Video Summarization with SOMs

Jorma Laaksonen, Markus Koskela, Mats Sjöberg, Ville Viitaniemi, and Hannes Muurinen
Helsinki University of Technology
P.O.Box 5400, FI-02015 TKK, Finland
email: {jorma.laaksonen,markus.koskela,mats.sjoberg,ville.viitaniemi,hannes.muurinen}@tkk.fi

Keywords: multimedia processing, video content analysis

***Abstract*— Video summarization is a process where a long 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 of each video. What is essential is subjective and dependent on the use of the videos and the overall content of the collection. In this paper we present an overview of the SOM-based methodology we have used for video summarization. The method uses temporal trajectories of the best-matching units of frame-wise feature vectors for shot boundary detection and shot similarity assessment. The video material we have used in our experiments comes from NIST's annual TRECVID evaluation for content-based video retrieval systems.**

1 Introduction

Decreasing prices of digital video cameras and cell phones with video capturing have tremendously increased the amount of video footage. Unfortunately, most of the video search engines are still based on textual search and are thus dependent on manual annotation of the data. Manual annotation is, however, slow, expensive and often inaccurate and heterogeneous since the annotations are always subjective and dependent on the annotator's cultural background, language and opinions. Automatic annotation methods and video search engines are needed to really be able to access the huge amount of video data available.

In video summarization one tries to extract relevant parts from a longer video to create a shorter one so that its overall structure and content is preserved. In this paper we study how the Self-Organizing Map (SOM) can be used in efficient video summarization. Our approach analyzes the trajectories of the best-matching units (BMUs) of feature vectors calculated from the video frames. The evolution of the trajectory is used both in detecting shot boundaries and evaluating the novelty of each shot.

The rest of the paper is organized as follows. Video summarization is described in Section 2. Then, in Section 3, our SOM-based summarization method is presented. The video data collection used in our experiments is described in Section 4 and the performed experiments in Section 5. Conclusions and future directions are stated in Section 6.

2 Video Summarization

A useful video summary considerably reduces the length or size of the original material yet preserving its essential content. 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, but can be used as a baseline technique. 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.

Due to these challenges, automatic video summarization has emerged as an important application for multimedia analysis methods, and several approaches have been proposed (see e.g. [8], [2], [3]). In [14], the Self-Organizing Map was used to track moving object trajectories for video summarization. For an overview of automatic video summarization techniques, see [7].

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. The latter type consists of collections of selected 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 contain also dynamic content such as important events in the original video.

Regardless of the type of the summary, most of the existing approaches to video summarization consist of three distinct processing steps. The first task usually involves segmenting the video into scenes or shots. The second task is clip or keyframe selection, typically using some image and audio analysis techniques. In many approaches, this step is divided further to inter-shot selection and intra-shot analysis. At the latter stage, the selected shots are analyzed in more detail to decide on the most informative frames or sub-shots for the summary. The final step in the generation

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

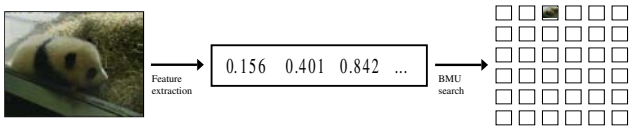

Figure 1: A feature vector is calculated from an image. Then the vector is mapped on the SOM by finding the best-matching map unit.

of the summary consist of assembling and presenting the summary, and deciding on various presentation details.

Different video summaries are needed for different purposes and applications. For example, media researchers analyzing archived material, TV producers skimming previously recorded rushes, and surveillance personnel monitoring CCTV recordings require detailed summaries. In designing such applications, one should strive to preserve the original content as accurately as possible. On the other hand, for entertainment purposes and consumer applications, such as personal video recorders (see e.g. [10]), the summaries need to be brief and cursory—too much detail can even be harmful as is the case with movie trailers that should not reveal any plot twists. Consequently, the perceived quality of a video summary depends heavily on its intended usage and any evaluations of different methods are necessarily subjective.

