INTRODUCING CAM - CONSTANT ACTION MOVIE

Thomas Stephan and Matthias Richter and Jürgen Beyerer

{thomas.stephan,matthias.richter,juergen.beyerer}@iosb.fraunhofer.de
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB,
Fraunhoferstr. 1, 76137 Karlsruhe (Germany)

1 INTRODUCTION

Many existing security-systems consist of CCTV-Cameras, which record data streams and store it on a large storage medium. After any security incident occurred, the video material has to be manually reviewed by a human operator. If the exact time of the incident is unknown, the material to be reviewed can be rather long and the examination becomes very time consuming. Typically the most interesting content is sparsely distributed along the video. Hence, most of the reviewing time is wasted. Moreover, through the long irrelevant parts of the video material the attention of the operator decreases and therefore important events could be missed. Video synopsis by constant action movie (CAM) offers a solution to this problem: The human operator only has to inspect (possibly) relevant parts of the video. The key idea of the proposed CAM approach is to time-compress the video stream, so that the interesting content is equally distributed (constant action) along the video.

1.1 Related Work

Video summarization has long been of great interest to the scientific community and many different approaches and methods have been tried over the years. Since a full review and classification of these methods (as can be found in [1]) is out of scope of this paper, we will focus on selected works that are closely related to our approach. In particular, state of the art methods that do not preserve chronological order of events (e.g. [5] and [6]) are not considered.

In 1999 Nam and Tewfik proposed to adjust the sample rate of a video in relation to the amount of “visual activity” in the video [2]. The video is divided into smaller units called sub-shots and the visual activity of each sub-shot is assessed using a temporal wavelet transform. Each sub-shot is sampled with a rate inversely proportional to the visual activity in the segment. This method is very similar to ours, but the dependency on shot boundary detection makes it difficult to apply the method in video surveillance: What should be considered a shot? In contrast, Petrovic et al. propose to adjust the sample-rate according to the likelihood of a generative model of the video content [3]. The model is learned from a query scene supplied by the user and the mapping can be applied either on a per-shot or per-frame basis. The result is that scenes similar to the query scene will be played back at normal speed, while dissimilar scenes will be fast-forwarded or skipped entirely. This approach is especially interesting if the model incorporates action recognition methods. However, a query scene might not always be available in a surveillance setting. Another more recent approach by Cong et al. formulates video synopsis as dictionary learning problem [4]. The core idea is to select a set of key frames so that the entire video can be reconstructed with minimal error. Experiments show that this method is superior to previous similar methods. However, in video surveillance such a synopsis is of limited use, since it (a) removes temporal context and (b) might exclude important events if the corresponding scenes can be described in terms of previously selected key frames.

It should be noted that the above methods can be retrofitted in the CAM framework, which will be discussed in the next section.

2 METHOD

In order map a given video $g(t)$ to a a time compressed video $\gamma(t)$ so that there is a constant amount of action over the temporal image sequence, one has to specify what is meant by the term *action*. For that we introduce an action density function $a_g(t)$ over the video $g(t)$,

$$\begin{aligned} a_g &: \mathbb{R}_+ \to \mathbb{R}_+ \\ t &\mapsto a_g(t), \end{aligned} \qquad (1)$$

so that (without loss of generality)

$$\int_0^\infty a_g(\tau)\, d\tau = 1. \qquad (2)$$

Action density functions can be defined in different forms and depend on the specific surveillance task and environment. It is the configurable degree of freedom of the proposed approach.

