DETECTION AND TRACKING OF MOVING OBJECTS BASED ON A STATISTICAL REGULARIZATION METHOD IN SPACE AND TIME

Patrick Bouthemy and Patrick Lalande
IRISA / INRIA-Rennes
Campus de Beaulieu, 35042 Rennes Cedex, France

1. Introduction

The function often devoted to a vision system is to detect and track moving objects appearing in its field of view, [1-3]. A variety of applications are concerned with this problem, such as traffic control, remote surveillance of industrial areas, biomedical studies or target tracking. When the camera is static, moving regions in the image plane necessarily correspond to moving objects in the scene. This paper addresses this basic issue of motion detection in an image sequence, and describes an original framework based on statistical models, namely *spatio-temporal Markov fields*. This method generalizes a first attempt of this kind we have recently presented in [4]. Substantial modifications have been introduced, which contribute to solve several shortcomings of the previous algorithm. This new version is able to handle textured moving objects and overlapping cases. By this last term, we mean both, situations where successive projections in time of a moving object overlap each other in the image plane, and occlusion situations between different moving objects. The ability to cope with overlapping cases also avoid to pay attention to the time sampling of the processed image sequence with respect to the size and speed of moving objects of interest. Besides the new way of modeling and using temporal contextual information enables to easily track moving regions through the image sequence.

2. Problem statement

If moving objects are present in the scene, obviously changes in time will occur in the image intensity array. In turn, when the camera is static, temporal intensity changes can be related mainly to motion. Nevertheless, motion detection cannot be reduced to temporal change detection. In particular a moving object gives raise to three kinds of change regions; first one corresponding to uncovered background, second to covered background, third to the overlap of object projections (by the way, this last sub-class is often very partially perceptible). As a matter of fact what is only but completely sought for in every image are projections of moving objects, (also called moving object masks). Usual methods first extract successive temporal change maps, then try to recover projections of moving objects by applying some heuristics, [2,3]. We have adopted a quite different approach. The motion detection issue is considered as a whole and is stated as a *statistical bayesian labeling* problem based on Markov field models. Such an approach has already been proved relevant to other issues related to motion analysis, as reported in [5] for scene segmentation according to motion information, and in [6,7] for optic flow estimation.

In a labeling problem, two sets of elements must be defined: *observations*, (i.e. data to be considered); *labels* (i.e. primitives to be extracted). The remarks of the paragraph

above naturally lead to take as observation the temporal derivative of the intensity function f. Let o_t denote the observation array at time t, and $p = (x, y)$ a pixel, we have: $o_t(p) = \partial f(p)/\partial t$. In fact temporal derivatives will be approximated by finite differences between time t and time $t - dt$, denoted by $\tilde{f}_t$. Moreover another set of information is taken into account: the binary logical map of temporal changes between time t and time $t - dt$: $\bar{o}_t$. It is obtained by an operator able to detect even weak temporal changes; the intensity is locally modeled by a linear function with an additive gaussian noise of constant variance and changes in the model parameter values are validated by a likelihood test as in [8].

On the other hand we consider primitives directly tied to the type of image content we want to delineate. Therefore the label set consists of two symbols, $\Omega = \{a, b\}$, a for moving object masks, b for static background. A priori spatio-temporal models are associated with these primitives; these models must express what properties the solution is supposed to have (that forms the *regularization effect*). Let us first give some intuitive insights to the modeling step. All masks of moving objects obviously share some intrinsic basic spatio-temporal properties. They must show sufficient spatial coherence and their successive positions in time obey a certain law. This can be expressed in terms of required spatio-temporal contextual configurations. Then Markov field models represent very efficient and well-posed means to mathematically formulate this problem, [9]. Let us denote the label field at time t by e_t. This field is modeled as a Markov field in space *and* time; this will be explained in details in the next section.

3. The modeling step and the decision criterion

The markovian property in time is assumed in both directions along the time axis. Indeed we need contextual information from the close past and from the near future to identify the label field at time t. That is the reason why we consider label fields in pairs in the identification process. The solution to the labeling problem is formulated according to the maximum a posteriori (MAP) criterion:

$$max_{e_{t-dt}, e_t} \quad P(e_{t-dt}, e_t / o_t, \bar{o}_t) \tag{1}$$

The best interpretation in terms of moving object projections must have the greatest a posteriori probability given the observations at hand. This statistical approach also permits to properly deal with noise-corrupted observations. Maximizing expression (1) with respect to e_{t-dt} and e_t is equivalent to maximizing the joint probability $\xi = P(e_{t-dt}, e_t, o_t, \bar{o}_t)$. One attractive aspect of this approach is that we can build an explicit version of ξ using the equivalence between Gibbs distributions and Markov fields, as primarily emphasized in the context of image processing in [9].