3 A SOM-based Method

In this paper, we propose a technique for video summarization using Self-Organizing Maps trained with standard visual features that have been applied in various multimedia analysis tasks. The method is based on observing the trajectories that the best-matching units (BMUs) of the feature vectors, extracted from the video frames, traverse on the SOM surfaces. The SOMs and trajectories are used in two stages: first in automatic shot boundary detection and then in selection of representative clips.

3.1 Video Trajectories on SOMs

Self-Organizing Maps can be used to visualize high-dimensional data on a two-dimensional lattice. Plotting a *trajectory* is a highly useful visualization technique that can be used when handling data vectors that vary as a function of some scalar variable. In video analysis, the variable is typically time, but other variables like spatial coordinates of detected objects can be used as well. In this work, only temporal trajectories describing changes in the entire video frame are discussed.

A temporal SOM trajectory is the "path" that the data mapped on a Self-Organizing Map traverses as time advances. At each time step a feature vector is first calculated from the data, and then the vector is mapped to the best-matching unit on the map. The process is illustrated in Figure 1. When visualizing the trajectory, the consecutive BMUs are connected with a line to form the full path. An example of such a trajectory is shown in Figure 2. In both figures we have used the original video frames as visual labels on the SOM units.

Figure 2: An example of a video trajectory on a SOM. Consecutive frame BMUs are connected to form the trajectory.

Trajectories can be used to monitor changes in the input data over time. Similar vectors are mapped spatially close to each other on the map, and thus the trajectory should hover over the same area on the map if the input vectors do not change much. If there is a sudden large change between the values of consecutive input vectors, there should be a relatively large leap in the trajectory on the map from some region to another. If the input vector values drift slowly over time in the input vector space from some region to another, there should also occur an analogous drift in the SOM trajectory from one region to another.

The use of SOM trajectories is not a novel idea. Indeed, in [5] a speech recognition method using SOM trajectories was proposed. In the method each SOM unit corresponds to a particular phoneme, and the trajectory can be used to combine the consecutive phonemes to form words. Later, SOM trajectories were proposed for chemical process monitoring [15]. In [1] trajectories were used to visualize the operation of a continuous pulp digester and a hot rolling steel mill. In these applications, the visual trajectory information can be used to aid operative decision making.

In this work the SOMs do not contain labels that correspond to phonemes or process states like in the aforementioned papers. Neither can the created SOMs be used as classifiers in the sense that some regions in the map would classify a frame to belong to a specific class. The evolution of the trajectory is instead used to monitor the rate of change in consecutive frames by observing the distances between the consecutive trajectory points. In a latter stage, the points of trajectory are also clustered on the SOMs to evaluate the novelty of each shot in a video.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

3.2 Shot Boundary Detection with SOMs

The SOM-based automatic shot boundary detection (SBD) method used in this study was published in [9]. The main idea is to spot the discontinuities in the visual stream by monitoring video frame trajectories of the BMUs of frame-wise feature vectors on Self-Organizing Maps. The SOM mapping compensates for the probability density differences in the feature space, and consequently distances between SOM coordinates are more informative than distances between plain feature vectors.

The method compares two sliding best-matching unit windows instead of just measuring distances between two trajectory points, which increases the robustness of the detector. The window technique can be seen as a variant of adaptive threshold SBD methods [16]. Furthermore, the robustness is increased by using a committee machine of multiple SOM-based detectors. Experimental evaluation made by NIST in the TRECVID 2006 evaluations [12] confirms that the SOM-based SBD method works comparatively well in news video segmentation, especially in gradual transition detection.

3.3 Detection of Special Contents

As discussed in Section 2, a video summary should be concise, but still contain the most noteworthy or special contents of the original video. A common method is to detect certain important contents by using specialized detectors. The type of content considered important is naturally domain-specific. In surveillance video, for example, all "abnormal" events are likely to be important and should be included in the summary. In movie summaries, the special contents may include events such as dialogs, close-ups of lead actors, gunfire, explosions, etc.

In the experiments described in this paper we have looked at raw unedited footage from several TV productions (see Section 4), with many scenes containing one or more persons talking. Due to the properties of the material, we have used special detectors for *color bar test screens*, *black frames*, *white frames* and *faces*.

