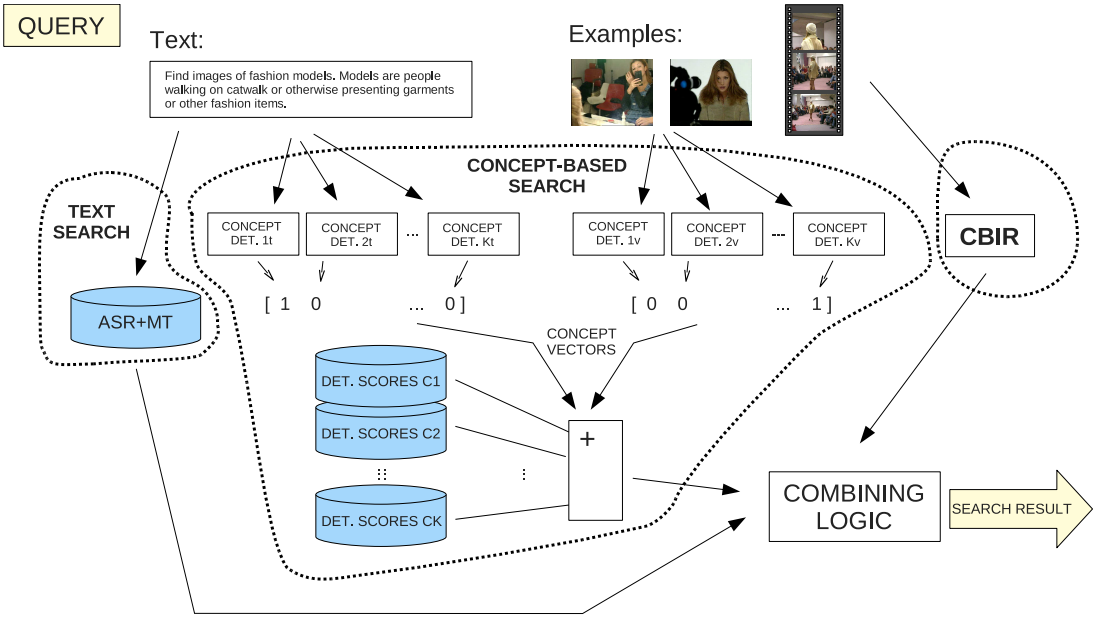

Figure 7.2: General architecture of the PicSOM search module.

co-occurrence analysis has the potential to improve the performance if good concept-wise post-processors can be chosen [13]. The use of audio features and face detection resulted in minor improvements. In TRECVID 2009 we used greatly improved Support Vector Machine (SVM) detectors for high-level feature extraction task, and our group had the sixth best result among the 20 participanting groups. These results could then be used in the search task, where we achieved the third best result.

7.4 Video analysis applications

In this section, two further applications of the content-based video analysis framework are described. These applications are *automatic video summarization* and *analysis of sign-language*.

Video summarization is a process where an original video file is converted to a considerably shorter form, which can then be used to facilitate efficient searching and browsing in large video collections. The aim of automatic summarization is to preserve as much as possible from the essential content and overall structure.

We have developed a technique for video summarization [14] using SOMs trained with standard visual features that have been applied in various multimedia analysis tasks. The produced summaries consist of collections of selected video clips from the original material. The method is based on initial shot segmentation, with the shots used in the following stages as basic units of processing. We then detect and remove unwanted "junk" shots, and apply face detection, speech detection, and motion activity estimation. Next, we compute the visual similarities between all pairs of shots and remove overly similar shots. We trace the trajectory of the frames within the shot in question and use the trajectory as a signature for the shot, which can then be compared to other shots' signatures to determine whether a shot is visually unique or similar to some other shots. Each remaining shot is then represented in the summary with a separately selected one-second clip.

We participated in the TRECVID 2008 rushes summarization task [2] and obtained

very promising results. Our summarization algorithm obtained average ground-truth inclusion performance with the shortest overall summaries over all the submissions.

