Preprint

Information-Based Adaptive Fast-Forward for Visual Surveillance

Benjamin Höferlin, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann

Information-Based Adaptive Fast-Forward for Visual Surveillance

**Benjamin Höferlin · Markus Höferlin ·
Daniel Weiskopf · Gunther Heidemann**

Abstract Automated video analysis lacks reliability when searching for unknown events in video data. The practical approach is to watch all the recorded video data, if applicable in fast-forward mode. In this paper we present a method to adapt the playback velocity of the video to the temporal information density, so that the users can explore the video under controlled cognitive load. The proposed approach can cope with static changes and is robust to video noise. First, we formulate temporal information as symmetrized Rényi divergence, deriving this measure from signal coding theory. Further, we discuss the animated visualization of accelerated video sequences and propose a physiologically motivated blending approach to cope with arbitrary playback velocities. Finally, we compare the proposed method with the current approaches in this field by experiments and a qualitative user study, and show its advantages over motion-based measures.

Keywords Information Theory · Adaptive Fast-Forward · Video Browsing · Video Summarization · Visual Surveillance

1 Introduction

A recent challenge in video surveillance is the efficient analysis and browsing of recorded video footage. Often the automatic analysis of the video data is not possible due to missing assumptions and constraints to the search problem. An example of such an ill-posed problem is the video analysis mini challenge of the IEEE VAST Challenge 2009[1]. In this case ten hours of video surveillance footage were provided. The task was

---

Benjamin Höferlin (✉) · Gunther Heidemann
Intelligent Systems Group, Universität Stuttgart, Stuttgart, Germany
Tel.: +49-711-7816334
Fax: +49-711-7816334
E-mail: benjamin.hoeferlin@vis.uni-stuttgart.de

Markus Höferlin · Daniel Weiskopf
VISUS, Universität Stuttgart, Stuttgart, Germany

[1] IEEE Symposium on Visual Analytics Science and Technology 2009 challenge, `http://hcil.cs.umd.edu/localphp/hcil/vast/index.php`
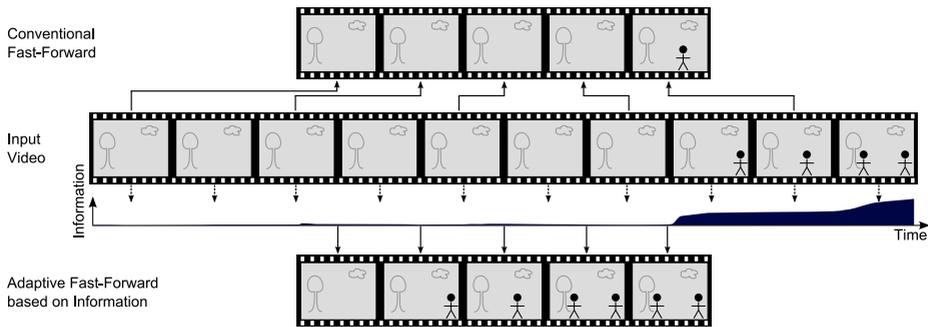
Fig. 1 Difference between traditional cue-play (top) and adaptive fast-forward (bottom). Both sequences are scaled to half the duration of the input sequence (mid). Fast-forward play based on information theory emphasizes the parts of the sequence with high activity while the conventional approach samples the sequence at a constant rate.

to search for suspicious events with the hint that one or more encounters of persons involved in a criminal case took place at locations captured by the camera. Such a problem is very hard to solve in an automatic manner, since the detection of suspicious meetings by several persons is a task that requires analyzing the behavior and intent of the persons. If even less information is provided (i.e. constraints/assumptions to the problem), the issue will be increased further. Actually, there is no applicable solution to such a problem, other than watching and browsing the video data manually. However, this is annoying due to the often overwhelming amount of data and largely monotonous sequences with short moments of activity. Finally, the problem may lead to a reduced willingness of manually analyzing the sequences[2]. Hence, the recorded surveillance data is often monitored in cue-play mode to reduce the time requirement. Here, the problem arises that the playback speed is either low in scenes with little changes or too high so that in periods with much activity important events might be missed. Finally, the users are kept busy by manually rewinding and adapting the video playback speed.

In this paper, we propose a novel adaptive video fast-forward technique that covers the issues mentioned above. Our approach is to adapt the playback speed of the video relative to the temporal information communicated to the viewers. This approach could be deemed an animated video summarization enabling the users to adjust the information load according to their personal abilities. For this reason, we accelerate the video playback during periods of little temporal information, and decelerate it when there are many changes. The resulting playback speeds do not necessarily relate to the semantic relevance (e.g., suspicious event) of the surveillance video data but rather support the visual analysis, by conveying a constant amount of information to the users. The qualitative user study in Section 6.3 evaluates how strong the amount of information correlates with the semantic relevance.

Note that in this paper we restrict the term "visual surveillance" not solely to video data originating from CCTV cameras, but also consider sequences that come from related domains like scientific video sequences, biological surveillance (e.g., animal studies), or digital life streams (e.g., webcams). These domains have to deal with ill-posed search targets and therefore involve the user in the analysis (user in-the-loop).

---

[2] "1,000 CCTV cameras to solve just one crime, Met Police admits", 08/25/2009, www.telegraph.co.uk/news/uknews/crime/6082530/1000-CCTV-cameras-to-solve-just-one-crime-Met-Police-admits.html

Applications that support such an analysis process have to cope with the typical issues arising from imaging and storage limitations. These issues include sensor noise, encoding noise, and time-lapse data. Additionally, surveillance video data offers some properties that can be exploited by an analysis approach. On the one hand, the video data is not artificially manipulated by the introduction of shot and scene changes or transitions like cuts, cross-fades, and dissolves. On the other hand, frequent changes between periods of high and low activity are common, whereas the camera rarely moves. Additionally, audio tracks are in general not recorded, hence issues such as how to handle them when changing the playback speed do not arise. This makes adaptive video fast-forward a suitable approach to support the user-centered video analysis. In contrast to surveillance video sequences, common movies or TV broadcasts do not usually meet these criteria, since they are largely edited to appropriately condense the content, for example for the narration of a story. Hence, watching such video data in fast-forward mode is uncommon, and neither suitable nor intended by the author of the material, even if it is possible.

Our first contribution is the formulation of the temporal information of a video sequence as symmetrized Rényi divergence between the temporal noise distribution and the frame difference distribution. Thus, the proposed approach is able to handle static changes and video noise. Therefore, the playback speed is adapted correctly even in cases of video footage with poor temporal resolution or large video noise, where other approaches become incorrect. By deriving the temporal information from Rényi divergence we are able to provide an additional parameter to the user that steers the information measure by emphasizing certain parts of the distribution ratios. As a second contribution, we adapt the playback duration of every frame and visualize it according to the human visual system. This way, we meet the requirement of visual surveillance that relevant information must not be discarded. Since the proposed technique is computationally cheap, real-time calculation can be achieved. Hence, we are independent of the choice of video compression contrary to other approaches.

