Surveillance Centric Coding

T. Zgaljic*, N. Ramzan*, M. Akram*, E. Izquierdo*, R. Caballero†, A. Finn†, H. Wang†, Z. Xiong†

*Multimedia and Vision Research Group, Queen Mary University of London, UK
{toni.zgaljic, naeem.ramzan, muhammad.akram, ebroul.izquierdo}@elec.qmul.ac.uk
†United Technologies Research Center, East Hartford, Connecticut, USA
{CaballRE, FinnAM, WangH1, XiongZ}@utrc.utc.com

Keywords: surveillance, surveillance centric coding, scalable coding.

Abstract

In this paper we introduce the paradigm of Surveillance Centric Coding (SCC), in which coding aims to achieve bit-rate optimisation and adaptation of surveillance videos for storing and transmission purposes. In the proposed approach the SCC encoder communicates with Video Content Analysis (VCA) module that detects events of interests in video captured by CCTV. Bit-rate optimization and adaptation is achieved by exploiting the scalability properties of the employed codec. Time segments containing events relevant to surveillance application are encoded using high spatio-temporal resolution and quality while the irrelevant portions from the surveillance standpoint are encoded at low spatio-temporal resolution and / or quality. Thanks to the scalability of the produced compressed bit-stream, additional bit-rate adaptation is possible, for instance for the transmission purposes. Experimental evaluation shows that significant reduction in bit-rate can be achieved by the proposed approach without loss of information relevant to surveillance applications.

1 Introduction

In surveillance applications video captured by CCTV is usually encoded using conventional compression technology, such as MPEG-1/2 or H.264/AVC. These systems encode the video signal regardless the video's significance as determined by video content analysis (VCA) modules used for automatic recognition, detection, tracking, etc. The reason is obvious - when using a conventional coder it is almost impossible to adapt the compressed stream to the scene properties in terms of activity and degree of event significance. Thus, the output of VCA cannot be used directly after or during signal encoding and transmission, unless an expensive and in several cases unfeasible transcoding process is carried out. As an example, in many surveillance situations the scene remains essentially static for seconds and even minutes in some cases. During these periods of time nothing interesting happens from the surveillance standpoint, and the video resembles a still picture for long periods of time with no other activity than random environmental motion. This is the case of surveillance in metro stations during night hours, or private car parking where the usual events are cars coming and leaving from time to time. These events, relevant for surveillance applications, are detectable by the VCA acting on the output signal from the cameras.

Several attempts have been made in the past to produce compressed content that retains only information relevant to surveillance applications. In [1] only those spatial segments of the video that are identified as the foreground were encoded using MPEG-4 Part 2 object based coding. The background was encoded as a static (or periodically refreshed) image and repeated throughout the sequence. Here, foreground objects are obtained by background subtraction algorithms and refined by performing analysis of their motion. In [2] this approach was further extended by deriving the compression efficiency model that considers the number and size of foreground objects. This work was motivated by the observation that if the size and number of foreground objects is high, the object based coding may produce worse results than conventional frame-based coding. Therefore, by using the derived compression model the encoder can adaptively chose whether to perform frame-based or object based coding for specific time segment of a surveillance video. In [3] a similar approach to [1] and [2] was proposed, which combines tracking and an own-developed object-based coding framework.

In this paper we propose an alternative approach to reduce the bit-rate of the encoded video segments that are irrelevant from the surveillance standpoint. The proposed approach combines background subtraction and wavelet-based scalable video coding. This produces a single scalable bit-stream that contains segments of video encoded at different qualities and / or spatio temporal resolutions. The irrelevant segments are encoded using low resolution / quality while the relevant segments are encoded at high resolution / quality. Additionally, the produced scalable bit-stream can easily be adapted for transmission purposes, without the need for computationally expensive transcoding.

The remainder of this paper is organised as follows. Section 2 provides an overview of surveillance coding framework to which the work presented in this paper has been tailored. Section 3 discusses the proposed approach for event-based encoding of surveillance videos. Section 4 provides experimental evaluation of the proposed approach. Finally, section 5 concludes this paper.