Color bar test screens typically appear in the beginning and the end of raw production material and should be cropped off from the summaries. They are characterized by vertical color bars and are thus easy to detect with color features such as the MPEG-7 standard Color Layout descriptor [4]. Totally black or white frames can be detected and cropped off in a similar fashion.

Faces were detected in our experiments by a face detector included in Intel's OpenCV Library[1]. The detector is based on Haar-like features and a cascade of boosted tree classifiers. Furthermore, the face candidates returned by the OpenCV detector were pruned by using a simple skin color detector in the YCbCr color space.

[1] http://www.intel.com/technology/computing/opencv/

3.4 Shot Selection

After the shot boundaries have been detected, we proceed to select which shots will be included in the summary. We trace the trajectory of the frames within a shot and record the corresponding BMUs on a SOM. The set of BMUs constitutes a SOM-based *signature* for the shot, which can then be compared to those of other shots to determine whether a shot is visually unique or similar to some other shots.

There are two distinct schemes to construct these shot-wise signatures. First, we could preserve the trajectory of the consecutive BMUs during the shot and obtain a temporal signature representing the dynamic structure of the shot. Comparing two shots then involves comparing their respective BMU trajectories on the surface of the SOM in question. This approach might be needed, for example, to distinguish a scene where a man is walking into a room from another one where he is walking out of the room, i.e. it takes into account the temporal or causal aspect of the video. The usefulness of this dynamic structure is, however, questionable and domain-dependent. It may also lead to large differences between shots that are overall similar, but have different temporal signatures for some reason, e.g. due to failed shot boundary detection.

In the second approach, which is the one used in these experiments, we discard the temporal element and treat all the frames of the shot equally as separate unordered data items. The set of frame-wise BMUs representing the shot can then be treated as any data subset whose distribution on a SOM surface can be obtained by counting the number of BMU hits for each map unit. Normalized to unit sum, the hit frequencies give a discrete histogram which is a sample estimate of the probability distribution of the subset on the SOM surface. Due to SOM's topology preservation, one may now force the neighboring units to interact by low-pass filtering or convolving the hit distributions on the SOM surface. When the surface is convolved, the one-to-one relationship between input vectors' SOM indices and hits on the SOM surface is broken. Instead, each hit results in a spread point response around the BMU. This process closely resembles how SOMs are being used for content-based information retrieval in our PicSOM system [6].

As the last step, by enumerating the units of all the two-dimensional SOM grids, we can represent the distribution of the mth shot as a vector $P_m \in \mathbb{R}^k$, where k is the total number of SOM units, concatenated over all feature spaces, and use a suitable metric to measure the distance between any two shots. The pruning of the most similar shots was done as follows. First, we calculate the pairwise Euclidean distances of all shot pairs. We then start deleting shots beginning from the most similar pair until the total number of remaining shots is below the allowed limit and an experimentally set minimum pair-wise similarity threshold is exceeded. This additional threshold was included based on a subjective analysis of the repetition remaining in the summaries created for the development set videos.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

3.5 Selection of Representative Clips

The next stage after deciding the shots to be included in the summary is the selection of the most informative clips or subshots from within the selected shots. In this task, we utilize the content detectors described in Section 3.3 in addition to the MPEG-7 Motion Activity descriptor [4] that has been computed using one second intervals from the entire video. Initially, we favor frames near the center of the shot, and award additional score for frames containing faces and for frames with or immediately after increased motion activity. Correspondingly, scores for any frames detected as color bars or black or white frames are reduced.

It should be highlighted that the emphasis can vary between the shot-level evaluation and intra-shot analysis, depending on the input video. With a video containing only a small number of long shots, the subshot selection becomes the dominant step, whereas the shot-level selection is crucial for the summarization of a video with numerous short shots. An effective summarization algorithm should be able to handle both these cases equally well.

3.6 Summary Presentation

After the clips constituting the final summary have been determined, the final stage is the generation of the final summary video. Several important parameters, such as the number of clips included and the clip length, are decided based on the results of the shot boundary detection and overall statistics of the video.

