Chapter 7

Content-based information retrieval and analysis

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Zhirong Yang, Mats Sjöberg, Hannes Muurinen

7.1 Introduction

Content-based image or information retrieval (CBIR) has been a subject of intensive research effort for more than a decade now. Content-based retrieval of images differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems.

In our PicSOM[1] CBIR system, parallel Self-Organizing Maps (SOMs) have been trained with separate data sets obtained from the multimodal object data with different feature extraction techniques. The different SOMs 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 SOMs that provide the most valuable information for retrieving relevant objects in each particular query.

7.2 Benchmark tasks of natural image content analysis

In the course of previous years we have outlined and implemented our generic PicSOM system architecture for image and information retrieval tasks. The architecture is based on extraction of numerous different features from the feature descriptors from the information objects, performing inference separately based on each feature, and fusing the partial inferences. The architecture supports hierarchical organization of the information objects. In the case of image analysis, the hierarchy is used to describe the decomposition of images into segments.

We have investigated how our architecture can be applied to various benchmark tasks concerning generic domain photographic images. While individual components of the architecture have been improved during the studies, the general architecture has proven to be successful. The improvements include the incorporation of new feature extraction methods, most notably the Scale-Invariant Feature Transform (SIFT) features calculated from interest points, the use of Support Vector Machines (SVMs) as an alternative to SOMs as the classification method, and alternative early and late feature feature fusion methods.

Our group has participated in the annual PASCAL FP6 NoE Visual Object Classes (VOC) Challenges [1, 2]. The material of the Challenges consists of photographic images of natural scenes containing objects from predefined object classes. In 2006 there were approximately 5000 images and ten object classes, including objects such as “bicycle”,”bus”,”cat” and “cow”. For the 2007 Challenge the number of images and object classes were both doubled. The Challenge included the classification task, ie. the determination whether an object of a particular class appears in the image, and the detection task for the object’s bounding box. In addition, the 2007 Challenge also included a novel competition of pixel-wise object segmentation. Our performance in the Challenge has been satisfactory, the highlights being the best segmentation accuracy and the fourth best classification performance in the 2007 Challenge.

For the VOC benchmarks we have investigated and analyzed techniques of automatic image segmentation, especially in [3]. The devised techniques have been fundamental for performing the bounding box detection tasks. However, for the classification task the usefulness of segmented images does not currently seem to be competitive against state-of-the-art global image analysis techniques. Partly this is due to the strong correlation of

