DETECTION OF MULTIPLE MOVING OBJECTS USING MULTISCALE MRF WITH CAMERA MOTION COMPENSATION

J.M. Odobez and P. Bouthemy

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

ABSTRACT

In this paper, we address the problem of detecting moving objects from a moving camera. The apparent flow field induced by the camera motion is modeled by a 2D parametric motion model and compensated for using the values of the parameters estimated by a multiresolution robust method. Motion detection is achieved through a statistical regularization approach based on multiscale Markov random field (MRF) models. Particular attention has been paid to the definition of the energy function involved and to the considered observations. This method has been validated by experiments carried out on different real image sequences.

1. INTRODUCTION

Detection and location of moving objects in an image sequence is a basic task in numerous applications, and useful as an initial step in many scene analysis schemes. It is therefore one of the main topics addressed in motion analysis. In the case of a static camera, different efficient solutions for the detection of moving objects have been developed ([1, 3, 6]). Although these methods are sometimes robust to small camera motion perturbations [6], they cannot be used when we are dealing with a mobile camera. In that case, among others, two classes of solutions can be found : the first one uses (*a priori* or sometimes estimated) information on the 3D camera motion, in order to derive constraints on the 2D apparent flow field of static objects in the image sequence induced by the movement of the camera [7, 10]; regions where these constraints are violated are then identified as projections of moving objects. The second class of solutions [5, 11] is based on the assumption that the apparent motion of the static background (due to camera motion) can be modeled by a 2D parametric motion model and is considered as the *dominant motion*. Such a motion model is estimated from frame to frame, and then used to compute a compensated sequence in which the background then appears as static ; thus, non-static regions in this sequence can be considered as moving objects (or as objects located at a significantly different depth compared to the background). Projections of moving objects are then obtained by simply thresholding some local error or statistical function at each pixel [11], which usually leads to noisy detection maps for most real sequences. In [5], the average of a few successive images registered (or compensated) by the computed dominant motion is used as a reference map. However, this "integration step" assumes that the regions, whose motion correspond to the computed dominant motion, are perfectly registered over the integration duration. If not, this temporal integration blurs the reference map, and the next motion estimations, which use the reference map as first image, can be greatly affected.

In this paper, we propose a motion detection algorithm belonging to the second class which can tolerate noisy data and imprecise registration (since the motion model is only an approximation to the dominant motion). This algorithm uses interframe observations rather than observations related to a reference frame, and relies on a statistical regularization approach based on MRF models.

2. MOTION ESTIMATION

The dominant motion is computed using a gradient-based robust estimation method that we have described in [8]. Any 2D polynomial motion model can be considered. This method 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. We have chosen the affine mo-

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.

tion model $\vec{w}_{\Theta_t}$ defined by:

$$\vec{w}_{\Theta_t}(r) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix}, r = (x, y) , \quad (1)$$

and we estimate the parameter vector $\Theta = (a_i), i = 1..6$ between frame I_t and I_{t+1} as follows :

$$\widehat{\Theta_t} = \underset{\Theta_t}{\operatorname{argmin}} \sum_{r \in R(t)} \rho\left(\mathrm{DFD}_{\Theta_t}(r)\right)$$

$$with \quad \mathrm{DFD}_{\Theta_t}(r) = I_{t+1}(r + \vec{w}_{\Theta_t}(r)) - I_t(r) , \quad (2)$$

and $\rho(x)$ is a function which is bounded for high values of x (more precisely, we use Tukey's biweight function). This estimator allows us to get an accurate estimation of the dominant motion (i.e., background apparent motion) between two images, which is of key interest for the subsequent steps.