Still, there remain multiple options that have an effect on the usability and subjective quality of the summary. Two such options are the ordering of the clips and the transitions (cuts, linear interpolations, etc.) between the clips. For simplicity, we use a temporal ordering and simple fade-outs to and fade-ins from black in our current system. More advanced effects, such as filler clips, frame rate changes, and image mosaics, have not yet been employed.

4 TRECVID BBC Rushes

The TRECVID workshop[2] [13] is an annual event for video retrieval evaluation, sponsored by the National Institute of Standards and Technology (NIST). In 2007 TRECVID included a summarization task [11] of video rushes provided by the BBC Archive. Rushes are raw video material from the shooting of a TV production, including failed shots, retakes and other extra footage—of which usually only 2–5% is used in the final product.

The video data provided by the BBC consists of about 40 hours of totally unedited material from five dramatic series divided into 89 rushes videos of about 5–40 minutes in length. The videos are divided into 47 development and 42 test videos. The frame rate is always 25 frames per second.

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

The goal of the summarization task is to generate shorter versions of the 42 test videos so that the main objects and events of the video are shown in a way that maximizes the usability and speed of recognition. The maximum length of the each summary is 4% of the original video length, usually 30–90 seconds.

The quality of the summaries of the participating systems were evaluated at NIST by a human judge, based on a set of pre-specified criteria. The evaluation criteria include the duration of the summaries, how many objects and events from a pre-defined ground truth set can be found and how quickly these can be recognized. It must be noted that the videos contain some events multiple times, but it is sufficient to find only one example of each event.

5 Experiments

We begin with a qualitative study of one 10'30" long development video as an example case. Figure 3 shows the special content detections for this video file, together with the detected shot boundaries and motion activity.

Figure 4 shows frames from three shots and the SOM-based trajectory signatures of those shots. The three signatures correspond to our Color Layout, Edge Co-occurrence and Edge Histogram features. All the SOMs have been of size 64×64 units and trained with feature vectors of the frames from that specific video file only. The diameter of the low-pass convolution kernel has been 15 units. One can easily see that for each of the three features, the SOM signatures of all the shots are clearly different.

Following the 4% summary length rule leads to a maximum of 629 frames long summary. As we have required that each clip in the final summary should last one second, we end in this case in selecting 25 most representative shots, each of them represented by one clip.

The ground truth provided by NIST associated with the example video lists the following 14 events:

1. police man setting up security tape
2. police woman and man in suit walking
3. people on lawn
4. man in white with camera in front of face walks across screen
5. back view police woman and man in suit approaching man in white
6. back view of man in white with camera, police woman and man in suit investigating at tree
7. back view of police man and man in white with camera at tree
8. closeup of bald man's head/face
9. closeup of bald man's head from behind
10. bald man's torso from behind
11. bald man's torso from front
12. close up of mans feet standing on black pouch in grass
13. bald man bends down and picks up a black pouch
14. close up of hands pouring jewellery from black pouch

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

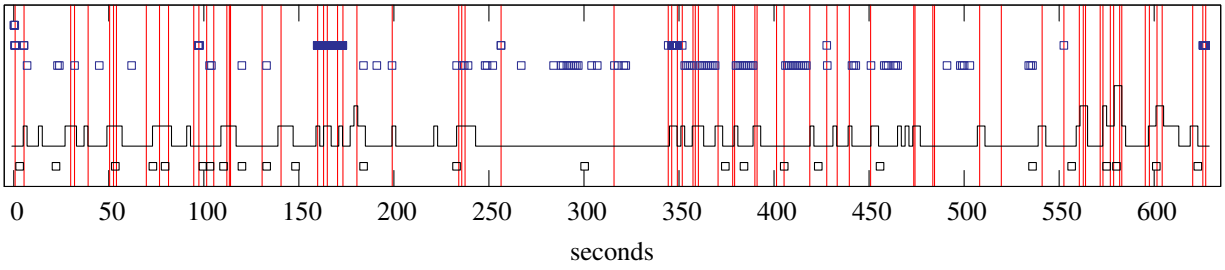

Figure 3: An example video analyzed with shot boundary detection (red vertical lines), motion activity (black staircase line), and specialized detectors (blue boxes). The detector results correspond to test screens, black/white frames, and faces (from top to bottom). The one-second clips selected to the summary are shown as black boxes on the bottommost row.

