MRF-BASED MOTION SEGMENTATION EXPLOITING A 2D MOTION MODEL ROBUST ESTIMATION

Jean-Marc Odobez and Patrick Bouthemy

IRISA/INRIA, Campus universitaire de Beaulieu
35042 Rennes Cedex, France
e-mail : odobez@irisa.fr bouthemy@irisa.fr

ABSTRACT

This paper is dealing with motion-segmentation, that is, with the partitioning of the image into regions of homogeneous motion. Here, homogeneous means that in each region a 2D polynomial model (e.g. an affine one) is able to describe at each location the underlying "true" motion with a predefined precision η. However, no estimation of this true motion field is required. The motion models are computed using a multiresolution robust estimator [1]. Therefore, as opposed to almost all other motion-segmentation scheme (for instance [2, 3]), the motion model of a given region only needs to be estimated once at a given time instant. Moreover, the determination of the boundaries between the different regions, which is stated as a statistical regularization based on a multiscale Markov Random Field (MRF) modeling, only requires one pass. Finally, thanks to the definition of an explicit detection step of areas where the error between the underlying motion and the one given by the estimated models is not whithin the precision η, we are able to get a good segmentation from the very beginning of the sequence, and to manage the appearance of new objects in the scene, as well as the momentary increase in complexity of motion in already existing regions. Results obtained on many real image sequences have validated our approach.

1. INTRODUCTION

Motion-based segmentation [3, 2, 4, 5] is an essential tool in dynamic scene analysis applications, like computation of time to collision, obstacle avoidance, detection and tracking of moving objects... It is also of great importance in image coding, where block matching schemes have shown their limit at yielding a good visual quality for the image sequences reconstructed at the receiver, whenever the allowed bitrate is very low or motions are complex. Moreover, in several image transmission applications, a spatially-constant reconstruction quality may not be required ("Head and shoulders" scenes, remote surveillance video systems, etc...). In this context, motion-based segmentation plays two essential roles. First, it is useful to analyse and interpret the scene, as well as to extract its different and visually important moving parts [6]. The choice of motion as a segmentation criterion allows us to deal with a small number of regions, as opposed to grey-level based segmentation approach. Second, the motion-based segmentation can be used directly for motion compensated coding purposes.
The motion-based segmentation algorithm that we propose, its features as well as its assets are described in the following sections.

2. GENERAL APPROACH

The motion-based segmentation algorithm relies on the use of 2D parametric motion models, robust estimation, and multiscale Markov Random Field (MRF) models. The goal of the segmentation is to jointly estimate the motion models $\{(\Theta_k)_t^{t+1}\}_{k\in\{1,\ldots,N_r\}}$ and the associated region partition $e(t) = \{R_k(t)\}_{k\in\{1,\ldots,N_r\}}$ at time t[1], whose labels are in the set $\{1,\ldots,N_r\}$. N_r stands for the number of regions in the image, and has to be estimated on-line also. Though any 2D polynomial motion model could be considered, we mainly dealt with the affine one. With this model, the displacement $\vec{d}_{\Theta_k}$ at point $s=(x,y)$ in region k is described by:

$$\vec{d}_{\Theta_k}(s) = \begin{pmatrix} a_1^k + a_2^k x + a_3^k y \\ a_4^k + a_5^k x + a_6^k y \end{pmatrix} \text{ with } \Theta_k = (a_i^k)_{i=1\ldots6} \quad (1)$$

Most approaches generally proceed in two steps that are iterated until convergence [2, 3]. The first one consists in estimating the motion models given the current partition; the second one in determining the optimal partition, the motion models being kept unchanged. In our case, those iterations that can be computationaly expensive, are avoided thanks to the use of a robust multiresolution motion estimator described in [1]. Let us note that robust estimation has already been used in [4] for segmentation purpose, but in a quite different manner. A least-median-square technique was used along with temporal decomposition for the motion model estimation, and segmentation was based on a merging procedure from an initial spatial segmentation, both steps leading to a far more complex solution.

The determination of the motion segmentation at a current intant t, given the segmentation at the previous instant, is basically made out of four main steps: the prediction of an initial partition map, the motion model estimation, the updating of the partition given the computed motion models, the detection of new regions. In the next sections, we describe each of these different steps.
The initialization step, i.e. the search for the motion-based segmentation corresponding to the two first images of the sequence, is conducted in the same way, starting with a "predicted" segmentation map composed of one single region.

This study was supported in part by the French Ministry of Research in the context of the GDR-PRC "Man-Machine Communications"(Vision research program, MRT contract 91S269), and by "Région Bretagne" (Brittany Council) through a contribution to student grant.