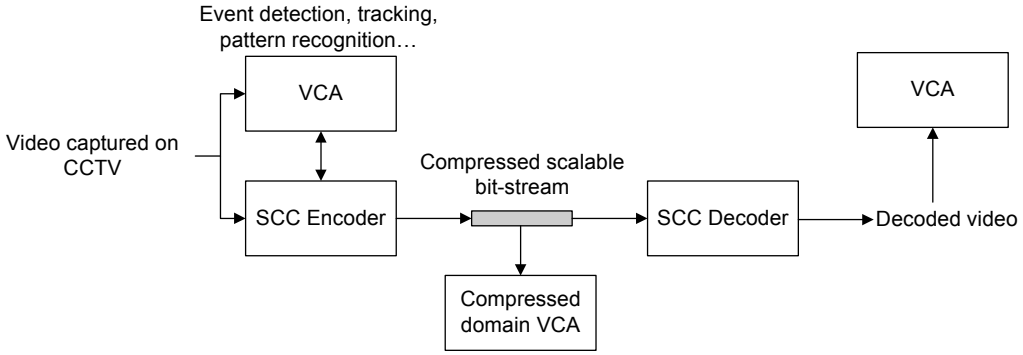

Figure 1: A generic SCC system.

2 Surveillance centric coding

In this section, we introduce the coding paradigm of Surveillance Centric Coding (SCC), in which coding for specific surveillance applications is targeted. SCC aims at exploiting specific properties of surveillance video in a comprehensive application framework including coding adaptation to surveillance, rate distortion optimisation according to VCA, and other related concepts. The architecture of a generic SCC system is outlined in Figure 1. The work presented in this paper focuses on the use of VCA to drive encoding, transmission, and streaming. By this approach the use of available resources is optimised according to the requirements of surveillance applications. Here, the term resources refers to storage systems, bandwidth, and trade-off between received signal and display devices. No analysis of the actual encoder for quantization and rate-distortion optimization according to VCA is presented. The latter is critical to the SCC concept but out of the scope of the work reported here. In the proposed approach, the SCC encoder communicates with the VCA modules and performs encoding by rate-optimisation according to events as specified by the VCA. The VCA can also be used at the decoder-side for off-line processing, e.g., car plate recognition, face detection, etc. Therefore, the question behind the work presented in this paper is how to exploit the information resulting from the VCA to tailor the coding and transmission or streaming of the video signal. Clearly, this cannot be achieved with conventional coding technology without complex transcoders. It can however be achieved if fine granularity scalable coding technology is applied.

3 Event-based encoding of surveillance video

As mentioned in the introductory section the basic principle behind the presented work is to use different encoding settings for segments of surveillance video that show different level of activity. For this purpose we classify the surveillance video into temporal segments that contain essentially static scene and segments that show some level of motion activity. To perform this classification, background subtraction and tracking module from [4] is used as VCA. The output of this module dictates the quality / spatio-temporal resolution of the encoded content. For actual encoding the wavelet-based scalable video codec – aceSVC [5] is employed. As these two modules represent main building blocks of the proposed framework we explain them briefly in the following two subsections and then we introduce the proposed system for event-based encoding of surveillance videos.

3.1 Background subtraction for motion tracking

Adaptive background subtraction method based on mixture of Gaussians [4] is used. This method is able to deal robustly with lightning changes, bimodal background like swaying trees and introduction or removal of objects from the scene. The value of each pixel is matched against weighted Gaussians of the mixture. If the pixel value is within 2.5 standard deviations of any Gaussian distribution then the mean value and standard deviation of the corresponding Gaussian are updated. If the pixel value is not within 2.5 standard deviations of any distribution then the least probable distribution is replaced by the new distribution. The mean value of the new distribution is set as the value of the current pixel and its initial variance is set to a high value. Weights are continuously updated for each distribution of the mixture. At each time instance Gaussians of the mixture that represent the background are identified according to the predefined threshold. The pixels whose value is not within 2.5 standard deviations of the Gaussians representing the background are declared as the foreground. Foreground pixels can then be segmented into regions and tracked throughout the sequence. The output of background subtraction and tracking module is illustrated in Figure 2.

