Appeared as Chapter (Chapter 8, pp.283-311) in :
"Video Data Compression for Multimedia Computing",
H. H. Li, S. Sun, and H. Derin (eds.), Kluwer Academic Publisher, 1997

Separation of moving regions from background in an image sequence acquired with a mobile camera

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:

We present a statistical method to detect regions whose apparent motion in the image is not conforming to the dominant motion of the background resulting from the camera movement. Alternatively, the same scheme can be used to track a particular region of interest of the scene. The apparent motion induced by the camera motion is represented by a 2D parametric motion model, and compensated for using the values of the motion model parameters estimated by a multiresolution robust statistical technique. Then, regions whose motion cannot be described by this global model estimated over the entire image, are extracted. The detection of these non conforming regions is achieved through a statistical regularization approach based on multiscale Markov random field (MRF) models. We have paid a particular attention to the definition of the energy function involved and to the observations taken into account. To gain robustness, information is integrated over time. This method has been validated by experiments carried out on many real image sequences.

1 Introduction

The design of efficient image compression techniques is essential to the developement of the future video sequence transmission networks, especially when very low bit rate compression is required [17]. Image sequence coding using motion compensation is a common tool to extract and subsequently remove the huge temporal correlation (and thus redundancy) that exists in image sequences. For instance, in the well known block-matching algorithms, a constant displacement model is estimated from frame to frame within each block of an image partitionning. However, despite many improvements, like for instance the use of an adaptative tiling, or the use of hierarchical motion models [20], the visual quality of the reconstructed images provided by these algorithms rapidly decreases at the receiver whenever the available bit rate becomes very low or the motion becomes complex. Nevertheless, in numerous applications like in remote surveillance video systems, all parts of the processed images are not of equal importance. Thus, a spatially constant reconstruction quality level is not required. This is due either to the specificity of the task definition, or to psychovisual considerations [10].

In this contex, motion segmentation can play two important roles. First, it can be used to analyse the dynamic content of the scene and to extract the different regions visually important for the task at hand. The choice of motion as a cue for image segmentation allows us to restrain the number of extracted regions, as opposed to an approach based on spatial image segmentation. Secondly, motion segmentation naturally leads to motion compensation, since it usually involves the estimation of motion models, and is therefore appropriate for coding purposes. The block effect observed with the more traditional schemes is much reduced with such a region-based approach.

In this paper, we will concentrate on the special case of binary segmentation, that is, the partitionning of the image into regions whose motion is, or is not, well compensated for using a globally estimated motion model. When the estimated motion represents the background apparent motion, it corresponds to the detection and location of independently moving objects when the camera is moving also. This is a basic task that can be useful as an initial step in many applications, like in [24]. Alternatively, when the estimated motion model represents the motion of a given region of interest (ROI), our algorithm behaves like a tracking method. In both cases, it achieves the goal of focusing on region(s) of interest of the scene, where most of the allowed bit rate should be allocated, [19]. Indeed, either the background or the ROI can be registered using the estimated motion model, which saves bit rate.

In the case of a static camera, different efficient solutions for the detection of moving objects have been developed ([1, 5, 6, 16]). For instance, in [1, 6], statistical tests, as well as a regularization step based on Markov Random Field models, are used. However, in all the works cited above, the solution is devoted to the rather simple situation of a static camera. It may even be reduced to intensity temporal change detection, like in [1]. Therefore, they cannot be used

when we are dealing with a truly mobile camera. In the latter case, every pixel may undergo apparent motion, and the resulting apparent motion field -the optical flow- can be complex. Then, detection of moving objects requires that we are able to make the distinction between apparent motion that is only due to the movement of the camera, and apparent motion that arises from the relative movement of an independently moving object. Among the studies that have concentrated on the detection of moving objects in the scene with a moving observer, we can find two broad classes of solutions.

The first one uses (*a priori* or sometimes estimated) information on the 3D camera motion, in order to derive constraints on the 2D image flow field of the projection of static objects, induced by the movement of the camera [18, 25, 26]. Regions where these constraints are violated are then identified as projections of moving objects. In [18], Nelson demonstrates a direct qualitative method running in real time and based on the motion epipolar constraint. Thompson and Pong [26] use a wider variety of principles, that can be applied in the case of a general motion, but leave open how they can be used in practice. In [25], a robust estimator is used to detect deviation from rigid motion associated to the camera movement. One drawback of these methods is that they are usually efficient when the 2D image displacements due to camera and motions of objects are important, which is for instance not the case around the focus of expansion, or when camera motion is small. Some of the techniques described in [26] work only if the camera is moving, requiring an additional first step for the determination of camera state. Besides, most of these methods do not deliver a partition (segmentation) of the images, but a pixel-based sparse decision map. Sparse motion estimates (based on token) used in [26, 25] are not appropriate for motion compensation too. These methods, usually requiring 3D motion parameter estimation that are not straightforwardly available, are therefore not quite adapted in most of the coding applications of interest.

In contrast, the second class of methods can be applied to the coding application we consider. In these methods, one of the following assumptions usually holds. Either the camera is only rotating, or the depth variation in the scene is small compared to the distance between camera and objects, or the visible surfaces of the static world are approximately located in the same 3D plane. In these cases, we can assume that the 2D apparent motion (due to camera motion) of the static background can be modeled by a 2D parametric motion model and can be considered as the *dominant motion*. Such a motion model is estimated from frame to frame, and then used in a warping procedure to compute a compensated sequence in which the background is supposed to appear as static. Thus, non-static regions in this sequence can be considered as moving objects. If the assumptions are not fulfilled, they may also comprise objects located at a significantly different depth compared to the background. Projections of moving objects can then be obtained by simply thresholding some local error or statistical function at each pixel [27], which usually leads to noisy detection maps for most real sequences. In [13], the

average of a few successive images registered (or compensated for) using the computed dominant motion is used as a reference map. However, this "integration step" assumes that the regions, whose motion corresponds to the computed dominant motion, are *perfectly* registered over the integration duration. If not, this temporal integration blurs the reference map. Furthermore, the subsequent motion estimations which use the reference map as first image, can be greatly affected. Interestingly, however, this method can be applied for the tracking of the objects themselves. Once an object is detected, the warping is applied using the estimated object motion model. This time, the static part in the compensated sequence corresponds to the object projection whereas the moving areas reveal the background as well as other moving objects.

In this paper, we propose a motion detection algorithm belonging to the second class. It can tolerate very noisy data as well as imprecise registration (since the 2D motion model used is only an approximation of the dominant motion). This algorithm uses interframe observations rather than observations related to a reference frame, since the maintenance of such a frame is usually a difficult task. To avoid false alarms, due for instance to acquisition noise, or to local unexpected intensity variations, and to increase the detection rate in low-contrasted regions, the algorithm relies on a spatio-temporal statistical regularization approach based on multiscale MRF models.

Section 2 introduces the 2D motion models which are considered, and briefly describes the robust multiresolution method used for the estimation of these models. In Section 3, forming the main part of this paper, we describe the original method developed for the detection -or tracking- of the moving objects in the image sequence. Section 4 deals with some important computational issues. Several results that validate our approach are reported in Section 5, and Section 6 contains concluding remarks.