The remainder of this paper is organized as follows. First a brief overview of related approaches is provided. In Section 3, we derive the temporal information of a video sequence as symmetric relative entropy between the noise distribution and the frame difference distribution based on Rényi's entropy measure. In Section 4, we discuss how the temporal noise distribution can be estimated. According to the temporal information measure of a video sequence, an appropriate method for fast-forward visualization of surveillance videos is proposed in Section 5. Finally, we evaluate our method in Section 6 and compare the results to other adaptive fast-forward measures. This section also provides a qualitative user study by means of expert interviews to investigate the applicability of playback speed adaption in video surveillance.

2 Previous Work

Three different classes of approaches are known from the literature to facilitate the fast analysis of unconstraint video data: video abstraction, video browsing, and adaptive video fast-forward.

Video abstraction techniques aim at the creation of short and meaningful video summaries. These methods can be further distinguished by the form of output they generate: still image abstraction techniques provide images while video skimming methods produce summarized sequences of shorter duration than the input video. Common

to video skimming techniques is the selection of important temporal parts, while the others are skipped. Often the selected parts are further condensed [28]. Such abstraction techniques cannot guarantee that suspicious events are always detectable. Video abstraction methods that retain all important information rearrange spatio-temporal parts of the sequence temporally [23,19,20] or spatio-temporally [10] to condense the information and decrease the video duration. Hence, several events are displayed at once, even if they occur at different times. This complicates the identification of the chronological context, which is important for surveillance applications.

Video browsing techniques ease fast video exploration by enabling the user to seek for distinct events. Almost every video player contains video browsing controls. For example, the common seek bar in the Windows Media Player allows the user to drag the current time marker to explore the video. Such standard controls can also be adopted to improve video navigation. In the case of seek bars, possible enhancements include the twist lens [22] or navigation summaries [25].

Adaptive video fast-forward techniques accelerate the playback speed of the video somehow related to the content. In contrast, the conventional video fast-forward technique just plays the video sequence multiple times faster than normal. In Figure 1, an exemplary comparison is depicted between traditional cue-play and information-based adaptive fast-forward.

Peker *et al.* proposed an adaptive video fast-forward technique that adapts the playback speed of the video sequence relative to the present motion [16] and the visual complexity [15] (as combination of the spatial complexity and the motion) of the scene. The idea is to watch the video at a constant pace. Another method by Petrovic *et al.* [17] is to adapt the playback speed with respect to a similarity measure between the video footage and a provided target clip. Cheng *et al.* [3] designed an adaptive video player called SmartPlayer which adjusts the playback speed according to three factors: motion, manually defined semantic rules, and former playback preferences of the user.

If we take a closer look at the requirements of surveillance applications, it becomes obvious that surveillance systems have to enable the detection of suspicious events. Since the system does not know what to search for, it is not possible to adapt the playback speed according to the similarity to target clips [17] or previous user preferences [3]. Manually defined semantic rules and temporal regions of interest [3] are also only feasible if interesting events have already been detected. These adaptive fast-forward schemes were proposed for other domains than visual surveillance.

So far, the adaption of the playback speed according to the motion magnitude best meets the needs of visual surveillance. Nevertheless, the use of motion features cannot satisfy all needs of surveillance applications. For example, motion features do not cover static changes like blinking lights. A big problem is that CCTV is often recorded at low frame rates. Gill and Spriggs report [7] that 9 of 13 evaluated control rooms capture their video footage with less than 1.5 fps. Two of the control rooms even have a frame rate of 0.2 to 0.33 fps (that is one frame every 3 to 5 seconds). These frame rates inhibit the calculation of reliable motion features and lead to static changes between frames. *Static changes* are scene changes that are uncorrelated with any motion and often cause change blindness (cf. Section 5). Another issue arises with temporal noise in video sequences, which occurs especially in badly illuminated scenes or due to video coding artifacts. The video noise induces wrong motion vectors and thus reduces the reliability of such a measure.

3 Temporal Information of Videos

The goal of our approach is to adapt the playback speed to the temporal information of a video. Therefore, we need an appropriate measure for the temporal information provided by video data. To meet the users' expectation we demand the information measure to correlate with the magnitude of changes, but to be independent of the level of noise. Additionally, we want to enable the users to control the measure by emphasizing the type of change they are interested in, for example to accentuate strong changes. In this section, we develop such a measure based on Shannon's information theory.

We start with a naïve approach, considering the video data as pure signal containing only relevant changes. Therefore, we regard the sum of changed pixels or the absolute luminance change as indicator of this kind of temporal "information". Obviously, this approach is inappropriate, since it violates our inherent sense of an information measure. For example, a continuously increasing scene illumination would result in high temporal information measures, even if the structural content of the scene remains unchanged. In contrast, small moving objects would be almost disregarded due to noise in the video. This naïve approach is quite related to background subtraction or foreground segmentation, which additionally try to suppress undesired effects (e.g., detection of shadows, noise, variations in illumination, or periodic changes) by the introduction of a background model.

Adopting Shannon's information theory, we could use the inverse mutual information measure to compare the common information of two frames F_1 and F_2. This measure is based on the discrete luminance distribution F_1^{hist}, F_2^{hist} of the frames and their probabilities $p(f_1)$, $p(f_2)$, as well as on the joint probability $p(f_1, f_2)$. Mutual information [4] expresses the interdependence between the normalized luminance distributions of two frames:

$$I(F_1; F_2) = \sum_{f_1 \in F_1^{hist}} \sum_{f_2 \in F_2^{hist}} p(f_1, f_2) \log_2 \left(\frac{p(f_1, f_2)}{p(f_1)p(f_2)} \right) \tag{1}$$

If the two luminance distributions are independent, $I(F_1; F_2)$ reaches its lower bound and thus this information measure between both frames is unlimited. This measure for the temporal information of a video sequence empowers us to adapt the playback speed according to the inverse mutual information coefficient. It has to be noticed that mutual information provides a foundation for the term "information" from the perspective of signal coding theory rather than human recognition.

Unfortunately, the assumption that subsequent frames remain constant (i.e. $|F_1 - F_2| = 0$) in a scene where no changes appear does not hold for real video data, since the sensor as well as the encoding process introduce noise to the signal. To cope with these effects, we model video data as additive combination of the signal S carrying the actual information and the temporal noise N. Then, the absolute frame difference is:

$$D = |F_1 - F_2| = |\Delta S + \Delta N| \tag{2}$$

For the sake of simplicity we assume that suitable noise estimation is provided. In Section 4, we discuss approaches to temporal noise estimation. For now we assume the noise to be independent and identically-distributed in time and space. In Section 4, we will also see that for some video sequences these assumptions are not tenable and more powerful models are needed.

The signal change and hence the temporal information provided by the video data can be considered as the dissimilarity of the estimated noise distribution and the distribution of the absolute frame difference D (cf. 2). We normalize both distributions to 1 to receive probability density functions.

An appropriate measure to compare both probability distributions is the Kullback-Leibler (KL) divergence [12], also known as relative entropy:

$$\mathcal{D}_{KL}(D||N) = \sum_i p(d_i) \log_2 \left(\frac{p(d_i)}{p(n_i)} \right) \tag{3}$$

Based on Shannon entropy, the KL divergence can be interpreted as the expected additional binary message length that is required if the discrete difference image distribution $p(d)$ is encoded using the alphabet of the estimated noise probability distribution $p(n)$. The index i denotes a particular bin of the distribution histogram.

