

REAL-TIME GENERATION OF ANNOTATED VIDEO FOR SURVEILLANCE

Olivier Steiger, Touradj Ebrahimi

Andrea Cavallaro

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute
CH-1015 Lausanne, Switzerland

Multimedia and Vision Laboratory
Queen Mary, University of London
London E1 4NS, United Kingdom

ABSTRACT

We present a system that generates annotated video in real-time. Moving objects are first automatically extracted by means of video analysis and then coded using MPEG-4. Moreover, an MPEG-7 description of object features is generated. This description is used for automated event detection as well as for rendering. Rendering enhances important objects in the scene on the receiver side, thus facilitating the task of a surveillance operator. The performance of the system is demonstrated using four surveillance videos. Experiments show that objects enhancement can be achieved at low additional cost and illustrate how automated event detection is obtained by taking advantage of the physical scene description based on MPEG-7.

1. INTRODUCTION

The problem of remote visual surveillance of unattended environments has received growing attention in recent years. Nowadays, applications include monitoring of indoor and outdoor environments, quality control in industrial applications, and military applications. However, event monitoring by human operators is rather boring, tedious and error-prone. Thus, advanced surveillance systems aim at employing video analysis to automatically select, enhance and interpret visual information [5]. Several approaches make use of artificial intelligence for incident detection [10], activity recognition [7], and personal identification [9]. These methods are usually limited to specific situations due to the use of machine learning. Selective enhancement is another interesting feature which highlights important image regions (e.g., moving objects) by using visual markers [1], or by selective coding (i.e., important regions are coded at a higher quality than the background [4]). However, semantic information extracted by video analysis is not available individually at the receiver's side. In order to make up for this drawback, recent approaches summarize semantics in a content description. This can be used alone or in conjunction with the coded video [3].

In this paper, we discuss a system that generates annotated video in real-time. In the block diagram in Figure 1, video objects (VO) are first automatically extracted by means of video analysis, and then coded with MPEG-4. Moreover, an MPEG-7 description of object features is generated. In surveillance, the description enables video enhancement in order to put important objects in a conspicuous position for the monitoring personnel. Moreover, descriptors are used for automated event detection. Also, the description can be stored in a database for further processing (video indexing). Our solution operates in cluttered environment and enables interoperability with third-party applications by making use of MPEG standards for video coding and description. While in similar work [3] content annotation summarizes *events* detected by means of artificial intelligence, the proposed MPEG-7 description captures high-level object features. Therefore, our solution is not bound to any particular setup.

The remainder of this paper is organized as follows. Video analysis and content description are addressed in Section 2 and in Section 3, respectively. A multi-level, region-object model is used to segment and track multiple and deforming objects. The performance of the system is demonstrated in Section 4. Finally, Section 5 concludes the paper and discusses future work.

2. VIDEO ANALYSIS

In order to extract meaningful objects from video, the analysis is based on interactions between a high-level and a low-level representation of visual information [2]. Following preprocessing, moving objects are segmented from the background by change detection. The pixel-wise difference between each color channel of the input frame and of the background model is thresholded to produce the *object partition* Π_o^n (high-level representation) at frame n . Π_o^n is regularized by eliminating small connected sets of pixels, and by suppressing small holes using morphology. After regularization, Π_o^n is further decomposed into a set of non-

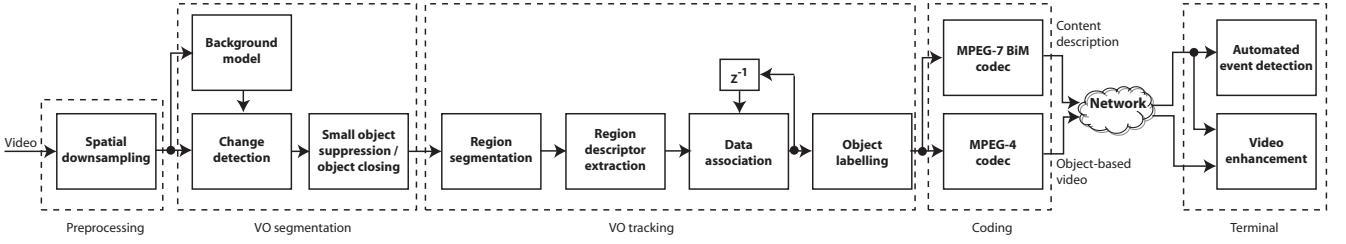


Fig. 1. Block diagram of the system for real-time generation of annotated video.

overlapping, homogeneous regions by region segmentation. Each region in the *region partition* Π_r^n (low-level representation) is represented by a descriptor that summarizes region's features:

$$\Phi_i(n) = \left(\phi_i^1(n), \phi_i^2(n), \dots, \phi_i^{K_i(n)}(n) \right)^T, \quad (1)$$

where $K_i(n)$ is the number of features in frame n . Examples of features are position, motion, color and texture. Video object tracking then operates on region descriptors instead of objects. Tracking region descriptors rather than an entire object is a simple and effective strategy. The simplicity comes from the fact that instead of projecting the entire object into the next frame, only region descriptors need to be processed. Therefore, there is no need for computationally expensive motion models. In addition, region descriptor tracking is effective, since it can cope with deformation, complex motion and occlusion.