2 Motion model and motion estimation

To detect moving objects between two images, we have first to estimate a motion model that describes the image motion of the background, i.e. the apparent motion due to camera movement. Since at the beginning of the sequence we have no information about the location of the moving objects, the estimation process will involve data taken over the whole image, and then must be robust to the presence of moving objects. Moreover, in order to consider the *dominant* motion model resulting from this estimation step as being the 2D motion due to camera movement, we must assume that the projections of the static components of the scene occupy the main part of the image (with the additional condition that they supply sufficient image spatial gradient information).

2.1 The parametric motion model

Usually, the choice of a motion model depends on considerations related to the kind of 3D motion undergone by the camera and objects, the type of transformation sufficient to account for the projection of the scene into the image plane, and the analytic description of the viewed surfaces. For instance, if the depth variation over the background is small with respect to the distance to the camera, then a planar surface constitutes a good approximation.
A first approach would be to use a full 3D model (3D motion and depth) in the analysis process. However, this leads to solve the general -but highly complex- problem of 3D reconstruction in the presence of moving objects. Let us note that such an attempt is described in [2], involving a global 2D motion model and additional local depth parameters without any prior camera calibration.

We prefer to use 2D parametric motion models $\vec{w}_\Theta$ to represent the projection of the 3D motion field of the static background, where $\vec{w}_\Theta$ denotes the modeled velocity vector field and Θ the set of model parameters. Such models are globally valid when either the camera translation magnitude is small with respect to the depth of the objects, or when there is not too much depth variation in the scene. Though less general than the full 3D case, the choice of 2D models leads to an efficient motion computation. In all the experiments we have carried out so far, the 2D affine motion model proved to be a good compromise between its relevance as a motion descriptor and the efficiency of its estimation. It is defined at pixel $\mathrm{p} = (x, y)$ by:

$$\vec{w}_\Theta(\mathrm{p}) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \tag{1}$$

where $\Theta = (a_i), i = 1..6$, is the parameter vector to be estimated. However, when the background is approximately located in a plane whose slant is important, a particular quadratic motion model should be used [14, 15].

2.2 Motion estimation

To estimate the dominant motion model between two successive frames I_t and I_{t+1}, we have developed a gradient-based multiresolution robust estimation method described in [22]. To ensure the goal of robustness, we minimize an M-estimator criterion with a hard-redescending function [12]. The constraint is given by the usual assumption of brightness constancy of a projected surface element over its 2D trajectory [11]. As in the considered experiments, the displacements between two frames can be very large, we use a discrete formulation of this constraint. Thus, the estimated parameter vector is defined as:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\ E(\Theta) = \underset{\Theta}{\operatorname{argmin}} \sum_{\mathrm{p} \in R(t)} \rho\left(\mathrm{DFD}_\Theta(\mathrm{p})\right) \tag{2}$$

$$with \quad \mathrm{DFD}_\Theta(\mathrm{p}) = I_{t+1}(\mathrm{p} + \vec{w}_\Theta(\mathrm{p})) - I_t(\mathrm{p})\,. \tag{3}$$

$\rho(x)$ is a function which is bounded for high values of x (more precisely, we use Tukey's biweight function). If we want to detect moving objects, the estimation support $R(t)$ consists of the whole image. However, in the case when we are tracking a region of interest, the support $R(t)$ is formed by the area corresponding to the prediction of the ROI location at time t, given the ROI location at time $t-1$ and the estimated motion model at time $t-1$. The minimization takes advantage of a multiresolution framework and an incremental scheme based on the Gauss-Newton method.

More precisely, at each incremental step k (at a given resolution level, or from a resolution level to a finer one), we have: $\Theta = \widehat{\Theta}_k + \Delta\Theta_k$. Then, a linearization of $\mathrm{DFD}_\Theta(\mathrm{p})$ around $\widehat{\Theta}_k$ is performed, leading to a residual quantity $r_{\Delta\Theta_k}(\mathrm{p})$ linear with respect to $\Delta\Theta_k$:

$$r_{\Delta\Theta_k}(\mathrm{p}) = \vec{\nabla} I_t(\mathrm{p} + \vec{w}_{\widehat{\Theta}_k}(\mathrm{p})).\vec{w}_{\Delta\Theta_k}(\mathrm{p}) + I_{t+1}(\mathrm{p} + \vec{w}_{\widehat{\Theta}_k}(\mathrm{p})) - I_t(\mathrm{p}) \tag{4}$$

where $\vec{\nabla} I_t(\mathrm{p})$ denotes the spatial gradient of the intensity function at location p and at time t. Finally, we substitute for the minimization of $E(\Theta_k)$ the minimization of an approximate expression E_a, which is given by $E_a(\Delta\Theta_k) = \sum \rho(r_{\Delta\Theta_k}(\mathrm{p}))$. This error function is minimized using an Iterative-Reweighted-Least-Squares procedure, with 0 as an initial value for $\Delta\Theta_k$. For more details about the method and its performances, the reader is referred to [22].

This estimation algorithm allows us to get a robust and accurate estimation of the dominant motion model (i.e., background apparent motion) between two images, which is of key interest for the subsequent steps. Nevertheless, as it will be shown in the next section, due to the fact that the motion model is only an approximation of the true motion, the outliers areas issued from this robust estimation process using a simple thresholding step, cannot correctly account for areas corresponding to moving objects. Too many false alarms and missing detections are obtained.

3 Motion detection algorithm

3.1 Outline of the approach

Once we have estimated the dominant motion model $\widehat{\Theta}_t$ between images at t and $t+1$, the problem that arises can be stated as follows: find the set of all image points p whose true 2D motion vector $\vec{w}_{\mathrm{true}}(\mathrm{p})$ does not conform to the modeled flow vector $\vec{w}_{\widehat{\Theta}_t}(\mathrm{p})$. As explained in the introductory section, this set may include more elements than the projections of the moving objects depending on the kind of motion model used with respect to the scene content and to the camera movement. Therefore, instead of calling these points "mobile", which is a scene-

related term, we will call them "non-conforming" points or "not compensated" point, which is more appropriate to our 2D image motion representation.

The question is now how to define a non-conforming point. One simple but meaningful method consists in performing some thresholding on the motion estimation error:

$$\left\{ \begin{array}{c} \text{if } \|\vec{w}_{\text{err}}(\text{p})\| > \delta \text{ then p is stated as a non-conforming point} \\ \text{with } \vec{w}_{\text{err}}(\text{p}) = \vec{w}_{\text{true}}(\text{p}) - \vec{w}_{\widehat{\Theta}_t}(\text{p}) \end{array} \right. \tag{5}$$

However, we would like to avoid the estimation of the 2D dense motion field, which is a complex and error-prone problem. An attractive alternative way is to rely only on locally measured quantities that supply valuable information accounting for the motion compensation accuracy obtained using the estimated motion model. However, as such local measurements can be too noisy or insufficient to reach a correct decision, leading to false alarms or reducing the detection rate, we state the detection problem as a labeling problem, and perform a statistical regularization. Moreover, thanks to this regularization scheme, information on motion compensation errors at reliable points (e.g., corners) will be propagated to points where ambiguities might exist (e.g., straight edge line) or points with no information (uniform areas), as will be explained below.

In the next subsection, we define the local measurements or observations we use. We describe the statistical regularisation framework and the related energy function designed for performing motion detection. Two versions will be presented: in subsection 3.3 a two-frame method and in subsection 3.4, a method exploiting several successive images.