[1] We will drop the time indexes when there is no ambiguity.

3. PREDICTION

The prediction of the partition at time t, denoted $\tilde{e}(t)$, is determined using the segmentation map along with the estimated motion models obtained at time $t-1$. This allows us to supply a coherent labeling of the same motion entity in the successive partitions over time. More precisely, the label k in $\hat{e}(t-1)$ at each pixel s is affected to each point on the grid around $s + \vec{d}_{(\hat{\Theta}_k)_{t-1}^{t}}(s)$ in $\tilde{e}(t)$. Pixels that receive no label (discovered regions between $t-1$ and t) are given a special label, as well as pixels that receive multiple labels (occlusion areas).

4. ROBUST MOTION ESTIMATION

The motion models $(\Theta_k)_t^{t+1}$ between frame at time t, I_t and frame at time $t+1$, I_{t+1} are estimated using the initial partition $\tilde{e}(t) = \{\tilde{R}_k(t)\}$. Owing to the robustness of our estimator, an imprecise predicted map or the appearance of new objects should not perturb the estimation process. More precisely, our estimator takes advantage of a multiresolution framework and an incremental scheme based on the Gauss-Newton method. It minimizes an M-estimator criterion with a hard-redescending function to ensure the goal of robustness. i.e. $(\Theta_k)_t^{t+1}$ is computed as follows [1]:

$$(\hat{\Theta}_k)_t^{t+1} = \underset{\Theta_k}{\operatorname{argmin}} \sum_{s \in \tilde{R}_k(t)} \rho\left(\mathrm{DFD}(s,k)\right)$$

$$\textit{with} \quad \mathrm{DFD}(s,k) = I_{t+1}(s + \vec{d}_{\Theta_k}(s)) - I_t(s), \qquad (2)$$

and $\rho(x)$ is the Tuckey's function, bounded for high values of x.

5. UPDATING OF THE PARTITION

The updating of the partition $e(t)$ using the estimated models $(\hat{\Theta}_k)_t^{t+1}$ is achieved through a statistical regularization approach based on multiscale MRF. More precisely, we use a MAP criterion, which, taking advantage of the MRF modeling, leads to the minimisation of an energy $U(e, o, \tilde{e})$ (where o is the field of observations and is composed of the images I_t and I_{t+1}):

$$U(e, o, \tilde{e}) = U_1(e, o) + U_2(e) + U_3(e, \tilde{e})$$

• We have paid particular attention to energy term U_1, which expresses the adequacy between the labels and the observations. Since we are dealing with *motion* segmentation, we prefer to rely our analysis on local motion estimates rather than on the usual displaced frame difference: $DFD(s,k)$. However, as the computation of a flow field is an intricate problem, as difficult to solve as the segmentation one, we only use partial motion measures $o_s(k)$ that can be straightforwardly computed from the images [7]: a weighted average of the residual -i.e. after motion compensation- normal flows $|DFD(p,k)|/\|\vec{\nabla} I(p)\|$, where $\vec{\nabla} I(p)$ is the spatial intensity gradient at point p, and p are points of a small neighborhood $N(s)$ of site s:

$$o_s(k) = \frac{\sum_{p \in N(s)} |DFD(p,k)| \times \|\vec{\nabla} I(p)\|}{\max(m, \sum_{p \in N(s)} \|\vec{\nabla} I(p)\|^2)} \qquad (3)$$

where m is a predetermined positive constant to account for noise in the uniform areas. An interesting property of this measure is the following. We can derive two bounds l_s and L_s, which only depend on the local distribution of directions of the spatial intensity gradient and on a preset factor η, such that:

$$\begin{cases} o_s(k) < l_s & \Rightarrow \ \|\Delta\vec{d}(s,k)\| < \eta \\ o_s(k) > L_s & \Rightarrow \ \|\Delta\vec{d}(s,k)\| > \eta \end{cases}$$

with $\Delta\vec{d}(s,k) = \vec{d}_{\mathrm{true}}(s) - \vec{d}_{\hat{\Theta}_k}(s)$. The potentials involved in the energy term U_1 reflect the "energetic translation" of the above inequalities:

$$U_1(e,o) = \sum_{s \in S} \alpha F_s \times V_1(e_s, o) \text{ with } F_s = A_{G,1}(\|\vec{\nabla} I(s)\|)$$