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

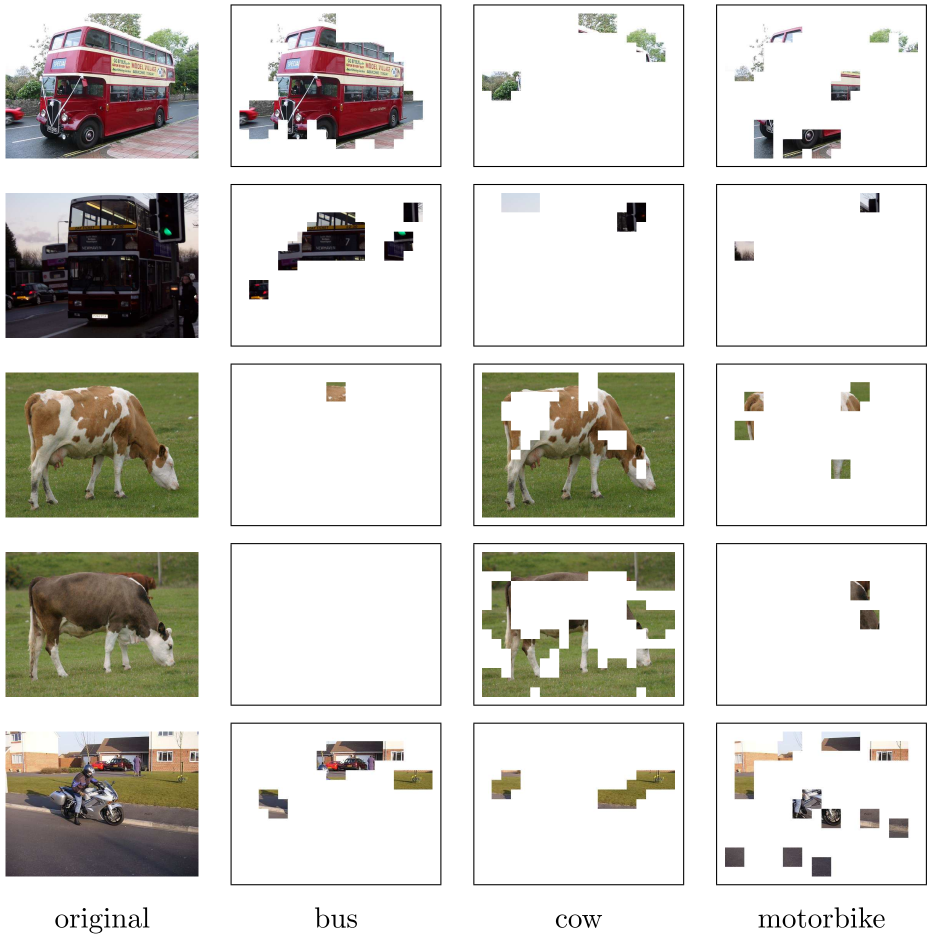


Figure 7.1: Images of the VOC2006 image collection shown together with those patches that on the collection level contribute most to the classification of the images as a “bus”, “cow” and “motorbike”.

actual target objects and the background both in the challenge databases and in natural images in general, which diminishes the advantage from focusing analysis exclusively to specific image locations. This effect is illustrated in Figure 7.1 where we have highlighted the image patches that contribute most to the decision of the image containing a particular object in the classification task.

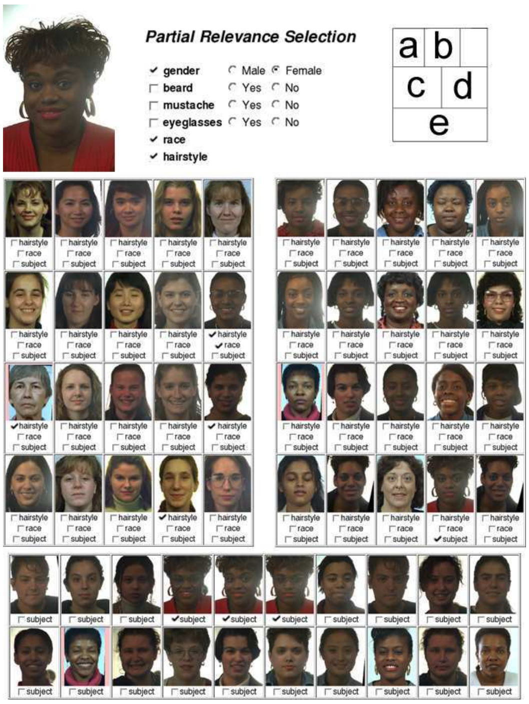
Other benchmark tasks we have studied include the ImageCLEF 2006 object annotation task, which we analyzed outside the competition, and the ImageCLEF 2007 object retrieval task, in which our results were clearly the best of the campaign submissions. We have also applied our CBIR system to benchmark tasks of automatic image annotation, performing clearly better than numerous state-of-the-art methods reported in literature [4, 5].

7.3 Interactive facial image retrieval

It is often desired to search for an image depicting a person only through an eyewitness' recalling about the appearance. Interactive computer-based systems for this purpose, however, confront the problem of evaluation fatigue due to time-consuming retrieval. We have addressed this problem by extending our PicSOM CBIR system to emphasize the early occurrence of the first subject image. Partial relevance criteria provide a common language understood by both the human user and the computer system. In addition to filtering by ground truth and hard classifier predictions, we have proposed Discriminative Self-Organizing Maps (DSOMs) [6] to adaptively learn the partial relevances.

