# SEGMENTING MOVING OBJECTS: THE Modest VIDEO OBJECT KERNEL

*Andrea Cavallaro*[*], *Damien Douxchamps*[**], *Touradj Ebrahimi*[*], *Benoit Macq*[**]
[*] Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland
[**] Université Catholique de Louvain, Louvain-la-Neuve, Belgium
Tel: +41 21 693 2708; fax: +41 21 693 7600
e-mail: andrea.cavallaro@epfl.ch

## ABSTRACT

A system separating objects moving within a slow changing background is presented. The originality of the approach resides in two related components. First, the change detection robust to camera noise which does not require any sophisticated parametric tuning as it is based on a probabilistic method. Second, the change is detected between a video frame representing a scene at a given time, and reference that is updated continuously to take into account slow variation in the background. The system is particularly suitable for indoor and outdoor surveillance. Simulation results show that the proposed scheme performs rather well in extracting video objects, with stability and good accuracy, while being of a relatively reduced complexity.

## 1 INTRODUCTION

Advances in micro-processors, software design and networking have made it possible to rely on more sophisticated machines capable of performing more complex operations, based on richer information. As a consequence, image understanding and computer vision is reaching a certain maturity so as to be considered for applications in every day life.

Emerging standards such as MPEG-4 and MPEG-7 have contributed to accelerate this trend. MPEG-4, as an object-based coding algorithm, allows manipulation of audiovisual objects in a compressed video, in a similar way one interacts with physical objects in the real life. This brings enhanced functionalities and applications such as efficient very low bit rate video coding where only the objects of interest are coded. Similarly, it is possible to reconstruct a photo-realistic virtual scene by taking objects from other real scenes and rendering them together in a manner similar to special visual effects in the movie industry. Such applications are obviously possible only when objects can be detected and extracted from natural scenes, either manually, in a semi-automatic way, or even in a fully automatic fashion. MPEG-7 the emerging standard for representation of audiovisual information based on a content-based approach allows for simple to sophisticated description of such content. This enables applications such as search and filtering where information with a specific content is (or is not) of interest. Video surveillance is another typical application where content of a scene has to be examined to decide if any abnormal behaviour has occured. Abnormal can vary from simple motion of certain object, to more sophisticated patterns in their behaviour.

Segmentation is one of the fundamental problems in image processing. Although human beings and most animals perform this task in a relatively straightforward manner, years of research and developments in machine vision have not yet succeeded to match the same performance. The problem of segmentation is difficult not only because of the complexity of mechanisms involved in it, but also because it is ill posed and in this sense, no unique solution exists to segment a scene. In most situations, a priori knowledge on the nature of the problem (or its solution) is needed, often as a function of the specific application in which the segmentation tool is to be used. A segmentation process leads to a partition of an image or a video sequence into regions according to a given criterion. Many of the above-mentioned applications aim at locating moving objects in the observed scene, thus a change detector can naturally drive the segmentation in a more efficient way. Change detection analysis provides a classification of the pixels in the video sequence into one out of two classes: foreground (moving objects) and background.

To this end, we combine a change detector with a background updating technique. On one hand, the change detector is designed to precisely detect object contours and to be robust to camera noise. On the other hand, the adaptive background scheme accounts for slow environmental light changes. The combination of the two (called the Video Object Kernel) allows to automatically detect multiple moving objects in long video sequences recorded by a monocular static camera. The foreground objects identified by the Video Object Kernel are then tracked along time. A successive step tranforms the 2D tracked shapes in 3D shapes. The description of the 3D shapes is finally given to the content understanding module which derives decisions on the observed scene. The complete system is depicted in Fig. 1.

The paper is organized as follows. Section 1 describes the Video Object Kernel. Section 2 presents the results of the extraction of foreground objects and their use in the advanced video surveillance system. Finally, in Sec. 5, we draw the conclusions.

## 2 THE VIDEO OBJECT KERNEL

The task of the Video Object Kernel is the identification of the areas in the video sequence corresponding to moving objects. Motion cannot be directly measured in video sequences. A related measure is the luminance intensity function and its variations in time. For this reason, a simple change detection technique consists in subtracting two images. A threshold operation is then applied on the difference image. The threshold is fixed empirically, and all pixels presenting a value larger than the threshold are considered as belonging to a moving object. The threshold has to be tuned manually according to the scene characteristics [12]. This approach is therefore not suitable for automatic applications. Various methods have been proposed in the literature to automatically extract objects [3, 7, 8, 10]. A review of these method can be found in [4].