3.2 Choice of the local measurements

One simple usual way to evaluate the adequacy of the estimated dominant motion model consists in warping one image to the other using this model. More precisely, let us define $\tilde{I}_t$ and $\tilde{I}_{t+1}$ as follows:

$$\tilde{I}_t(\text{p}) = I_t(\text{p}) \text{ and } \tilde{I}_{t+1}(\text{p}) = I_{t+1}(\text{p} + \vec{w}_{\widehat{\Theta}_t}(\text{p}))$$

Values of $I_{t+1}(\text{p}+\vec{w}_{\widehat{\Theta}_t}(\text{p}))$ are obtained through a bilinear interpolation since $\text{p}+\vec{w}_{\widehat{\Theta}_t}(\text{p})$ usually does not fall on the image grid. Thus, the motion field between images $\tilde{I}_t$ and $\tilde{I}_{t+1}$ is exactly $\vec{w}_{\text{err}}$. If we assimilate the temporal derivative $\frac{\partial \tilde{I}}{\partial t}$ with the finite difference $\tilde{I}_{t+1} - \tilde{I}_t$, the well-known brightness constraint equation [11] gives us:

$$FD_t(\text{p}) = \tilde{I}_{t+1}(\text{p}) - \tilde{I}_t(\text{p}) \simeq -\vec{\nabla}\tilde{I}_t(\text{p}).\vec{w}_{\text{err}}(\text{p}) \tag{6}$$

Let us point out why the frame difference FD_t is in fact not an appropriate local measurement for motion detection. An adequate measure should be defined in such a way as being high in non-conforming regions and low in well compensated ones, with a rather continuous behaviour in-between. It is easy to realize, when considering relation (6), that using the frame difference:

1. the response is low in uniform intensity areas, whatever the type of region it is;

2. the measure is generally large[1] along highly contrasted edges, whenever there exists even a small compensation error;

This indicates that the response of this measure is mainly driven by the spatial intensity gradient rather than by the presence of small or large residual motion.

It is therefore necessary to consider another local measurement more directly related to the underlying motion compensation error. In that vein, let us consider the normal flow of the residual error displacement at a given point p:

$$v_n(\mathrm{p}) = \frac{FD_t(\mathrm{p})}{\|\vec{\nabla}\tilde{I}_t(\mathrm{p})\|} \tag{7}$$

Two main shortcomings remain with this local measure. First, it is still unreliable in uniform regions. Secondly, if the residual displacement $\vec{w}_{\mathrm{err}}(\mathrm{p})$ is perpendicular to the spatial intensity gradient direction, then, this measure is equal to zero (aperture problem), even though there is a motion. However, these two difficulties can be alleviated as explained hereafter.

In [4], it is shown through the analysis of the results of different kinds of optic flow estimation algorithms, that $\|\vec{\nabla}\tilde{I}(\mathrm{p})\|$ is indeed a proper measure of the reliability of the estimation of the normal flow v_n. Thus, instead of performing a simple local averaging of the normal flow, we use the following weighted averaging, which is also proposed in [14]:

$$\begin{aligned} \mathrm{Mes}_{\widehat{\Theta}_t}(\mathrm{p}) &= \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\left(\|\vec{\nabla}\tilde{I}(\mathrm{q})\|^2 \times v_n(\mathrm{q})\right)}{Max(\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\|\vec{\nabla}\tilde{I}(\mathrm{q})\|^2, n\times G_m^2)} \\ &= \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\left(\|\vec{\nabla}\tilde{I}(\mathrm{q})\| \times |FD_t(\mathrm{q})|\right)}{Max(\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\|\vec{\nabla}\tilde{I}(\mathrm{q})\|^2, n\times G_m^2)} \end{aligned} \tag{8}$$

where $\mathcal{F}(\mathrm{p})$ is a small neighborhood (typically 3×3) around p which contains n points, and G_m is a constant which accounts for noise in the uniform areas. An interesting property of this local measure is the following. Let us suppose that the pixel p and its neighborhood undergoes the same displacement of magnitude δ and direction $\vec{u}$. By studying how the measure varies with respect to the direction $\vec{u}$, we can derive two bounds $l(\mathrm{p})$ and $L(\mathrm{p})$ such that, whatever the direction $\vec{u}$ might be, the following inequality holds:

$$0 \leq l(\mathrm{p}) \leq \mathrm{Mes}_{\widehat{\Theta}_t}(\mathrm{p}) \leq L(\mathrm{p}) \tag{9}$$

[1] Another problem arises around highly contrasted edges, though less important. Even if the residual motion $\vec{w}_{\mathrm{err}}(\mathrm{p})$ is null, interpolation errors might give rise to a strong FD_t value, wrongly indicating the presence of a non-conforming point.

In [15], we have determined two pairs of bounds. Assuming that the brightness constraint equation (6) is valid[2], the first one is given by:

$$\begin{cases} L(\mathrm{p}) = \delta \\ l(\mathrm{p}) = \eta \times \delta \times \lambda'_{min} \end{cases} \quad \text{with } \eta = \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \|\vec{\nabla}\tilde{I}(\mathrm{q})\|^2}{Max(n \times G_m^2, \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \|\vec{\nabla}\tilde{I}(\mathrm{q})\|^2)} \text{ and } \lambda'_{min} = \frac{\lambda_{min}}{\lambda_{max} + \lambda_{min}} \tag{10}$$

where λ_{min} and λ_{max} are respectively the smallest and highest eigenvalues of the following matrix (with $\nabla\tilde{I}(\mathrm{q}) = (\tilde{I}_x(\mathrm{q}), \tilde{I}_y(\mathrm{q}))$:

$$M = \begin{pmatrix} \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \tilde{I}_x(\mathrm{q})^2 & \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \tilde{I}_x(\mathrm{q})\tilde{I}_y(\mathrm{q}) \\ \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \tilde{I}_x(\mathrm{q})\tilde{I}_y(\mathrm{q}) & \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \tilde{I}_y(\mathrm{q})^2 \end{pmatrix} \tag{11}$$

To obtain tighter bounds (especially for the upper bound), we have modeled the local iso-intensity contour passing through the pixel p as the union of two segments joining at p[3]. Using this model, we have been able to obtain a second set of bounds that are actually reached for specific values of the displacement direction $\vec{u}$:

$$\begin{cases} l(\mathrm{p}) = \eta\delta \sqrt{\lambda'_{min}(1 - \lambda'_{min})} \\ L(\mathrm{p}) = \delta \sqrt{1 - \lambda'_{min}} \end{cases} \tag{12}$$

In [15], experiments have been carried out with simulated displacements between two images. They show that the inequality (9) is indeed well verified using the second set of bounds, and rather immune to a large amount of noise by setting G_m to a sufficiently high value. We will use this set of bounds in the experiment.

Let us note that l and L depend on the local distribution of the directions of the local spatial intensity gradients, which reflects the local intensity structure in the image and allow us to take into account the aperture problem. If a linear iso-intensity contour is sliding along itself, the measure (8) will be nearly zero, though there is really motion. However, in that case, the bound l is also equal to zero. We can conclude that a low measurement value will indicate a conforming site with no doubt only if it is lower than the bound l.

If we want to detect residual motion $\vec{w}_{\mathrm{err}}$ of magnitude greater than a preset value δ, the local measurements given by (8) can be classified in three classes (see Fig 1). First, an observation lower than $l(\mathrm{p})$ proves with certitude that the underlying residual displacement magnitude $\|\vec{w}_{\mathrm{err}}\|$ is lower than δ. Second, when the local observation is greater than L,

[2]In order relation (6) to be valid for the given displacement δ, the image intensity must be sufficiently smooth. If we use a Gaussian filter of variance σ^2 for instance, σ must depend on the actual (preset) value δ.