A straightforward method to obtain DSOMs is to employ discriminant analysis as a preprocessing step before normal SOM training. We have applied the widely used method, PCA+LDA, in pattern recognition as our baseline. Furthermore, we have adapted the Informative Discriminant Analysis (IDA) to maximize the discrimination for more complicated distributions. Our Parzen Discriminant Analysis [7] regularizes the IDA objective by emphasizing the prior of piecewise smoothness in images. Both LDA and our PDA have been extended for handling fuzzy cases. The original IDA optimization algorithm is computationally expensive. We have presented three acceleration strategies [8]: First, the computation cost of batch gradients is reduced by using matrix multiplication. Second, the updates follow an geodesic flow in the Stiefel manifold without Givens reparameterization. Third, a more efficient leading direction is calculated by preserving only the principal whitened components of the batch gradient at each iteration.

Simulations have been performed on the FERET database. We have provided a query example (Figure 7.2) and also presented a quantitative study on the advantage in terms of the first subject hit and retrieval precisions at various recall levels [6].



Partial Relevance Selection

✓ gender ○ Male ● Female
□ beard ○ Yes ○ No
□ mustache ○ Yes ○ No
□ eyeglasses ○ Yes ○ No
✓ race
✓ hairstyle

Figure 7.2: A query example using the PicSOM system: (a) the target person; (b) specifying partial relevance; the displayed images in the first phase, with (c) the first round and (d) the second round; (e) the images displayed in the first round of the second phase.

7.4 Content analysis and change detection in earth observation images

Earth observation (EO) data volumes are growing rapidly, with an increase in both the number of satellite sensors and in their resolutions. Yet, it is estimated that only 5% of all EO data collected up to now has been used. Therefore, traditional remote sensing archiving systems – with queries made typically on sensor type, geographical extents or acquisition date – could become obsolete as the amount of data to be stored, accessed and processed explodes. Using image content indexing would allow a more efficient use of these databases. This has led to the emergence of content-based image retrieval systems for archive management of remote sensing images and for annotation or interpretation of satellite images. In co-operation with the VTT Technical Research Centre of Finland, we have applied the PicSOM system for analysis of multispectral and polarimetric radar (PolSAR) satellite images divided in small patches or *imagelets*.

With the high-resolution optical images the aim has been to detect man-made structures and changes on the studied land cover. Fusion of panchromatic and multispectral information was done conveniently within the PicSOM framework, in which several SOMs are trained in parallel, one SOM per feature. Qualitative and quantitative evaluation of the methods were carried out for man-made structure detection and change detection, using partially labeled datasets. The results were encouraging, considering that a totally new approach was presented to the challenging problem of change detection in very high-resolution images [9]. Possible applications of this work are high-resolution satellite image annotation and monitoring of sensitive areas for undeclared human activity, both in an interactive way.

With the radar images, the availability of dual-polarization and fully-polarimetric data, instead of earlier single-polarization data, will in the near future enable a deeper analysis of backscattering processes. This development will in turn pave the way for many new applications for spaceborne SAR data. At the same time, these satellite missions generate a huge amount of data at a higher resolution than previous spaceborne SAR sensors. It is still quite unclear what low-level features will be the most efficient ones for the automatic content analysis of the satellite polarimetric SAR data. In our research [10] we have compared six different types of polarimetric features and their different postprocessings, including averages and histograms, to gain quantitative knowledge of their suitability for the land cover classification and change detection tasks. The results proved that different features are most discriminative for different land cover types, and the best overall performance can be obtained by using a proper combination of them.