We have also applied our methods for video content analysis in a multidisciplinary research project for the recognition and analysis of recorded Finnish Sign Language [15]. Automatic and semi-automatic computer vision techniques are used to recognize and analyze gestures and facial expressions in sign language videos (see Figure 7.3). The aim is to identify linguistic sign and gesture boundaries and to indicate which video sequences correspond to specific signs and gestures. This will facilitate indexing and the construction of an example-based open-access visual corpus of the Finnish Sign Language for which there already exists large amounts of non-indexed video material.

Figure 7.3: An example frame from the sign language video material. Left: Skin-color filtering. Right: Motion tracking.

References

[1] Ville Viitaniemi and Jorma Laaksonen. Techniques for image classification, object detection and object segmentation. In Monica Sebillo, Giuliana Vitiello, and Gerald Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 231–234, Salerno, Italy, September 2008. Springer.

[2] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008.

[3] Mats Sjöberg, Ville Viitaniemi, Markus Koskela, and Jorma Laaksonen. PicSOM experiments in TRECVID 2009. In *Proceedings of the TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[4] Ville Viitaniemi and Jorma Laaksonen. Experiments on selection of codebooks for local image feature histograms. In Monica Sebillo, Giuliana Vitiello, and Gerald Schaefer, editors, *Proceedings of the 10th International Conference on Visual Information Systems (VISUAL 2008)*, volume 5188 of *Lecture Notes in Computer Science*, pages 126–137, Salerno, Italy, September 2008. Springer Verlag.

[5] Marcin Blachnik and Jorma Laaksonen. Image classification by histogram features created with learning vector quantization. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'08)*, pages 827–836, September 2008.

[6] Ville Viitaniemi and Jorma Laaksonen. Combining local feature histograms of different granularities. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 636–645, Oslo, Norway, 2009. Springer Verlag.

[7] Ville Viitaniemi and Jorma Laaksonen. Representing images with χ^2 distance based histograms of SIFT descriptors. In *Proceedings of 19th International Conference on Artificial Neural Networks (ICANN 2009)*, volume 5768 of *Lecture Notes in Computer Science*, pages 636–645, Limassol, Cyprus, 2009. Springer Verlag.

[8] Ville Viitaniemi and Jorma Laaksonen. Spatial extensions to bag of visual words. In *Proceedings of ACM International Conference on Image and Video Retrieval (CIVR 2009)*, Fira, Greece, July 2009.

[9] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, Austria, May 2008.

[10] Alexander G. Hauptmann, Michael G. Christel, and Rong Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, April 2008.

[11] Markus Koskela, Mats Sjöberg, and Jorma Laaksonen. Improving automatic video retrieval with semantic concept detection. In *Proceedings of 16th Scandinavian Conference on Image Analysis (SCIA 2009)*, volume 5575 of *Lecture Notes in Computer Science*, pages 480–489, Oslo, Norway, 2009. Springer Verlag.

[12] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. PicSOM experiments in TRECVID 2008. In *Proceedings of the TRECVID 2008 Workshop*, pages 408–416, Gaithersburg, 2008.

[13] Ville Viitaniemi, Mats Sjöberg, Markus Koskela, and Jorma Laaksonen. Exploiting temporal and inter-concept co-occurrence structure to detect high-level features in broadcast videos. In *Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008)*, pages 12–15, Klagenfurt, 2008.

[14] Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Ville Viitaniemi, and Hannes Muurinen. Rushes summarization with self-organizing maps. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 45–49, Augsburg, Germany, September 2007. ACM Press. Available online at `http://dx.doi.org/10.1145/1290031.1290039`.

[15] Markus Koskela, Jorma Laaksonen, Tommi Jantunen, Ritva Takkinen, Päivi Rainò, and Antti Raike. Content-based video analysis and access for finnish sign language – a multidisciplinary research project. In *Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages at 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 101–104, Marrakech, Morocco, May-June 2008.