[3]Note that, because of the definition of the measure in terms of normal flow, what matters is only the local distribution of the directions of the local spatial intensity gradient. Our model assumes that this distribution has locally two main modes (which can be identical).

we can infer that the residual displacement magnitude is greater than δ. Finally, when the observation value falls between the two bounds, it is not possible to conclude. In such a case, we would like to use some spatial and temporal context to correctly classify the corresponding pixel. This can be done using the statistical regularization scheme that we now describe.

3.3 Two-frame statistical detection scheme

Let S denote the set of sites s (here, pixels p) and $\mathcal{C}$ the set of cliques of two elements associated to a second-order neighbourhood system ν. In this section, we shall write any quantity related to site s with a subscript s. We formulate the motion detection issue as the estimation of a binary label field (also called detection map) $d = \{d_s, s \in S\}$ which is the most likely to have produced the field of observations at time t, $o^t = \{o_s^t, s \in S\}$. The two considered label values are defined as "conforming" and "non-conforming". The observations are composed of the local measurements defined previously: $o_s^t = \mathrm{Mes}_{\widehat{\Theta}_t}(s)$. To solve this problem, we adopt the Maximum A Posteriori (MAP) criterion, i.e., we maximize the *a-posteriori* distribution of the labels given the observations. If we use Markov Random Field (MRF) to model the sets of observed and hidden variables, and due to the equivalence between MRF and Gibbs distribution ($p(x) = \frac{1}{Z}e^{-U(x)}$), [9], this is equivalent to minimizing an energy function $U(d, o)$, which in turn can be written as the sum of the so-called local potential functions. We consider an energy function of the form:

$$U(d, o) = U_1(d) + U_2(d, o^t) \tag{13}$$

• U_1 is the regularization term which accounts for the expected spatial properties (homogeneity) of the label field. It is defined as the sum of local potentials:

$$U_1(d) = \sum_{\{s,u\} \in \mathcal{C}} V_1(d_s, d_u) \quad \text{with} \tag{14}$$

$$V_1(d_s, d_u) = \begin{cases} -\beta_m & \text{if } d_u = d_s = \text{"non-conforming"} \\ 0 & \text{if } d_u = d_s = \text{"conforming"} \\ \beta_d & \text{if } d_u \neq d_s \end{cases} \tag{15}$$

β_d is the cost to pay to get neighbours with different labels. β_m ($0 < \beta_m \ll \beta_d$) is a potential value which favors the spatial grouping of "non-conforming" labeled points. We have introduced the term β_m in order to counter-balance the fact that in uniform areas, the observations (and the second energy term U_2) globally tend to favor the "conforming" label, as will be seen below. Therefore, β_m will help in "filling" with the right label the -usually almost uniform- inside of the non compensated regions.

• $U_2(d, o^t)$ is the data-driven energy term expressing the adequacy between observations and

labels. We have chosen to construct this energy term as follows:

$$U_2(d, o^t) = \sum_{s \in S} V_2(o_s^t, d_s) \tag{16}$$

This means that we make the potential V_2 at a given site s only depend on the value of the label d_s and of the observation o_s^t at this site. However, since o_s^t is computed using information defined on the window $\mathcal{F}$, it should also depend on the labels in this window. Also, observations at neighbouring sites are likely to be slightly correlated. Nevertheless, as such a dependence and correlation is hard to model in practice, we have retained the above formula for simplicity.

Since our goal is to detect points undergoing a residual motion greater than a predefined value δ, the role of the potential V_2 is to convert in some way the inequalities (9) in an energy-based formulation. This can be achieved by defining V_2 as follows:

$$V_2(d_s, o_s^t) = \begin{cases} \alpha_c \times F_s^t \times A_{l_s,k_c}(o_s^t) & \text{if } \ d_s = \text{“conforming”} \\ \\ \alpha_{nc} \times F_s^t \times (1 - A_{L_s,k_{nc}}(o_s^t)) & \text{if } \ d_s = \text{“non-conforming”} \end{cases} \tag{17}$$

where:

- $A_{tr,k}(x)$ is a smoother version of a step edge, i.e., an increasing function from 0 to 1 such that the transition occurs at tr (we have $A_{tr,k}(tr) = 0.5$). The smoothness of the transition is controlled by k (we have $\frac{\mathrm{d}A_{tr,k}}{\mathrm{d}x}(tr) = k$). We have chosen the inverse tangent function:
$$A_{tr,k}(x) = \frac{1}{\pi} \arctan(k\pi(x - tr)) + 0.5 \quad . \tag{18}$$
This function is preferable to a sigmoide, because it reaches the saturation levels (0 or 1) less rapidly. Let us point out that, since the potential energy function $V_2$is bounded, it behaves similarly to a “robust estimator”. It avoids a strong erroneous observation to have a sufficient influence to locally impose the wrong label even if all the neighbors disagree.

- α_c and α_{nc} are the maximal values that the potentials can take ($\alpha_c \simeq \alpha_{nc}$); k_c and k_{nc} regulate the transition around the bounds.

- $F_s^t = F(\|\vec{\nabla}\tilde{I}_t(s)\|) = \max\left(A_{G,1}(\|\vec{\nabla}\tilde{I}_t(s)\|), \mathrm{At}_{max}\right)$ is a damping factor. As already mentioned, a site with low image spatial gradient usually carries poor and unreliable information about the presence of motion. By diminishing the values of the potential V_2 for the observations coming from uniform areas, we *conversely* increase the relative contribution of the regularization term. For instance, if F_s^t were 0, V_2 would be zero whatever the label is. Thus, the decision would be based only on the local context. To avoid such an extreme case, the parameter At_{max} fixes the maximal allowed damping factor. The parameter G controls the separation between pixels supposed to carry information (i.e. with $\|\vec{\nabla}I_t(s)\| > G$), and those which do not.

Figure 2 illustrates the form of the potential V_2, and highlights the role of the bounds l_s and L_s in two specific situations. In these plots, we can recognize the curve representing the potential function V_2 associated with the non-conforming label, because it exhibits low potential values for high observations, and conversely. More precisely, as long as the observation is greater than the upper bound L_s, the potential value is low since we are rather confident that the residual velocity is larger than δ. However, as o_s^t becomes lower than L_s, the potential function increases, indicating that the label might not be appropriate. For the curve representing the potential function V_2 associated with the conforming label, the behaviour is symmetrical and involves the lower bound l_s.

Fig. 2.a displays the potential function V_2 for a site where the local distribution of the directions of the spatial intensity gradients exhibits two strong modes corresponding to perpendicular directions (for instance, if the site lies at a corner). In Fig. 2.b, there is only one single mode (typically, the site s lies on a straight edge). In those plots, the uncertainty interval of Fig. 1 qualitatively corresponds to the interval of observations values for which the gap between the potential values associated with the non-conforming label and the conforming one is small, meaning that both labels could be convenient (or indifferent !) to the given observation. Thus, for observations falling in this interval, the regularization term, which brings contextual information, will dominate and will strongly influence the choice of the label at that site. As desired, this interval is bigger in the case of a linear structure since, as previously pointed out, low observations do not then necessarily characterize the absence of motion.

3.4 Time-extended detection scheme

It is well known that the information that one can extract from two successive images can be ambiguous, in the sense that they may fit several different interpretations of the scene structure and of the 3D motion. Those ambiguities can usually be removed by accumulating information over time. This aspect can be incorporated in our scheme by merely adding supplementary terms to the energy function defined in (13).