The KL divergence is related to the calculation of the sum of self-information of the temporal change distribution $p(d)$ if the probabilities of the noise distribution $p(n)$ are used. With the self-information formulated in Equation 4, Equation 5 illustrates this relationship using the cross-entropy $H(D, N)$ as an intermediate step.

$$I(p(x)) = \log_2 \left(\frac{1}{p(x)} \right) \tag{4}$$

$$\begin{aligned} \sum_i I(p(n_i)^{p(d_i)}) &= \sum_i \log_2 \left(\frac{1}{p(n_i)^{p(d_i)}} \right) \\ &= -\sum_i p(d_i) \log_2(p(n_i)) = H(D, N) \\ &= \mathcal{D}_{KL}(D||N) + H(D) \end{aligned} \tag{5}$$

It becomes obvious that the sum of self-information and the KL divergence behave very similarly, except for the additional entropy term $H(D)$.

Now we can cope with temporal noise in video sequences. To give the user additional control over the accentuation of this information measure, we enhance this measure by using a generalization of the KL divergence proposed by Rényi [24] for generalized probability density functions. He defines the relative entropy $\mathcal{D}_\alpha$ (or information gain, α-divergence) of order α for $\alpha > 0$ and $\alpha \neq 1$:

$$\mathcal{D}_\alpha(D||N) = \frac{1}{\alpha - 1} \log_2 \left(\sum_i \frac{p(d_i)^\alpha}{p(n_i)^{\alpha-1}} \right) \tag{6}$$

As α approaches 1, the limit of $\mathcal{D}_\alpha$ is the KL divergence (Equation 3). The α-divergence takes on its minimum value $\mathcal{D}_\alpha(D||N) = 0$ if and only if $p(d)$ is equal to the temporal noise distribution $p(n)$. Rényi describes the measure $\mathcal{D}_\alpha$ as "the information of order α obtained if the distribution of N is replaced by the distribution of D."

The α-divergence provides an additional parameter α that allows to place emphasis on certain parts of the distribution relation. Large values for α amplify high probability ratios, while the α-divergence tends to treat all probability ratios in an equal manner for α approaching zero. By letting the users choose the α parameter we enable them to place emphasis on distributions according to their interest and application (e.g., Hero *et al.* proposes the Hellinger affinity distance ($\alpha = 0.5$) for classification tasks of a hardly discriminable set [8]). Thus, they are not only able to define the information gain (i.e.
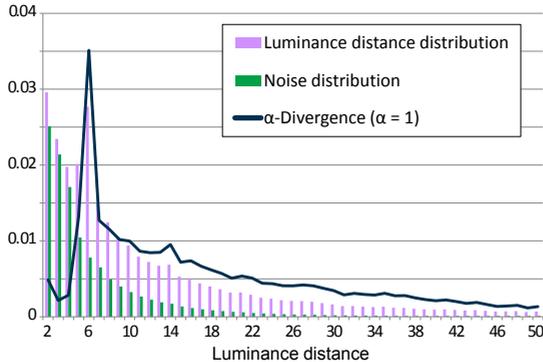
Fig. 2 Terms of Rényi divergence (α-divergence) between the noise distribution and absolute frame difference distribution. The first two bins as well as the last 200 bins were omitted.

the information throughput) for steering the playback speed, but they may also adapt the acceleration accentuating the higher probability density ratios, for example. The Rényi divergences for certain α parameters are strongly related to other divergences, see [1,2] for details. The effect of the α parameter in our approach is empirically evaluated in Section 6. The behavior of the Rényi divergence of an estimated noise distribution and the absolute frame difference distribution is illustrated in Figure 2. For an appropriate presentation, parts of the probability density functions are omitted.

In the case of strong noise, the noise distribution may dominate over the absolute frame difference distribution. This results in low values for the α-divergence, even when large image changes are present. Hence, playback acceleration performs contrary to the expected behavior. This issue can be tackled by symmetrizing the α-divergence according to the Jensen-Shannon divergence with equal weights. It can be defined for the α-divergence as

$$\hat{\mathcal{D}}_\alpha(D||N) = \frac{1}{2}\mathcal{D}_\alpha(D||M) + \frac{1}{2}\mathcal{D}_\alpha(N||M) \tag{7}$$

with $M = \frac{1}{2}(D + N)$. Properties of the Jensen-Shannon divergence and its relation to other measures can be found in the work of Lin [13]. In the remainder of this paper, we use the symmetrized α-divergence $\hat{\mathcal{D}}_\alpha(D||N)$ for all cases independent of the signal-to-noise ratio.

Please note that α-entropy measure was previously used for image matching or image registration (e.g., [14]). In contrast to these approaches, we do not operate on two images (or successive frames of a video), but use α-divergence measure to quantify the information between noise distribution and luminance distribution of difference images. If we would just use α-divergence on successive frames, changes in the image that do not affect luminance distribution would be disregarded (e.g., a moving object in front of homogeneous background).

4 Temporal Noise Estimation

For the estimation of the temporal noise in video sequences, we rely on techniques known from the literature, since noise estimation is not the focus of this paper. For more background information and more sophisticated approaches, we refer to recent noise estimation and noise reduction literature (e.g., [27,6]).

Assuming additive temporal noise that is independent and identically-distributed in time and space, we can estimate the noise distribution as luminance differences in static areas. If the video sequence includes temporal parts without any moving objects, these parts can be used to derive the motion distribution. In some cases, there are special calibration or training sequences that are suitable for noise estimation. In the absence of such sequences or static periods, it is possible to extract the noise distribution out of smaller areas that do not change during some frames. This introduces the need for motion detection. For the evaluation of the proposed information-adapted video fast-forward, motion estimation based on the global differential Horn-Schunck method [9] is applied. For sequences with a high noise ratio, we use the mentioned fall-back to noise estimation on static temporal parts.

We average the estimated distribution over several frames to cope with errors that originate from pixel saturation and lighting conditions of the scene. Finally, the estimated noise distribution is normalized to 1, to obtain a proper probability density function.

Noise in video sequences is most commonly Gaussian distributed [6]. This means that large luminance distances are less likely to be covered by our estimation process since we consider only a small number of estimation samples. Although the absolute error is very small, the effect on the information measure is severe. If we consider for example the relative entropy divergence (cf. Equation 3) from a signal coding or communication theoretical point of view, we might try to encode an arbitrary frame difference distribution by the estimated noise distribution whose "alphabet" lacks some "characters". Hence, the information divergence approaches infinity, since receiving such a message is unexpected. We handle this issue by adding a small offset to the estimated noise distribution. A more sophisticated approach is to estimate the parameters of the underlying noise distribution for the restoration of the missing values.