The purpose of CAM is to derive a time distortion function $\sigma(t)$, so that $\gamma(t) = g(\sigma(t))$, where the corresponding action density should be constant, i.e. $a_g\big(\sigma\,(t)\big) = a_\gamma(t) \equiv c$. Here, $c = {}^{1}\!/_{t_{max}}$ is the reciprocal of the desired length of the time compressed video. To derive the time distortion function $\sigma(t)$, the cumulative action distribution function

$$\mathcal{A}_g(t) = \int_0^t a_g(\tau) d\tau \qquad (3)$$

is introduced. Note that because of eq. (1) and (2), $\mathcal{A}_g(t)$ is limited and monotonically increasing. Now $\sigma(t)$ can be derived using the following relation:

$$\mathcal{A}_\gamma(t) = \mathcal{A}_g\big(\sigma(t)\big) \qquad (4)$$

If the action density function is strictly greater than zero for all t, $\mathcal{A}_g(t)$ is strictly monotonically increasing. Therefore $\mathcal{A}_g(t)$ is invertible, so that

$$\sigma(t) = \mathcal{A}_g^{-1}\Big(\mathcal{A}_\gamma(t)\Big) \Rightarrow \sigma(t) = \mathcal{A}_g^{-1}(t/t_{max}) \qquad (5)$$

Otherwise $\sigma(t)$ can be calculated by minimizing $\big|t/t_{max} - \mathcal{A}_g\big(\sigma(t)\big)\big|$. Finally, the time compressed constant action movie (CAM) is calculated as

$$\gamma(t) = g(\mathcal{A}_g^{-1}(t/t_{max}))\ . \qquad (6)$$

2.1 Action density examples

The choice of action density function $a_g(t)$ highly depends on the context of video surveillance; in different domains, action can be defined in different manner. But even simple action density functions like the absolute difference of sequential images

$$a_g(t) = \lambda\, \|g(t) - g(t - \Delta t)\|_{1,} \qquad (7)$$

with a normalization constant λ, lead to CAMs which are practicable for many different surveillance tasks. A similar action density can be calculated by adaptive background subtraction or summed magnitude of optical flow.

Another advantage is that existing approaches for semantic video analysis, like person tracking, face detection and action recognition can be used to generate action density functions that reflect that semantic content. Hence, CAM can be easily integrated into existing surveillance technologies and frameworks to condense video data.

3 IMPLEMENTATION

For a practical implementation, several implementation details have to be considered. Depending on the action measure, the measurements should be smoothed over time to avoid jittery and “unnatural” video output. In our implementation we use simple Gaussian smoothing, which acts as a low-pass filter on the action signal. More elaborate methods (e.g. Kalman filters) could be used as well, but we found that this simple treatment generally yields good results. A related issue arises when $\sigma(t)$ locally requires more frames than are contained in the original video, i.e. when oversampling occurs. Simple selection of the nearest frame often results in uncanny, jittery videos. We solve this problem by simple linear interpolation between frames. Another extension is motivated by the observation that one is generally interested in only parts of the scene (e.g. entryway, part of a road). Effective video synopsis should submit to this requirement. Therefore, we constrict computation of the action measure to a region of interest (ROI) in the scene. A simple ROI is defined by a rectangle around the interesting parts of the scene, while a complex method defines weights to certain part of the scene.

4 RESULTS

CAM was qualitatively evaluated on different surveillance videos with different action density functions. In this section, results of a time compressed video are summarized. The originating video stems from the CAVIAR dataset [7] and shows a corridor with people walking up and down. The action density was calculated using image difference as defined in eq. (7) in a rectangular ROI placed at the foreground of the scene. The Figure below shows the action density with a few selected key-frames and the resulting mapping of the video.