The role of the temporal aspect is twofold:

- first, it should ensure the coherence of the detection maps at successive instants;
- secondly, it should be used to filter the extracted motion information, thus reducing decision errors due to noisy measurements (acquisition noise, errors in computation of the intensity gradients and of intensity interpolations, illumination variations,...).

3.4.1 Using the past detection map

As the motion of objects in the scene is smooth, the projections of those objects in the image plane at successive instants usually overlap[4] each other. Therefore, it would be useful to exploit this correlation, and use the previous detection maps in the labeling process at time t. A first straightforward procedure is the following. For obvious computational reasons, the iterative relaxation algorithm used to solve the minimization problem at hand is a deterministic one. Then, it converges to a local minimum depending on the initialization. Since the detection maps are supposed to vary smoothly, the map obtained at the preceding instant should be relatively close to the optimal solution at time t, and can be used as the initial map in the relaxation algorithm. However, more elaborate use of temporal integration can be considered. In a second approach, the detection maps are considered as a temporal process. For instance, recursive filtering like Kalman algorithm [16] could be used. Alternatively, a subset of successive detection maps could be considered as a whole as a spatio-temporal Markov random field. However, in that case, the state space becomes huge. The minimiZation of the global energy defined in such a space would be computationally expensive, even with a deterministic algorithm, and can only be processed in a batch mode. We have chosen a simpler version of this scheme. Detection maps are considered as a first order Markov chain, and the minimization is performed over the current map only. This means that only the detection map at the preceding instant is involved. In practice, it is merely considered as additional observations for the labeling process at time t.

We therefore add a third energy term U_3 to the energy function U given in (13). It plays a conservative role, and is defined as follows:

$$U_3(d^t, \tilde{d}^{t-1}) = \sum_{s \in S} V_3(d_s^t, \tilde{d}_s^{t-1}) \quad \text{with} \tag{19}$$

$$V_3(d_s^t, \tilde{d}_s^{t-1}) = \begin{cases} 0 & \text{if } \tilde{d}_s^{t-1} = d_s^t \\ +\beta_{dt} & \text{if } \tilde{d}_s^{t-1} \neq d_s^t \end{cases} \tag{20}$$

where, to account for the estimated motion between frame I_{t-1} and I_t, $\tilde{d}^{t-1}$ is a transformed version of d^{t-1} given by:

$$\tilde{d}^{t-1} = \mathrm{reg}_{t-1}^{t}(d^{t-1}) \tag{21}$$

where $\mathrm{reg}_i^j(X)$ consists in transforming the map X at time i into a map at time j by combining the motion models estimated from frame to frame between i and j.

[4] Assuming that the image temporal sampling is small enough with respect to the image motion and object size.

3.4.2 Filtering motion information over time

Increasing too much the relative contribution of the term U_3 in the global energy U in order to impose a stronger temporal constraint may result in undesirable behaviours. For instance, in large uniform areas, a "chain reaction" can occur, that is, a given label is imposed and transmitted from frame to frame despite the incoming of contradictory observations. One way to overcome this problem consists in combining the benefits of a "reasonable" temporal regularization term U_3 with the use of past local motion measurements, the latter allowing us to introduce a temporal smoothness in a more flexible manner. The direct filtering of the observations is not a good approach in our case, since we have no model for the time evolution of this observation. Moreover, the analysis of subsection (3.2) resulting in the inequalities (9) would not be valid anymore. We prefer to consider independently each observation with its associated bounds. This leads to the following refinement and improvement of the energy term U_2.

Let us denote o^{t-q} the observation field between image $t-q$ and $t-q+1$, and $\tilde{o}^{t-q} = \mathrm{reg}_{t-q}^{t}(o^{t-q})$ the motion-oriented projected version using the motion models estimated within the interval $[t-q, t]$. Assuming that these temporal observations are independent, we can derive the expression of the energy term U_2' that we actually use instead of U_2 :

$$U_2'(d^t, \tilde{o}^{t-q}, q \in \{0, \ldots, T\}) = \frac{1-\gamma}{1-\gamma^{T+1}} \sum_{q=0}^{T} \gamma^q \, U_2(d^t, \tilde{o}^{t-q}) \tag{22}$$

where $\gamma \in [0, 1[$ is a damping factor, which expresses that, further in the past an observation is, the less relevant to the determination of the detection map at time t it is. If $\gamma = 0$, only the current observation field is taken into account ; if γ is close to 1, the T past observation fields are considered almost equivalently. From a computational point of view, it is important to note that the determination of U_2' can be obtained in a recursive manner. In fact, only one registration step (from $t-1$ to t) is performed, and it is applied not to the observation field, but to a smoothed potential map derived from the function V_2 at the different instants.

4 Computational issues

4.1 Minimization algorithm

The global minimization of the energy function is performed using the multiscale MRF modeling approach described in [23]. First, it allows us to derive in a consistent mathematical way the expression of the energy function, parameters included, at every scale given the one at the finest scale, the observations being only considered at the same original resolution. Then, it consists in starting the minimization 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 very initial detection map is obtained by maximizing the conditional likelihood, i.e. by locally minimizing the energy term U_2'. Then, the Highest Confidence First [8] minimization procedure is used to compute the solution at that scale.

Minimization is performed from scale to scale, using the projection onto scale $l-1$ of the detection map obtained at coarser scale l as an initial solution, until the finest scale is reached. At each scale l, the solution is constrained to be constant within blocks of size $2^l \times 2^l$. Thus, at the finest scale, i.e. $l = 0$, the "blocks" are of pixel size, meaning that there is no more constraint on the solution. As illustrated in the experimental results, this minimization scheme quite improves the results. Especially, it reinforces in some way the homogeneity constraint without having to overweight the corresponding energy term U_1. In [23], it is shown that the multiscale approach gives results very similar to the stochastic relaxation, but is far much faster. Indeed, the minimization itself is very fast: around $0.4s$ on a SPARC-10 workstation for a 256×256 image. In comparison, the whole algorithm has a computational cost of 6 to 7 seconds for the same images. Approximately 2 seconds are spent in the image and gradient pyramid building (for the motion estimation process), and 1.6 second in computation of the observations, the bound and the potential V_2. These processing could be greatly reduced with specialized hardware, since they are local and uniform.

4.2 Parameter setting

Parameter setting is an important issue in any computer vision technique. Results should not exhibit a high sensitivity to the choice of parameter values. For a given application, it is usually possible and preferable to realize a learning step to appropriately set the parameter values, either empirically, or stated as an estimation problem [7]. However, an *a priori* analysis of the influence of the parameters should allow us to derive adequate values, prior to any experimentation. We recall below the main properties of the parameters we have introduced.

1. $\underline{G_m}$: This parameter depends mainly on the noise level in the image sequence that can affect the computation of the observations in low intensity gradient areas. Thus, a too low value for this parameter increases the false-alarm rate in a noisy sequence, which may be very critical. In contrast, too high values for G_m only tend to diminish motion information in -unnoisy- regions of medium intensity gradient, which is not critical since high intensity contrast areas usually provide enough information. Values for G_m range from 3 for nearly noise-free images (e.g., corresponding to TV broadcast image quality) to 8 for images with a low signal-to-noise ratio.

