2D motion description and contextual motion analysis: issues and new models

P. Bouthemy

IRISA / INRIA
Campus universitaire de Beaulieu
35042 Rennes cedex, France

Abstract. In this paper, several important issues related to visual motion analysis are addressed with a focus on the type of motion information to be estimated and the way contextual information is expressed and exploited. Assumptions (i.e., data models) must be formulated to relate the observed image intensities with motion, and other constraints (i.e., motion models) must be added to solve problems like motion segmentation, optical flow computation, or motion recognition. The motion models are supposed to capture known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information. The latter can be formalized in a probabilistic way with local conditional densities as in Markov models. It can also rely on predefined spatial supports (e.g., blocks or pre-segmented regions). The classic mathematical expressions associated with the visual motion information are of two types. Some are continuous variables to represent velocity vectors or parametric motion models. The other are discrete variables or symbolic labels to code motion detection output (binary labels) or motion segmentation output (numbers of the motion regions or layers). We introduce new models, called mixed-state auto-models, whose variables belong to a domain formed by the union of discrete and continuous values, and which include local spatial contextual information. We describe how such models can be specified and exploited in the motion recognition problem. Finally, we present a new way of investigating the motion detection problem with spatial coherence being associated to a perceptual grouping principle.

1 Introduction

Motion is seamlessly perceived by human beings when directly observing a day-life scene, but also when watching films, videos or TV programs, or even various domain-specific image sequences such as meteorological or heart ultrasound ones. However, motion information is hidden in the image sequences supplied by image sensors. It has to be recovered from the observations formed by the image intensities in the successive frames of the sequence.

Assumptions (i.e., *data models*) must be formulated to relate the observed image intensities with motion. When dealing with video, the commonly used data model is the brightness constancy constraint which states that the intensity does not change along the trajectory of the moving point in the image plane (at least, to a short time extent). The motion constraint equation can then be expressed in a differential form that relates the 2D velocity vector, the spatial image gradient and the temporal intensity derivative at any point p in the image. Nevertheless, this enables to locally retrieve one component of the velocity vector only, the so-called normal flow, which corresponds to the aperture problem. Then, other constraints (i.e., *motion models*) must be added. They are supposed to formalize known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information.

In this paper, several important issues related to visual motion analysis are addressed with a focus on the type of motion information to be estimated and the way contextual information is formulated and exploited. Visual motion information can involve different kinds of mathematical variables. First, we can deal with *continuous variables* to represent the motion field : velocity vectors $\mathbf{w}(p)$ with $\mathbf{w}(p) \in \mathbb{R}^2$, or parametric motion models with parameters $\theta \in \mathbb{R}^d$ with d denoting the number of parameters. Let us note that the latter can be equivalently represented by the model flow vectors $\{\mathbf{w}_\theta(p)\}$ with $\mathbf{w}_\theta(p) \in \mathbb{R}^2$. Second, we can consider *discrete values or symbolic labels* to code motion detection output: binary values $\{0, 1\}$, or motion segmentation output: number n of the motion region or layer with $n \in \{1, ..., N\}$. Furthermore, we will introduce new models, called *mixed-state auto-models*, whose variables belong to a domain formed by the union of discrete and continuous values, and which include local spatial contextual information too. We will describe how such models can be specified and exploited in the motion recognition problem.

Spatial coherence can be formalized by conditional densities defined on local neighborhoods as in Markov Random Field (MRF) models, or equivalently by potentials on cliques as in Gibbs distributions. Another way is to first segment each image into spatial regions according to a given criterion (grey level, colour, texture) and to analyse the motion information over these regions. Perceptual grouping schemes can also be envisaged.

The remainder of the paper is organized as follows. In Section 2, the motion measurements that can be locally computed are briefly recalled and the subsequent needs for complementary constraints or motion models are outlined. Section 3 reviews briefly several MRF-based approaches we developed in the past to deal with the motion segmentation issue stated as a contextual labeling problem involving discrete variables. Section 4 is concerned with the main aspects of optical flow computation using MRF

models or more generally relying on energy minimization methods. In that case, continuous motion variables are considered. Motion recognition or classification, and more specifically event detection in video, is addressed in Section 5, requiring the introduction of new contextual models with mixed states. Section 6 describes a new way to address motion detection based on a perceptual grouping principle.