Local contextual models can then be related to potential functions defined on so-called cliques; cliques are subsets of sites (here sites are pixels) which are mutual neighbors. We consider the following spatio-temporal neighborhood system: a 3x3 spatial neighborhood centered in (p, t) and one-to-one connections from (p, t) towards $(p, t-dt)$ and $(p, t+dt)$. As far as spatial cliques c_s are concerned, we only take into account the four ones comprising two sites (horizontal, vertical and two diagonal cliques). Spatial potentials V_{c_s} have been defined in such a way as to favouring homogeneity of the label field, that is to have

a spatial regularization effect, (e.g. to eliminate isolated points). They are of logistic kind; that is equal to a predefined level, (resp. $-\beta_s$ and β_s), according to the label configuration of the clique at hand, (resp. same labels and different labels), knowing that a negative value encourages the corresponding configuration. It is not necessary here to introduce a complementary edge-site system to take into account discontinuities as in [9], because of the existence of the temporal cliques. These potentials summed over spatial cliques form a first energy function W_s. Of more specificity to the problem at hand, are potential functions tied to the temporal clique also containing two sites. Again they are of logistic kind. They contribute to determine which temporal configurations between (a,a), (a,b), (b,a), (b,b), at pixel p, at two successive times, are encouraged and which are discouraged, according to $\bar{o}_t(p)$ considered as a deterministic external information. They are described in Table 1. In particular overlapping situations are thus correctly handled. These temporal potentials V_{c_τ} lead to a second energy function W_τ. The third energy function W_ϵ will express the adequacy between observations and current estimates of the primitives. We assume that the relation between these two sets is defined by:

$$o_t(p) = \psi(e_{t-dt}(p), e_t(p)) + n(p) \tag{2}$$

where ψ can take a value among three possible ones which can be either predefined or locally estimated on-line; and n is a white (in space and time) zero-mean Gaussian noise of constant variance σ^2. The corresponding energy function W_ϵ is then given by:

$$W_\epsilon = \frac{1}{2\sigma^2} \sum_p [\tilde{f}_t(p) - \psi(e_{t-dt}(p), e_t(p))]^2 \tag{3}$$

The noise variance is estimated once at the beginning of the processed image sequence. Finally we get the total energy function

$$W = W_s + W_\tau + W_\epsilon \tag{4}$$

We can assume that ξ is proportional to $\exp(-W)$. The optimal map of moving object masks will correspond to the lowest value of energy W.

4. The optimization procedure, the tracking scheme, and results

To minimize W, we use an iterative deterministic method as in the early version described in [4]. It yields a very good trade-off in our case between computation speed and result quality. Moreover it is completed by an efficient procedure for iteratively selecting sites to be visited, as suggested in [10]. Let us outline that all computations are very local. The optimization of every label field e_t is done in a two-pass manner. First we derive a first estimate $\check{e}_t$ when considering pair (e_{t-dt}, e_t); second, considering pair (e_t, e_{t+dt}), we update $\check{e}_t$ and we get the final estimate $\hat{e}_t$; of course the first estimate $\check{e}_{t+dt}$ is simultaneously obtained. Afterwards the same holds for e_{t+dt} and so on. This prediction scheme, associated with the markovian property of the label field in time, explains that we can now easily implement a tracking procedure of the detected moving regions through the image sequence.

This is essentially a matter of recursive number allocation. First, an initializing step is needed; that is, given the first estimated label field $\hat{e_{t_0}}$ a different number is assigned to

each connected subset of a-labeled points. Then let us assume that the tracking process has been achieved until time $t-dt$, it is pursued at time t as follows. Let q_k be a reference point among points with number k in the image at time $t-dt$. The same number k is assigned to the point p_q in the image t belonging to the temporal clique of q_k. This assumes that the intersection of two successive projections in time of a given moving object is not empty. Then starting from point p_q, the assignment of number k in the image t is propagated step-by-step to the a-labeled points connected through the spatial cliques. When this is achieved, number $k+1$ is taken into account, and so on. Two specific cases may happen: the merging and the splitting of moving masks. They can correspond for instance to the crossing of two different moving objects. The first case will be detected when two different numbers k and k' are present in a spatial clique. This conflict situation is easily solved by "equaling" k' to k (i.e. by creating a link). The second case will provide the same situation as the one encountered when a new moving object is appearing. A subset of connected a-labeled points in the image at time t will remain without any number assigned at the end of the process. It will then receive a new number. This tracking stage has been separately presented to make the explanation easier. Actually numbers k can be considered as supplementary primitives which can be straightforwardly included in the modeling step. Therefore detection and tracking can be nearly simultaneously carried out.