2. $\underline{\alpha_c, \alpha_{nc}, k_c, k_{nc}, G, \mathrm{At}_{max}, T, \gamma}$: The parameters that model the shape of the energy function are defined once and kept unchanged ($\alpha_c = 200, \alpha_{nc} = 206, k_c = k_{nc} = 4$). The parameter

G concerns the definition of the damping factor F_s^t introduced to reduce the strong bias in favor of the conforming label encountered in large uniform areas. More precisely, for a pixel s located in such an area, we usually have $o_s^t \simeq l_s^t \simeq 0$. Thus, as seen from Fig. 2.b, the potential value U_2 associated with the label "conforming" or "compensated" is lower than the energy value associated with the label "non-conforming" or "not compensated" by an amount of $\Delta_0 = F_s^t \times \alpha_c/2$. Let us point out that, usually, the large uniform areas mainly belong to the static background (sky, road, walls, ...). If the label "compensated" is the one associated with this background (this is the case when we want to detect moving objects), then the bias will help the labeling process, and G can be set to a small value (0 to 1), depending on the image noise level. However, when we track a particular region of the image, the estimated motion represents the region motion, and the background is therefore associated with the label "non-compensated". Then, the bias wrongfully favors the label "compensated" in uniform areas. To alleviate this difficulty, we can set G to a higher value (2 for instance), which leads to a smaller value for Δ_0, and therefore a smaller bias, in the uniform areas.

At_{max} is determined in such a way that, in the absence of any spatial regularization, the labeling decision can always override (with an appropriate observation) the label selection induced by the energy term U_3 taken alone. A necessary and sufficient condition is: $At_{max} \geq \frac{2\beta_{dt}}{\alpha_c}$.

T and γ rule the temporal integration of the observations. In our recursive implementation of the energy computation, only the value of 2 (rarely used in practice) and ∞ can be given to T. γ (as well as β_{dt}) is related to the temporal change rate of the image content, tied to the temporal sampling and to objects size and motion. Increasing γ gives more weight to the past. A value of 0.5 for γ means that all the past observation is given the same weight than the current observations. Thus, a high value (above 0.5) of γ can be used only if the overlapping between regions of same labels at two successive instants is always expected to be high.

3. $\underline{\beta_m, \beta_d, \beta_{dt}, L}$: We usually take $\beta_{dt} = \beta_d$. Values of β_m and β_d parameterize the *a priori* homogeneity of the detection map, and depend on the expected minimum size of the moving objects. The multiscale minimization scheme reinforces this *a priori*. Thus, if objects to be detected are really small, the regularization term β_d should be low, and we use a small number of scales. The method of so-called qualitative boxes [3] has been used to delimit appropriate sets of values for β_m and β_d, [15]. More precisely, a set of meaningful and relevant local configurations of observations and labels are chosen at and around a given site. A qualitative behaviour is conveyed (off-line) to the Markov field by promoting the desirable label at this site for each considered configuration. We express for each configuration that the probability to select a given label rather than the other one

should be greater than a preset value. This is derived through the local evaluation of the energy function for the different considered situations. This results in a set of inequalities, linear with respect to the parameters, that restricts the set of admissible values for the regularization parameters to a "box".

Finally, let us note that δ is obviously the main parameter. In fact, it is not difficult to set it, since it defines the threshold under which residual motion (after compensation) will be considered as not significant. It is typically an application-driven parameter which can be set according to the needs and requirements. Its meaning is explicitely taken into account by our method, in the definition of U_2'. It allows us to adapt the algorithm to different situations, as demonstrated by the results.

5 Experimental results

The algorithm has been successfully tested on many different sequences. Here, we report two examples that illustrate its behaviour in two different situations. Parameter values used in the different experiments are given in Table 1. The motion model is the affine one in each case. The examples shown below have also been processed with the two-frame detection scheme, but using (and building) a reference frame as described in [14]. Results were always less accurate. Sometimes, no coherent results could be obtained (as in the second example). Indeed, if registration between images is not perfect over the temporal integration (3 to 4 images), even the conforming regions in the reference frame will be blurred. Since in the scheme described in [14], the motion is estimated between the current reference image and the next image, the blurring will greatly impair the motion estimation. Therefore, the loop composed of the motion estimation and the updating of the reference frame is unstable.
On the contrary, our algorithm only requires that the registration, *between two images only*, is *more precise* than the expected motion magnitude that we want to detect. Thus, with our scheme, we can tolerate large registration errors as far as the moving objects are moving more rapidly.

5.1 First example: the "interview" sequence

Figure 3a shows the first image of the Interview sequence (by courtesy of the BBC), which is often used to evaluate image coding algorithms. We have considered a spatially subsampled (by 2) version. Images are of quite good quality. In this sequence, the camera is tracking the woman who is standing up on the right of the scene, and whose left hand is initially hidden by a bouquet. Besides, casted shadows of this woman are moving over the sofa.
The time-extended version of the algorithm is utilized, the same holds for the next example

reported below. Figure 3b shows the 43^{th} image of this sequence, compensated for by the 2D motion models robustly estimated from frame to frame. As can be seen, the image is quite well registered. Figure 3c-f contains the detection maps obtained with the time extended version of the algorithm at different instants. In particular, note that the bouquet is well delimited, and that the non-conforming region comprises the whole woman (except at the beginning, where the legs are still -almost- static).

5.2 Second example: the "roundabout" sequence

Figures 4a-b-c present three images from a real sequence (digitized at 5 frames per second). The camera is mounted on the left side of a car approaching a roundabout. The dominant motion in the image sequence is due to the camera movement. It is conveyed by the scene background, that is, mainly the areas comprising the houses. Hence, regions corresponding to moving objects in the scene, the car, and to static entities in the near foreground due to significant difference in depth, the marks on the road and the sign, are expected to be detected as "non-conforming". These two classes of objects could be further discriminated, but this is beyond the scope of this paper. Let us point out that this sequence is really very noisy. Moreover, the low sampling rate and the high frequency content of the image projections of the tiles produced temporal aliasing, leading to an apparent motion in the roof area (this occured at some other places too).
Figures 4d-e-f contain the corresponding detection label fields $\widehat{d}$, where "conforming" regions are in black, and the original intensity information has been kept inside the regions labeled as "non-conforming". Let us note that "non-conforming" regions are quite correctly segmented, and that there is no spurious detection within the "conforming" parts. On the other hand, Fig. 5b proves that simply thresholding the observations gives a very noisy detection map. We can observe a large amount of false alarms, and masks of moving regions are quite partially recovered. However, the result is better than when thresholding the displaced frame difference (Fig. 5a).
Fig. 5c and 5d both report the detection map obtained at time t_{62} using only observations between two frames ($T = 1$) and without considering the energy term U_3. Fig. 5d is still obtained using the multiscale minimization procedure, while Fig. 5c is derived using a single scale scheme. The multiscale algorithm obviously outperforms the other one. It can also be pointed out that, although the two-frame detection scheme gives good results in that case, the use of past observations helps in recovering the complete masks of the car and of the sign (compare Fig. 4e and Fig. 5d). This was even more obvious in other parts of the sequence, as well as in other sequences.