2 Local motion measurements

The brightness constancy assumption along the trajectory of a moving point $p(t)$ in the image plane, with $p(t) = (x(t), y(t))$, can be expressed as $dI(x(t), y(t), t)/dt = 0$, with I denoting the image intensity function. By applying the chain rule, we get the well-known motion constraint equation [22,32]:

$$r(p,t) = \mathbf{w}(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{1}$$

where ∇I denotes the spatial gradient of the intensity, with $\nabla I = (I_x, I_y)$, and I_t its partial temporal derivative. The above equation can be straightforwardly extended to the case where a parametric motion model is considered, and we can write:

$$r_\theta(p,t) = \mathbf{w}_\theta(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{2}$$

where θ denotes the vector of motion model parameters. It can be easily derived from equation (1) that the motion information which can be locally recovered at a pixel p is contained in the *normal flow* given by:

$$\nu(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|} \ . \tag{3}$$

It can also be written in a vectorial form: $\boldsymbol{\nu}(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|}\boldsymbol{\omega}_{\nabla I}(p,t)$, where $\boldsymbol{\omega}_{\nabla I}$ denotes the unit vector parallel to the intensity spatial gradient. However, it should be clear that the orientation of the normal flow vector does not convey any information on the motion direction, but implicitly on the object texture (for inner points) or on the object shape (for points on the object border). Besides, the normal flow can be computed at the right scale to enforce reliability as explained in [15].

In case of a moving camera and assuming that the dominant image motion is due to the camera motion and can be correctly described by a 2D parametric motion model, we can exhibit the *residual normal flow* given by:

$$\nu_{res}(p,t) = \frac{-DFD_{\hat{\theta}}(p,t)}{\|\nabla I(p,t)\|} \ , \tag{4}$$

where $DFD_{\hat{\theta}}(p,t) = I(p + \mathbf{w}_{\hat{\theta}}, t+1) - I(p,t)$ is the displaced frame difference corresponding to the compensation of the dominant motion described by the estimated motion model parameters $\hat{\theta}$.

Since the computation of intensity derivatives is usually affected by noise and can be unreliable in nearly uniform areas, it may be preferable to consider the local mean

of the absolute magnitude of normal residual flows weighted by the square of the norm of the spatial intensity gradient (as proposed in [23,36]):

$$\bar{\nu}_{res}(p,t) = \frac{\sum_{q \in \mathcal{F}(p)} \|\nabla I(q,t)\|.|DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q \in \mathcal{F}(p)} \|\nabla I(q,t)\|^2\right)} \ , \tag{5}$$

where $\mathcal{F}(p)$ is a local spatial window centered in pixel p (typically a 3×3 window), and η^2 is a predetermined constant related to the noise level. An interesting property of the local motion quantity $\bar{\nu}_{res}(p)$ is that the reliability of the conveyed motion information can be locally evaluated. Given the lowest motion magnitude δ to be detected, we can derive two bounds, $l_\delta(p)$ and $L_\delta(p)$, verifying the following properties [36]. If $\bar{\nu}_{res}(p) < l_\delta(p)$, the magnitude of the (unknown) true velocity vector $\mathbf{w}(p)$ is necessarily lower than δ. Conversely, if $\bar{\nu}_{res}(p) > L_\delta(p)$, $\|\mathbf{w}(p)\|$ is necessarily greater than δ. The two bounds l_δ and L_δ can be directly computed from the spatial derivatives of the intensity function within the window $\mathcal{F}(p)$.

By defining the motion quantity $\bar{\nu}_{res}(p)$, we already advocate the interest of considering spatial coherence to compute motion information. Here, it simply amounts to a weighted averaging over a small spatial support and it only concerns the data model. In the same vein, more information can be locally extracted by considering small spatio-temporal supports, either through spatio-temporal (frequency-based) velocity-tuned filters as in [16] or using 3D orientation tensors [4,33]. On the other hand, more benefit can be gained by introducing contextual information through the motion models.

3 Discrete motion labels and motion segmentation

One important step ahead in solving the motion segmentation problem was to formulate the motion segmentation problem as a statistical contextual labeling problem or in other words as a discrete Bayesian inference problem [7,31]. Segmenting the moving objects is then equivalent to assigning the proper (symbolic) label (i.e., the region number) to each pixel in the image. The advantages are mainly two-fold. Determining the support of each region is then implicit and easy to handle: it merely results from extracting the connected components of pixels with the same label. Introducing spatial coherence can be straightforwardly (and locally) expressed by exploiting MRF models.

Here, by motion segmentation, we mean the competitive partitioning of the image into motion-based homogeneous regions. Motion detection can be viewed as a simplified case where two labels only are considered: static background versus moving object, either with a static camera [1,30,39], or a mobile one [36]. The latter assumes that the camera motion (or more specifically, the dominant global motion) can be computed and somehow canceled, usually requiring to resort to robust estimation as we proposed in [35] (joint work with Jean-Marc Odobez). This formulation can also encompass the determination of motion layers by assuming that the regions of same label are not necessarily connected.