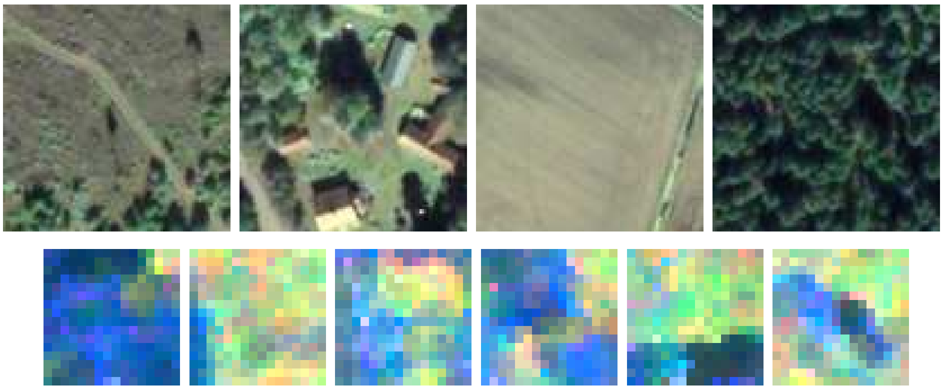


Figure 7.3: 100×100-pixel optical and 16×16-pixel SAR (Pauli decomposition) imagelets.

7.5 Multimodal hierarchical objects in video retrieval

The basic ideas of content-based retrieval of visual data can be expanded to multimodal data, where we consider multimodal objects, for example video or images with textual metadata. The PicSOM system has been extended to support general multimodal hierarchical objects and to provide a method for relevance sharing between these objects [11]. For example a web page with text, embedded images and links to other web pages can be modeled as a hierarchical object tree with the web page as the parent object and the text, links and images as children objects. The relevance assessments originally received from user feedback will then be transferred from the object to its parents, children and siblings. For example, if we want to search for an image of a cat from a multimedia message database, we can let the system compare not only the images, but also the related textual objects. If the reference message text contains the word “cat” we can find images which are not necessarily visually similar, but have related texts containing the same keyword.

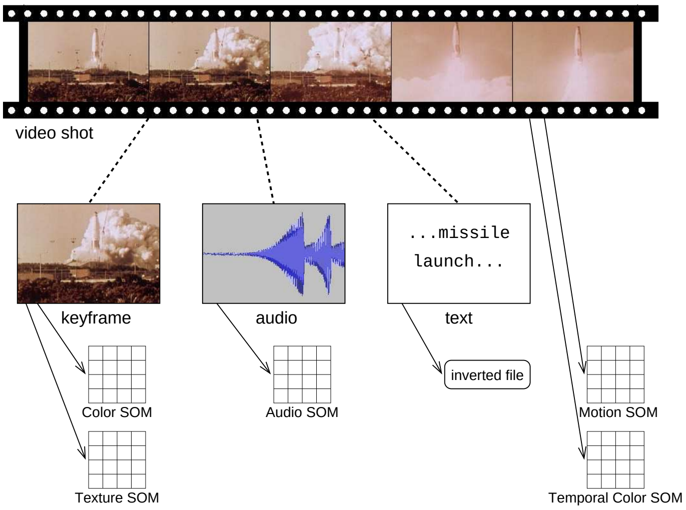


Figure 7.4: The hierarchy of video and multimodal SOMs.

The multimodal hierarchy used for indexing video shots and supporting multimodal fusion between the different modalities is illustrated in Fig. 7.4. The video shot itself is considered as the main or parent object in the tree structure. The keyframes (one or more) associated with the shot, the audio track, and text obtained with automatic speech recognition are linked as children of the parent object. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy can have links to a set of associated feature indices.

A common approach to semantic video retrieval is to combine separate retrieval results obtained with low-level visual features and text-based search. The relative weights of these sub-results are specified based on e.g. validation queries or query categorization.

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 information 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 [12, 13].

7.6 Semantic concept detection

Extracting semantic concepts from visual data has attracted a lot of research attention recently. The aim of the research has been to facilitate semantic indexing and concept-based retrieval of unannotated visual content. The leading principle has been to build semantic representations by obtaining intermediate semantic levels (objects, locations, events, activities, people, etc.) from automatically extracted low-level features. The modeling of mid-level semantic concepts can be useful 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.