The data association stage considers the proximity between region descriptors in Π_r^{n+1} and in Π_r^n , and thus establishes a correspondence between the region partition in the current frame n and the region partition in the new frame $n+1$. The proximity is computed by first measuring the Mahalanobis distance D_f in the feature space between region descriptors in frame $n+1$ and in frame n . For each feature category f , we have

$$D_f(\Phi_i(n+1), \Phi_j(n)) = \sqrt{\sum_{s=1}^K \frac{(\Phi_i(n+1)^s - \Phi_j(n)^s)^2}{\sigma_s^2}}, \quad (2)$$

where σ_s^2 is the variance of the s^{th} feature over the entire feature space. The complete point-to-point similarity measure is then obtained by fusing the distances computed within each category:

$$D(\Phi_i(n+1), \Phi_j(n)) = \frac{1}{F} \sum_{f=1}^F w_f D_f(\Phi_i(n+1)^s, \Phi_j(n)^s), \quad (3)$$

where F is the number of feature categories, and w_f is the weight which accounts for the reliability of each feature category. Reliability is lowered for those features that have similar values in adjacent objects so as to favor stability.

The result of the distance computation can be represented as a distance matrix $\mathbf{D} = \{d_{p,q}\}$, where each row, p , corresponds to a region descriptor in frame $n+1$, and each column, q , corresponds to a region descriptor in frame n . A correspondence between the p^{th} region descriptor in frame $n+1$ and the q^{th} region descriptor in frame n is confirmed if

$$d_{\bar{p},\bar{q}} = \min_q(d_{p,q}) = \min_p(d_{p,q}). \quad (4)$$

If the condition in Equation (4) is respected, the track is updated. Otherwise, region descriptors are iteratively paired based on their distance. At last, the track of objects in Π_o^{n+1} is updated as a consequence of region descriptor tracking.

3. MPEG-7 DESCRIPTION

Video object's location, shape and color are summarized by MPEG-7 *Visual Descriptors* [6]. In Table 1, *Motion trajectory* gives the spatial location of objects (e.g., gravity center). *Region locator* approximates the shape of objects by their bounding box. Accurate shape is given by contour shape. Dominant color at last defines salient object colors. The descriptors are organized so as to provide a layered description of the scene [8]. That is, location, shape and color can be defined in any desired combination and order.

FEATURE	DESCRIPTOR	PURPOSE
Location	Motion trajectory	Spatial location
Shape	Region locator	Bounding box
	Contour shape	Closed contour shape
Color	Dominant color	Salient color

Table 1. MPEG-7 Descriptors for object features.

