Video object extraction based on adaptive background and statistical change detection

Andrea Cavallaro and Touradj Ebrahimi

Signal Processing Laboratory (LTS)
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

ABSTRACT

This paper introduces a system for video object extraction useful for general applications where foreground objects move within a slow changing background. Surveillance of indoor and outdoor sequences is a typical example. The originality of the approach resides in two related components. First, the statistical change detection used in the system does not require any sophisticated parametric tuning as it is based on a probabilistic method. Second, the change is detected between a current instance of the scene and a reference that is updated continuously to take into account slow variation of the background. Simulation results show that the proposed scheme performs well in extracting video objects, with stability and good accuracy, while being of relative reduced complexity.

Keywords: segmentation, video object, change detection, reference background

1. INTRODUCTION

Advances in microelectronics, software engineering and telecommunications have made it possible to rely on more sophisticated machines capable of performing more complex operations. As a consequence, image understanding and computer vision is becoming more of a reality in the every day life.

Machines that see and understand their environment already exist in research laboratories, and even start to make their appearance in many professional and consumer applications.

Emerging standards such as MPEG-4 and MPEG-7 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 effects in Hollywood movies. 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 audio visual information in a content-based manner allows for simple to sophisticated description of the content of audiovisual information. This enables applications such as search and filtering where information with a specific content is of interest. Video surveillance is another typical application where content of a scene has to be examined to decide if any abnormal behaviour is 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. Change detection analysis provides a classification of the pixels in the video sequence into one out of two classes: foreground (moving objects) and background.

Correspondence: Andrea Cavallaro, E-mail: andrea.cavallaro@epfl.ch

To this end, a novel automatic method for the extraction of moving objects in video sequences recorded by a monocular static camera is proposed in this paper. The procedure is fully automatic, it discriminates moving foreground from static background by integrating the information provided by a motion detector robust to camera noise, relying on a continuously updated reference frame. This takes into account slow environmental light changes. The use of a reference image representing the background reduces the uncovered background effects and improves the exact detection of low textured moving objects.

The complete system is depicted in Fig. 1 and described in detail in the following sections. Section 2 presents the state of the art in the framework of extraction of moving objects by means of background updating and motion detection techniques. The proposed method is presented in Sec. 3. Experimental results are shown in Sec. 4, and finally, in Sec. 5, we draw the conclusions.

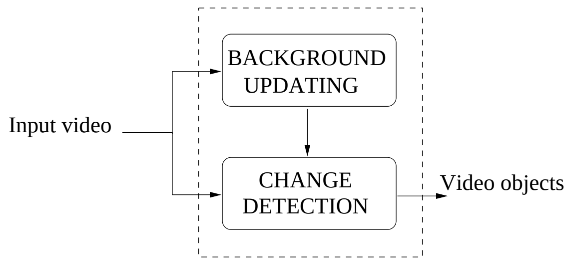

Figure 1. Block diagram of the proposed video object extraction system.

2. STATE OF THE ART

As stated in the previous section, many computer vision applications require 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.[1] This approach is therefore not suitable for automatic applications. Various methods have been proposed in literature to automatically extract objects.[2–5] They will be discussed in this section.

An important choice for a change detection technique is the selection of a reference image. This image is compared to a new frame in order to detect the structural changes. The reference can be a previous frame in the sequence or an image representing the background of the scene. The first frame of the sequence, when no foreground objects appear in the scene, is often chosen as reference background.

On one hand, the use of such a fixed reference image has the advantage of being a simple solution, and it allows the detection of objects in the scene even if they become immobile. In addition, a reference image representing the background reduces the uncovered background effects and improves the detection accuracy of low textured moving objects.

On the other hand, this solution is not always practical. An initial frame of a sequence with no foreground objects is, in fact, not available in every application. Moreover, if such a reference frame is fixed, it is not suitable for long outdoor sequences. In such a case, the illumination conditions gradually change, and the method cannot adapt to the lighting variations that naturally happen.

For this reason, techniques aiming at reconstructing the reference frame by integrating the background information from following images of the sequences have been proposed. The updated background is then compared to the current frame to detect the moving objects.

Nakanishi[2] considers the case of an outdoor scenario for automatically extracting vehicles. The technique extracts the information of occlusion, velocity and length and updates periodically the background. A major drawback is that this method is not general and requires hypothesis on the vehicles motion and camera's view. It is not applicable then to other kind of scenes.