3. MOTION DETECTION ALGORITHM

3.1. Detection problem

Let S denote the set of sites s (pixels) and $\mathcal{C}$ the set of cliques of two elements associated to a second-order neighbourhood system ν. The issue of motion detection will consist in searching the binary label field (called detection map) $d = \{d_s, s \in S\}$ which is the most likely to have produced the field of observations at time t, $o^t = \{o_s^t, s \in S\}$. We would like to detect the set of regions whose motion does not conform to the estimated dominant motion. As explained in the introduction and as we will see in the experiments, this set may include more elements than the projections of the moving objects. However, for convenience, the labels will still be named "static" and "mobile" instead of "conforming" and "non-conforming". To solve this problem, we use the Maximum A Posteriori (MAP) criterion, i.e., we want to maximize the *a-posteriori* distribution of the labels given the observations. If we use MRF to model the sets of observed and hidden variables, and due to the equivalence between MRF and Gibbs distribution ($p(x) = \frac{1}{Z} e^{-U(x)}$), [4], this is equivalent to the minimization of an energy function $U(d, o)$, which is now described.

3.2. Energy function

We decompose $U(d, o)$ into three terms :

$$U(d, o) = U_1(d) + U_2(d, \tilde{d}) + U_3(d, o), \quad (3)$$

$$\text{with} \qquad \tilde{d} = \mathrm{reg}_{t-1}^t(d^{t-1}) \quad (4)$$

where d^{t-1} is the detection map at time $t-1$ and $\mathrm{reg}_i^j(X)$ consists in transforming the map X at time i into a map at time j using the motion model estimated between i and j.

• U_1 is the regularization term which accounts for the expected spatial properties (homogeneity) of the label field:

$$U_1(d) = \sum_{\{s,u\} \in \mathcal{C}} A_r(d_s, d_u) \quad \text{with} \quad (5)$$

$$A_r(d_s, d_u) = \begin{cases} -\beta_m & \text{if } d_u = d_s = \text{“mobile”} \\ 0 & \text{if } d_u = d_s = \text{“static”} \\ \beta_{\text{diff}} & \text{if } d_u \neq d_s \end{cases} \quad (6)$$

β_{diff} is the cost to pay to get neighbours with different labels, and β_m ($0 < \beta_m \ll \beta_{\text{diff}}$) is a potential value which facilitates the selection of the "mobile" label.

• Energy U_2 plays a "conservative" role and takes into account the detection map at the preceding instant :

$$U_2(d, \tilde{d}) = \sum_{s \in S} A_p(d_s, \tilde{d}_s) \quad \text{with} \quad (7)$$

$$A_p(d_s, \tilde{d}_s) = \begin{cases} 0 & \text{if } \tilde{d}_s = d_s \\ +\tilde{\beta}_{\text{diff}} & \text{if } \tilde{d}_s \neq d_s \end{cases} \quad (8)$$

• $U_3(d, o)$ is the energy term expressing the adequacy between observations and labels. The observation we use is a weighted average of the expression $\frac{|\mathrm{DFD}_{\widehat{\Theta}_t}(r)|}{\|\nabla I_t(r)\|}$, that is:

$$o_s^t = \frac{\sum_{r \in \nu(s) \cup \{s\}} |\mathrm{DFD}_{\widehat{\Theta}_t}(r)| \times \|\nabla I_t(r)\|}{Max(m, \sum_{r \in \nu(s) \cup \{s\}} \|\nabla I_t(r)\|^2)} \quad (9)$$

where m is a predetermined positive constant. It can be shown that if a site s is supposed to undergo a translation of magnitude δ, then we have the inequality : $o_s^t \geq l_s = \delta \times f(s)$, where $f(s)$ depends only on the local image gradient distribution. A large value of o_s^t means that the site is likely to have moved. Due to the well-known "aperture problem", if a straight edge is sliding along itself, observations will be nearly zero though there is motion. However, the bound l_s is equal to zero in that case. Hence, a low observation value will really indicate a static site only if it is lower than the bound l_s. Therefore, if we want to detect moving objects with a velocity magnitude greater than δ, the following energy term U_3 is appropriate:

$$U_3(d, o) = \sum_{s \in S} \alpha_s A_a(d_s, o_s) \text{ with } \alpha_s = \alpha\sigma(\|\nabla I_t(s)\|)$$