To cope with noise stemming from coding artifacts, we also have to consider noise that varies periodically over time. We expect only a small number of different noise distributions caused by coding schemes for distinct frame types (I-frame, P-frame, B-frame) or by the re-encoding of a video sequence. Punchihewa and Bailey provide an overview [21] on the different noise types and their origin in video processing. Based on the estimated noise distributions of several frames, we calculate a couple of noise probability functions. These are retrieved as the cluster centroids after k-means clustering is applied. Finally, the information gain ΔI we use to adopt the playback speed is obtained as the minimum information measure between the absolute frame difference distribution and the i noise distribution estimations:

$$\Delta I = \min_i \left(\hat{\mathcal{D}}_\alpha(D||N_i) \right) \tag{8}$$

5 Fast-Forward Visualization

Our adaptive fast-forward approach necessitates the possibility to play the video faster, but the frame rate cannot be increased arbitrarily. Reasons are hardware constraints like the refresh rate of the monitor. Liquid crystal displays typically have a refresh rate of approximately 60 Hz. The maximum frame rate is therefore limited. Typical video fast-forward visualization discards frames to boost the playback speed. Instead of presenting every frame for a shorter duration, the requested frame rate is reached

by displaying every n-th frame. In the context of visual surveillance, this is unsuitable since important events may be skipped.

Another issue for fast-forward visualization for visual surveillance is indicated by Scott-Brown and Cronin [26]. They describe *change blindness*, which occurs from interruptions in consequence of discarded frames. Change blindness is the surprising inability to detect large changes due to short visual interruptions. Examples are the inability to recognize the disappearance of buildings or the movement of large objects for a long period of time if motion in the video is interrupted. They additionally report that such interruptions are omnipresent in the context of CCTV footage, due to low frame rates. Discarding frames will cause interruptions of motion and finally induces change blindness.

To overcome this problem, we provide a visualization that does not skip any frames. We blend frames according to the information they convey. The blended frames can be regarded as images taken by a virtual camera with varying exposure time. All information over a small period is combined into a single image. In contrast to a real-world camera, we preserve constant luminance and support non-uniform weights of the original frames. We weight the original frames by their significance (i.e. the information gain) calculated by the proposed approach. For slow motion scenes, we limit the playback speed — this as well as the blending approach aids in tracking motion.

The effect of the proposed fast-forward visualization is similar to the image integration of the human eye when playing the video at higher speed. The human visual system requires an appropriate integration time to distinguish between temporally separated light flashes [11]. The integration time depends on several conditions like the luminance of the stimulus, contrast, spectral composition, the area of the retina stimulated, and the retinal position. The integration time varies between 10 ms and 100 ms (i.e. 100 – 10 Hz). Depending on these physiological constraints, the eye is integrating the light stimuli for different durations. We adapted the visualization to do the same: we only present the original frames as long as the hardware constraints do not limit the intended frame rate (i.e. 60 fps for LCD). In this case, the human visual system integrates the single frames on its own. Beyond this limit, we integrate the light stimuli similar as the human eye by blending the original frames according to their information gain.

We need to consider that recorded video footage is typically gamma-encoded. The original reason for the gamma correction is the non-linearity of CRT monitors: if we double the voltage of the signal sent to the display device, the radiometric (physical) brightness does not double. To address this problem, most of the video and image capture devices internally gamma encode the signal. The linearly scaled chromaticity stimuli (linear RGB) are non-linearly transformed to sRGB. If those non-linear values are now presented to the user on a CRT monitor, the intrinsic gamma decoding characteristic automatically re-transforms the signal to linear RGB (cf. Figure 3(b)). For this purpose, a gamma pre-correction is included in other display devices like LCDs. In the simplest case, the gamma-corrected (R', G', B') are calculated by $(R', G', B') = (R^\gamma, G^\gamma, B^\gamma)$. Images are typically encoded by the camera with $\gamma = \frac{1}{2.2}$ and decoded by the display device with $\gamma = 2.2$. A similar argument can be made for other color systems like YIQ or YUV, known from video systems. For more background information on gamma correction and color systems, we refer to [18,5].

Originally, the observed scene is not changed by an imaging–displaying process as depicted in Figure 3(a). To achieve physiologically correct integration we blend the
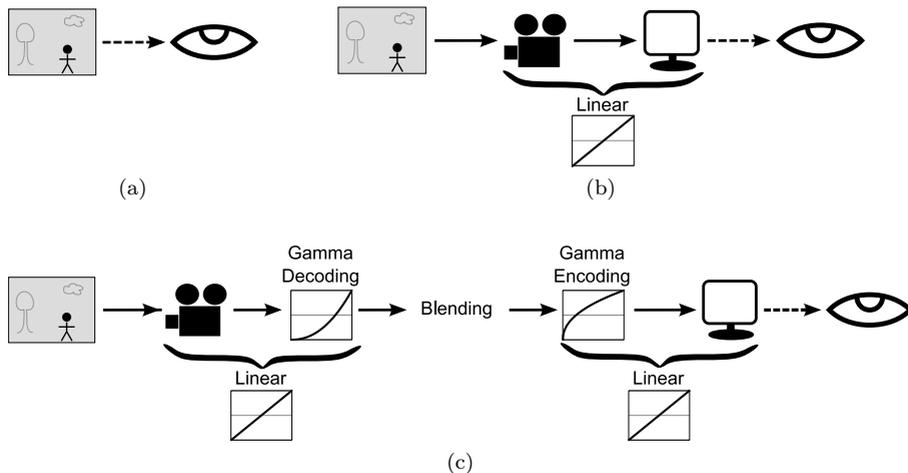
(a)

(b)

Linear

Gamma
Decoding

Gamma
Encoding

Blending

Linear

Linear

(c)

Fig. 3 (a) Visual stimuli arrive at the eye unchanged. (b) Videos and images captured by a camera are gamma-encoded. The monitor reverses the non-linear transformation (gamma decoding). (c) Artificial integration of visual stimuli according to the human vision system. Before blending, the image is transformed to linear RGB and finally gamma-encoded again.

frames in linear RGB space. Therefore, the input frames (which are already gamma-encoded) are firstly gamma-decoded back to linear RGB. Now correct blending is possible. After this step, the resulting images are gamma-encoded again (cf. Figure 3(c)).

Pixels in linear RGB are blended according to:

$$p^{out}_{(x,y)} = \sum_i w_i \cdot p^{in}_{(x,y),i} \tag{9}$$

A pixel of the blended frame (output frame, still in linear RGB) is represented by $p^{out}_{(x,y)}$; $p^{in}_{(x,y),i}$ denotes a pixel of an input frame, (x,y) is the position of the pixel and w_i is the weight factor. The index i refers to an input frame and its weight within the bounds of the first and the last input frame considered for the blended image.

The weights w_i are calculated by dividing the display time t_i by the output frame duration t_{out}:

$$w_i = \frac{t_i}{t_{out}} \tag{10}$$

Time t_i is the duration a frame should be displayed with respect to the information it conveys.

The number of frames blended into a single frame depends on their weights. The sum of the weights for each output frame should be 1 to preserve the luminance.

If $\sum_{i=j..k} w_i > 1$ and $\sum_{i=j..k-1} w_i < 1$, then weight w_k will be split and image k will be considered for two (or even more) output frames. Hence, the weight of frame k for the first output frame is $w_k^* = 1 - \sum_{i=j..k-1} w_i$. For the second output frame input frame k will be considered according to weight $w_k' = w_k - w_k^*$. A blending example is shown in Figure 4.