Aubert[3] presented a method based on level lines image representation which is used to automatic measure passengers queue length. The advantage of this method is that a global illumination change does not affect the geometry of the level lines, and will just create or remove some level lines. A drawback is represented by its computational complexity. To overcome this problem a simplification of the computations is proposed which severely decreases the system performance. The contours of the extracted objects are therefore not accurate.

An intrusion detection algorithm based on moving edges is presented by Makarov.[4] The method is characterized by a low computational cost, but has the limitation of dealing with only one moving object in the scene because of the procedure used to close the contours. Scene with multiple moving objects cannot therefore be analyzed.

A simple segmentation algorithm for on-line internet applications is presented by Marichal.[5] The method is based on a combination of change detection masks to detect the moving objects. Moreover, a comparison with a reference updated background is performed to cope with slow changes in the illumination conditions. The process for updating the background compares the updated background with the new input frames and it is based on a mobile median filter. The algorithm has a very low computational complexity, but this constraint limits the precision of the detection. The contour of the extracted objects is, in fact, not accurate due to the use of the previous and next frame in the change detection process.

In order to overcome the limitations of the previous methods, we propose to combine a change detection robust to camera noise with a background updating technique in order to get precise contours and detect multiple moving objects in long sequences.

3. PROPOSED METHOD

Since moving objects generate changes in the image intensity, motion detection is related to temporal change detection. However, 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.[6] 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.[7,8]

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 (as described in the previous section) of a fixed reference frame, we use an adaptive background reference frame.

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 (or not) to the background (i.e. inlier/outlier discrimination). This is achieved by computing an error map that takes into account both changes with respect to the previously computed background and with respect to the previous frame.[9] The block diagram describing the background updating module is depicted in Fig. 2.

This adaptive refreshment of the background has two main advantages: it allows the change detection algorithm to rely on an effective reference frame even if a frame without foreground objects is not available, and it increases significantly the robustness to slow changes in the environmental and illumination conditions (e.g. clouds occluding the sun light or sunsets).

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 being independent from the video frame rate.

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, early change detection techniques perform a threshold operation on the difference image. A threshold is fixed empirically, and all pixels presenting a difference larger than the threshold are considered

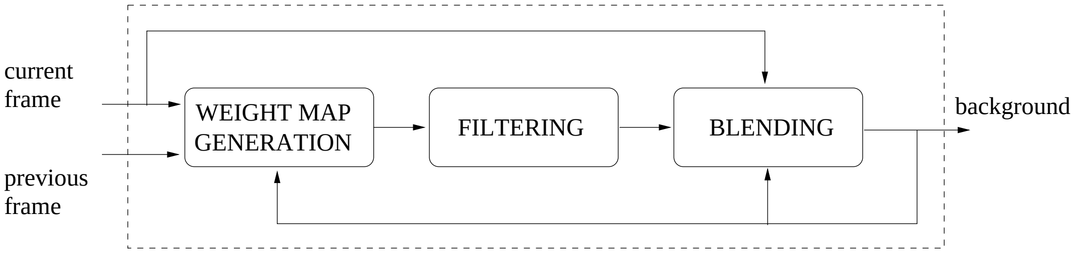

Figure 2. Block diagram of the iterative background adaptation.

as belonging to a moving object. This approach performs sufficiently well on sequences where moving objects are highly contrasted. However, thresholds have to be tuned manually according to the sequence properties and often need an update along the sequence itself. These major drawbacks limit this approach for a fully automatic application.

If it is not possible to tune the threshold, in fact, a large number of pixels that do not correspond to a structural change in the real world appear as changed in the sequence. The difference between two images does not correctly define the areas corresponding to moving objects. In Fig. 3 an example of what we obtain by computing the difference between a current frame and a reference is presented. The resulting values are different from zero in those areas interested by motion, but also where motion is not present.

Figure 3. (a) Current frame; (b) reference frame; (c) The absolute difference in luminance intensities are different from zeros (white pixels) also in some areas where no real motion is present.