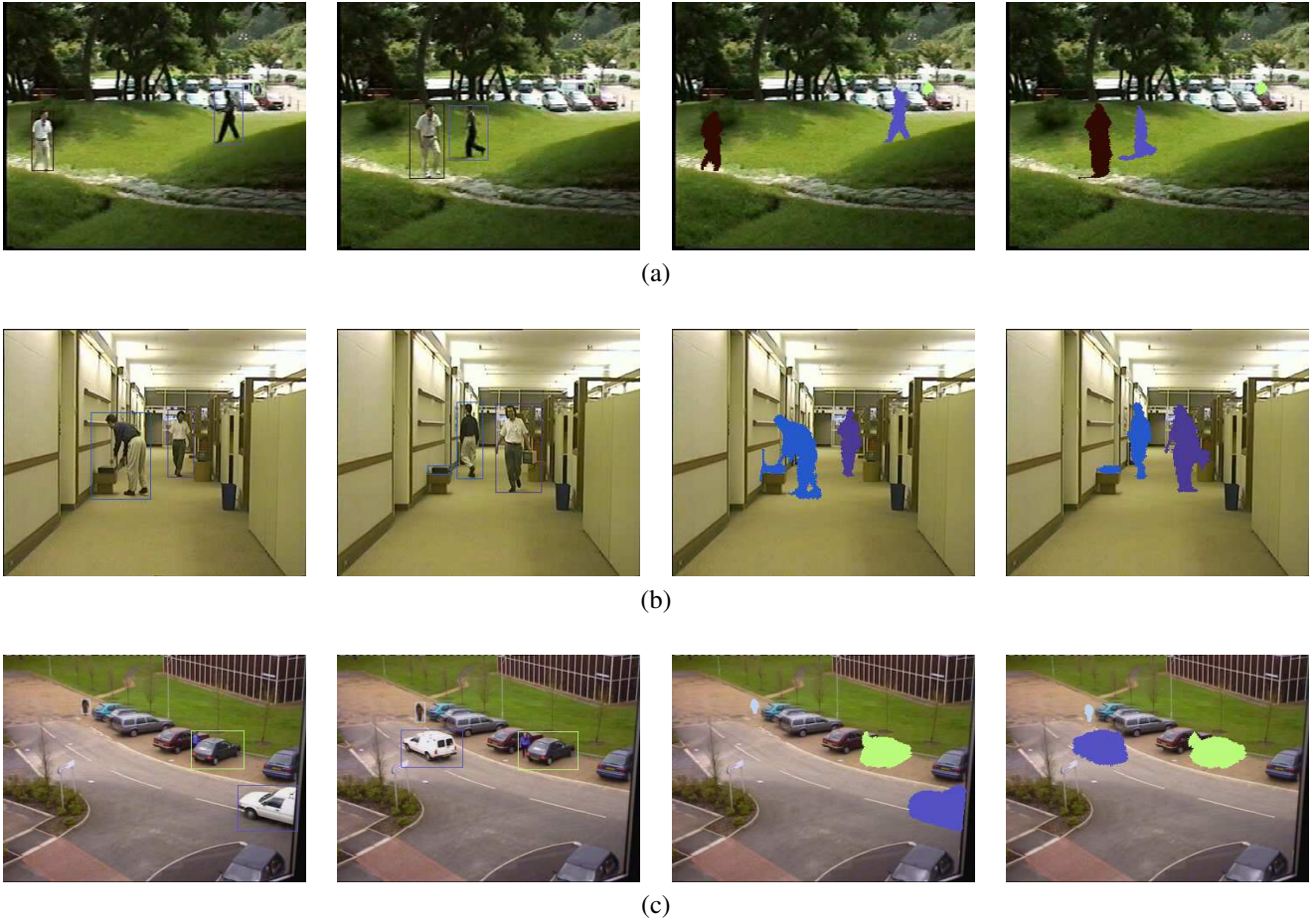


Fig. 2. Enhancement of surveillance video by MPEG-7 descriptors. (Left) Moving objects are highlighted by their bounding box; (Right) contour shape of objects is rendered on a static background. The sample frames stem from the test sequences (a) *Surveillance*; (b) *Hall monitor*; (c) *PETS'2000*.

4. VISUAL SURVEILLANCE APPLICATIONS

The performance of the system in Figure 1 is demonstrated using four surveillance videos: (1) *Surveillance*, from the MPEG-7 Video Content Set; (2) *Hall monitor*, from the MPEG-4 Video Content Set; (3) *PETS'2000*; (4) *PETS'2001*, from the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. In our experiments, the sequences were processed in real-time (25 frames/s) on a 2.8 GHz Pentium 4 PC. The MoMuSys MPEG-4 VM reference software version 1.0 video encoder is used. Binary MPEG-7 is generated by the Expway MPEG-7 *BiM* Payload encoder/decoder version 02/11/07. Note that the same set of parameters was used to generate all the results.

4.1. Video enhancement

Video enhancement is illustrated in Figure 2. In the left part, moving objects are enhanced by their bounding box. Boxes

DESCRIPTION	Location	Box	Shape
<i>Surveillance</i>	19	52	87
<i>Hall monitor</i>	16	45	98
<i>PETS'2000</i>	18	49	85

Table 2. Average bitrate (Kbit/s) required to transmit MPEG-7 *BiM* description at different levels of details.

are defined by MPEG-7 *Region locator* and rendered by the terminal. Video enhancement helps lowering fatigue of the monitoring personnel that is due to extended concentration, since relevant objects are put in a conspicuous situation. By comparison with the common approach that consists in producing enhanced video at the encoder's side, enhancement by the receiver requires only low additional resources for transmission and provides additional flexibility. For instance, the receiver might switch amongst avail-