$$A_a(d_s, o_s) = \begin{cases} -\tan^{-1}(k_1(o_s^t - \delta)) & \text{if } d_s = \text{“mobile”}, \\ +\tan^{-1}(k_2(o_s^t - l_s)) & \text{if } d_s = \text{“static”}, \end{cases} \quad (10)$$

where α is a constant, and $\sigma()$ transforms the gradient norm into a multiplicative weight between 0 and 1. Since a site with low image gradient usually carries poor and unreliable information about the presence of motion, the role of $\sigma()$ is to lower the gap at such a

site between the value of energy term U_3 if that site is labeled as "mobile" and the value if it is labeled as "static". Thus, $\sigma()$ is defined as:

$$\sigma(x) = \frac{1}{1+\exp^{-(x-G)}} \qquad (11)$$

where G is a parameter which is set according to the noise level in the image. In the design of the potential $A_a()$, we have chosen the inverse tangent function for three reasons : first, it is easy to locate the transition separating relevant from irrelevant information given a label (around δ and l_s respectively); second, the shape of this transition is easy to control (through parameters k_1 and k_2) and independent of the transition location; third, energy functions are bounded, and thus behave similarly to a "robust estimator", by avoiding erroneous information to locally impose the wrong label even if all the neighbors disagree. Fig. 1 illustrates the form of potentials $A_a()$ and highlights the role of the bound l_s. In this plot, we can recognize the energy curve associated with the "mobile" label, because it exhibits a low energy for high observations, and conversely. The two other curves are associated with the "static" label, but with different value of the bound l_s: the maximal value l_{max} (equal to $\frac{\delta}{2}$), and the lowest l_{min}(equal to 0). As expected, these curves do not show much difference for high observation values. However, for small observation values, the influence of the bound is obvious and corresponds to the qualitative analysis we have reported earlier: for a given observation (of low value), the gap between the energies associated with the "mobile" label and the "static" one is smaller for a site with a low bound (i.e. a site located on an edge or a site lying in a uniform region) than for a pixel with a high bound, (i.e. a corner). Therefore, for the same small observation value, it will be easier to force a site to be labeled as "mobile" if it has a low bound.

The energy term U_3 can be further improved by considering K past observations at a site s. Denoting $\tilde{o}^{t-q} = \mathrm{reg}^t_{t-q}(o^{t-q})$ and assuming that the temporal information are independent, we can derive the expression of the energy U_3 that we actually use :

$$U_3(d,o) = \sum_{q=0}^{K} \frac{\gamma_q}{\sum_{q=0}^{K}\gamma_q} \left(\sum_{s\in S} \alpha_s A_a(d_s, \tilde{o}^{t-q}(s)) \right) \qquad (12)$$

where $\gamma_q = \gamma^q$, and $\gamma \in [0,1]$ is a damping factor. If $\gamma = 0$, only the current observation is taken into account ; if $\gamma = 1$, the K past observations are considered equivalently. Because here we consider misregistration of observations between two successive images of the sequence (and not between an average image and an image of the sequence as in [5]), the only assumption we make is that our motion model is valid between two frames.

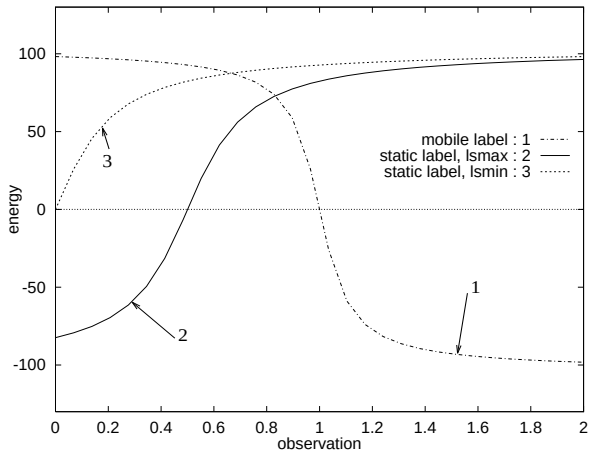

Fig 1: energy U_3 in function of the observation o_s, and according to the label and the bound l_s.