To overcome these problems and to obtain a much more flexible procedure, a method that models the noise statistics based on a statistical decision rule is adopted. According to a model proposed by Aach,[10] 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 with mean μ and standard deviation σ_c, $N(\mu_c, \sigma_c)$. It is also assumed that there is no correlation between the noise affecting successive frames of the sequence. These hypothesis are sufficiently realistic and extensively used in literature. They allow us to define the intensity value of a generic pixel (i, j) of an image of the sequence as:

$$\hat{I}_{(i,j)} = I_{(i,j)} + \mathbf{x} \tag{1}$$

Where $I_{(i,j)}$ is the real intensity value of an image at pixel position (i, j), and $\hat{I}_{(i,j)}$ is the corresponding observed value. The noise affecting the image is represented by $\mathbf{x}$ that is a random variable with distribution $N(\mu_c, \sigma_c)$. The value σ_c can be obtained by computing the statistics of the difference $d_{(i,j)} = \hat{I}^r_{(i,j)} - \hat{I}^c_{(i,j)}$ between a reference image, I^r, and the current image, I^c. Given the hypothesis H_0, that $I^r_{(i,j)} = I^c_{(i,j)}$, then the conditional probability density function $f(d_{(i,j)}|H_0)$ is defined by:

$$f(d_{(i,j)}|H_0) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{d^2_{(i,j)}}{2\sigma^2}\} \tag{2}$$

which still represents a Gaussian distribution with zero mean and variance $\sigma^2 = 2\sigma_c^2$.[11] Note that the hypothesis H_0 corresponds to the hypothesis that $d_{(i,j)}$ is different from zero because of the camera noise and not because of other reasons like moving objects (structural change). To get a more robust detection, it is possible to extend this result over a group of pixels belonging to a certain window. From the results of Eq.(2), we can model the conditional density $f_{\mathbf{\Delta^2}}(\Delta^2_{(i,j)}|H_0)$ where $\Delta^2_{(i,j)} = \sum_{(k,l)\in W^n_{(i,j)}} d^2_{(k,l)}$ and where H_0 is valid for all the pixels in $W^n_{(i,j)}$ *. Here $W^n_{(i,j)}$ is a square window centered in (i,j) and containing n pixels. The corresponding random variable $\mathbf{\Delta^2}$ is the summation of the square values of n random variables whose conditional density respects Eq.(2). Thus, according to Wilson,[12] $\mathbf{\Delta^2}$ follows a χ^2 distribution defined by:

$$f_{\mathbf{\Delta^2}}(\Delta^2_{(i,j)}|H_0) = \frac{1}{2^{n/2}\sigma^n\Gamma(n/2)}(\Delta^2_{(i,j)})^{(n-2)/2}e^{-\frac{\Delta^2_{(i,j)}}{2\sigma^2}} \tag{3}$$

Where $\Gamma(n/2)$ is the Gamma function as it is defined in.[13] It is now possible to derive the conditional probability $P(\mathbf{\Delta^2} \geq \Delta^2_{(i,j)}|H_0)$:

$$P(\mathbf{\Delta^2} \geq \Delta^2_{(i,j)}|H_0) = \frac{\Gamma(n/2, \frac{\Delta^2_{(i,j)}}{2\sigma^2})}{\Gamma(n/2)} \tag{4}$$

Equation(4) is a function of 3 variables $n, \sigma, \Delta^2_{(i,j)}$ that can be efficiently computed. The variable n represents the number of pixels we use to compute the statistics. By increasing n, the statistics computed on $W^n_{(i,j)}$ will be more reliable. This results in a reduced sensitivity to noise. However, the probability that the hypothesis H_0 remains valid on all the pixels belonging to $W^n_{(i,j)}$ decreases. This leads to a wrong labeling along the edge of moving objects and the corresponding blocking effects. A good compromise is to choose the value n=25. This value allows us to obtain semantic video objects with fine boundaries and without holes. A larger value introduces blocking artifacts and a smaller value does not allow for a good detection of the objets as shown in Fig. 4. The variable σ characterizes the statistical behavior of a zero mean Gaussian noise. Since it is related to the σ_c of the camera noise, it can be computed directly from the difference image on those areas where the hypothesis H_0 is valid. Finally $\Delta^2_{(i,j)}$ represents a realization of the random variable $\mathbf{\Delta^2}$. It is directly computed on the difference image over the window $W^n_{(i,j)}$.

