OBJECT TRACKING BASED ON TEMPORAL AND SPATIAL INFORMATION

Fabrice Moscheni[†], *Frédéric Dufaux*[‡][1] *and Murat Kunt*[†]

[†]Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

[‡]Digital Equipment Corporation, Cambridge Research Lab
Cambridge, MA 02139, USA

ABSTRACT

This paper presents an object tracking algorithm. The objects are characterized through their temporal and spatial features so as to identify them and carry out the tracking procedure. The proposed tracking algorithm helps to detect the objects present in the current frame by supplying previous spatio-temporal information to the spatio-temporal segmentation procedure. In addition, the proposed algorithm tackles the correspondence problem. This is achieved through the use of a multiple hypotheses framework, the latter tests being based on both temporal and spatial characterizations of the objects.

1. INTRODUCTION

This paper addresses the problem of segmenting an image sequence in terms of multiple moving objects and tracking them through time. Such a spatio-temporal segmentation assures a thorough understanding of dynamic scenes and enables to derive visually meaningful results. This semantic approach is furthermore very well suited to content-based video coding and to the new functionalities which are promoted in the framework of the MPEG-4 activities.

The spatio-temporal segmentation is generally carried out between two consecutive frames and as such suffers from many drawbacks. In particular, estimation inaccuracies, noise, as well as the lack of decisive information may alter the interpretation of the scene. Furthermore, as the coherence of the spatio-temporal segmentation is not guaranteed through time, no temporal evolution of the segmentation and identification of the different objects present in the scene can be robustly derived.

In order to tackle these problems, a tracking procedure has to be applied [1, 2]. It allows using the information obtained over past images. The spatio-temporal segmentation can thus be performed more robustly. By its mere definition, the tracking may also be a key element to dynamic scene analysis. The correspondence problem [3] can be solved by defining the spatio-temporal trajectories [4].

Among the many types of features which can undergo tracking, the success of the tracking algorithm depends heavily on the intrinsic meaning of those features. In other words, the tracked features should carry a natural meaning in order to allow for a tracking procedure to be successful. Due to their intrinsic meaning, the moving objects present in the scene are very well suited to this task. In such a framework, techniques have been proposed to solve the correspondence problem by 3D segmentation [5]. The latter, however, only rely on spatial information. On the other hand, techniques such as [6] only use the motion information, which is merely used to initialize the spatio-temporal procedure.

In this paper, a tracking algorithm for image sequences is proposed. The algorithm is unsupervised and does not require any a priori information. The tracked features are chosen to be the moving objects present in the scene. These objects are provided by the spatio-temporal segmentation technique described in [7] and [8]. The proposed algorithm interacts with the segmentation algorithm so as to carry out in a first stage the moving objects detection process and, in a second stage, to address the temporal linkage of the objects. This is achieved by exploiting both the objects spatial and temporal information so as to fully characterize each of them. Based on the latter identification, the correspondence problem is solved in the statistical framework of multiple hypotheses testing. The proposed tracking algorithm is thus able to stabilize the spatio-temporal segmentation procedure and to define trajectories for the moving objects present in the scene.

The paper is structured as follows. The spatio-temporal method used in conjunction with the proposed tracking algorithm is described in Sec. 2. Sec. 3 presents the tracking algorithm, while experimental results are shown in Sec. 4. Finally, Sec. 5 draws the conclusions and describes the future work.

2. SPATIO-TEMPORAL SEGMENTATION

As the tracking is performed on the moving objects present in the scene, the proposed tracking algorithm has to be used in conjunction with a technique yielding a spatio-temporal segmentation. The latter describes the scene in terms of coherently moving entities which can be seen as objects. In this paper, the spatio-temporal segmentation is performed by the algorithm presented in [7], with the exception of the regions merging which uses the technique presented in [8]. In a first step, the algorithm removes the

[1] Part of this work was done when the author was with the MIT Media Laboratory

camera motion through global motion estimation and compensation. Then, starting from a static segmentation, the algorithm iteratively merges or splits regions on the basis of the motion information. The latter is expressed in the form of an affine motion model whose parameters are robustly estimated by a matching technique. Regions with similar motions are merged based on the information given by a robust nonparametric test applied on the residues. In parallel, regions which are not well compensated are split. The above spatio-temporal segmentation algorithm is characterized by its efficient use of both motion and spatial information.

As the spatio-temporal segmentation is applied between two consecutive frames, the coherence of the segmentation is however not granted through time. Inaccuracies in the motion estimation, noise in the luminance as well as the lack of decisive information may alter the segmentation process. Objects present in past segmentations may suddenly disappear or change their shape drastically. Moreover, the successive segmentations are not linked temporally. The temporal evolution of the moving objects is thus not available.