$$\text{and } V_1(e_s, o) = A_{l_s,2}(o_s(e_s)) - (1 - A_{L_s,2}(o_s(e_s))$$

where $A_{tr,r}(x)$ is a smoother version of a step edge. Thus, it is an increasing function from 0 to 1, such that the step occurs at tr ($A_{tr,r}(tr) = 0.5$) and the slope of the transition is controlled by r ($\mathrm{d}A_{tr,r}/\mathrm{d}x(tr) = r$). We use the normalized arctangent: $\frac{1}{\pi}\arctan(r\pi(x-tr)) + 0.5$ instead of a sigmoide, because it reaches values 0 or 1 less rapidly. A site with low image gradient usually carries poor and unreliable information about the adequation of a given motion model; therefore, the role of the damping factor F_s is to reduce the amplitude of the energy term provided by the observations at such a site (approximately characterized by $\|\vec{\nabla} I(s)\| < G$), which *conversely* increases the relative contributions of the regularisation terms.

The figure 1 shows how these bounds and inequalities allow us to take into account the well known "aperture problem". More precisely, they allow us to tell, *given the local intensity structure*, which motion model is "adequate" (i.e., describes the underlying true motion with precision η), and which is not. When there exists an ambiguity (for instance, if there is not enough structure, like in uniform areas, or if several motion models are convenient,...), energy terms U_2 and U_3 that we now briefly describe introduce the contextual information necessary to remove it and perform a correct labeling.

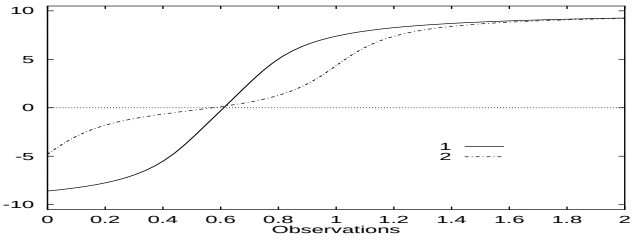

Figure 1: Potentiels V_1 in two extrem cases: case of a site located on a corner (curve 1), and on straight edge (curve 2). The curves, through the bounds, reflect the local intensity structure, and thus, the aperture problem. For instance, a very low energy indicates that the residual motion magnitude is below a given predefined value η. In the straight edge case, because of the aperture, an observation, however low it is, cannot tell with certainty whether the residual motion magnitude is below η; therefore, the energy does not fall as low as in other cases, like in the corner case, where the aperture problem is less important.

• Energy term U_2 accounts for the expected spatial properties (homogeneity) of the label field, and has the following

usual expression:

$$U_2(e) = \sum_{(s,t)\in\mathcal{C}} \beta_d(1 - \delta_{e_s=e_t})$$

where $\mathcal{C}$ represents the set of cliques of two elements associated to a second order neighbourhood system ν, and δ is the Kronecker function.

• U_3 favours the conservation of labels over time (except in occluded and uncovered regions between I_{t-1} and I_t where no label is favoured), when the newly estimated motion models $(\hat{\Theta}_k)_t^{t+1}$ are still able to correctly describe the motion inside their corresponding region. It is given by:

$$U_3(e, \tilde{e}) = \sum_{s\in S} F_s \times \beta_{dt}(1 - \delta_{e_s=\tilde{e}_s})$$

where δ is again the Kronecker function. Note that this energy term also plays a second role. Suppose that a new region l is created (see next section) in an area inside the region k, because the motion estimate $\hat{\Theta}_k$ is not adequate anymore to describe the motion in this area. This new region l, of course not existing in $\tilde{e}$, will therefore have to bring a significant gain in the description of motion in this area -through $\hat{\Theta}_l$ and thus, U_1-, to compensate for the penalty introduced by the potential V_3. As the possible gain is weighted by F_s in U_1, the penalty term V_3 must also be weighted by F_s.

The global minimization of the energy function is performed using a multiscale approach, [8], where at a given scale the solution is computed with the Highest Confidence First minimization procedure [9].

In this 3^{rd} step of the algorithm, the areas labeled "occluded" or "uncovered" in the prediction step (section 3) are assigned the labels that suit them. The updating of the boundaries between two regions k and k', according to the new motion estimates, is also performed. Thus, at the end of this step, the segmentation into regions with homogeneous motion, using the number of motion models intervening in the segmentation at the previous instant, is achieved. The purpose of the step that follows is therefore to test: a) if new mobile objects have appeared in the scene; b) if the current number of motion models is still enough to provide a good description of the apparent motion in the image.