Equation(4) represents the probability that $\Delta^2_{(i,j)}$ respects the hypothesis H_0. When this probability is smaller then a certain significance level, α, we consider that H_0 is not satisfied at the pixel position (i,j). Thus we label that pixel as changed. If the probability is higher than the significance level, the pixel is labeled as unchanged.

It is worth to notice that the only threshold whose value needs to be defined is the significance level α. This is an extremely stable parameter that does not need manual tuning along a sequence or for different sequences. Experimental results show that valid values might range from 10^{-2} to 10^{-6}.

This method does not require therefore any manual threshold tuning and does not severely increase the computational load compared to simple threshold techniques. The results of its application to long sequences are presented in the following section.

*Other test statistics can be defined. For example, it is possible to determine the $f_{\mathbf{\Delta}}(\Delta_{(i,j)}|H_0)$ where Δ is the sum over a certain window of the absolute values of the difference image. If Δ^2 obeys to a Gaussian distribution, Δ follows a Laplacian distribution.

Figure 4. Statistical change detection mask obtained with different values of n

4. RESULTS

Several simulations were run to evaluate the performance of the proposed algorithm. Typical video sequences from the MPEG-4 and MPEG-7 test set, as well as test sequences from the European IST project *art.live*†, have been used. Both indoor and outdoor sequences, and small and large foreground objects have been considered. The spatial resolution of the test sequences is 288×352 pixels (CIF format) and the temporal resolution is 30 images per second.

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. The window $W^n_{(i,j)}$ for the statistical change detection is 5×5 ($n = 25$).

The visualisation of the results is organized as follows. Two columns are displayed: on the left hand side, sample frames of an original sequence are shown. On the righ hand side, the extraction of the video objects corresponding to the original frame in the same row is visualized by superposing the resulting change detection mask over the original sequence.

In Fig. 5 the indoor sequence *Hall Monitor* is considered. It is possible to notice that the contours of the extracted objects are correctly defined and they are stable over time. An example of video objects extraction at the beginning of an outdoor sequence (*Highway*, from the MPEG-7 test set) is provided in Fig. 6. The first frame of the sequence containing foreground objects is used to initialize the background frame. The reference frame is then progressively updated by eliminating the moving objects, thus allowing a correct extraction of the video objects. The areas corresponding to the cars in the first frame disappear progressively over time. Figure 7 shows the results of the extraction once the background frame is stable. The sequence *Group* belonging to the test set of the European IST project *art.live* is finally considered. The results show a correct extraction of the foreground also for this indoor scene with large objects close to the camera (Fig. 8).

All the sequences have been processed without changing the parameters of the extraction algorithm. The obtained results demonstrate that the performance of the proposed method does not vary if the scene content changes (indoor/outdoor scenario, small/large objects). 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 robust change detector. This is due to the use of a 5×5 window as compromise between robustness to noise and accuracy of detection. For this reason the extracted contours are slightly larger then the real ones. This error can be easily corrected by a postprocessing module if the application requires contours exactly fitting the objects. The second

†The goal of *art.live* (Architecture and authoring tools for prototype for Living Images and new Video Experiments) is to develop an architecture and a set of tools for the enhancement of narrative spaces. Two aspects of this goal will be tested: telepresence (inclusion of real objects into artificial worlds) and augmented reality (inclusion of artificial objects into real worlds).