3.2 Scalable coding

The employed wavelet-based scalable video coding architecture – aceSVC – features spatial, temporal, quality and combined scalability. Temporal scalability is achieved by applying temporal wavelet filtering in direction of motion vectors. In the first level of temporal decomposition the odd input frames are decomposed into high-pass frames (predicted frames) and the even input frames are decomposed into low-pass frames (averaged frames or approximations). The second level of temporal decomposition is performed by applying temporal filtering on low-pass frames resulting from the

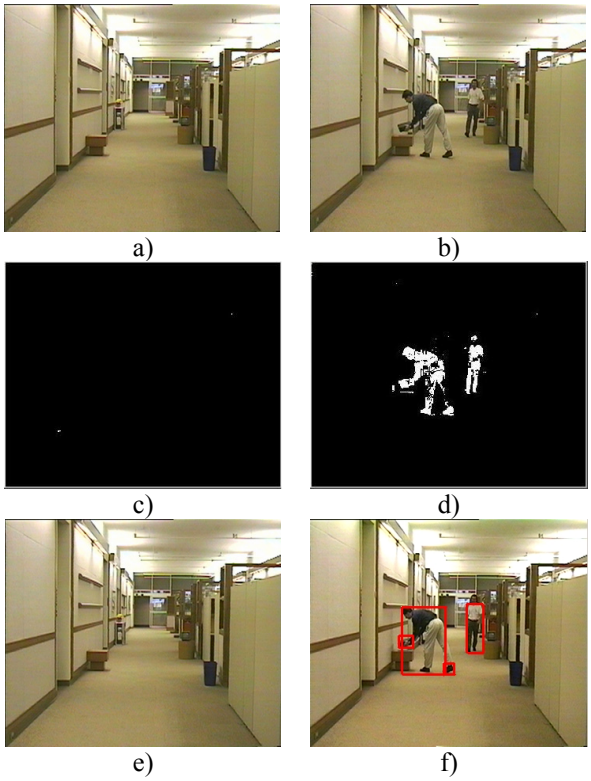

Figure 2: Background subtraction. a) 3-rd and b) 110-th frame of the hall sequence. Result of the background subtraction for the c) 3-rd and d) 110-th frame. Result of tracking for the e) 3-rd and f) 110-th frame.

first level of temporal decomposition. The video decomposition continues in such manner until desired number of temporal decomposition levels has been reached. The required temporal resolution of the decoded video can be achieved by discarding high-pass frames at the appropriate levels of temporal decomposition directly from the compressed scalable bit-stream.

To achieve spatial scalability each frame is decomposed using a 2D wavelet transform. The result after first level of spatial decomposition is low-pass subband and three high pass-subbands, each of them covering the quarter of the original frame. To reach the second level of spatial decomposition 2D wavelet transform is performed on low-pass subband resulting form the first level of spatial decomposition. Again, the desired spatial resolution of the decoded video can be achieved by discarding the appropriate high-pass subbands.

Quality scalability is achieved through bit-plane coding of wavelet coefficients resulting form spatio-temporal wavelet transform explained in the previous two paragraphs. During the bit-plane coding most significant bits of all wavelet coefficients are encoded first. Encoding continues with the bit positioned next to the most significant bit and progresses towards the least significant bit of all wavelet coefficients. In this way precision of wavelet coefficients increases in a progressive way, thus providing fine granular quality scalability of the compressed content.

The other features of the employed coded are [5]: hierarchical variable size block matching motion estimation, flexible selection of wavelet filters for both spatial and temporal wavelet transform on each level of decomposition, including the 2D adaptive wavelet transform in lifting implementation and embedded zeroblock coder with binary arithmetic coding.