Input Frames

5 ms   5 ms   40 ms   5 ms   5 ms

5/30 = 1/6   1/6   4/6   4/6   1/6   1/6
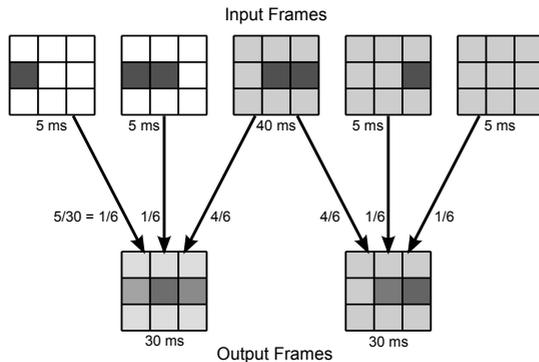
30 ms   30 ms

Output Frames

Fig. 4 Blending of frames in regard to their conveyed information: In the upper row, the input frames (3x3 pixel each, already gamma decoded) are displayed. The time each frame has to be shown is noted below. It depends on the information gain of the frame. The input pixel values are weighted by the fraction of the display time depending on the information gain and the output frame rate. The 3rd original frame is split and affects both blended frames.

6 Results and Discussion

The method proposed in this paper formulates the information gain ΔI as symmetric Rényi divergence of the absolute frame difference and a previously known (or estimated) noise distribution. Thus we derive the term "temporal information" from signal coding theory. The information gain is further used to adapt the playback speed of the video sequence to the amount of information communicated. Please note that information based adaptive video fast-forward is not capable of pointing out relevant events on its own, but more generally enables the user to keep track of the things going on in the video.

In this section, we evaluate the properties and the performance of the proposed approach. To illustrate the strengths and weaknesses of information-adapted video fast-forward for visual surveillance we will first evaluate the proposed technique using 3 different video sequences, then summarize the theoretical aspects of this method, and finally compare the applicability of different fast-forward techniques to the task of video surveillance conducting a qualitative user study. For a visual impression of the results we provide comparative evaluation sequences on our website[3] (a quick overview is provided in Table 1).

6.1 Evaluation of Playback-Speed Adaption

The first of the three sequences used for evaluation is provided by the video analysis mini challenge of the IEEE VAST 2009 challenge, referred to as *VAST Sequence*. This sequence is captured by a surveillance camera that periodically pans to four different locations. Additionally, the material was re-encoded and, thus, includes common coding artifacts (blocking, etc.), temporal shifts, and interlacing artifacts. Both other sequences are gathered from the i-LIDS multi-camera tracking scenario[4]. These sequences were collected by surveillance cameras on an airport. Both sequences were captured and encoded with high quality settings, i.e. large spatial and temporal resolution and high bitrates. While we use the first airport video as it is, the second

---

[3] www.vis.uni-stuttgart.de/~hoeferbn/AdaptiveFastForward/index.html

[4] Imagery library for intelligent detection systems (i-LIDS). http://scienceandresearch. homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/ i-lids/ilids-datasets-pricing/545840

12

Table 1 Overview of the examples provided on the website including their relation to the experiments of Section 6.1.

Name	Methods compared	Relation to experiments
Example 1 *Comparison to constant fast-forward*	Fast-forward at constant speed (no blending – discard frames)	Figure 6 shows acceleration profiles of various approaches on a part of the sequence used in this example.
	Information-based adaption (biologically corrected blending)	
Example 2 *Performance comparison under noise*	Motion-based adaption (biologically corrected blending)	Example illustrates some results of the experiment depicted by Figure 7.
	Information-based adaption (biologically corrected blending)	
Example 3 *Comparison of visualization techniques*	Information-based adaption (biologically corrected blending)	This example provides visual impression of different fast-forward visualization techniques.
	Information-based adaption (no blending – discard frames)	

airport sequence is degraded by adding Gaussian noise (50% normal distributed luminance changes). This was done to show the effects of different adaptive fast-forward techniques to videos with a high amount of noise. We call these sequences the *Airport Sequence* and the *Noisy Airport Sequence*, respectively.

To show the strength of information theoretic adaptive fast-forward, two excerpts of the *VAST Sequence* are presented in Figure 5. Both show the inverse velocity of the accelerated sequence, i.e. the information gain ΔI of several frame transitions. The first one (Figure 5 (a)) depicts a detailed view on some frames revealing temporal lags of the video sequence due to re-encoding. These lags originate from a different frame rate of the sequence before being re-encoded and introduce intermediate frames with no changes except for coding artifacts. The second chart (Figure 5 (b)) illustrates the sudden panning of the surveillance camera that generates pixel changes and is also overlaid with temporal re-encoding artifacts. The traditional fast-forward mode would increase the playback speed by a constant factor. This would result in complex analysis conditions since the temporal lags as well as the camera pan would be ignored in the playback acceleration.

In Figure 6, several approaches (number of pixels changed, mutual information according to Equation 1, average motion magnitude derived from Horn-Schunck [9],
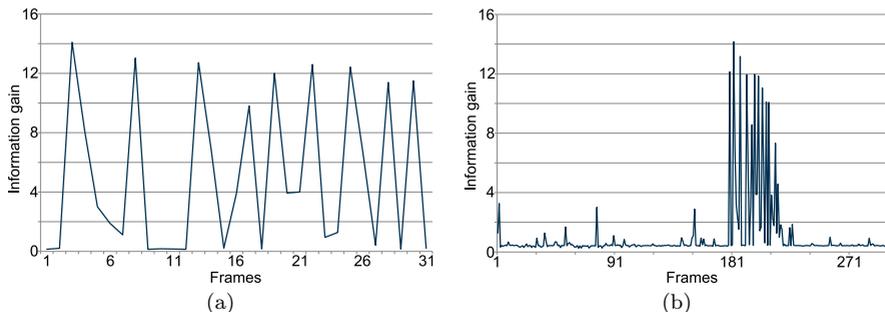
Fig. 5 Information gain on parts of the *VAST Sequence*: (a) detailed view of some frames to show lags due to re-encoding of the sequence, (b) increase of information gain during camera pan.
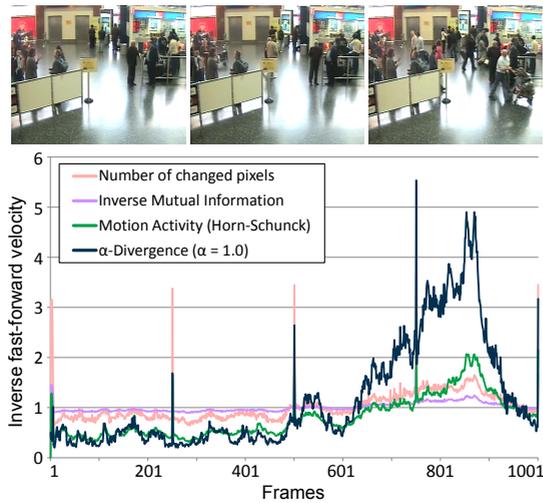
Fig. 6 Different measures (inverse velocity normalized to an expectation value of 1, i.e. the output sequences accelerated by this measure have equal duration) for adaptive fast-forward, applied to the *Airport Sequence*. Top row shows three keyframes of this sequence for rough visual evaluation of the results.