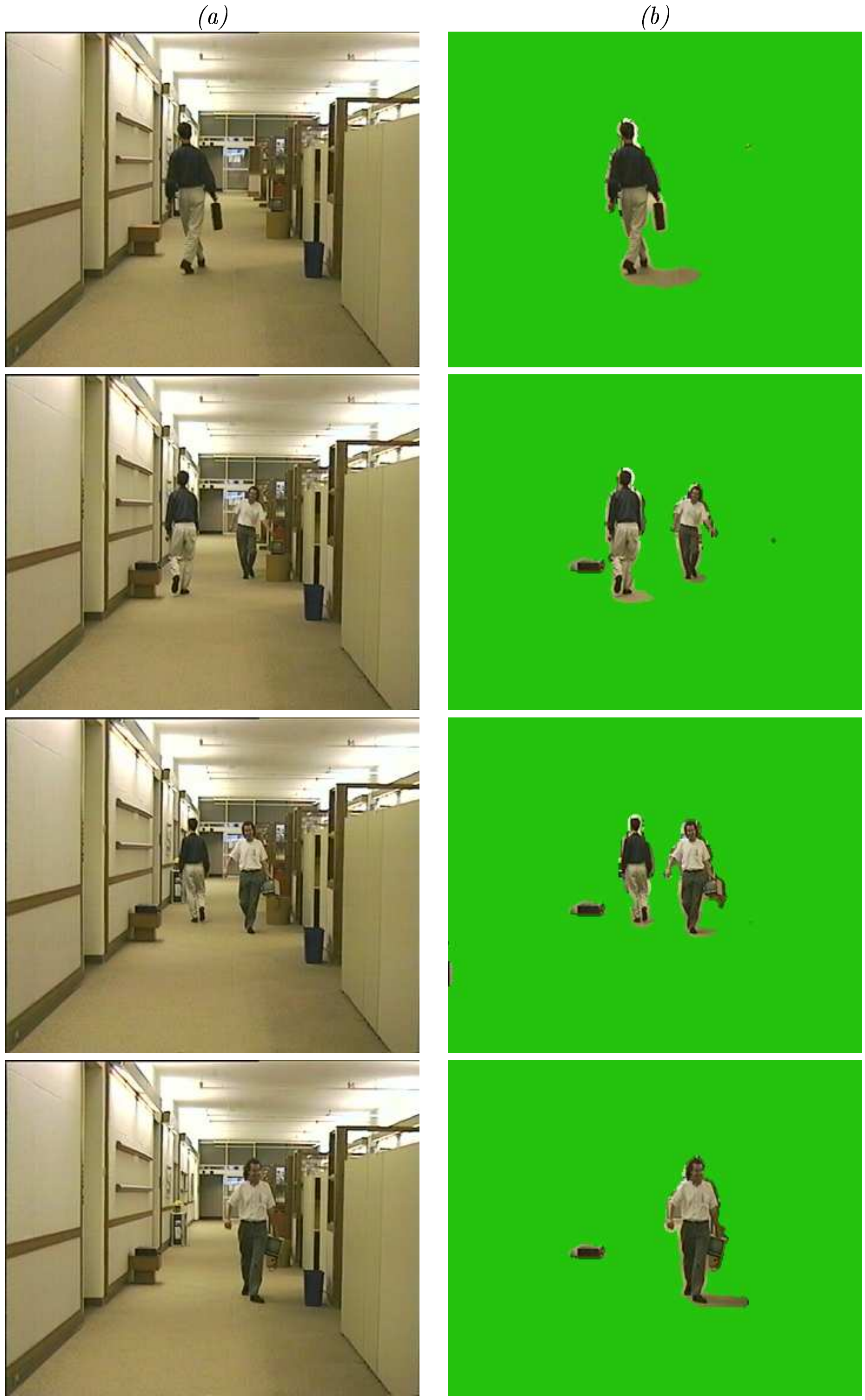

Figure 5. Automatic video objects extraction for the sequence *Hall Monitor*: (a) original frames, (b) video objects.