The relationship between motion and temporal changes is not unique. Temporal changes in two successive images occur not only in the area corresponding to moving objects, but also in two additional areas referred to as uncovered background and overlap of the same object [9]. The uncovered background area does not belong to a moving object, but it is detected as temporally changed. The overlap of two successive instances of the same object is hard to be detected as changed when the object is not sufficiently textured. These problems are less critical when the temporal changes are computed between the current image and a reference frame that represents the scene background [5, 6]. For this reason, we have chosen to detect changes in the current image with respect to a reference background. In addition, in order to avoid the drawbacks of a fixed reference frame, we use an adaptive background reference frame.

### 2.1 Adaptive background

A reliable reference frame is fondamental for the identification of moving areas through change detection. A frame captured when no objects are present in the scene is used when short sequences are analysed. However, such a frame is not always available. In addition, a fixed background image is not suitable for long sequences. In this case, changes in the enviromental illumination lead to misdetections.

For these reasons, we use an adaptive background scheme. The scheme allows to begin the detection of moving object from any time instant in the sequence, even if foreground objects are present. In this case a short set-up time is necessary to create the reference image.

In the Video Object Kernel, the computation of the adaptive background frame is an iterative process that refreshes, at an instant n+1, the background obtained from n previous frames of the sequence with the incoming n+1 frame. The background updating method uses a blending formula that weights pixels of the incoming frame, according to their chances to belong to the background. This is achieved by computing an error map. The error map takes into account both changes with respect to the previously computed background and with respect to the previous frame. A detailed description of this adaptive process is given in [11]. The block diagram describing the background updating module is depicted in Fig. 2.

This adaptive refreshment of the background brings two main advantages. First, it allows the change detection algorithm to rely on an effective reference frame even if a frame without foreground objects is not available. Second, it increases significantly the robustness to slow changes in the environmental and illumination conditions (e.g. clouds occluding the sun light or sunsets). For long sequences, indeed, when considering the first frame as reference, changes in daylight are detected as structural changes.

From a computational point of view, it is important to note that all the frames of the video sequence do not need to be used in this process. The refreshment rate is independent from the video frame rate and it can be set according to the application and the available hardware.

### 2.2 Change detector

The reference background frame computed and updated as described in the previous section is given as input to the change detector. The second input is the current frame of the sequence under analysis. The goal is the detection of moving objects. Since moving objects generate changes in the image intensity, motion detection is related to temporal change detection. However, besides the perturbation in the temporal changes introduced by a moving object, camera noise also heavily influences the results of the segmentation. In fact, a large number of pixels that do not correspond to a change in the real world appear as changed in the sequence due to the noise introduced by the acquisition process. To discount the effect of noise, simple change detection techniques perform a threshold operation on the difference image. The threshold is fixed empirically. All pixels presenting a difference larger than the threshold are considered as belonging to a moving object. This approach performs well only on sequences where moving objects are highly contrasted. However, thresholds have to be tuned manually according to the sequence properties. In addition, thresholds need an update along the sequence itself. These major drawbacks limit this approach for a fully automatic application.

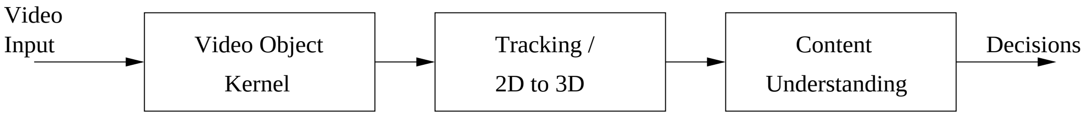


Figure 1: Block diagram of the MODEST video surveillance system. The foreground objects extracted by the Video Object Kernel are first tracked and then transformed in 3D shapes. Finally, a content understanding module derives decisions from the 3D description.