Numerous experiments with several image sequences depicting outdoor scenes have been processed. Results are quite fine. We present here one example. Fig.1 shows three (not successive) images from the input sequence taken at times t_1, t_2, t_3. The camera is static and some cars are moving down and another one is moving up. The binary map of the change regions supplied by a simple temporal intensity change detector depicts very incomplete moving object masks and a lot of spurious isolated points all over the image. In the results obtained with the markovian approach, Fig.2, the stationary background is free of spuriously detected points and the extracted regions rather well correspond to the real masks of the moving objects. This is confirmed throughout the entire processed sequence. Let us outline that the images are not of high quality and that the part missing in the moving object mask at the bottom of the image is due to the complete lack of contrast between background intensity and car windscreen intensity. It has also been found that the parametrization of the model is not a critical problem for this motion detection issue; the same set of parameter values has been used for different image sequences. Moreover the number of parameters is rather small.

We have described a general, model-based and robust method for motion detection in an image sequence, which besides leads to a simple tracking procedure. This method does not suppose any a priori knowledge on the respective intensity values of moving object projections and background; it does not require anymore any identification of the background intensity distribution. As all the computations are local, an efficient fast implementation is indeed reachable, which allows an effective use in practical situations.

References:

[1] **H.-H. Nagel**, From image sequences towards conceptual descriptions, *Image and Vision Computing*, Vol. 6, No 2, May 1988, pp.59–74

[2] **R. Jain**, Dynamic scene analysis, in *Progress in Pattern Recognition 2*, L. Kanal and A.

Rosenfeld (eds.), Mach. Intell. and Patt. Rec. Series, Vol. 1, North–Holland, 1985, pp.125–167
[3] **J. Wiklund** and **G.H. Granlund**, Image sequence analysis for object tracking, *Proc. 5th Scandinavian Conf. on Image Analysis*, Stockholm, June 1987, pp.641–648
[4] **P. Bouthemy** and **P. Lalande**, Determination of apparent mobile areas in an image sequence for underwater robot navigation, *Proc. IAPR Workshop on Computer Vision: Spec. Hardw. and Ind. Applic.*, Tokyo, Oct. 1988, pp.409–412
[5] **D.W. Murray** and **B.F. Buxton**, Scene segmentation from visual motion using global optimization, *IEEE Trans. on PAMI*, Vol. 9, No 2, March 1987, pp.220–228
[6] **J. Konrad** and **E. Dubois**, Multigrid Bayesian estimation of image motion fields using stochastic relaxation, *Proc. 2nd Int. Conf. on Computer Vision*, Dec. 1988, pp.354–362
[7] **F. Heitz** and **P. Bouthemy**, Motion estimation and segmentation using a global Bayesian approach, *Int. Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, April 1990
[8] **Y.Z. Hsu, H.–H. Nagel** and **G. Rekers**, New likelihood test methods for change detection in image sequences, *Computer Vision, Graphics and Image Processing*, Vol. 26, 1984, pp.73–106
[9] **S. Geman** and **D. Geman**, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. on PAMI*, Vol. 6, No 6, Nov. 1984, pp.721–741
[10] **P.B. Chou** and **R. Raman**, On relaxation algorithms based on Markov random fields, *Technical Report* No 212, Computer Science Dpt, Univ. of Rochester, July 1987, 28p.

This work was partly supported by the French CNRS Program, "PRC Communications Homme–Machine, pôle Vision". The image sequence was provided by the Laboratoire d'Electronique, University of Clermont-Ferrand

Figure 1: *Three original images (not successive) out of the sequence at times t_1, t_2, t_3.*

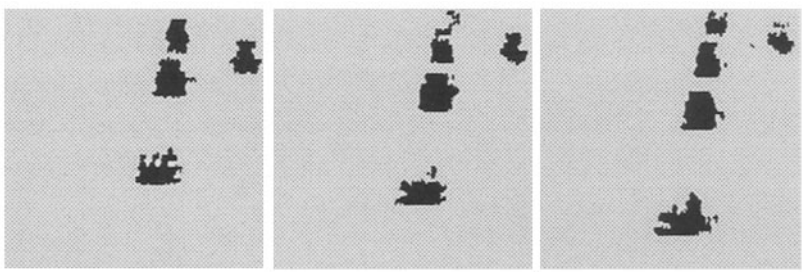

Figure 2: *The moving object masks at times t_1, t_2, t_3; ($\beta_\tau = 100$, $\beta'_\tau = 1000$, $\beta_s = 10$) (The time interval dt indeed corresponds to the video rate).*

$(e_{t-dt}, e_t, \bar{o}_t)$	(b,b,0)	(b,b,1)	(a,b,0)	(a,b,1)	(b,a,0)	(b,a,1)	(a,a,0)	(a,a,1)
$V_\tau(e_{t-dt}, e_t, \bar{o}_t)$	$-\beta_\tau$	$+\beta_\tau$	$+\beta'_\tau$	$+\beta'_\tau$	$+\beta_\tau$	$-\beta_\tau$	$+\beta_\tau$	$-\beta_\tau$

Table 1: *The temporal potentials ($\bar{o}_t = 1$, means temporal change, $\bar{o}_t = 0$ no change)*