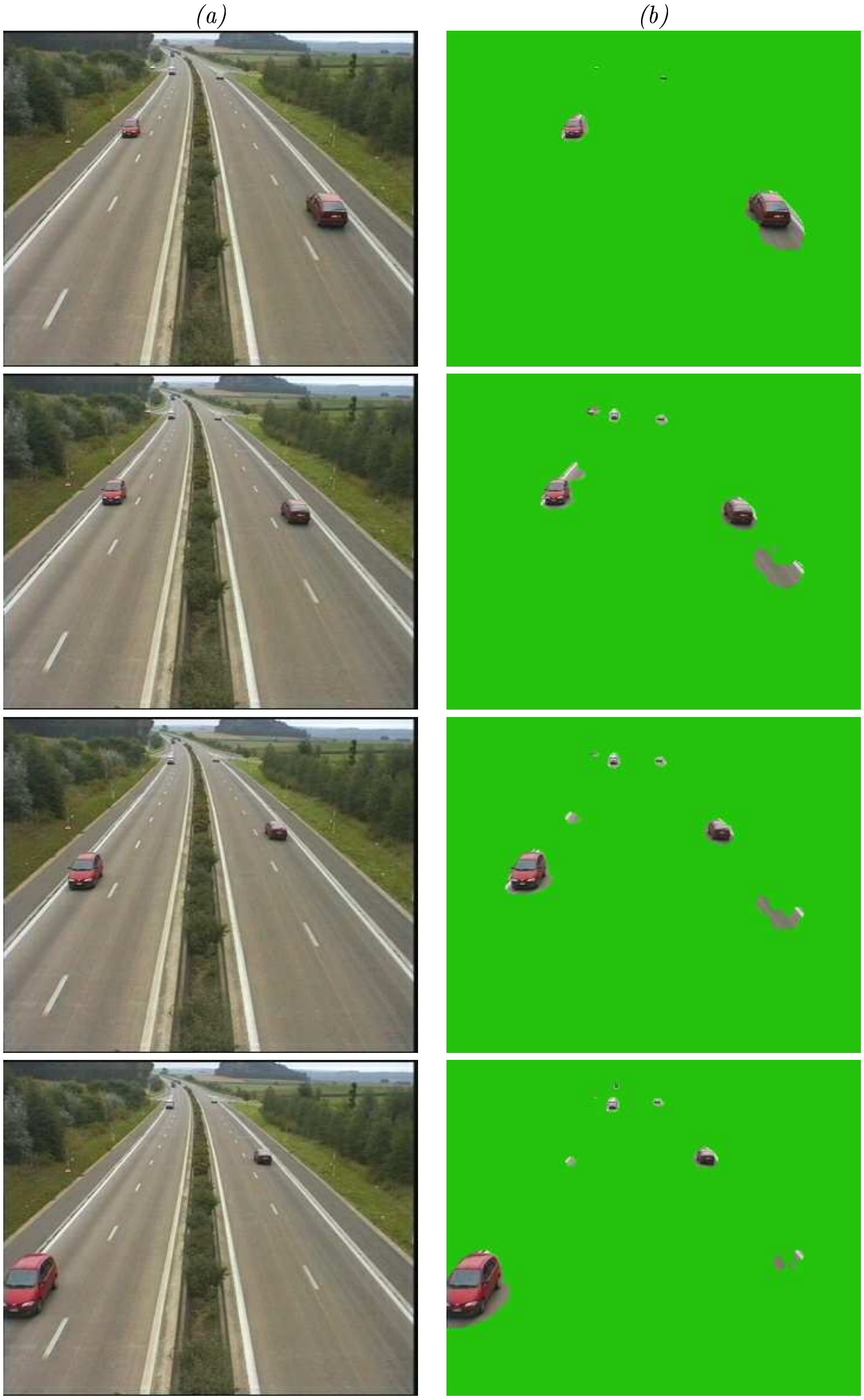

Figure 6. Example of automatic video objects extraction at the beginning of a sequence. (a) original frames, (b) video objects. The background reference image is being generated and spurious areas are detected. The first background image is the first frame of the sequence that contains foreground objects. The information provided by the following frames is then used to update the background allowing a correct extraction of the moving objects. It is possible to notice that the size of the spurious areas corresponding to the cars is getting smaller and smaller. It will disappear in the following frames as shown in Fig. 7.

(a) (b)