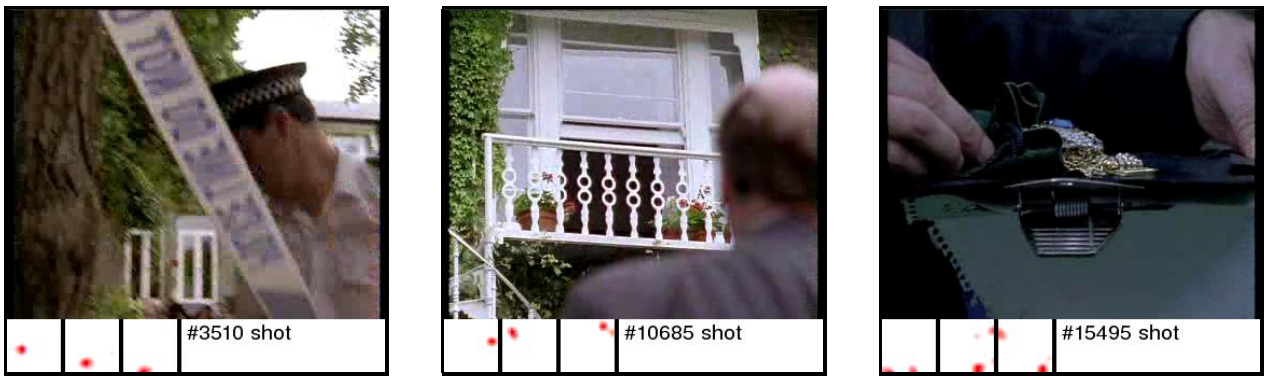

Figure 4: Frames and SOM signatures of three video shots.

The shots selected by the system are shown by the bottom row boxes in Figure 3. Visual examination of the created summary reveals that the system was able to include 12 of the listed 14 events. Events 7 and 14 were left missing, whereas six of the ground truth events were selected more than once. Five such clips were selected that did not depict any of the listed events. Due to the amount of duplicated clips in this and other development set summaries, we decided to reduce the summary durations from the maximum of 4% by introducing an additional similarity threshold (Section 3.4) if the included clips were overly similar.

The evaluation of the 42 test set summaries at NIST provides quantitative results for our summarization method. Table 1 gives an overview by listing some of the standard measures provided by NIST [11]. In brief, the two topmost results (DU and TT) are measures of time (in seconds), corresponding to the average duration of the 42 summaries and the average time spent by the assessors judging the ground-truth inclusions, respectively. IN lists the fraction of ground-truth inclusions found in the summaries, and the two remaining results are from an assessor questionnaire with the range of 1–5, where 5 is the desired value.

The most striking characteristics of our summaries, when compared to the other submissions, are both the duration of and total time spend assessing the summaries—in both these senses our submitted summaries were the shortest. This is also reflected in the assessments of the amount of duplicate video present as our mean result is clearly above the median. The average duration (in seconds) also equals the average number of distinct clips in our summaries since we used fixed one-second clips.

Table 1: An overview of the summarization results. The values are averages over all 42 test videos.

	ours	max	median	min
duration (DU)	26.1	64.2	53.6	26.1
total time (TT)	61.7	119.3	94.2	61.7
inclusion (IN)	0.45	0.68	0.49	0.25
understandability (EA)	3.23	3.60	3.33	1.97
duplicate video (RE)	3.87	3.98	3.66	3.02

On the other hand, the fraction of ground-truth inclusions found in our summaries (45% on average) was slightly below the median of all scores. This is undoubtedly directly affected by the comparatively short durations of our summaries. To analyze this relation further, Figure 5 shows for all submissions the average fraction of inclusions found over the average summary durations.

Figure 5 shows a clear relation between these measures as the summaries with high fractions of inclusions found tend to have high time expenditure values, as was to be expected. In fact, the summary with the longest duration and most time spent has also the highest fraction of inclusions found. The duration of our summaries is directly related

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

to the shot similarity threshold used to prune overly similar shots. By controlling the threshold value, we can tune the durations of the resulting summaries. The exact effect of the threshold is, however, yet unknown as each team was allowed to submit only a single summary per test video.