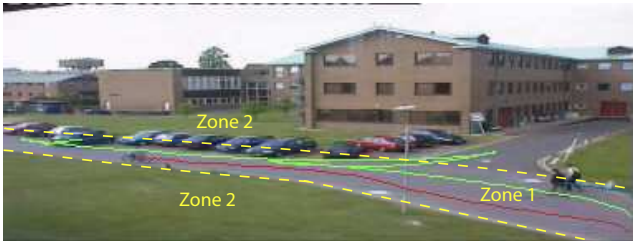


Fig. 3. Automated event detection. The setup in *PETS'2001* has been subdivided in two distinct zones. Each time the trajectory of an object (solid line) enters Zone 2 for more than one second, an intrusion alarm is generated.

able features or enhance individual objects. In the right part of Figure 2, `contour_shape` is rendered on a static background shot. With this representation, the identity of moving objects (e.g., people) remains hidden, thereby enabling privacy preservation. The representation of object's shape suffices to convey the meaning and action of a scene when the objects are familiar.

The average bitrate for binary MPEG-7 description at different levels of details is shown in Table 2. These figures are low as compared to medium-quality MPEG-4 video coding (500 Kbit/s). Despite its low cost, video description permits automated event detection, as discussed next.

4.2. Automated event detection

To perform automated event detection, the setup in *PETS'2001* has been divided in two distinct zones, as shown in Figure 3. Zone 1 corresponds to the authorized area, whereas Zone 2 is restricted. The goal is to automatically generate an intrusion alarm each time an object enters Zone 2. To arrive at this, the location of object's gravity center in successive frames is described by `Motion trajectory`. When an object enters Zone 2 for more than one second, an alarm is generated. To evaluate detection performance, we compare the number of automatic alarms to groundtruth. For the entire sequence (1780 frames; 10 moving objects), the system has generated three correct alarms, zero false alarms, and zero misses. This simple experiment illustrates how automated event detection is achieved by taking advantage of the physical scene description provided by MPEG-7.

5. CONCLUSIONS

We have discussed a system that generates object-based video and content description in real-time. This is achieved by combining multi-level video object extraction with MPEG-4 coding and MPEG-7 description. Our approach exhibits several features that help improve visual surveillance. Enhancement of video by descriptors is used to put objects in

a conspicuous situation. This helps lowering fatigue of the monitoring personnel that is due to extended concentration. The privacy of monitored people is preserved by rendering object descriptors on a static background. The description also enables automated event detection. By monitoring objects' trajectories in outdoor surveillance video, intrusion alarms are generated.

Despite its simplicity, the proposed approach applies to various situations and can be extended in several ways. Video enhancement might be used to highlight small objects, such as a football, in sports broadcasting. Additional functionality for privacy preservation is provided by scrambling image regions corresponding to moving objects. Also, additional features and object recognition might be added for automated event detection.

6. REFERENCES

- [1] M. Bramberger, J. Brunner, B. Rinner, and H. Schwabach. Real-time video analysis on an embedded smart camera for traffic surveillance. In *Proc. IEEE Real-Time and Embedded Technology and Applications Symposium*, pages 174–181, May 2004.
- [2] A. Cavallaro, O. Steiger, and T. Ebrahimi. Multiple video object tracking in complex scenes. In *Proc. ACM Multimedia Conference*, pages 523–532, Dec. 2002.
- [3] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani. Computer vision techniques for PDA accessibility of in-house video surveillance. In *ACM SIGMM Intl. Workshop on Video Surveillance*, pages 87–97, Nov. 2003.
- [4] A. Del Bue, D. Comaniciu, V. Ramesh, and C. Regazzoni. Smart cameras with real-time video object generation. In *Proc. Intl. Conf. on Image Processing*, volume 3, pages 429–432, June 2002.
- [5] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man and Cybernetics, Part C*, 34(3):334–352, Aug. 2004.
- [6] ISO/IEC. Information technology – multimedia content description interface – part 3 visual. Technical Report ISO/IEC 15938-3/FDIS, ISO/IEC JTC1/SC29/WG11, 2001.
- [7] J. Owens and A. Hunter. Application of the self-organising map to trajectory classification. In *Proc. Third IEEE Intl. Workshop on Visual Surveillance*, pages 77–83, July 2000.
- [8] O. Steiger, A. Cavallaro, and T. Ebrahimi. MPEG-7 description of generic video objects for scene reconstruction. In *Proc. SPIE Conf. on Visual Communications and Image Processing*, volume 4671, pages 947–958, Jan. 2002.
- [9] L. Wang, T. Tan, W. Hu, and H. Ning. Automatic gait recognition based on statistical shape analysis. *IEEE Trans. on Image Processing*, 12:1120–1131, Sept. 2003.
- [10] Y. Wu, L. Jiao, G. Wu, E. Chang, and Y.-F. Wang. Invariant feature extraction and biased statistical inference for video surveillance. In *Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 284–289, July 2003.