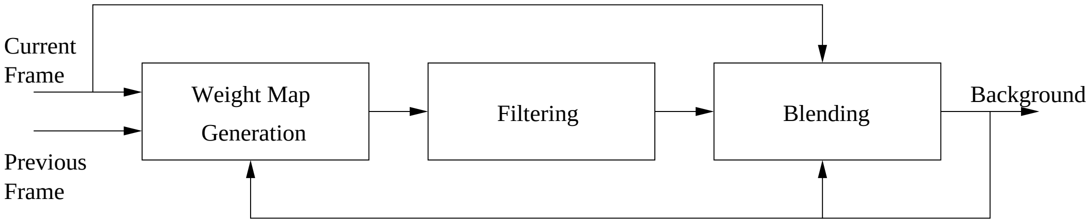


Figure 2: Block diagram of the iterative background adaptation. A background frame is generated by integrating information from the previous frame of the sequence and the already computed background.

To overcome these problems and to obtain a more flexible procedure, we adopt a method that models the noise statistics. The method is based on a statistical decision rule. According to this model proposed by Aach [1], it is possible to assess what is the probability that the value at a given position in the image difference is due to noise instead of other causes. This procedure is based on the hypothesis that the additive noise affecting each image of the sequence respects a Gaussian distribution. It is also assumed that there is no correlation between the noise affecting successive frames of the sequence. These hypotheses are sufficiently realistic and extensively used in literature. The results of the change detector is a classification of the image pixels into two groups: changed, and not changed. The classification is performed according to a significance test, after windowing the difference image. The dimension of the window can be chosen according to the application. The method and the parameter selection strategy are described in details in [4]. It is worth to notice that the only parameter whose value needs to be defined is a significance level. This is a stable parameter that is not dependent on the sequence, but on the error rate that it is tolerate for the application. This method does not require therefore any manual threshold tuning and does not severely increase the computational load compared to simple threshold techniques. An implementation of the method on a Pentium II, 300MHz processor, performs close to real time (6 frames per second, CIF format).

Since the change detector does not necessarely provide close contours, a hole filling procedure is added at the end of the scheme (Fig. 3). The results of the Video Object Kernel are then passed to the tracking module. The 2D object shapes are then translated into 3D, and their description is finally used by the content understanding module. These modules are described in the following section.

## 3 THE Modest SURVEILLANCE SYSTEM

The segmentation and background adaptation techniques implemented in the Video Object Kernel are integrated within the MODEST surveillance platform. Their cooperation provided satisfactory results at a reasonable computational cost. More details about the MODEST system can be found in [2]. The architecture of this system is described on Fig. 4. The sensors used are digital cameras overlooking the surveilled scene. Although several cameras are used, their field of view do not overlap and the traffic is thus analysed at a number of sparse areas. This is typical to commonly installed video surveillance systems and allows the MODEST platform to be installed without excessive hardware investments. In order to further cut investments and enhance performance, the MODEST Video Object Kernel isq designed to be placed close to the camera, leaving the scene descriptors as sole output on an IP network. This contrasts with current system that often require optical fibre to convey multiple video streams.

Besides the lower data rate at the segmention output, a first higher-level semantic information is available: the idea of object, defined at this level as an area of connected and segmented image pixels.

Once masks of objects are generated, they are used by a 3D reconstructor. This reconstractor computes position, sizes, orientation and speed of the objects as metric values. The images of objects have thus been further reduced to a small number of values of a higher semantic level, gaining substantial signification for the content understanding platform and for the final user. The geometric representation of objects is then packed

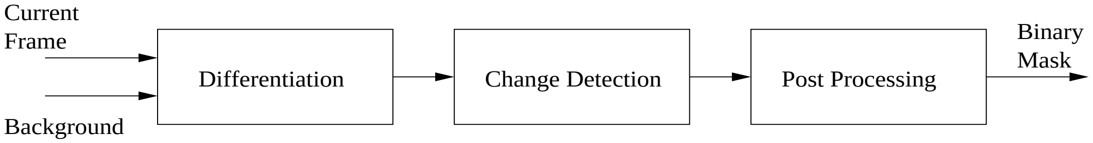


Figure 3: Block diagram of the change detector. The changes detected in the difference between a current frame and the background frame are then postprocessed to obtain masks of objects without holes.