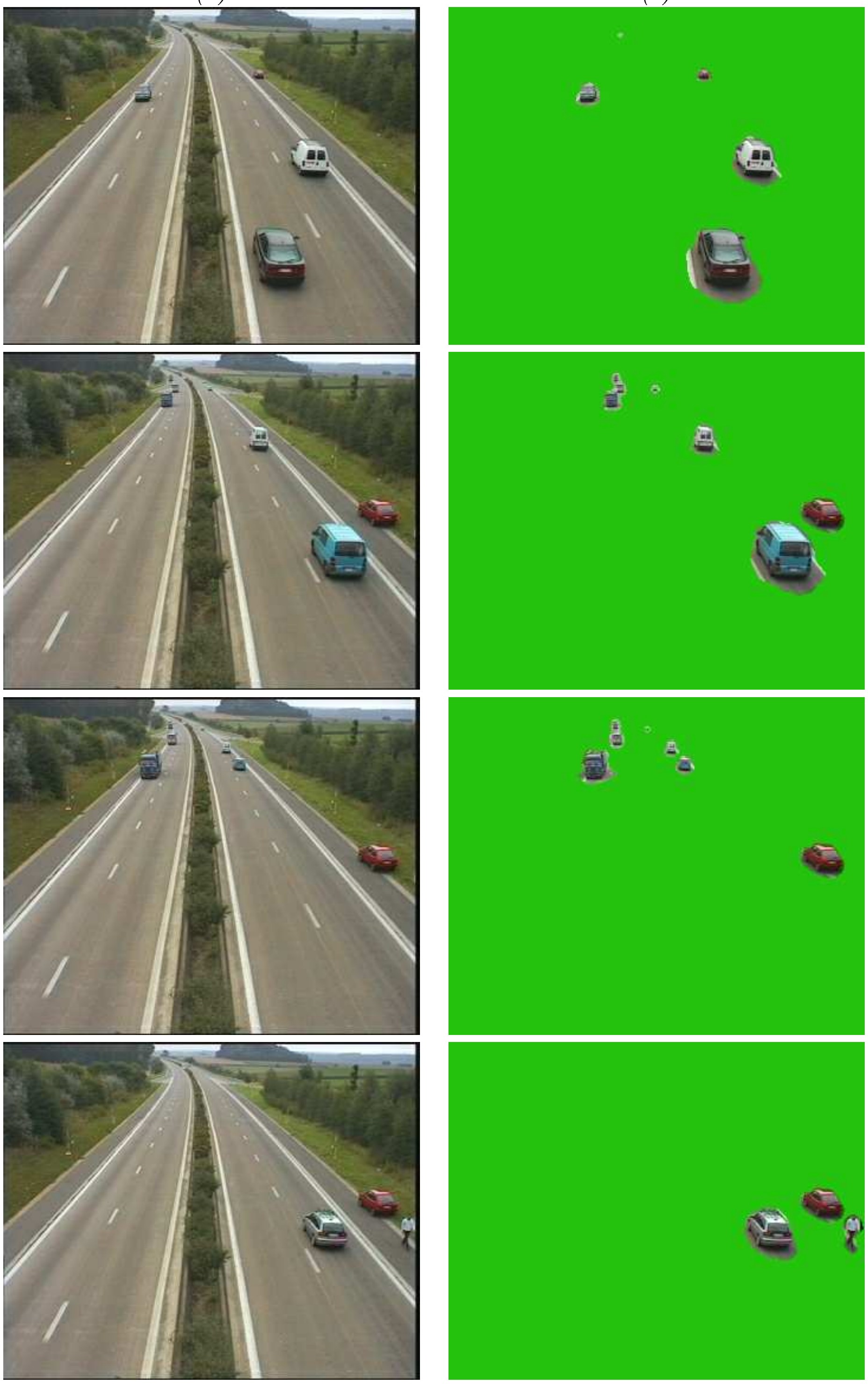

Figure 7. Automatic video objects extraction for the sequence *Highway*: (a) original frames, (b) video objects.