3.3 Proposed system

The proposed system for application in SCC coding is outlined in Figure 3. At each time instance the encoder communicates with VCA module (background subtraction and tracking). When the input video is essentially static the output of the background subtraction does not contain foreground regions. This can be used to signal to the encoder to encode captured video at low spatio-temporal resolution and quality. Encoding parameters in this case can be defined by the user or they can be chosen automatically by the system. This allows, for instance, encoding and / or transmitting the portions of the video containing long, boring, static scenes using low quality frame-rate and spatial resolution. On the other hand, when some level of activity in the captured video is detected, the VCA module notifies the encoder to automatically switch encoding to a desired much higher spatio-temporal resolution and quality video. Therefore, decoding and use of the video at different

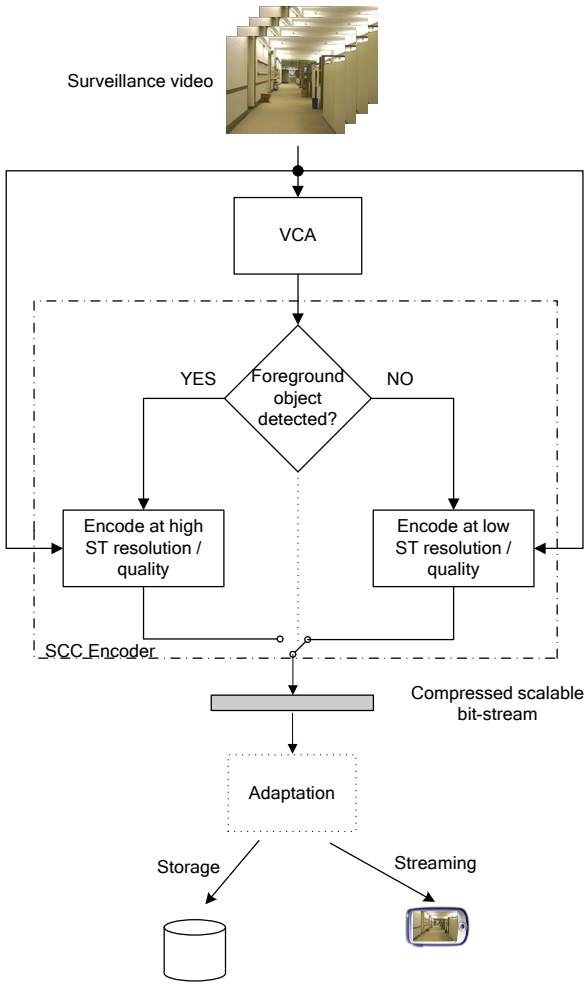

Figure 3: The workflow of the event-based encoding.

spatio-temporal resolutions and qualities corresponding to different events is achieved from a single bit-stream, without multicasting or complex transcoding. Moreover, additional optional adaptation to lower bit-rate is also possible without decoding the video. This is, for instance, very useful in cases where video has to be delivered to a device with a low display capability. Using this approach, the bit-rate of parts of the video that are of low interest is kept low while the bit-rate of important parts is kept high. In many realistic applications it can be expected that large portions of the captured video have no events of interest. Thus, the proposed model leads to significant reduction of resources without jeopardizing the quality of any off-line event detection module.

As added side effect, critical metadata encoding events detected by the VCA can be interleaved with the bit-stream. Therefore events of interests can easily be identified in the compressed domain just by reading the corresponding metadata extracted directly from the video stream.

4 Experimental results

The performance of the proposed SCC framework has been evaluated using three different typical surveillance sequences: the "Parking" sequence at resolution of 352×288 pixels (CIF) and frame-rate of 25 Hz and "Bridge-far" and "Hall" sequences at CIF resolution and frame-rate of 30 Hz. The length of the "Parking" and "Bridge-far" sequences were 2100 frames and the length of the "Hall" sequence was 300 frames. Encoding was performed on group of pictures (GOP) by GOP basis, i.e. the smallest group of frames that can be encoded / adapted according to the output of VCA is defined by the GOP size. Here, we use the GOP of 16 frames.