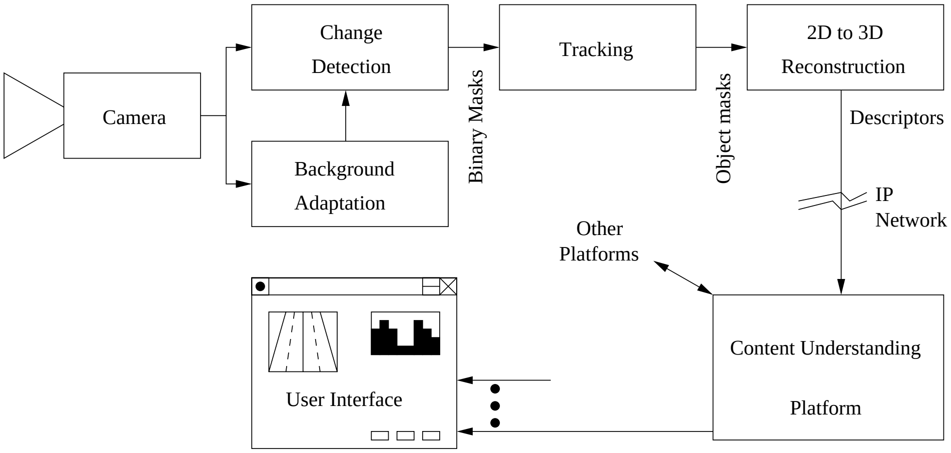


Figure 4: The architecture of the Modest surveillance system.

with other descriptors (such as color) and sent to the content understanding platform. This platform is the only part of the MODEST system to be dependent on the application. The platform is composed by a set of application-specific agents. By processing the 3D information, the agents derive statistics, analyse the behavior of the objects, and track them through the different camera sites. The data is then displayed to the user via an appropriate interface. An example is given in the next section.

## 4 RESULTS

The results of the Video Object Kernel are presented in this section. Since the module is addressed to both indoor and outdoor surveillance, sequences with very different characteristics have been considered. The sequences selected to present the results are the following. *Hall Monitor*, a typical example of indoor surveillance scene, from the MPEG-4 data set. *Group*, an indoor sequences characterized by many interactions and occlusions between the objects. The sequence belongs to the test set of the European IST project *art.live*. Finally, *Highway*, a typical traffic surveillance sequence from the MPEG-7 data set, is considered. The sequences contain both small and large foreground objects. The spatial resolution of the test sequences is $288 \times 352$ pixels (CIF format) and the temporal resolution is 30 images per second for *Hall Monitor* and 25 images per second for *Group* and *Highway*.

The same set of parameters has been used for all the sequences. The background refresh rate has been selected as half the original sequence frame rate.

Figure 5 presents the input and the output of the Video Object Kernel for the three sequences considered in this section. The results show a correct extraction of the foreground for both small and large objects. In addition the contours of the extracted objects are correctly defined and they are stable over time.

It is important to stress that all the sequences have been processed without changing the parameters of the Video Object Kernel. The obtained results demonstrate that the performance of the proposed method does not vary if the scene content changes. However, the results shown in this section differ from the ones of an ideal object extractor for two aspects. The first aspect is the low-pass filter effect introduced by the windowing in the change detector. The extracted contours are slightly larger then the real ones. This error is acceptable for surveillance applications. It could be corrected by a postprocessing module, if another application requires contours exactly fitting the objects. The second deviation from an ideal extraction is the presence of shadows in the change detection mask. Shadows are in fact detected as moving objects since they possess the same characteristics. The 2D to 3D conversion module in the MODEST system takes care of this problem and provides a correct description of the 3D shapes. An example of 3D shapes and object tracking is given in Fig. 6.