6. DETECTION OF NEW REGIONS

Within each region k, sub-areas whose motion do not conform to the estimated motion model $\hat{\Theta}_k$ are detected. This is achieved using a scheme similar to the one described in [7], which was concerned with the detection of moving regions not conforming with the global motion model estimated in the whole image. The significant connected components are extracted from the set of all those sub-regions where the existing motion models are not valid, and the number of region N_r is updated accordingly. If there is no significant new regions, the final partition at time t is that obtained at the end of the last relaxation performed at step 3 . If there is, the motion models in the newly created regions are estimated still using the multiresolution robust estimator, and the partition is updated again according to the relaxation scheme described in step 3. It is important to note here, that the relaxation now involves only few computations, since it is only concerned with the adjustment of the boundaries of the created regions.

7. RESULTS AND CONCLUSION

Figures 2a-b, 3a-b and 4a-b illustrate the performance of our algorithm, where all parameters defining the algorithm are the same, except for the value of η, which represents the precision with which a motion model in a given region represents the underlying true motion. Indeed, Fig. 2a and 2b emphasis the role of this parameter, and displays results obtained with two different values of η. As expected, a larger value of η is more suitable for motion analysis, as in that case the algorithm correctly captures only the few principal motion components of the scene. A smaller value can be used if a better description of the motion field is preferred, for motion-compensation purposes in image coding for instance, and usually leads to the recovery of more regions. The results presented here, as well as others not reported here, validate the robustness and the accuracy of our approach.

8. REFERENCES

[1] J-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *To appear in International Journal of Visual Communication and Image Representation*, 1995.

[2] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Int. Journal of Computer Vision*, Vol.10, No 2:157–182, April 1993.

[3] E.H. Adelson and J.Y.A Wang. Representing moving images with layers. *IEEE Trans. Image Proc.*, 3(5):625–638, Sept. 1994.

[4] S. Ayer, P. Schroeter, and J. Bigün. Segmentation of moving objects by robust motion parameter estimation over multiple frames. In *Proc. of the 3rd ECCV*, vol. 2, pp 316–327, Stockholm, May 1994.

[5] S. F. Wu and J. Kittler. A gradient-based method for general motion estimation and segmentation. *Jal of Visual Communication and Image Representation*, 4(1):25–38, March 1993.

[6] E. Nguyen, C. Labit, and J-M. Odobez. A ROI approach to hybrid image sequence coding. In *Proc.* 1^{st} *IEEE ICIP*, vol. 3, pp 245–249, Austin (TX), Nov. 1994.

[7] J-M. Odobez and P. Bouthemy. Detection of multiple moving objects using multiscale MRF with camera motion compensation. In *Proc.* 1^{st} *IEEE ICIP*, vol. 2, pp 257–261, Austin (TX), Nov. 1994.

[8] P. Perez, F. Heitz, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP : Image Understanding*, 59(1):125–134, January 1994.

[9] P.B. Chou and C.M. Brown. The theory and practice of Bayesian image modeling. *Int. Jal of Computer Vision*, Vol.4:185–210, 1990.

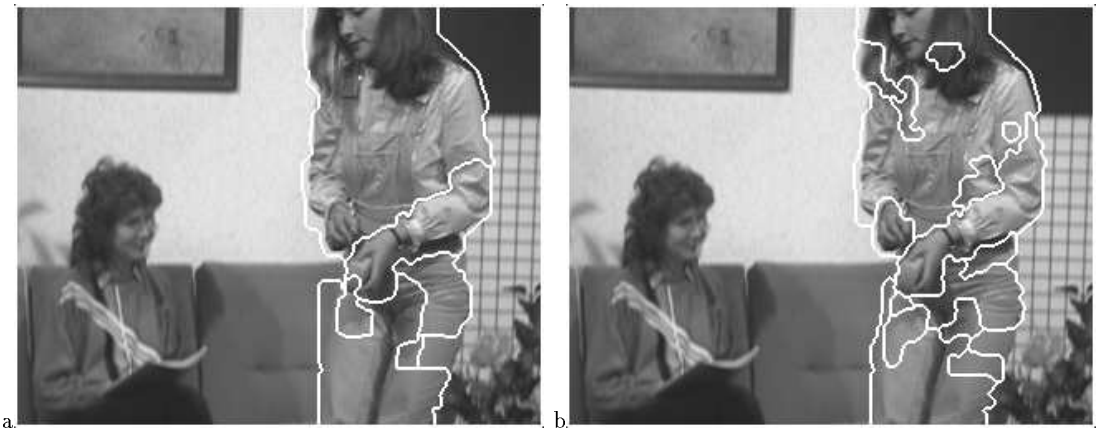

Figure 2: "Interview" sequence. The camera is tracking the woman on the right, who is standing up while moving her arm. Motion-based segmentation map at time t_{37} with two different "precision" levels: a) $\eta = 1.25$ and b) $\eta = 0.75$. Boundaries are overprinted in white. In Fig. 2b, the higher precision required (w.r.t. Fig. 2a) implies that the algorithm creates regions to account for the motion of the right hand, the hairs,...