In order to stabilize the segmentation procedure, all the available information should be used. Not only the information present between the two considered frames, but also the information extracted from the previous frames should be exploited. This procedure is referred to as tracking and is closely related to the data association and correspondence problems [2]. Moreover, the tracking procedure should take into account the inaccuracies which are inherently present in the data.

Unlike the segmentation procedure, the tracking procedure also permits to address the correspondence problem. A temporal linkage between successive segmentations can be obtained. Each object undergoes a pursuit procedure which results in an increased semantic understanding of the scene.

3. OBJECT TRACKING

The proposed tracking algorithm takes as role model the Human Visual System (HVS). The latter is indeed able to define spatio-temporal coherent entities and track them robustly. In order to achieve this, the HVS works in terms of objects and uses both the spatial and temporal information. For these reasons, the proposed tracking algorithm relies on the objects present in the scene. It identifies each of them through both their spatial and temporal characteristics so as to exploit all the available information. No a priori knowledge or any user's input are required. The proposed approach is based on two distinct successive steps. In a first stage, the tracking algorithm helps the spatio-temporal segmentation at defining the moving objects present in the scene. In a second stage, the tracking algorithm puts the obtained segmentation in correspondence with past ones.

3.1. First Stage of the Tracking

The likelihood that an object found in the previous frames will appear in the current one depends on the intrinsic meaning of the object. If the object corresponds to a well defined spatio-temporal entity, it is very likely to show up, while badly defined objects are bound to vanish. The proposed tracking algorithm takes this phenomenon into account and,as such, helps the segmentation process in robustly detecting the moving objects.

First, the tracking algorithm provides the predicted location of each object in the current frame. This task is accomplished by projecting each object of the previous frame onto the current frame. So as to precisely define the shape of the projected objects, the spatial information of the current frame has to be exploited. The projection is thus performed onto a over-segmented label image of the current frame, each label being assigned to the projected object which most covers it. The motion information used for the projection procedure is estimated by means of a Kalman filter [2]. It acts as a temporal filter which takes into account the inherent accuracies of each motion measurement. Moreover, the estimate relies not only on the last motion measurement, but also on all the motion measurements extracted from the previous frames. It therefore uses all the available motion information.

Such a projection allows to use the segmentations obtained previously in order to initialize the current spatio-temporal segmentation procedure. However, it would be a waste of time to allow the latter to examine the objects which are already well defined and which have a natural correspondent in the former segmentation. The tracking algorithm detects them in two steps. First, it checks whether the Mean Square Error (MSE) obtained after motion compensation is lower than a preset threshold. The objects which satisfy this requirement are classified as valid. For each valid object, the hypothesis of whether it corresponds both spatially and temporally to an object in the previous frame is tested. In case the answer is positive, the object is kept unsplit in the spatio-temporal segmentation procedure. However, merging is still permitted.

By means of the projection procedure as well as the detection of objects already well defined, the tracking algorithm is able to input previous spatial and temporal information and interacts with the segmentation, making the moving objects detection more robust and stable.

3.2. Second Stage of the Tracking

In a second step, the proposed tracking algorithm permits to tackle the correspondence problem and hereby to perform a pursuit of each detected object [4]. More precisely, it provides a temporal linkage between successive spatio-temporal segmentations. This defines a trajectory for each object and thus allows a thorough semantic understanding of the scene. Using the statistical framework of multiple hypotheses testing [9], the proposed tracking algorithm addresses the correspondence problem through the use of both the spatial and temporal information of the object. The latter object identification permits to check different correspondence hypotheses. More precisely, each hypothesis relevance is obtained through a test on the temporal information and a test on the spatial information. This decoupling derives from the fact that the motion of an object and its spatial characteristics are clearly not correlated. Sec. 3.2.1 and Sec. 3.2.2 present respectively the temporal test statistic and the spatial test statistic. The different hypotheses tested in the framework of the correspondence problem are developed in Sec. 3.2.3.

3.2.1. Test Statistic for Temporal Information

The determination of the temporal information requires a motion model. In our case, a fully parametric model is used. More precisely, the objects are assumed to undergo temporal changes which can be represented by an affine transformation. With regard to the proposed tracking algorithm, a temporal test statistic is needed to decide whether an object **A** of the current frame can be seen as having the same motion as an object **B** from the previous frame, when the latter motion is extrapolated to the current frame by Kalman filtering. Due to its robustness, the modified Kolmogoroff-Smirnov test presented in [8] is chosen. It is a nonparametric test statistic which exploits all the available motion information.