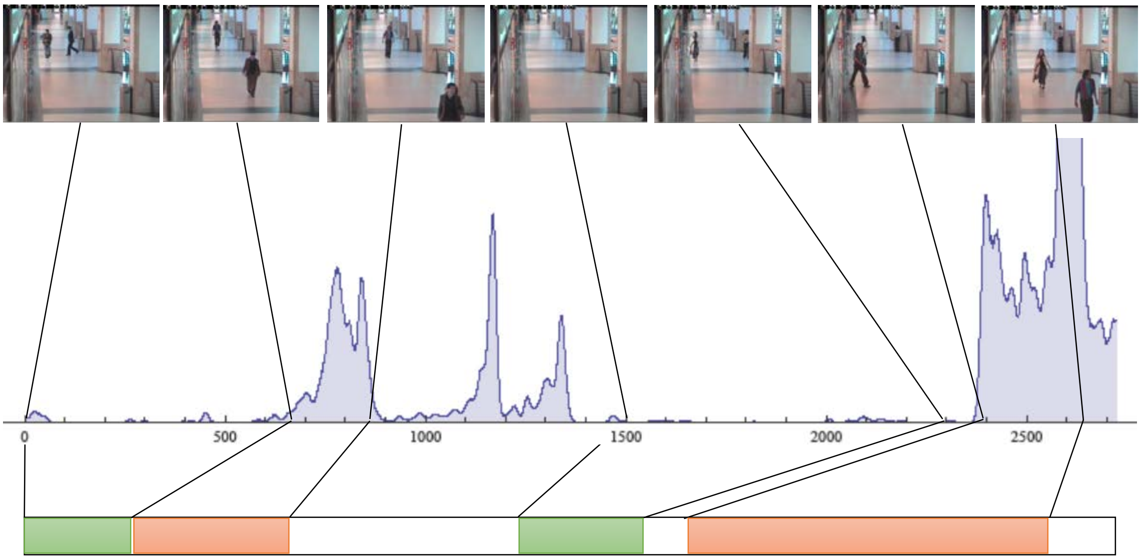

Figure 1: The first row shows some key-frames at different timestamps of the original video from CAVIAR dataset. The action density (plot at second row) increases when people walk through the ROI in the foreground area. The last row shows compressed (green) or spread (red) parts of the resulting constant action movie.

5 CONCLUSIONS

With CAM, we have introduced a generic framework for video synopsis that preserves temporal relationships between scenes. The key idea is to measure the action content over time and construct a mapping to resample the video so it has constant action. The action measure can be defined almost arbitrarily. In fact many existing methods such as [3], [4], and [5] can be retrofitted in the CAM framework. An online version of CAM could be integrated in smart cameras that only record interesting scenes and discard non-relevant material.

Aside from video surveillance, the CAM framework finds applications in the evaluation of research videos, for example to automatically skip to the significant part in slow motion videos and long-time recordings of rare events. Given an action density that reacts to subjectively interesting scenes, e.g. similar to the approach of Cong et al. [4], CAM could also be used to automatically generate movie trailers or a “best of” video of personal videos. Moving in that direction, it might be interesting to not equalize the action content, but instead produce a mapping that lifts the action to a target curve, i.e. $a_g\big(\sigma\,(t)\big) = a_\gamma(t) = d(t)$, that can be constructed according to a predefined dramaturgy or to match the progression of a music piece.

REFERENCES

[1] Truong, B. and Venkatesh, S. (2007). *Video abstraction: A systematic review and classification*. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 3(1), 3.

[2] Nam, J. and Tewfik, A. H. (1999). *Video abstract of video*. IEEE 3rd Workshop on Multimedia Signal Processing (pp. 117-122).

[3] Petrovic, N., Jojic, N. and Huang, T. S. (2005). *Adaptive video fast forward*. Multimedia Tools and Applications, 26(3), 327-344.

[4] Cong, Y., Yuan, J. and Luo, J. (2012). *Towards scalable summarization of consumer videos via sparse dictionary selection*. Multimedia, IEEE Transactions on, 14(1), 66-75.

[5] Pritch, Y., Rav-Acha, A. and Peleg, S. (2008). *Nonchronological video synopsis and indexing*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30(11), 1971-1984.

[6] Li, Z., Ishwar, P. and Konrad, J. (2009). *Video condensation by ribbon carving*. Image Processing, IEEE Transactions on, 18(11), 2572-2583.

[7] Fisher, R., Santos-Victor, J. and Crowley, J. (2004). *CAVIAR test case scenarios*.