Formally, we have to determine the hidden discrete motion variables (i.e., region numbers) $l(i)$ where i denotes a site (usually, a pixel of the image grid; it could be also an elementary block [7,13]). Let $l = \{l(i), i \in S\}$. Each label $l(i)$ takes its value in the

set $\Lambda = \{1, .., N_{reg}\}$ where N_{reg} is also unknown. Moreover, the motion of each region is represented by a motion model (usually, a 2D affine motion model of parameters θ which have to be conjointly estimated; we have also explored a non-parametric motion modeling in [13], joint work with Ronan Fablet). Let $\Theta = \{\theta_k, k = 1, .., N_{reg}\}$. The data model of relation (2) is used. The *a priori* on the motion label field (i.e., spatial coherence) is expressed by specifying a MRF model (the simplest choice is to favour the configuration of the same two labels on the two-site cliques so as to yield compact regions with regular boundaries). Adopting the Bayesian MAP criterion is then equivalent to minimizing an energy function E whose expression can be written in the general following form:

$$E(l, \Theta, N_{reg}) = \sum_{i \in S} \rho_1[r_{\theta_{l(i)}}(i)] + \sum_{i \sim j} \rho_2[l(i), l(j)] \ , \tag{6}$$

where $i \sim j$ designates a two-site clique. In [7] (joint work with Edouard François), we considered the quadratic function $\rho_1(x) = x^2$ for the data-driven term in (6). The minimization of the energy function E was carried out on l and Θ in an iterative alternate way, and the number of regions N_{reg} was determined by introducing an extraneous label and using an appropriate statistical test. In [37] (joint work with Jean-Marc Odobez), we instead chose a robust estimator for ρ_1. This allowed us to avoid the alternate minimization procedure and to determine or update the number of regions through an outlier process in every region.

Specifying (simple) MRF models at a pixel level (i.e., sites are pixels and a 4- or 8-neighbour system is considered) is efficient, but remains limited to express more sophisticated properties on region geometry (e.g., more global shape information [10]) or to handle extended spatial interaction. Multigrid MRF models [21] (as used in [36,37]) is a means to address somewhat the second concern (and also to speed up the minimization process while usually supplying better results). An alternative is to first segment the image into spatial regions (based on grey level, colour or texture) and to specify a MRF model on the resulting graph of adjacent regions as we did in [17] (joint work with Marc Gelgon). The motion region labels are then assigned to the nodes of the graph (which are the sites considered in that case). This allowed us to exploit more elaborated and less local *a priori* information on the geometry of the regions and their motion [17]. However, the spatial segmentation stage is often time consuming, and getting an effective improvement on the final motion segmentation accuracy remains questionable. Using the level-set framework is another way to precisely locate region boundaries while dealing with topology changes [38,39], but handling a competitive motion partioning of the image (with the number of regions *a priori* unknown) remains an open issue in that context even if recent attempts have been reported [11,26].

Finally, let us mention other recent work on Bayesian motion segmentation, exploring the use of edge motion [41], offering extension to spatio-temporal models [11], or introducing (two-step) hidden Markov measure field (HMMF) models [27]. Tensor voting could also be considered as an implicit way to enforce spatial coherence [34].

4 Continuous motion information and optical flow computation

By definition, the velocity field formed by continuous vector variables is a complete representation of the motion information. Computing optical flow based on the data model

of equation (1) requires to add a motion model enforcing the expected spatial properties of the motion field, that is, to resort to a regularization method. Such properties of spatial coherence (more specifically, piecewise continuity of the motion field) can be expressed on local spatial neighborhoods. First methods to estimate discontinuous optical flow fields were based on MRF models associated with Bayesian inference [20,30,42] (i.e., minimization of a discretized energy function). Then, continuous-domain models were designed based on PDE formalism [2,8,25,45]. Spatial coherence can also be explicitly formulated by first segmenting the image in spatial regions forming the delimited domains where motion models, either dense or parametric ones, can be defined and estimated [6,17].