3.2.2. Test Statistic for Spatial Information

The spatial information carries information about the shape and the texture of the object. In the framework of the tracking algorithm, the chosen spatial features have to be invariant under the transformation induced by motion. In our case, affine invariants are thus required. In [10], a moment-based approach to 2D and 3D object recognition is presented. In particular, affine moment invariants are derived which are perfectly suited to the task of identifying spatially an object undergoing an affine transformation [11]. In our experiment, objects are described by five affine moment invariants. For each object, these invariants are combined in the vector $\vec{I}$. Considering two objects **A** and **B** with their respective spatial characterizations $\vec{I}_A$ and $\vec{I}_B$, the hypothesis whether the two objects are spatially similar is given the significance level α defined by:

$$\alpha \;=\; \min_i \left(P(|q(i)| \geq \left|\vec{I}_A(i) - \vec{I}_B(i)\right|), i = 1, \ldots, 5 \right),$$

where $q(i)$ is the maximum likelihood estimator of the hypothesis that $\vec{I}_A(i)$ and $\vec{I}_B(i)$ are equal. Assuming $\vec{I}(i)$ to be Gaussian with variance $\sigma(i)$, it can be shown that $q(i)$ is a random variable such as $q(i) \sim N(0, \sqrt{2}\sigma(i))$. Practically, the variance $\sigma(i)$ is estimated on the population composed by the parameters $\vec{I}(i)$ of all the detected objects.

The necessary condition for accepting the hypothesis of spatial coherence is that the significance level α is higher than a preset threshold. The other requirement is that the total surface disparity between the objects A and B is lower than a preset threshold. This additional condition is necessary due to the affine invariance of our object spatial characterization.

3.2.3. Multiple Hypotheses Testing

The hypotheses to be tested have to be chosen so as to cover the whole range of possible situations. Six cases are considered, whose significance is computed in the presented order. Once a current object has been put into correspondence with a previous object, it is no longer examined by successive testings. In order to speed up the correspondence problem, a selection in terms of covered area is performed. More precisely, each object in the previous frame is projected in the current one and the portion of its surface covered by each object in the current frame is computed. The reciprocal measure is obtained for each object in the current frame. This procedure is used for the cases 1, 2, 3 and 5 presented below.

Case 1: Spatial and temporal hypotheses are both accepted. The two objects share temporal and spatial characteristics and are therefore put into correspondence.

Case 2: While the temporal hypothesis is accepted, the spatial one is not. Nevertheless, the current object is detected as being covered by the projection of the previous object onto the current frame. This obviously corresponds to an over-segmentation or an occlusion in the current frame. The current object can thus be seen as a part of the previous object and be put into correspondence with it.

Case 3: Similarly to case 2, the temporal hypothesis is accepted, while the spatial one is not. However and unlike case 2, the projection of the previous object onto the current frame is detected as being covered by the current object. At this stage, three explanations can be foreseen. One possibility is that the current spatio-temporal segmentation is too rough and has to be refined based on the previous segmentation. The second explanation is that the previous frame is over-segmented and that the current segmentation has to be trusted. The last explanation is that a disocclusion is taking place. Further testings should be carried out to determine which explanation is more likely. In our experiments, however, the decision is based on the size ratio between the projection of the previous object and the current object. If the ratio is smaller than a preset threshold, the hypothesis of over-segmentation in the previous frame is accepted. In case the ratio is bigger than the threshold, the hypothesis that the current segmentation is too rough is tested. The previous segmentation is projected onto the current one through the projection procedure detailed in Sec 3.1. The newly defined object undergoes the spatial hypothesis testing and, in case of acceptation, is kept as a refinement of the current segmentation. This allows to keep track of objects even though they stop moving and are therefore not detected by the spatio-temporal segmentation.

Case 4: While the hypothesis testing on shape is accepted, the one on motion is not. This case clearly corresponds to an object having performed a maneuver. The brisk change in its motion has fooled the hypothesis testing on motion. However, the hypothesis testing on shape is able to recognize it.

Case 5: Neither the temporal hypothesis nor the spatial hypothesis are accepted. Nevertheless, the projection of the previous object onto the current frame is detected as being covered by the current object. The hypothesis that the current segmentation is too rough has to be checked. This is achieved by projecting the previous object onto the current frame as in case 3. At this stage, the newly defined object serves as the basis for the spatial hypothesis. If the latter is accepted, the newly defined object is kept as a refinement of the current segmentation. This corrects the situation where an object is totally lost by the current spatio-temporal segmentation.

Case 6: Following the above suite of hypotheses testings, it may occur that a previous object has not found any corresponding object in the current frame. This characterizes the disappearance of an object or its total occlusion. Conversely, a current object may not have found any corresponding object in the previous frame. This can be seen as the appearance of a new object or the disocclusion of an object which was formerly present in the scene.