6 Conclusion

We have described an algorithm which can be used both to detect parts of the scene whose apparent motion is not conforming to the dominant background one in a sequence acquired with a moving camera, and to track a region of interest. Obviously, it can also be used in the case of a static camera, which is not the case of some schemes devoted to a moving sensor situation [26].
Our algorithm decouples the detection problem in two natural components. First, the identification of some global parameters (the motion model), regardless of the presence of non-consistent data. This is reached due to the robustness property of the estimator we have designed. In fact, since the model is global, we don't need all the data in the support region to be valid points in order to estimate it. Thus, we don't mind if there are outliers, and which they are. Second, the localization of regions whose motion is not compensated for by the estimated global motion model in the first step. This second step uses only local *motion* measurements instead of intensity change measurements, and is embedded in a multiscale MRF framework to avoid false alarms and to increase the detection rate. A key feature is that the aperture problem is explicitly and directly acknowledged in this framework, allowing us to differentiate between informative sites and non-informative ones. Another important consequence of our approach is that it explicitly deals with situations where inaccurate registration occurs, either due to a "rough" estimation, or more often, to an imperfect adequacy of the 2D motion model used to describe the background apparent motion induced by the camera movement. The use of an extended period of time in the MRF framework makes the algorithm even more robust to such kind of "noise". Experiments carried out on many different sequences have demonstrated the robustness and the validity of our approach.

Extensions to the presented algorithm can be considered. First, the algorithm involves a few parameters. This has allowed us to deal with very different situations. As indicated, most of these parameters can be set and kept unchanged for a given type of application, and no fine tuning of the other parameters was necessary to obtain good results in our experiments. However, estimation of these parameters could be performed on line, leading to a more adaptive algorithm. For instance, acquisition noise could be estimated and related to the corresponding parameters (see section 4.2). Also, the main parameter δ, which is a kind of upper bound on the accuracy of the motion estimation and registration, is directly related to the standard deviation of the residual normal flow. Finally, statistical methods for the estimation of the parameters of the Markov model have been proposed [7], but they remain computationally quite expensive.
The extension of the binary detection scheme to a more sophisticated step of segmentation is also an important issue. For instance, the described algorithm could be applied in a recursive fashion,

as described in [13]. More precisely, a dominant motion could be subsequently estimated in the regions which have not yet been classified as being well compensated for by already estimated motion models. Then, the areas whose motion is conforming with this model can be found according to our scheme and withdrawn from further processing. The algorithm is applied again on the remaining data, and so on. We have oriented ourselves towards a more direct scheme [21] where all motion models are considered equivalently, instead of one after the other. Finally, the development of efficient image sequence coding algorithms using such schemes is of course an essential issue, and are under investigation. In [19], an hybrid coding scheme is proposed to achieve selective compression based on inhomogeneous spatial reconstruction.

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.

References

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

[2] S. Ayer, H. S. Sawhney, and M. Gorkani. Model-based 2d and 3d dominant motion estimation for mosaicing and video representation. In *Proc. IEEE Int. Conf. Computer Vision*, pages 583–590, Boston, June 1995.

[3] R. Azencott. Image analysis and Markov fields. In *ICIAM87: First International Conference on Industrial and Applied Mathematics*, pages 53–61, Philadelphia, June 1987.

[4] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of Optical Flow techniques. *International Journal of Computer Vision*, 12(1):43–77, January 1994.

[5] M. Bichsel. Segmenting simply connected moving objects in a static scene. *IEEE Trans. Pattern Anal. Machine Intell.*, 11(16):1138–1142, November 1994.

[6] P. Bouthemy and P. Lalande. Recovery of moving object masks in an image sequence using local spatiotemporal contextual information. *Optical Engineering*, 32(6):1205–1212, June 1993.

[7] B. Chalmond. Image restoration using an estimated Markov model. *Signal Processing*, 15(2):115–129, September 1988.

[8] 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.

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

[10] B. Girod. Eyes movements and coding of video sequences. In *Proc. SPIE (Visual Communications and Image Process.) 88*, volume 1001, pages 398–405, 1988.

[11] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, Vol.17:185–203, 1981.

[12] P.J. Hubert. *Robust statistics*. Wiley, 1981.

[13] 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.

[14] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motion. *Intern. J. Comput. Vis.*, 12(1):5–16, 1994.

[15] Jean-Marc Odobez. *Estimation, détection et segmentation du mouvement: une approche robuste et markovienne*. PhD thesis, IRISA, University of Rennes I, N^o 1304, December 1994.

[16] K.-P. Karmann, A. v.Brandt, and R. Gerl. Moving object segmentation based on adaptive reference images. In *Signal Process. V: Theories and Applications (Proc. Fifth European Signal Process. Conf.)*, pages 951–954, Barcelona, September 1990.

[17] H. Li and R. Forchheimer. Image sequence coding at very low bitrates: a review. *IEEE Trans. on Image Processing*, 3(5):589–609, September 1994.

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

[19] E. Nguyen, C. Labit, and J-M. Odobez. A ROI approach to hybrid image sequence coding. In *1st Int. Conference on Image Processing*, volume 3, pages 245–249, Austin, Texas, November 1994.

[20] H. Nicolas and C. Labit. Global motion identification for image sequence analysis and coding. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume Vol. IV, pages 2825–2828, Toronto, May 1991.

[21] J-M. Odobez and P. Bouthemy. Direct model-based image motion segmentation for dynamic scene analysis. In *Proc. of 2nd Asian Conf. Computer Vision*, Singapore, December 1995.

[22] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.

[23] 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.

[24] R. Thoma and M. Bierling. Motion compensating interpolation considering covered and uncovered background. *Signal Processing : Image communication*, Vol.1(2):191–212, October 1989.

[25] W.B. Thompson, P. Lechleider, and E.R. Stuck. Detecting moving objects using the rigidity constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(15):162–166, February 1993.

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

[27] 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.

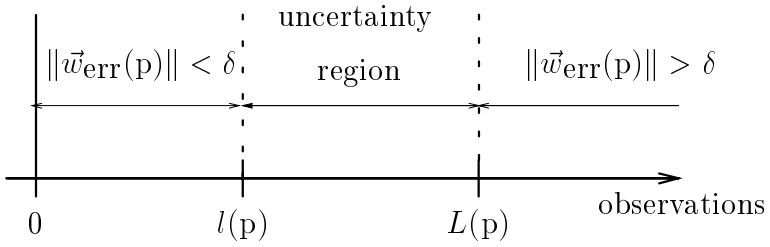

Figure 1: Classification of measurements.

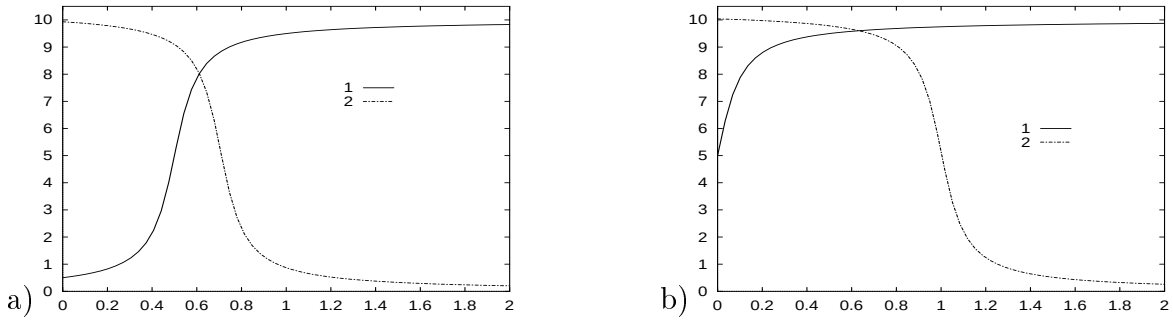

Figure 2: Potential V_2 associated with "conforming" label (solid lines) and "non-conforming" label (dashed lines) for two different local structure configurations: a) λ'_{min} value is maximal and is equal to $1/2$ (typically, at a corner site), $l_s = 1/2, L_s = \sqrt{2}/2$. b) λ'_{min} value is minimal and is equal to 0 (typically, on a straigth edge), $l_s = 0, L_s = 1$. Parameter values are $\delta = 1, k_c = k_{nc} = 4$.