3.3. Computational issues

The global minimization of the energy function is performed using a multiscale approach, [9]. More precisely, it consists in starting the minimisation process at the lowest scale L, where the solution is constrained to be constant within blocks of size $2^L \times 2^L$. At this scale, the initial detection map is obtained by maximizing the conditionnal likelihood, i.e. by locally minimizing the energy term U_3. Then, the Highest Confidence First [2] minimisation procedure is used to compute the solution at that scale. Finally, minimisation is performed from scale to scale, using the projection onto scale $l-1$ of the detection map obtained at scale l as an initial solution, until the finest scale is reached. In [9], it is shown that the multiscale approach gives results very similar to the stochastic relaxation, but is much faster. Indeed, our algorithm has a low computational cost (about 3 seconds on a SPARC 10 for a 256×256 image), which could be further reduced using the inherent parallelism of the MRF model, or if pixel-level precision is not required.

Parameter setting is not a crucial matter. Parameter values are chosen according to "the noise level" in the image sequence (for the value of m and G), the minimal motion magnitude to be detected (for the value of δ), the minimum size of the moving objects (for the value of β_{diff} and β_m), the temporal change rate of the image content (for the value of γ and $\tilde{\beta}_{\mathrm{diff}}$). Those related to the form of energy functions $A_a()$ are unchanged. Let us note that δ is the main parameter; however, it is not difficult to set it for a given applica-

tion since it indicates the boundary between residual motion (after compensation) that will be considered as not significant, and the motion of independent moving objects.

4. RESULTS

Figures 2a-b-c present three images of a real sequence. Here, the camera is mounted on the left side of a car approaching a roundabout; the dominant motion in the image sequence is due to its movement. It is conveyed by the background motion (mainly houses areas). Hence, regions corresponding to moving objects in the scene (here, the car) and to - even static - entities in the near foreground, due to significant difference in depth, (here, the marks on the road and the sign) are expected to be detected as "mobile" (i.e., regions not conforming with the dominant motion). These two classes of objects can be further discriminated, but this is beyond the scope of this paper. Figures 3a-b-c contain the corresponding detection label fields $\widehat{d}$ ("static" regions are in black, and the original intensity information has been kept inside the regions labeled as "mobile"). Let us note that "mobile" regions are quite correctly segmented, and that there is no spurious detection within the "static" parts. On the other hand, Fig. 4 shows that simply thresholding the observations gives a very noisy detection map (false alarm, and incomplete masks of moving regions). Fig. 5 and 6 report the detection map obtained at time t_{62} using only observations between two frames ($K = 1$) and without considering the energy term U_2. Besides, Fig. 5 results from the multiscale minimisation procedure, while Fig. 6 is obtained using a single scale scheme; the multiscale algorithm performs better. It can also be noted that although the detection between two frames gives good results, the use of past observations helps in recovering the complete mask of the car and of the sign (compare Fig. 3b and Fig. 5).
Fig. 7a-b-c show detected moving regions (with a white envelope) obtained in the "plane" sequence. In this sequence, the camera is moving to keep three approaching planes in its field of view. The relative motions of these planes with respect to the background range from 0.4 to 0.8 pixels per frame (on an average). However, as the projections of the planes principally slide along themselves in the image, they produce little interframe intensity difference, making them quite difficult to detect. At the very beginning of the sequence, the algorithm gives spurious detections (Fig. 7a), but the temporal integration of the observations improves the results very rapidly (Fig. 7b-c). Fig. 8a displays a detailed view of the plane in the middle of Fig. 7c, and Fig 8b shows the obtained result.

All these results (as well as others not reported here) validate the method proposed in this paper.

5. REFERENCES

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

[2] 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.