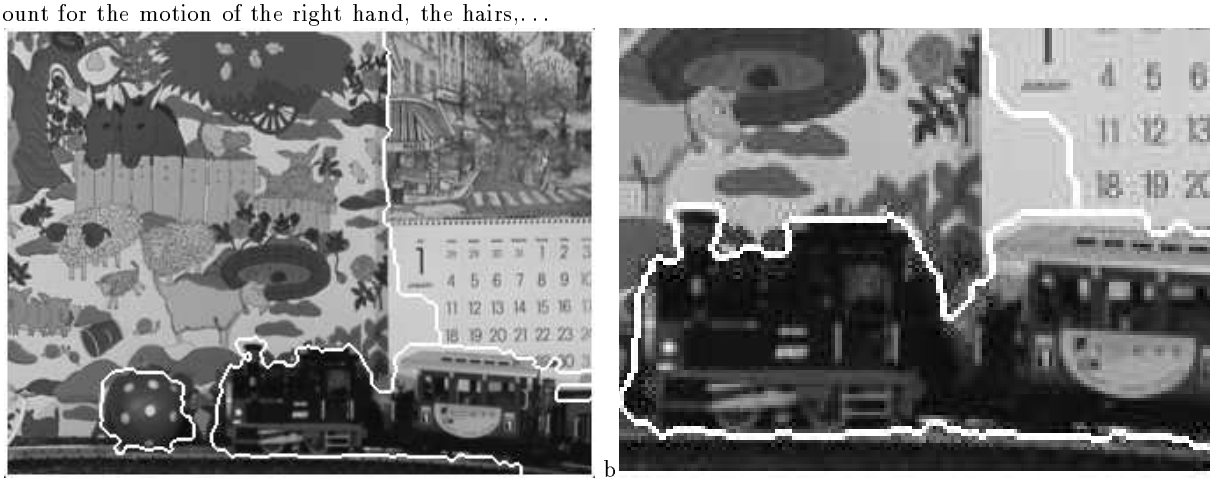

Figure 3: "Mobile" sequence. The camera is panning the scene while the calendar is sliding vertically, the ball is rolling and the train is moving forward. a) segmentation map obtained at time t_1. b) enlargement of one part of the segmentation map obtained at time t_4. $\eta = 0.6$. Similar results have been obtained with values ranging from 0.4 to 1.0 pixels.

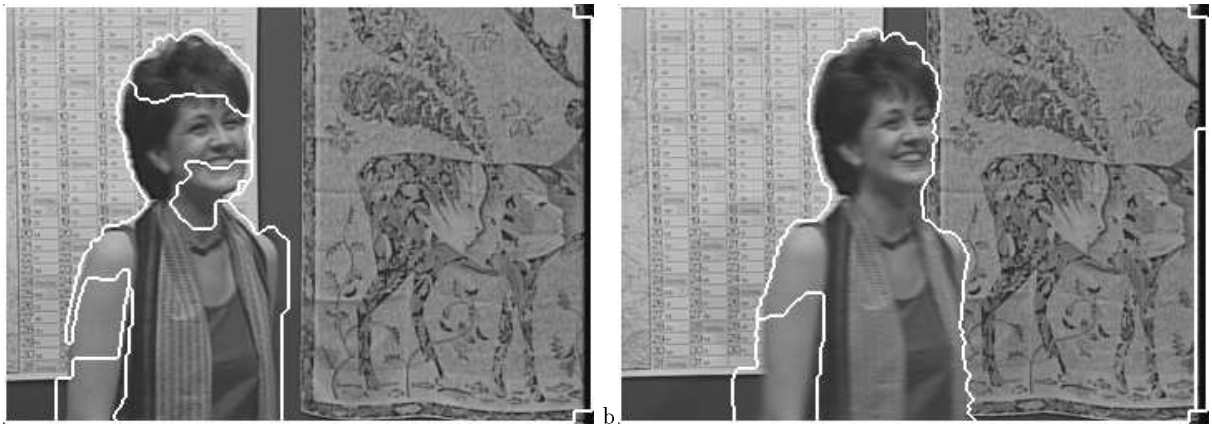

Figure 4: "Renata" sequence: segmentation map at time t_1 (a) and t_{15} (b). The small swing of the arm, as well as the slight nod of the head at the beginning of the sequence, are taken into account. The camera is also moving. $\eta = 0.5$. Similar results have been obtained with values ranging from 0.4 to 1.0 pixels.