A general formulation of the global (discretized) energy function to be minimized to estimate the velocity field $\mathbf{w}$ can be given by:

$$E(\mathbf{w},\zeta) = \sum_{p\in S} \rho_1[r(p)] \; + \sum_{p\sim q} \rho_2[\|\mathbf{w}(p)-\mathbf{w}(q)\|, \zeta(p'_{p\sim q})] + \sum_{A\in\chi} \rho_3(\zeta_A) \;\; , \quad (7)$$

where S designates the set of pixel sites, $r(p)$ is defined in (1), $S' = \{p'\}$ the set of discontinuity sites located midway between the pixel sites and χ is the set of cliques associated with the neighborhood system chosen on S'. In [20] (joint work with Fabrice Heitz), quadratic functions were used and the motion discontinuities were handled by introducing a binary line process ζ. Then, robust estimators were popularized [5,28] leading to the introduction of so-called auxiliary variables ζ now taking their values in $[0,1]$. Depending on the followed approach, the third term of the energy $E(\mathbf{w},\zeta)$ can be optional. Multigrid MRF are moreover involved in the scheme developed by Mémin and Pérez in [28]. Besides, multiresolution incremental schemes are required to compute optical flow in case of large displacements. Dense optical flow and parametric motion models can also be jointly considered and estimated, which enables to supply a segmented velocity field as designed by Mémin and Pérez [29].

Recent advances have dealt with the computation of fluid motion fields involving the definition of a new data model (derived from the continuity equation of the fluid mechanics) and of a motion model preserving the underlying physics of the visualized fluid flows (2^{nd} order div-curl constraint) as defined by Corpetti, Mémin and Pérez in [9]. A comprehensive investigation of physics-based data models is described in [19].

5 Motion recognition and mixed-state auto-models

5.1 Event detection in video and mixed-state probabilistic models

A big challenge in computer vision consists in approaching the "semantic" content of video documents while dealing with physical image signals and numerical measurements. Here, we consider the detection of relevant events (dynamic content). Therefore, we focuse on motion information and we propose new probabilistic image motion models. The motion information is captured through low-level motion measurements so that it can be efficiently and reliabily computed in any video whatever its genre and its content. Our approach (joint work with Gwénaëlle Piriou and Jian-Feng Yao [40]) consists in modeling separately the camera motion (i.e., the dominant image motion) and the

scene motion (i.e., the residual image motion) in a sequence, since these two sources of motion bring important and complementary information. The dominant image motion is represented by a deterministic 2D affine motion model (which is a usual choice):

$$\mathbf{w}_\theta(p) = (a_1 + a_2 x + a_3 y, a_4 + a_5 x + a_6 y)^T \ , \tag{8}$$

where $\theta = (a_i, i = 1, \ldots, 6)$ is the model parameter vector and $p = (x, y)$ is an image point. This simple motion model can handle different camera motions such as panning, zooming, tracking, (including of course static shots). To estimate the motion parameters θ, we employ the robust real-time multi-resolution algorithm[1] described in [35]. The motion model parameters are directly computed from the spatio-temporal derivatives of the intensity function. Consequently, the model motion vector $\mathbf{w}_{\theta_t}(p)$ is available at any pixel p and time t. The two components of $\mathbf{w}_{\theta_t}(p)$ are finely quantized, and we build the empirical 2D histogram of their distribution over the considered video segment. Finally, this histogram is represented by a mixture of 2D Gaussian distributions denoted γ^{cam}. The number of components of the mixture is determined with the Integrated Completed Likelihood criterion (ICL) and their parameters are estimated using the Expectation-Maximization (EM) algorithm [40].

The residual motion measurements are given by the $\bar{\nu}_{res}(p,t)$'s as defined in (5). The probabilistic model of scene motion is derived from global statistics on these measurements. The 1D histograms of $\bar{\nu}_{res}(p,t)$ which have been computed over different video segments, present usually a prominent peak at zero and a continuous component part. The latter can be modeled either by an exponential distribution or a zero-mean Gaussian distribution, both restricted to $]0, \infty[$ (since by definition $\bar{\nu}_{res}(p,t) \geq 0$). Therefore, we consider a specific mixture model to represent the distribution of the local residual motion measurements within a video segment with density [40]:

$$f(z) = \varrho\delta_0(z) + (1-\varrho)\phi_\kappa(z) \ , \tag{9}$$

where z holds for $\bar{\nu}_{res}(p,t)$, ϱ is the mixture weight, δ_0 denotes the Dirac function at 0, and ϕ_κ designates either the (restricted) Gaussian density function with variance $1/2\kappa$ or the exponential density function with mean $1/\kappa$, both with support $]0, \infty[$. Consequently, the proposed model has explicitly two degrees of freedom: ϱ handles the peak at zero and κ accounts for the continuous component of the distribution. ϱ and κ are estimated using the ML criterion. In order to capture not only the instantaneous motion information but also its temporal evolution over the video segment, the temporal contrasts $\Delta\bar{\nu}_{res}$ of the local residual motion measurements are also considered: $\Delta\bar{\nu}_{res}(p,t) = \bar{\nu}_{res}(p,t+1) - \bar{\nu}_{res}(p,t)$. They are modeled, in a similar manner as in (9), by a mixture model $g(z')$ of a Dirac function at 0 and a zero-mean Gaussian distribution, where z' holds for $\Delta\bar{\nu}_{res}(p,t)$. The mixture weight and the variance of the Gaussian distribution are again evaluated using the ML criterion. The full probabilistic residual motion model is then simply defined as the product of these two models: $h^{res}(z,z') = f(z).g(z')$.