We treat semantic concept detection from shot-segmented videos as a general supervised classification task by utilizing the hierarchical approach shown in Fig. 7.4 and by extracting multiple low-level features from the different data modalities [14]. A set of SOMs is trained on these features to provide a common indexing structure across the different modalities. The particular features used for each concept detector are obtained using sequential forward feature selection. The method has proven to be readily scalable to a large number of concepts, which has enabled us to model e.g. a total of 294 concepts from a large-scale multimedia ontology [15] and utilize these concept models in TRECVID video search experiments [12]. Figure 7.5 lists and exemplifies the 36 semantic concepts detected for the TRECVID 2007 high-level feature extraction task.

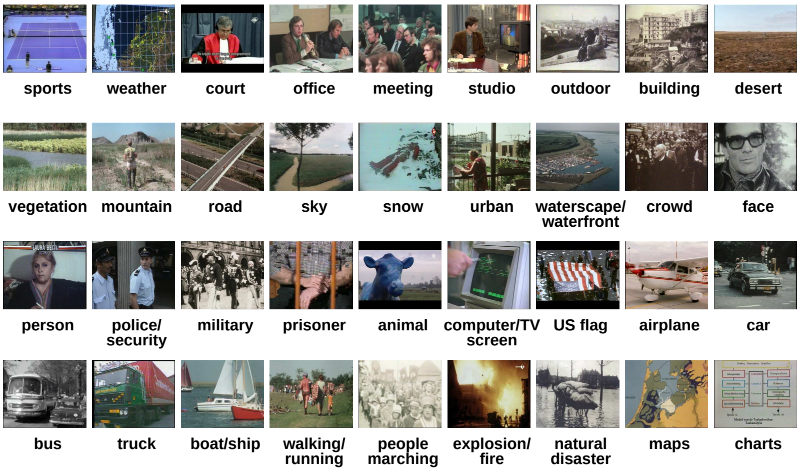


Figure 7.5: The set of 36 semantic concepts used in TRECVID 2007.

Semantic concepts do not exist in isolation, but have different relationships between each other, including similarities in their semantic and visual (low-level) characteristics, co-occurrence statistics, and different hierarchical relations if a taxonomy has been defined for the concepts. We have studied how multimedia concept models built over a general clustering method can be interpreted in terms of probability distributions and how the quality of such models can be assessed with entropy-based methods [16].

In addition we also explored the possibility of taking advantage of temporal and inter-concept co-occurrence patterns of the high-level features using n-gram models and clustering of temporal neighborhoods. The method was found to be very useful in our TRECVID 2007 experiments [13].

7.7 Shot boundary detection

We have applied our general multimedia analysis framework to shot boundary detection and summarization of video data. Our approach for shot boundary detection utilizes the topology preservation properties of SOMs in spotting the abrupt and gradual shot transitions. Multiple feature vectors calculated from consecutive frames are projected on two-dimensional feature-specific SOMs. The transitions are detected by observing the trajectories formed on the maps.

Due to the topology preservation, similar inputs are mapped close to one another on the SOMs. The trajectory of the best-matching map units of successive frames thus typically hovers around some region of a SOM during a shot, provided that the visual content of the video does not change too rapidly. Abrupt cuts are characterized by sudden trajectory leaps from one region on the map to another, and gradual transitions on the other hand are characterized by a somewhat rapid drift of the trajectory from one region to another. The detector tries to detect these kinds of characteristic phenomena.

To increase detector robustness and prevent false positive cut detection decisions, e.g. due to flashlights, we do not only monitor the rate of change of the map position between two consecutive frames, but take small frame windows from both sides of the current point of interest, and compare the two frame windows. A circular area with a constant radius is placed over each map point in the given frame window as illustrated in Figure 7.6. We call the union of these circular areas the area spanned by the frame window. If the areas spanned by the preceding and following frame windows overlap, there are some similar frames in both of them, and we decide that the current point of interest is not a boundary point. If there is no overlapping, the frames in the frame windows are clearly dissimilar, and we decide that we have found a boundary. The flashlights are characterized by sudden trajectory leaps to some region on the map followed by a leap back to the original region. If the duration of the flashlight is smaller than the frame window size, the proposed method helps to avoid false positives.