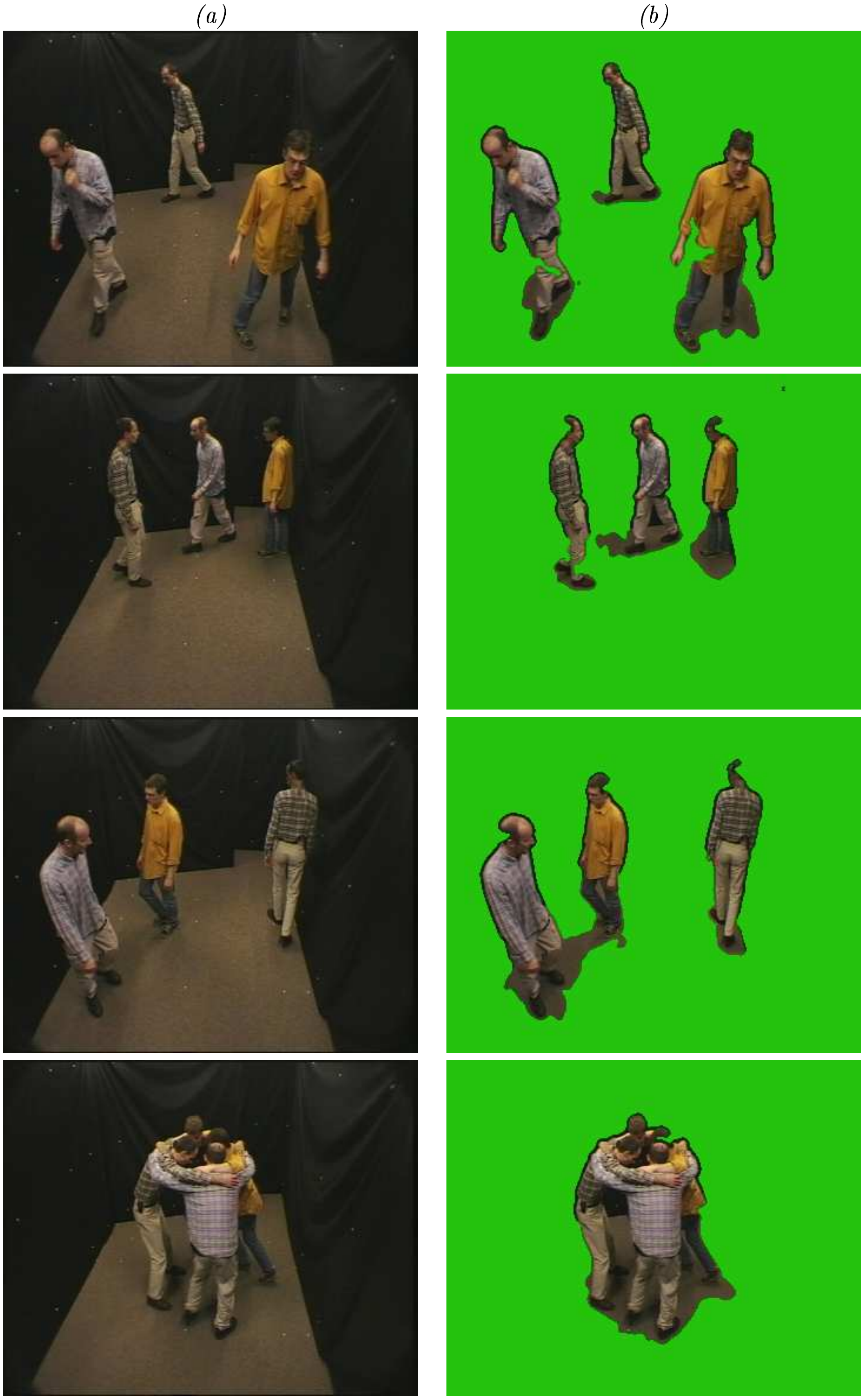

Figure 8. Automatic video objects extraction for the sequence *Group*: (a) original frames, (b) video objects.

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. If it is necessary to eliminate them, an ad hoc strategy can be employed to discount from the resulting mask the shadows cast by the foreground objects.[14]

5. CONCLUSION AND FUTURE WORK

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. Our future research direction is to deal with moving camera sequences by integrating the global motion information.

ACKNOWLEDGMENTS

The authors thank the members of the European ACTS project Modest for the fruitful dicussions that contributed to this work, and especially Benoit Mory from Laboratoires d'Electronique Philips (LEP) for providing the background updating software.

REFERENCES

1. P. L. Rosin, "Thresholding for change detection", In *Proc. of International Conference of Computer Vision (ICCV-98)*, pp. 274–279, 1998.
2. 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.
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.Makarov. "Comparison of background extraction based intrusion detection algorithms" In *Proc. of IEEE International Conference on Image Processing (ICIP)*, pages 521–524, 1996.
5. X. Marichal "On-line Web Application using Image Segmentation' ', In *Proc. of WIAMIS99*, Berlin, pp. 141–144, 1999.
6. 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.
7. 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.
8. 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.
9. 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.
10. T.Aach, A.Kaup, and R.Mester. "Statistical model-based change detection in moving video" *Signal Processing*, 31:165–180, 1993.
11. A. Papoulis. *Probability, Random Variables, and Stochastic Processes* McGraw-Hill, 1991.
12. S. G. Wilson. *Digital Modulation and Coding* Prentice Hall, 1996.
13. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flaunery. *Numerical Recipes in C* Cambridge, 1993.
14. J. Stauder, R. Mech, J. Ostermann, "Detection of Moving Cast Shadows for Object Segmentation", *IEEE Trans. on Multimedia*, Volume 1, No.1, March 1999, pp. 65-76.