Parameter	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{dt}	L
Interview	3.0	0.5	1.0	0.3	∞	0.4	2	30	27	5
Roundabout	10.0	1.0	1.0	0.4	2	0.4	2	36	36	5

Table 1: Values of the parameters used in the reported experiments for the image sequence "Interview" (Fig. 3) and "Roundabout" (Fig. 4).

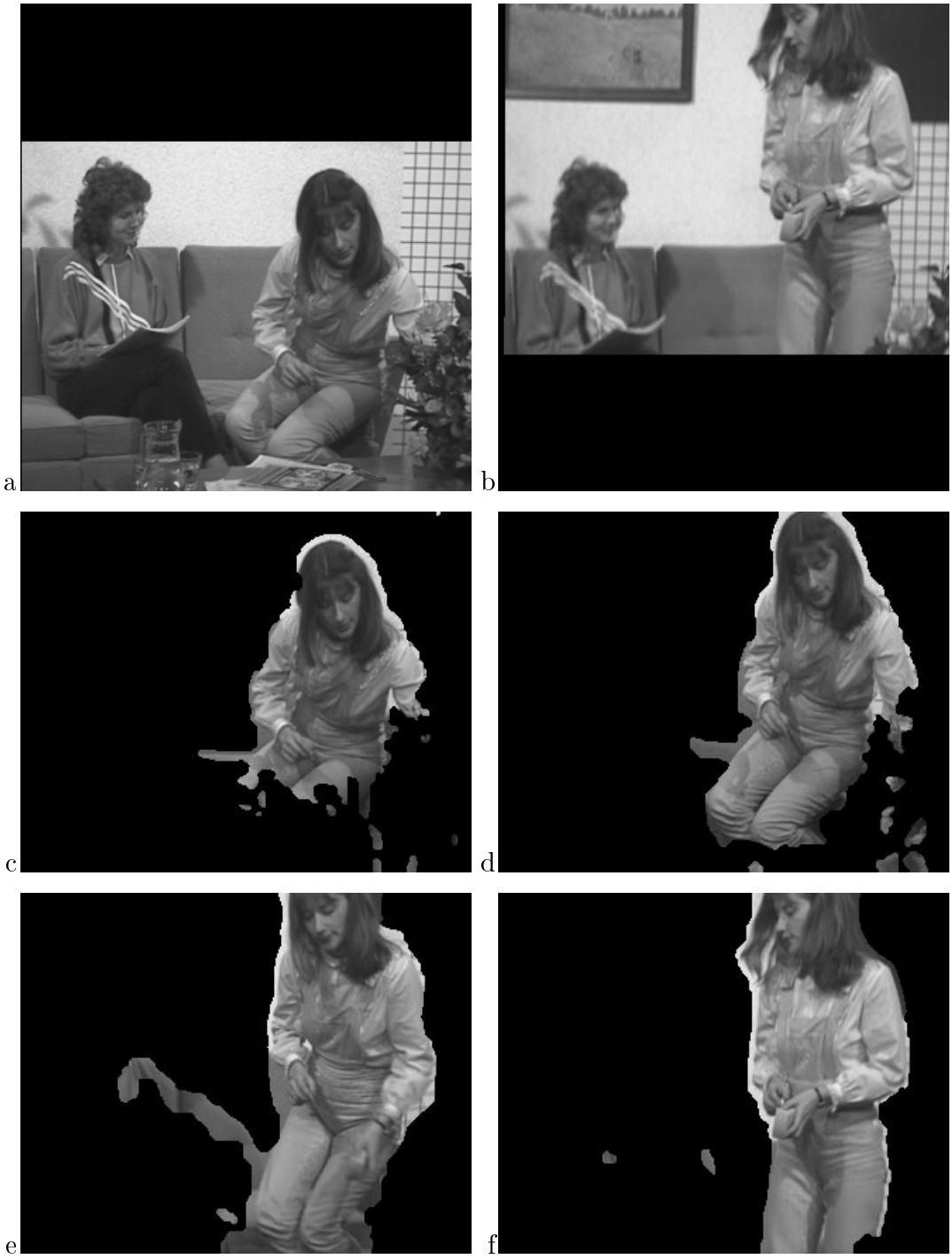

Figure 3: Sequence "interview" a) image at time t_1; b) image at time t_{43}, compensated by the estimated dominant motions. c-d-e-f) Detection maps obtained at time: c) t_1, d) t_{13}, e) t_{25} and f) t_{43}.

Figure 4: a) b) c) Roundabout sequence images at time: a) t_{58}, b) t_{62}, and c) t_{66}. d) e) f) Detection maps obtained at time: d) t_{58}, e) t_{62}, and f) t_{66}.

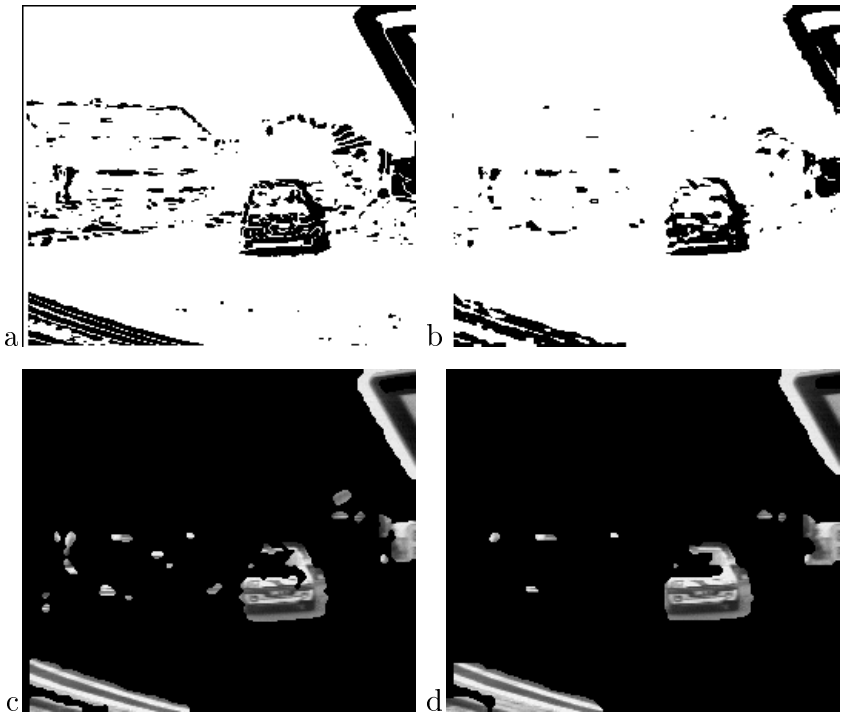

Figure 5: a) Thresholded displaced frame difference (the threshold is set to 8). b) Thresholded observation field o^{62} (threshold equal to 1.0). c) Detection maps at time t_{62} obtained using only two frames and a single scale minimization scheme ($\beta_{dt} = 0, \gamma = 0$). d) Same as in c), but with the multiscale minimization scheme ($L = 5$).