The proposed framework is evaluated with respect to its flexibility and efficiency. Criteria given in Table 1 are used to adapt the encoding of surveillance sequences with respect to different spatio-temporal resolutions and qualities. If the VCA detects no event in the particular portion of sequence, it passes the information to the SCC encoder that performs the encoding process either at low spatial resolution (spatial adaptation – QCIF), frame-rate (temporal adaptation – half of original frame-rate) or the combination of all three scalability directions (spatial, temporal, quality), as given in Table 1. When the event is detected by VCA, encoding is switched to original resolution and frame-rate and high quality.

Table 2 shows byte savings for each testing sequence, using the SCC encoder with different types of rate adaptation. The byte savings are shown relatively to the corresponding sequence encoded at full resolution, frame-rate and high quality for the whole length. Relative byte saving is calculated as:

$$rbs = \left(1 - \frac{NB_{\mathrm{SCC}}}{NB_{\mathrm{conv}}}\right) \cdot 100\% \qquad (1)$$

where NB_{SCC} represent the number of bytes of the compressed sequence using the SCC approach from Figure 3 and NB_{conv} represent number of bytes of the compressed sequence encoded at full resolution, frame-rate and quality for the whole length. From Table 2 can be observed that the compression gains for the "Hall" sequence are rather small. This is because throughout the whole sequence some level of activity is present and therefore almost whole sequence is encoded at original spatio-temporal resolution and high quality. For the other two sequences significant bit-rate savings can be observed.

Event	Adaptation	Bit-rate
Essentially static scene	Spatial	288 kbps
	Temporal	400 kbps
	Combined	128 kbps
Event detected (Man/truck/boat)	Full spatio-temporal resolution	512 kbps

Table 1: Adaptation bit-rates.

Sequence	Adaptation	*rbs* (%)
"Parking"	Spatial	25.6
	Temporal	11.3
	Combined	38.4
"Bridge-far"	Spatial	16.6
	Temporal	7.4
	Combined	26.5
"Hall"	Spatial	5.4
	Temporal	2.4
	Combined	8.0

Table 2: Relative byte savings.

Subjective results of the proposed SCC for the "Hall" sequence are presented in Figure 4. The first row shows original frames. The second row represents binary mask of the original video, which is the output of the background subtraction module. The third row shows the reconstructed sequence whose essentially static segments were encoded at lower spatial resolution. The fourth row represents the temporally adapted sequence. Note that only one of the two consecutive original frames is kept in the adapted portion of the sequence. The last row shows the combined scalability, i.e. reduction of spatio-temporal resolution and quality.

5 Conclusions

In this paper we proposed the coding system that performs rate optimisation and adaptation in surveillance applications. This is achieved by an interaction between Video Content Analysis (VCA) module and wavelet-based scalable video coding. Time segments containing events relevant to surveillance applications are detected by VCA and encoded using high spatio-temporal resolution and quality. The other portions of the video are encoded at low spatio-temporal resolution and / or quality. Experimental results showed that significant bit-rate reductions can be achieved by using the proposed approach.