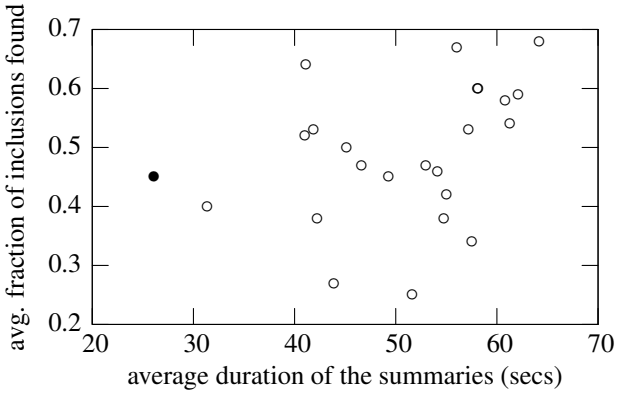

Figure 5: The average fraction of inclusions found vs. average duration of the summaries; our run shown as "•".

6 Conclusions and Future Prospects

In this paper we have presented a novel method for video summarization by using Self-Organizing Maps. The core idea is in observing the trajectories of the BMUs of feature vectors extracted sequentially from the video frames. SOM trajectories have been used before in applications where the SOMs contain known labels or regions and can thus be used as classifiers. In our case we do not have such information, instead we look at the rate of change of the trajectory and the clustering of the temporal signatures from the BMUs of the video shots on the maps. In this way the evolution of the trajectories can be used both in detecting the shot boundaries and in evaluating the uniqueness of distinct shots.

The results of our initial experiments are promising. In further studies we will tune the parameters of the system for finding an optimal operating point with respect to the duration–inclusion tradeoff.

Acknowledgements

Supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research*.

References

[1] Esa Alhoniemi, Jaakko Hollmén, Olli Simula, and Juha Vesanto. Process monitoring and modeling using the self-organizing map. *Integrated Computer Aided Engineering*, 6(1):3–14, 1999.

[2] Michael G. Christel, Michael A. Smith, C. Roy Taylor, and David B. Winkler. Evolving video skims into userful multimedia abstractions. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 171–178, Los Angeles, CA, USA, April 1998.

[3] Yihong Gong and Xin Liu. Generating optimal video summaries. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2000)*, volume 3, pages 1559–1562, New York City, NY, USA, July-August 2000.

[4] ISO/IEC. Information technology - Multimedia content description interface - Part 3: Visual, 2002. 15938-3:2002(E).

[5] Teuvo Kohonen. The 'neural' phonetic typewriter. *Computer*, 21(3):11–22, March 1988.

[6] Jorma Laaksonen, Markus Koskela, and Erkki Oja. Class distributions on SOM surfaces for feature extraction and object retrieval. *Neural Networks*, 17(8-9):1121–1133, October-November 2004.

[7] Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. Technical Report HPL-2001-191, HP Laboratories Palo Alto, July 2001.

[8] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstracting. *Communications of the ACM*, 40(12):55–62, December 1997.

[9] 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.

[10] Isao Otsuka, Regunathan Radhakrishnan, Michael Siracusa, Ajay Divakaran, and Hidetoshi Mishima. An enhanced video summarization system using audio features for a personal video recorder. *IEEE Transactions on Consumer Electronics*, 52(1):168–172, February 2006.

[11] Paul Over, Alan F. Smeaton, and Philip Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 1–15, New York, NY, September 2007. ACM Press.

[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] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[14] A. Stefanidis, P. Partsinevelos, P. Agouris, and P. Doucette. Summarizing video datasets in the spatiotemporal domain. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)*, pages 906–912, London, UK, September 2000.

[15] Viktor Tryba and Karl Goser. Self-organizing feature maps for process control in chemistry. In *Proceedings of International Conference on Artificial Neural Networks (ICANN-91)*, volume 1, pages 847–852, Espoo, Finland, June 1991.

[16] Boon-Lock Yeo and Bede Liu. Rapid scene analysis on compressed video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 5(6):533–544, 1995.