4. EXPERIMENTAL RESULTS

Experimental results are presented in this section. Figure 1 shows a frame of the sequence "Car" and the initial spatio-temporal segmentation. The efficiency of the tracking algorithm is demonstrated by comparing the subsequent spatio-temporal segmentations for the 6^{th} frame obtained respectively without and with the proposed tracking approach. The use of the latter entails a stabilized spatio-temporal segmentation. The car is much better segmented and there is less noise in the background. Moreover, a pursuit of the objects can be carried out. In particular, the car can be tracked from frame to frame.

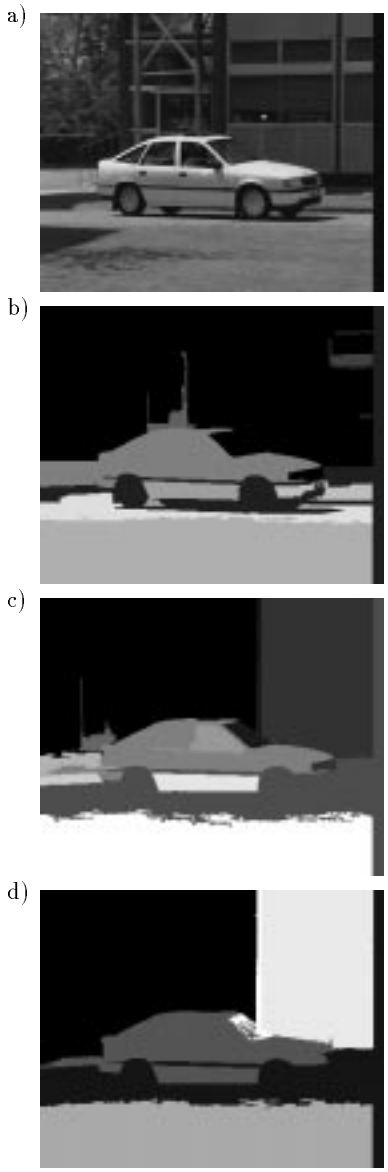

Figure 1: "Car": a) a frame, b) the initial spatio-temporal segmentation, c) subsequent spatio-temporal segmentation without tracking and d) idem with tracking.

5. CONCLUSIONS

This paper addresses the problem of segmenting an image sequence in terms of moving objects. An object tracking algorithm is presented which not only renders the spatio-temporal segmentation procedure more robust, but also tackles the correspondence problem. This is achieved by characterizing the objects both spatially and temporally in the framework of a multiple hypotheses testing. Experimental results show the efficiency of the the proposed algorithm at helping the segmentation procedure and at tracking the objects in the scene. Future work will aim at applying the proposed tracking technique to content-based coding.

6. REFERENCES

[1] F. Dufaux and F. Moscheni. Segmentation-based motion estimation for second generation video coding techniques. In L. Torres and M. Kunt, editors, *Video Coding: The Second Generation Approach*. Kluwer Academic Publishers, 1995 (to be published).

[2] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, Inc., 1988.

[3] J.K. Aggarwal, L.S. Davis, and W.N. Martin. Correspondence processes in dynamic scene analysis. *Proc. IEEE*, 69(5):562–572, May 1981.

[4] F. G. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long image sequences. *Computer Vision, Graphics, and Image Processing*, 60(2):119–140, 1994.

[5] M. Pardas, P. Salembier, and L. Torres. 3D morphological segmentation for image sequence processing. In *Proc. of IEEE Winter Workshop on Nonlinear Signal Processing*, pages 6.1/3.1–6.1/3.6, Tampere, Finland, January 1993.

[6] C. Gu, T. Ebrahimi, and M. Kunt. Morphological spatio-temporal segmentation for content-based video coding. In *VLBV'95*, Tokyo, Japan, November 1995.

[7] F. Dufaux, F. Moscheni, and A. Lippman. Spatio-temporal segmentation based on motion and static segmentation. In *IEEE Proc. ICIP'95*, volume 1, pages 306–309, Washington, DC, October 1995.

[8] F. Moscheni and F. Dufaux. Regions merging based on robust statistical testing. In *SPIE Proc. Visual Communications and Image Processing '96*, Orlando, Florida, March 1996.

[9] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 1991.

[10] G. Taubin and D.B. Cooper. Object recognition based on moment (or algebraic) invariants. In J.L. Mundy and A. Zisserman, editors, *Geometric Invariance in Computer Vision*, pages 375–397. MIT Press, 1992.

[11] C.Y. Lee and D.B. Cooper. Structure from motion: A region based approach using affine transformations and moment invariants. In *IEEE int. Conf. on Robotics and Automation*, volume 3, pages 120–127, Atlanta,GA, 1993.