[3] G.W. Donohoe, D.R. Hush, and N. Ahmed. Change detection for target detection and classification in video sequences. *Proc. Int. Conf. Acoustics, Speech and Signal Processing, New-York*, 1084–1087, 1988.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.6, No.6:721–741, Nov. 1984.

[5] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of 2nd ECCV-92, S.Margherita Ligure, Italy*, pages 282–287, Springer-Verlag, May 1992.

[6] J.M. Létang, V. Rebuffel, and P. Bouthemy. Motion detection robust to perturbations: a statistical regularization and temporal integration framework. In *Proc. 4th Int. Conf. Computer Vision, Berlin*, pages 21–30, May 1993.

[7] R.C. Nelson. Qualitative detection of motion by a moving observer. *Int. Journal of Computer Vision*, Vol.7, No.1:33–46, 1991.

[8] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models in complex image sequences. In *Proc. of* 7^{th} *Eusipco, Edinburgh, Scotland*, September 1994.

[9] 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.

[10] W.B. Thompson and T.-G. Pong. Detecting moving objects. *Int. Journal of Computer Vision*, Vol.4:39–57, 1990.

[11] P.H.S. Torr and D.W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 11(4):180–187, May 1993.

Fig. 2a, 2b and 2c : three images of the "roundabout" sequence at time t_{57}, t_{62} and t_{67}.
Fig. 3a, 3b and 3c : results with parameter values ($\alpha = 66$): $\delta = 1.0$, $\beta_{\text{diff}} = 40$, $\beta_m = 4$, $\tilde{\beta}_{\text{diff}} = 27$, $G = 0.5$, $m = 900$, $\gamma = 0.4$.
Fig. 4 : observation field o^{62} thresholded with $\lambda = 1.0$. Fig. 5 : results at time t_{62} using only two frames ($\tilde{\beta}_{\text{diff}} = 0$ and $\gamma = 0.0$; otherwise same parameters as in Fig. 3) and the multiscale minimization sheme (5 levels). Fig. 6 : same as in Fig. 5, but with a single scale minimization scheme.

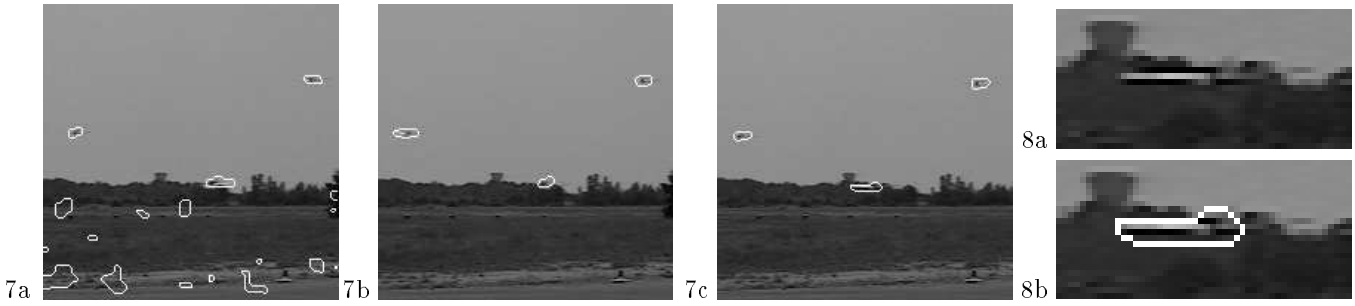

Fig. 7a-b-c: three images of the "plane" sequence (by courtesy of CEA-LETI, Grenoble) at time t_1, t_5 and t_{22} moving regions are shown with a white envelope. Parameters value are ($\alpha = 66$): $\delta = 0.3$, $\beta_{\text{diff}} = 25$, $\beta_m = 6$, $\tilde{\beta}_{\text{diff}} = 27$, $G = 0.5$, $m = 144$, $\gamma = 0.7$.
Fig. 8a,8b: detailed views of the plane in the middle of the image at time t_{22}.