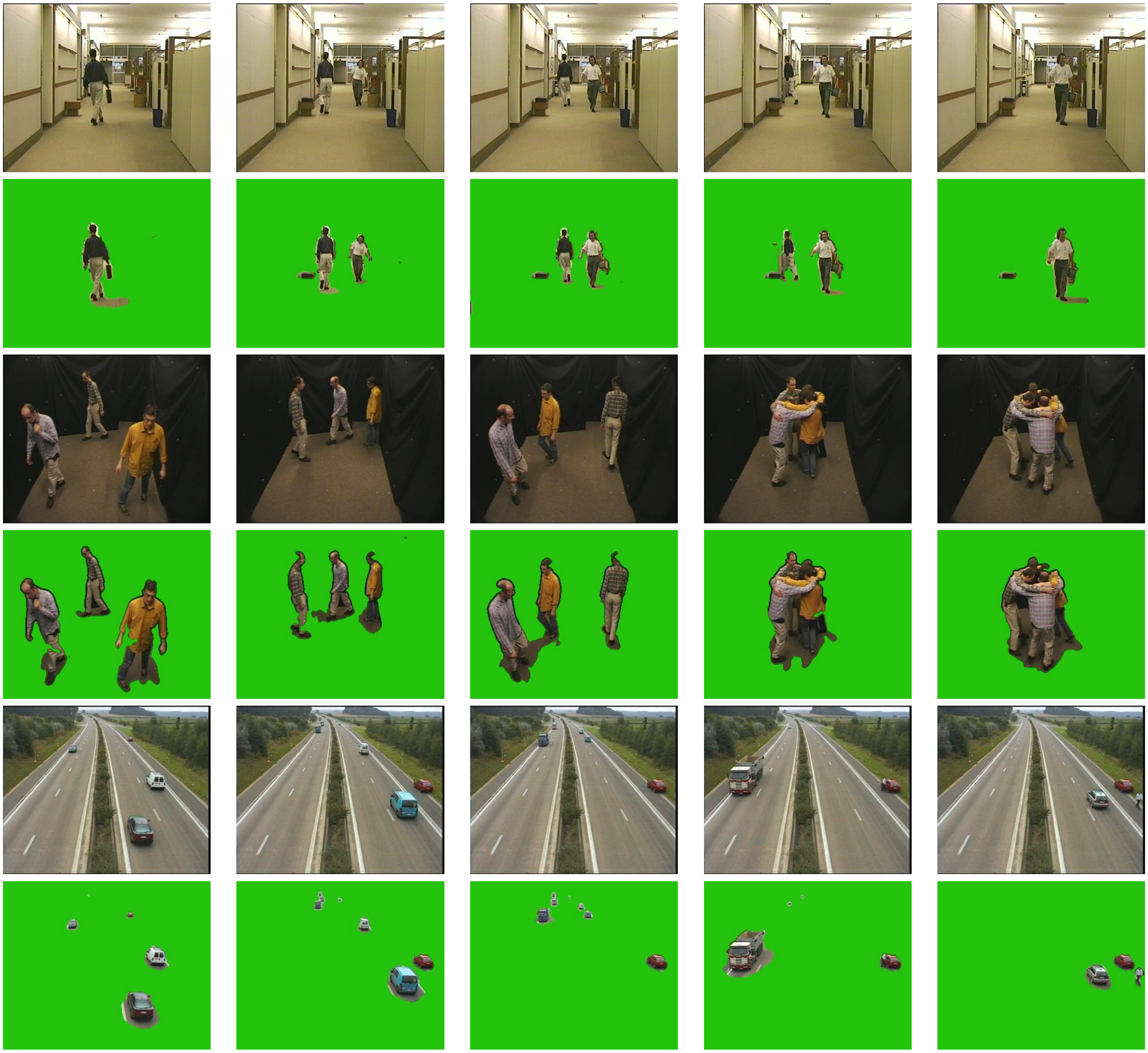
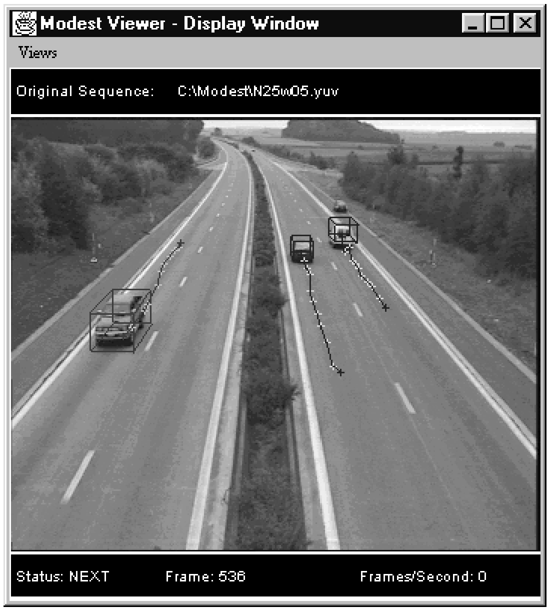


Figure 5: Original frames (first row:*Hall Monitor*, third row:*Group*, fifth row:*Highway*), and corresponding results (second, fourth, and sixth row) of the Video Object Kernel. The extraction of the video objects corresponding to the above original frame is visualized by superposing the resulting change detection mask over the original sequence. The complete sequences are available at http://ltswww.epfl.ch/∼andrea/vok.html .