The final boundary decision is done by a committee machine that consists of this kind of parallel classifiers. There is one classifier for each feature calculated from the frames, and each classifier has a weight value. The final decision is made by comparing the weighted vote result of the classifiers against a threshold value. Abrupt cuts and gradual transitions are detected using the same method. The detected boundary points that are close to one another are combined, and as the result we get the starting locations and lengths of the transitions. To facilitate detection of slow gradual transitions, our system also allows to use a frame gap of given length between the two frame windows. A more detailed description and quantitative results with the algorithm are given in [17].

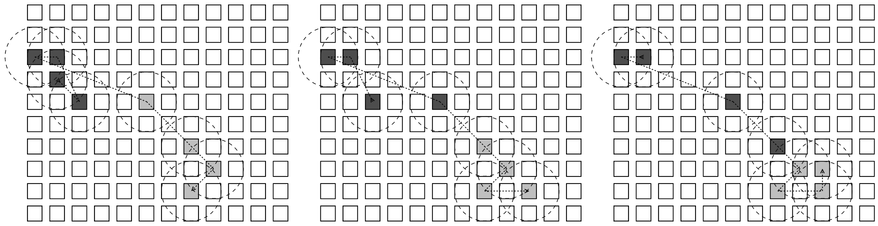


Figure 7.6: Segments of a trajectory at three consecutive time steps. The SOM cells marked with a dark gray color represent trajectory points belonging to the set of preceding frames, and light gray cells represent the following frames. The circles represent the area spanned by the preceding and following frame sets.

7.8 Video summarization

Video summarization is a process where an original video file is converted to a considerably shorter form. The video summary can then be used to facilitate efficient searching and browsing of video files in large video collections. The aim of successful automatic summarization is to preserve as much as possible from the essential content and overall structure. Straightforward methods such as frame subsampling and fast forwarding produce incoherent summaries that are strenuous to view and cannot usually be absorbed with a single viewing. The strategy of selecting parts of the video using a fixed interval can easily lose important information. More sophisticated summarization algorithms typically use shot-based segmentation and analysis. However, including each shot in the summary may not be optimal as certain shots may be almost duplicates of each other or there may be too many of them for a concise summary, depending on the original material.

There are two fundamental types of video summaries: *static abstracts or storyboards* and *video skims*. The former typically consist of collections of keyframes extracted from the video material and organized as a temporal timeline or as a two-dimensional display. Video skims consist of collections of selected video clips from the original material. Both these types of summaries can be useful, depending on the intended application. Storyboards provide static overviews that are easily presented and browsed in many environments, whereas skims preserve the original media type and can also contain dynamic content such as important events in the original video.

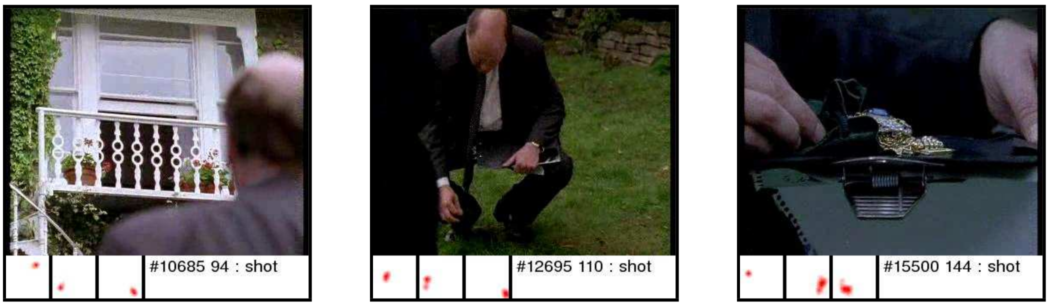


Figure 7.7: Representative frames and SOM signatures of three video shots.