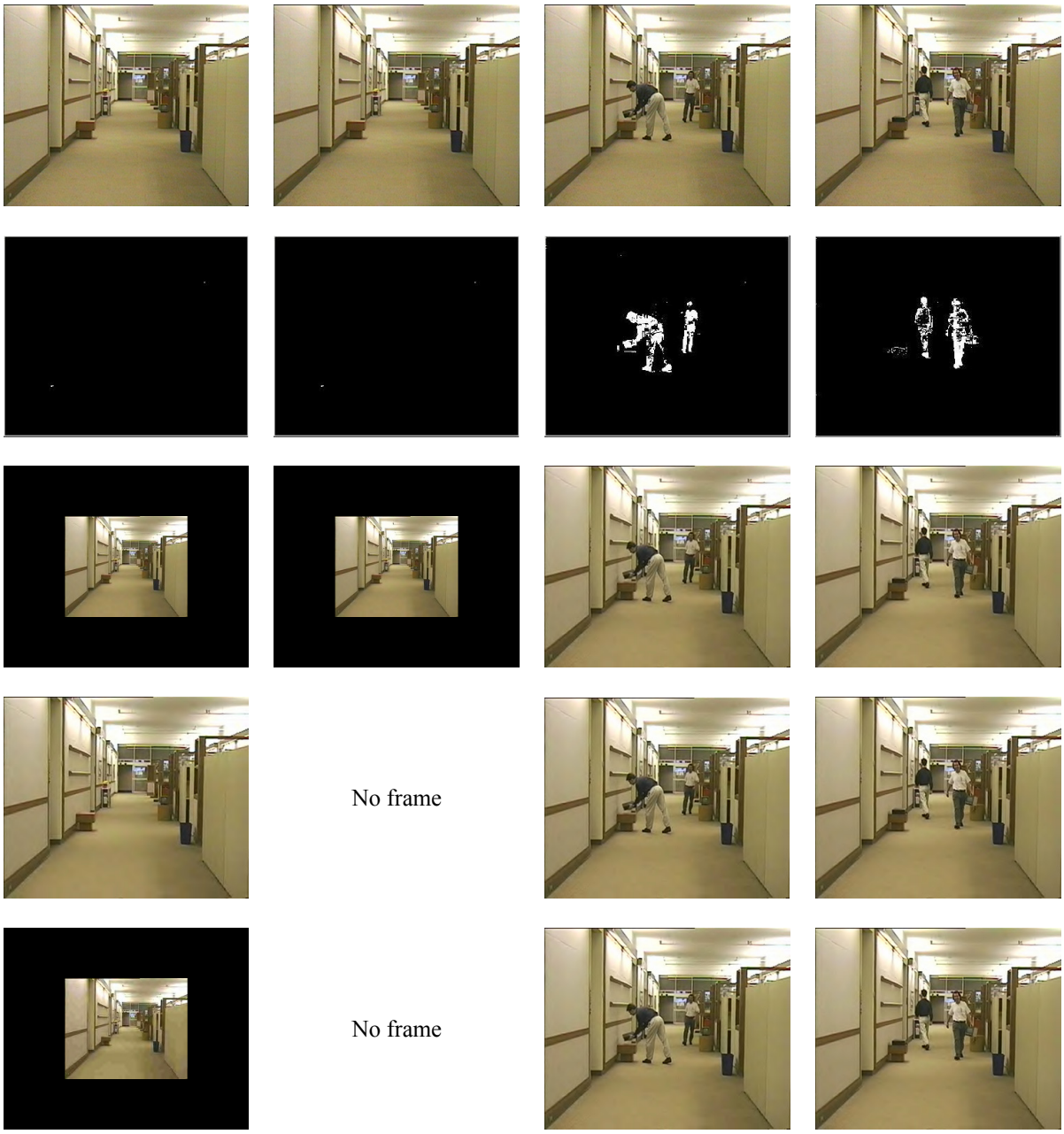

Figure 4: Frames of the reconstructed "Hall" sequence obtained. First row: the original frames. Second row: binary mask. Third row: spatial adaptation. Fourth row: temporal adaptation. Fifth row: combined adaptation.

References

[1] A. Vetro, T. Haga, K. Sumi, H. Sun. "Object-based coding for long-term archive of surveillance video", *Technical Report, TR-2003-98, MERL*, (2003).

[2] Y. Yu and D. Doermann. "Model of object-based coding for surveillance video", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, **2,** pp. 693-696, (2005).

[3] A. Hakeem, K. Shafique, M. Shah, "An object-based video coding framework for video sequences obtained from static cameras", *Proc. ACM International Conference on Multimedia*, pp. 608-617, (2005).

[4] C. Stauffer, W. E. L Grimson. "Learning patterns of activity using real-time tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, pp. 747-757, (2000).

[5] M. Mrak, N. Sprljan, T. Zgaljic, N. Ramzan, S. Wan, E. Izquierdo. "Performance evidence of software proposal for Wavelet Video Coding Exploration group", *Technical Report* ISO/IEC JTC1/SC29/WG11/MPEG2006/M13146, (2006).