symmetric α-divergence with $\alpha = 1$) are compared to each other using an excerpt of the *Airport Sequence*. Please note that we use the Horn-Schunck method for calculating the optical flow instead of block matching as it is intrinsically applied by Peker *et al.* [16]. This is done to obtain more precise motion vectors and a dense motion field due to Horn-Schunck's regularization term, increasing the quality of motion-based fast-forward. Block matching is generally designed for video coding applications and thus optimizes for PSNR instead for semantically correct motion vectors. We also tested more recent approaches to calculate the dense motion field, but since results were qualitatively equal regarding playback speed adaption, we decided to use the popular Horn-Schunck method for evaluation. The chart in Figure 6 shows the inverse fast-forward velocities (related to the amount of motion, information, etc.) normalized to an expectation value of one. In the first half of this excerpt, only a short moments of activity is observed; however, in the second half the activity level raises due to many persons moving. It is clearly visible that all methods compared, measure the activity level to some degree, i.e., all methods share the same qualitative and intended behavior. It can also be observed that for all evaluated methods there is a distinct peak at every 250th frame. These peaks are caused by the different encoding of keyframes, resulting in more pixel changes.

The advantage of the proposed method over recent other approaches in terms of robustness to noise is illustrated in Figure 7. The chart depicts the inverse fast-forward velocities, extracted from the *Noisy Airport Sequence*. Surveillance sequences with large amounts of noise often suffer from badly illuminated scenes or strong encoding artifacts. In this case, motion activity is completely dominated by video noise and the normalized playback velocity does not reflect the movement activity of people. Conversely, in periods with almost no scene activity, the motion magnitude is quite high, resulting from the wrong estimation of optical flow due to noise. Similar estimation errors occur in the case of static changes that are common in surveillance videos with low temporal resolution ($< 1\,\text{fps}$). Since correct optical flow estimation is not possible in those cases, a wrong motion field is calculated and thus, the motion magnitude does not reflect the movement in the video. An example of such a case is depicted in Figure 8. In the same way, measures based on the amount of changed pixels or on mutual information fail to
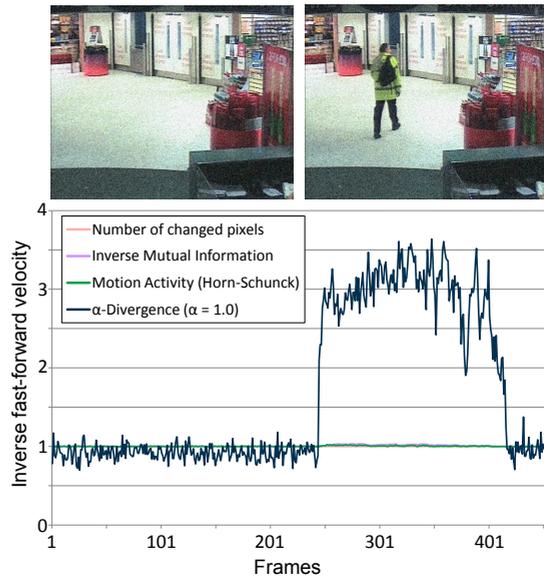
Fig. 7 Different measures applied to the *Noisy Airport Sequence*. All evaluated measures badly adapt the playback speed to the walking man in the second half of the sequence, except for our new method.

recognize the movement activity apart from noise. In those cases, our new approach is superior to the others, as long as a reasonable noise estimation is provided.

Peker and Divakaran [15] consider the spatio-temporal complexity (visual complexity) of a video sequence as cue for its playback velocity. Note that in most cases the proposed information measure intrinsically covers the spatio-temporal complexity of the video sequence, while neglecting the noise component. We show this for two typical examples: an object increasing in size, while moving towards the camera and an object that moves orthogonal to the camera view. While the first object approaches, its image enlarges. Thus, the information gain increases, since it is based on the relation between frame difference distribution and noise distribution. As a second example, we consider the information gain of a textured object (high spatial complexity) and a homogeneously colored object with low spatial complexity, both moving orthogonal to the camera direction. In the case of the object with low spatial complexity, only the borders orthogonal to the movement direction contribute to the frame difference distribution. Hence, the information gain is lower than for the textured object, where the difference image of the frames shows a greater number of changes.

Fig. 8 Three subsequent frames of a temporal subsampled surveillance sequence with arbitrary motion vectors due to static changes.

The α parameter of the Rényi divergence regulates the emphasis of the measure to certain parts of the probability density functions — as already mentioned in Section 3. This effect can be deemed an amplification of certain distribution ratios. While the Rényi divergence tends to treat all probability ratios as equal for $\lim_{\alpha \to 0}$, the α-divergence converges to the amplification of only the highest probability ratio for $\lim_{\alpha \to \infty}$. Within a certain range of α, the parameter can be considered an acceleration of the fast-forward velocity as it emphasizes the higher probability ratios. This effect is illustrated in Figure 9 for a short period of the *Airport Sequence*.

In Figure 10, we depict the adaption of the information gain with increasing number of noise distributions. The different noise distributions are calculated according to the method proposed in Section 4. As the number of noise distributions increases, the impact of wrong noise estimations, originating from coding artifacts in keyframes, is reduced. The noise distributions improve slightly even outside the keyframes, since the influence of the outliers is removed.

6.2 Theoretical Aspects of Playback Adaption

The findings in the above examples are backed by the following theoretical considerations. Assuming the true noise distribution N to be provided, the proposed approach calculates the temporal information gain ΔI based on the symmetric Rényi divergence $\hat{\mathcal{D}}_\alpha$, cf. Equation 6. By calculating the α-divergence as information "distance" to the noise distribution, we consider image changes originating from temporal noise to be irrelevant for the user. This property of the proposed information measure agrees with the intuitive human perception and its relevance rating of image changes to be valuable information or not. On the other hand, the proposed measure is generally (e.g., in the case of Gaussian noise distribution) sensitive to the absolute frame distance. That means that the amount of information retrieved by ΔI depends on the luminance difference of moving objects to the background.