We have developed a technique for video summarization as video skims [18] using SOMs trained with standard visual features that have been applied in various multimedia analysis tasks. The method is based on initial shot boundary detection providing us with lists of shots, which are used in the following stages as basic units of processing. We detect and remove unwanted "junk" shots (e.g. color bar test screens, empty frames) from the videos, and apply face 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 record the corresponding BMUs. The set of BMUs constitutes a SOM-based 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. Fig. 7.7 shows example frames from three shots and the convolved SOM-based trajectory signatures of those shots as red-colored responses on the SOM surfaces. Each remaining shot is then represented in the summary with a separately selected one-second clip. The selected clips are finally combined using temporal ordering and fade-outs and fade-ins from black.

We participated in the TRECVID 2007 rushes summarization task [18] and obtained very promising results. Our summarization algorithm obtained average ground-truth inclusion performance with the shortest overall summaries over all the submissions.

References

[1] Mark Everingham, Andrew Zisserman, Chris Williams, and Luc Van Gool. The Pascal Visual Object Classes Challenge 2006 (VOC2006) results. Technical report, 2006. Available on-line at `http://www.pascal-network.org/`.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] Ville Viitaniemi and Jorma Laaksonen. Techniques for still image scene classification and object detection. In *Proceedings of 16th International Conference on Artificial Neural Networks (ICANN 2006)*, volume 2, pages 35–44, Athens, Greece, September 2006. Springer.

[4] Ville Viitaniemi and Jorma Laaksonen. Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. *Signal Processing: Image Communications*, 22(6):557–568, July 2007.

[5] Ville Viitaniemi and Jorma Laaksonen. Improving the accuracy of global feature fusion based image categorisation. In Bianca Falcidieno, Michela Spagnuolo, Yannis S. Avrithis, Ioannis Kompatsiaris, and Paul Buitelaar, editors, *Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies (SAMT 2007)*, volume 4669 of *Lecture Notes in Computer Science*, pages 1–14, Genova, Italy, December 2007. Springer.

[6] Zhirong Yang and Jorma Laaksonen. Interactive content-based facial image retrieval with partial relevance and parzen discriminant analysis. *Pattern Recognition Letters*, 2008. In submission.

[7] Zhirong Yang and Jorma Laaksonen. Face recognition using Parzenfaces. In *Proceedings of International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *Lecture Notes in Computer Science*, pages 200–209, Porto, Portugal, September 2007. Springer.

[8] Zhirong Yang and Jorma Laaksonen. Principal whitened gradient for information geometry. *Neural Networks*, 2008. In press.

[9] Matthieu Molinier, Jorma Laaksonen, and Tuomas Häme. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):861–874, April 2007.

[10] Matthieu Molinier, Jorma Laaksonen, Yrjö Rauste, and Tuomas Häme. Detecting changes in polarimetric SAR data with content-based image retrieval. In *Proceedings of IEEE International Geoscience And Remote Sensing Symposium*, Barcelona, Spain, July 2007. IEEE.

[11] Erkki Oja, Mats Sjöberg, Ville Viitaniemi, and Jorma Laaksonen. Emergence of semantics from multimedia databases. In Gary Y. Yen and David B. Fogel, editors, *Computational Intelligence: Principles and Practice*, chapter 9. IEEE Computational Intelligence Society, 2006.

[12] Mats Sjöberg, Hannes Muurinen, Jorma Laaksonen, and Markus Koskela. PicSOM experiments in TRECVID 2006. In *Proceedings of the TRECVID 2006 Workshop*, Gaithersburg, MD, USA, November 2006.

[13] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[14] Markus Koskela and Jorma Laaksonen. Semantic concept detection from news videos with self-organizing maps. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Proceedings of 3rd IFIP Conference on Artificial Intelligence Applications and Innovations*, pages 591–599, Athens, Greece, June 2006. IFIP, Springer.

[15] Milind Naphade, John R. Smith, Jelena Tešić, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[16] Markus Koskela, Alan F. Smeaton, and Jorma Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Transactions on Multimedia*, 9(5):912–922, August 2007.

[17] Hannes Muurinen and Jorma Laaksonen. Video segmentation and shot boundary detection using self-organizing maps. In *Proceedings of 15th Scandinavian Conference on Image Analysis (SCIA 2007)*, pages 770–779, Aalborg, Denmark, June 2007.

[18] 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.