Original Sequence: C:\Modest\N25w05.yuv

Status: NEXT Frame: 536 Frames/Second: 0

Figure 6: Example of final result of the MODEST system. The moving objects are tracked and their 3D shapes are described.

## 5 CONCLUSIONS

We presented here a novel scheme for extraction of video objects in a scene generated from a single camera without any specific calibration. The scheme makes use of a statistical change detection where the only parameter to set is related to a probability of detection of a moving objects (structural change) versus that of the noise of the camera. Making use of a varying reference image, which approximates scenes background, further enhances this change detection. The update of the reference allows for a better adaptation of the scheme to slowly changing backgrounds, such as outdoor scenes where changes in illumination occur.

Experimental results on long sequences show that the proposed method provides stable results in terms of detection of moving objects and accuracy of their contours. This method is a component technology to be placed in a larger framework (e.g. object-based indexing and retrieval), and has been successfully applied within an advanced video surveillance system (European project ACTS304 Modest). In this application, the information extracted from the moving objects (such as motion, shape, colour) are provided to a content understanding module which is responsible for interpreting the monitored scene.

## ACKNOWLEDGMENTS

The authors whish to thank the MODEST team (http://www.tele.ucl.ac.be/MODEST/) at Laboratoires d'Electronique Philips (LEP), Instituto Superior de Ciencias do Trabalho e da Empresa (ADETTI), EPFL, and UCL. A special thanks to Benoit Mory for his work on background adaptation.

## References

[1] T.Aach, A.Kaup, and R.Mester. "Statistical model-based change detection in moving video" *Signal Processing*, 31:165–180, 1993.

[2] B. Abreu, L. Botelho, A. Cavallaro, et al., "Video-Based Multi-Agent Traffic Surveillance System", Proc. of IEEE Intelligent Vehicles Symposium (IV2000), Detroit (USA), pp. 457-462, 3-5 October 2000.

[3] D. Aubert, "Passengers Queue Measurement", In *Proc. of 10th International Conference on Image Analysis and Processing*, Venice (Italy), pp. 1132–1135, 27-29 September 1999.

[4] A. Cavallaro and T. Ebrahimi, "Video Object Extraction based on Adaptive Background and Statistical Change Detection", Proc. of SPIE Electronic Imaging 2001 - Visual Communications and Image Processing, San Jose' (California, USA), pp. 465-475, 21-26 January 2001

[5] G.W. Donohoe, D.R. Hush, and N.Ahmed. "Change detection for target detection and classification in video sequences" In *IEEE Proceedings of ICASSP*, pp. 1084–1087, New-York, 1988.

[6] K.P. Karmann, A.Brandt, and R.Gerl. "Moving object segmentation based on adaptive reference images" In *Proc. 5th European Signal Processing Conf.*, pp. 951–954, Barcelona, 1992.

[7] A.Makarov. "Comparison of background extraction based intrusion detection algorithms" In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 521–524, 1996.

[8] X. Marichal "On-line Web Application using Image Segmentation' ', In *Proc. of WIAMIS99*, Berlin, pp. 141–144, 1999.

[9] A.Mitiche and P.Bouthemy. "Computation and analysis of image motion: A synopsis of current problems and methods" *International Journal of Computer Vision*, 19(1):29–55, 1996.

[10] T. Nakanishi and K. Ishii. "Automatic vehicle image extraction based on spatio-temporal image analysis" In *Proc. of 11th International Conference on Pattern Recognition (ICPR)*, pp. 500–504, 1992.

[11] P. Piscaglia, A. Cavallaro, M. Bonnet and D. Douxchamps,"High Level Descriptors of Video Surveillance Sequences", In *Proc. of 4th European Conference on Multimedia Applications, Services and Techniques (ECMAST'99)*, Madrid (Spain), pp. 316-331, 26-28 May 1999.

[12] P. L. Rosin, "Thresholding for change detection", In *Proc. of International Conference of Computer Vision (ICCV-98)*, pp. 274–279, 1998.