Although human perception is to some degree robust to inaccurate playback velocity, the correctness of the calculated information gain depends on the quality of noise estimation. For theoretic considerations of the proposed approach, we expect a perfect noise distribution to be provided. However, our evaluation results were using an estimated noise distribution based on a simplified noise model. These practical and realistic tests have shown that even simplified noise models lead to appropriate play-
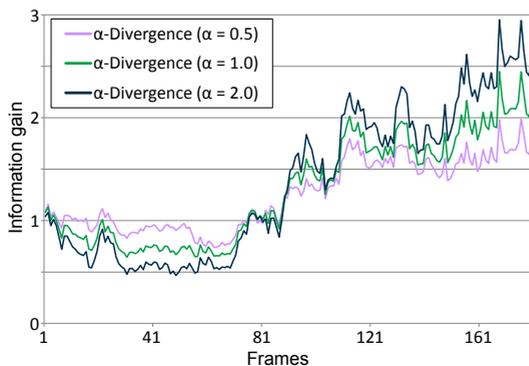

Fig. 9 Effect of the α parameter of the Rényi divergence: $\alpha > 0$ emphasizes the speed modification for fast-forward, whereas $\alpha \to 0$ levels out those modifications. The information gain is normalized to an expectation value of 1.
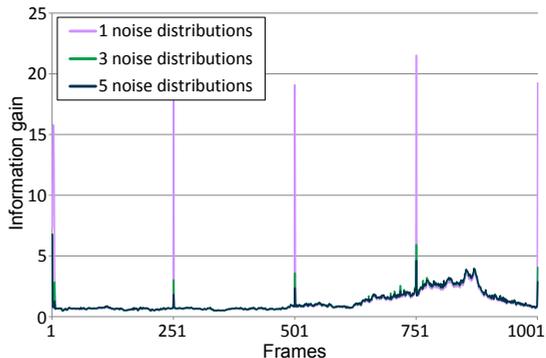
Fig. 10 Reduction of information gain error due to wrong noise estimation with the increase of noise distribution clusters. This chart was calculated on the *Airport Sequence.*

back velocities. A more sophisticated noise estimation would consider other types of noise, temporal changing noise (especially non-periodic changes due to lighting conditions), or spatially inhomogeneously distributed noise. Such approaches would result in more precise estimations of the information gain, with a possibly further improved adaptation of playback velocity.

The uniform time complexity of the proposed approach is in $O(n+k)$, where n is the number of pixels in the two frames and k represents the number of histogram bins of the discrete luminance distance distributions. The low complexity of this approach allows real-time processing of video sequences in contrast to other approaches that deal with more complex features and thus have non-linear algorithmic time complexity. For real-time processing, these other approaches often rely on the availability of preprocessed features and thus are less flexible. To get a rough sense for the different complexities of the activity measures, we provide a small benchmark using single-threaded, non-optimized C++ code on an Intel Core2 Quad CPU Q6600 (2.4 GHz): for the *Airport Sequence*, we achieve an average processing rate of 87 fps on PAL resolution (720×576) and 36 fps on a HD-1080 sequence (1920×1080), whereas the Horn-Schunck motion-based approach yields 12.5 fps / 1 fps and block matching 32 fps / 2.5 fps for PAL and HD, respectively. Note that the number of frames that have to be processed each second for real-time processing depends on the fast-forward properties and may approach hundreds of frames in sequences with little interframe changes.

6.3 Qualitative User Study

To evaluate the applicability of adaptive fast-forward systems in video surveillance, we conducted a qualitative user study by means of expert interviews. The goal of this study was to determine if playback adaption based on low-level features is generally capable of emphasizing the periods of surveillance footage that are perceived to be relevant for video analysis by security personnel. Further, we tested whether the proposed approach in particular outperforms existing approaches, as we expected from results of previous sections. Please note that we evaluate playback-adaption at a higher semantic level (video surveillance) than the approaches were immediately designed for (pixel change information). Hence, differences may sometimes appear between expected playback speed and the behavior of the adaption algorithm. This discrepancy is often called the *semantic gap.*

Sequence 1   Sequence 2   Sequence 4

Fig. 11 Frames of the scenarios used in expert interview.

Task Description & Experimental Protocol The study was introduced by explaining the idea of adapting playback speed. Then, an example of constant and adaptive fast-forward was provided. After this tutorial, we confronted the participants with 4 video sequences of different quality, duration, and activity level. After each presentation (each sequence was subsequently accelerated by the three methods: constant speed, adaption by Rényi entropy, and adaption according to motion), we asked the participants prepared questions. Among them, we asked to provide their opinion about each method's adaption efficiency ("Did the adaption of playback speed support monitoring by reducing tedious periods while keeping an overview of bustling parts of the video?") and their conformity with user expectation ("Does the acceleration/deceleration of playback speed meet your expectation?"). In addition, the experts were advised to *think-aloud* in order to investigate how they perceive particular situations (stress, boredom, etc.). Finally, we asked them to judge the acceptance of adaptive fast-forward methods in the field of video surveillance and to provide suggestions for the improvement of such methods. The interviews had an average duration of 45 minutes.
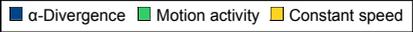
Participants The group of experts that participated in the qualitative user study was composed of visualization experts and domain experts. We chose visualization experts since they have knowledge about the visual communication process, while the domain experts are experienced in monitoring people and places and thus are able to rate the relevance of particular situations more reliable. The experts for visualization are research associates of the institute for visualization and interactive systems of the University of Stuttgart. This group included three males and one female. The domain experts, both male, are employees of a security company and work as CCTV operators. Both of them received reimbursement. All 6 participants had (corrected) normal vision.

Experimental Setup Participants were individually interviewed and the study was conducted by two interviewers, one asking questions and one preparing a protocol. Additionally, each participant was recorded by a webcam to backup the protocol and to capture facial expressions while watching the sequences. The video sequences were presented by a PC with a standard TFT-display.

Stimulus For the expert interviews, we chose 4 video sequences showing different situations that may emerge during surveillance monitoring. The first sequence covers a continuously crowded airport hall, to evaluate playback speed adaption in situations with high activity. We accelerated this sequence to be in average 5 times faster than the non-accelerated version (09:52 minutes duration). The second sequence is similar

Table 2 Results of the expert interview for the questions: a) "Did the adaption of playback speed support monitoring by reducing tedious periods while keeping an overview of bustling parts of the video?", b) "Does the acceleration/deceleration of playback speed meet your expectation?"

	Stimulus 1 *Crowded Airport*	Stimulus 2 *Airport*	Stimulus 3 *Noisy Airport*	Stimulus 4 *Night*
Duration	09:52 min	14:20 min	14:20 min	21:08 min
Mean Speed	5x	10x	10x	10x
a) Efficiency	1 / 1 / 5	0 / 1 / 5	- / 0 / 6	- / 2 / 3
b) Conformity with user expectation	1 / 4 / 1	0 / 1 / 5	- / 1 / 6	- / 1 / 2

■ α-Divergence ■ Motion activity ■ Constant speed

to the *Noisy Airport Sequence*, but without noise. It contains periods without activity as well as crowded situations. The video (14:20 minutes duration) was accelerated in average by factor 10. The third sequence we used was the *Noisy Airport Sequence* with the same speed-up settings. These three video sequences were obtained from i-LIDS multi-camera tracking scenario[5] and are sampled with 720×576 pixels at 25 fps. The last sequence was monochrome video captured at night. Hence, contrast is very low and sensor noise is dominant due to high gain settings. The sequence (656×494 pixels, 15 fps, 21:08 minutes duration) was also accelerated in average by factor 10. Keyframes of the scenarios used in expert interview are depicted in Figure 11. The playback speed of both adaptive fast-forward approaches was adjusted to match the duration of constant cue-play, in order to make them comparable. To counterbalance learning effects, we randomly permuted the presentation order of the three methods and displayed them anonymously.

Results The results of the qualitative user study based on the experts' comments and our observations during the study can be summarized as follows (for a comparative summarization see Table 2):

- Adapting the playback speed relative to the content in the video was appreciated by all participants. They mentioned that watching a video this way is more efficient and they felt more confident than watching the video at constant playback speed. They also felt it was more comfortable due to less stress during periods of high activity. The advantage of an adaptive playback speed depends highly on the activity in the video. For crowded scenes, such as sequence 1, the benefit was marginal.
- *Efficiency:* For each of the sequences, the majority of participants rated the proposed method to be most appropriate for monitoring.

---

[5] Imagery library for intelligent detection systems (i-LIDS).

- *Conformity with user expectation:* For sequences 2, 3, and 4, the proposed method met the users' expectation of speed adaption more closely than the motion-based method. There were different reasons for the poor impression of the motion-based method: in sequence 2, there are strong accelerations when people exit the room. In sequence 3 and 4, noise is present and inhibits correct motion calculation. For the first sequence, most of the participants preferred the approach based on motion. In this case, lots of highlights are present. A person moving across a highlight yields slower playback speeds compared to a person moving in the dark, which results in jerky playback speed changes. For both adaptive methods, users found it more difficult to estimate the speed of people compared to the constant method. They suggested to add visual feedback to increase awareness of playback speed.
- The participants did not recognize any differences between the adaption of playback speed in sequence 2 (without noise) and 3 (including noise) for the proposed method. Contrary, for the third sequence the motion-based approach was wrongly interpreted as having no adaption. Sequence 4 (low contrast, noise) was judged to be not optimally adapted by any adaptive methods, but the proposed method was slightly favored over the motion-based approach.
- The preferred acceleration varies from participant to participant. While some were bored during slow-motion playback of crowded periods, others were comfortable with the speed, and some desired an even slower playback. For simplification reasons, we did not provide a control to adjust the acceleration level for the qualitative user study. In practice, surveillance operators should be provided with an acceleration control to adjust the information load according to their personal abilities.

7 Conclusion And Future Work

In this paper, we have formulated the temporal information of a video sequence as symmetric information gain between an estimated noise distribution and the absolute frame difference distribution by means of the Rényi entropy. This has enabled us to adapt the playback velocity of video to this information theoretic measure. The proposed approach was evaluated under different sequences exhibiting common properties and difficulties of surveillance videos, like noise and static changes. We have compared the results to recent adaptive fast-forward approaches and have pointed out the advantage over conventional cue-play. The main advantages of our method are robustness against noise, suitability for low-framerate videos, and high computational efficiency. We have also conducted a qualitative user study which verifies the efficiency of adaptive fast-forward in visual surveillance and points out the advantage of the proposed method. Further, we have discussed possible fast-forward visualization methods and proposed a biologically motivated approach for information-preserving visualization that can handle arbitrary playback velocities. The combination of adaptive playback velocity and an appropriately animated display is the basis for user-centered browsing and the analysis of extensive surveillance video data.

In future work, noise estimation could be improved to cope with temporally varying and other types of noise arising from video processing. Additionally including color information, which is currently ignored, could improve the saliency of the temporal information measure. As further enhancement of fast-forward visualization, the introduction of movement interpolation should be considered to provide smoothed visual results in slow-motion periods in scenes with high activity. Further, the participants

of the conducted study asked for visual feedback on the playback speed and a smooth and consistent adaption to the amount of activity (i.e., by a low-pass filter). Finally, a quantitative controlled user study including recent adaptive fast-forward methods is required to judge the advantage of our method as well as the benefit of adaptive fast-forward in general.

References

1. M. Basseville. Distance measures for signal processing and pattern recognition. Technical Report 899, Institut de Recherche En Informatique Et Systemes Aleatoires (IRISA), 1988.
2. S. Blyth. Local divergence and association. *Biometrika*, 81(3):579–584, 1994.
3. K. Cheng, S. Luo, B. Chen, and H. Chu. Smartplayer: user-centric video fast-forwarding. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI)*, pages 789–798. ACM New York, NY, USA, 2009.
4. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.
5. M. D. Fairchild. *Color Appearance Models*. Wiley-IS&T, Chichester, UK, 2005.
6. M. Ghazal, A. Amer, and A. Ghrayeb. A real-time technique for spatio–temporal video noise estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(12):1690–1699, 2007.
7. M. Gill and A. Spriggs. *Assessing the impact of CCTV*. Home Office Research, Development and Statistics Directorate, 2005. Home Office Research Study 292.
8. A. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval. Technical Report CSPL-328, Communication and Signal Processing Laboratory, The University of Michigan, 2002.
9. B. Horn and B. Schunck. Determining optical flow. *Computer Vision*, 17:185–203, 1981.
10. H. Kang, X. Chen, Y. Matsushita, and X. Tang. Space-time video montage. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1331–1338, 2006.
11. H. Kolb, E. Fernandez, and R. Nelson. *Webvision: The Organization of the Retina and Visual System. Part IX. Psychophysics of Vision*. National Library of Medicine (US), NCBI, June 2007. Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid= webvision.part.4145.
12. S. Kullback. *Information Theory and Statistics*. Wiley Publication in Mathematical Statistics, 1959.
13. J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.
14. H. Neemuchwala, A. Hero, and P. Carson. Image matching using alpha-entropy measures and entropic graphs. *Signal Processing*, 85(2):277–296, 2005.
15. K. Peker and A. Divakaran. Adaptive fast playback-based video skimming using a compressed-domain visual complexity measure. In *2004 IEEE International Conference on Multimedia and Expo, 2004. ICME'04*, volume 3, pages 2055–2058, 2004.
16. K. Peker, A. Divakaran, and H. Sun. Constant pace skimming and temporal sub-sampling of video using motion activity. In *Proc. IEEE International Conference on Image Processing, Thessaloniki*, volume 3, pages 414–417, 2001.
17. N. Petrovic, N. Jojic, and T. Huang. Adaptive video fast forward. *Multimedia Tools and Applications*, 26(3):327–344, 2005.
18. C. Poynton. *Digital Video and HDTV: Algorithms and Interfaces*. San Francisco: Morgan Kaufmann Publishers, 2003.
19. Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *Proc. ICCV*, pages 1–8, 2007.
20. Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1971–1984, 2008.

21. A. Punchihewa and D. Bailey. Artefacts in image and video systems; classification and mitigation. In *Proceedings of Image and Vision Computing New Zealand*, pages 197–202, 2002.

22. G. Ramos and R. Balakrishnan. Fluid interaction techniques for the control and annotation of digital video. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 105–114. ACM New York, NY, USA, 2003.

23. A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 435–441, 2006.

24. A. Renyi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

25. K. Schoeffmann and L. Boeszoermenyi. Video browsing using interactive navigation summaries. *International Workshop on Content-Based Multimedia Indexing*, 7:243–248, 2009.

26. K. Scott-Brown and P. Cronin. An instinct for detection: Psychological perspectives on CCTV surveillance. *The Police Journal*, 80(4):287–305, 2007.

27. B. Song and K. Chun. Motion-compensated noise estimation for efficient pre-filtering in a video encoder. In *IEEE International Conference on Image Processing*, volume 2, pages 211–214, 2003.

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