Let us stress the peculiar nature of the probabilistic model introduced in relation (9). The value 0 plays a particular role since it accounts for no motion which is a clear

[1] The corresponding software called MOTION-2D can be downloaded at http://www.irisa.fr/vista/Motion2D.

semantic information. We can consider that it corresponds to a symbolic state defined by the discrete value $z = 0$ and that the other state is defined by $z > 0$. Therefore, the variable z takes its value in the set $\{0\} \cup]0, \infty[$. We call such a set a *mixed-state space*.

The event detection proceeds in two steps. The first step permits to eliminate the segments that are not likely to contain the searched relevant events. Typically, if we consider sports videos, we try to first distinguish between "play" and "no play" segments. This step is based on the residual motion only. The second step consists in retrieving several specific events among the candidate segments $\{s_0, \ldots, s_N\}$. Here, the two kinds of motion information (residual and camera motion) are required since the combination allows us to characterize more finely a specific event. A residual motion model with density h_j^{res} and a camera motion model with density γ_j^{cam} have to be previously estimated from a training set of video samples, for each type j of event to detect. The label l_i of each segment s_i is determined using the ML criterion:

$$l_i = \arg \max_{j=1,\ldots,J} \prod_{(p,t)\in s_i} h_j^{res}(z_{(p,t)}, z'_{(p,t)}) \prod_{(p,t)\in s_i} \gamma_j^{cam}(\mathbf{w}_{\hat{\theta}_t}(p,t)) \quad . \tag{10}$$

More details and results on sports videos can be found in [40].

5.2 Mixed-state auto-models and motion classification

Here, we describe joint work with Jian-Feng Yao and Gwénaëlle Piriou and report preliminary results. The scene motion model (to be learnt from image data) defined above only accounts for global (occurrence) statistics accumulated over both the image plane and time (i.e., over all the frames of the video segment). Obviously, it does not capture how the motion information is spatially (or temporally) organized. In [14,15] (joint work with Ronan Fablet and Patrick Pérez), we have proposed the design of causal Gibbs models from scale and temporal co-occurrences of quantized motion values $\bar{\nu}$. Here, we will extend the model (9) to take into account spatial interaction between neighbours, and define mixed-state auto-models (to follow the terminology introduced in [3]). We will consider the Gaussian case only, but mixed-state auto-models can be defined as well for any distribution from the exponential distribution family [18].

Let us first rewrite the mixed-state probabilistic model (9) in the following exponential family form:

$$f_\theta(z) = \exp\left[\langle \theta, B(z)\rangle - \psi(\theta)\right] \quad , \tag{11}$$
$$\text{with} \quad \theta = (\theta_1, \theta_2)^T = \left(\log \frac{(1-\varrho)\phi_\kappa(0)}{\varrho}, \kappa\right)^T \quad , \quad B(z) = (\delta^*(z), -z^2)^T \ ,$$

where $\delta^*(z) = 1 - \delta_0(z)$. Let us note that we can easily recover the original parameters ϱ and κ from the "natural" ones θ_1 and θ_2.

To build our mixed-sate auto-models for the field $(z_i, i \in S)$, we start by considering, as in [3], the family of conditional densities $\mu_i(z_i|\cdot) := \mu_i(z_i|z_j, j \neq i)$, that is the conditional distribution of z_i at a site i given its outside configuration $(\cdot) = (z_j, j \neq i)$. Because of the mixed-state nature of the observations at hand, namely the residual motion measurements, we require that all these conditional distributions are

of type defined in (9), or equivalently (11). Let us note that, for each i, the parameters $\theta_i(\cdot) = (\theta_{i,1}(\cdot), \theta_{i,2}(\cdot))$ of the conditional density $\mu_i(z_i|\cdot)$ (here, we use the representation (11)) depend on the spatial context $(\cdot) := (z_j, j \neq i)$. It can be shown [18] that there are vectors $\alpha_i = (a_i, b_i) \in \mathbb{R}^2$ and 2×2 matrices $\beta_{ij} = \begin{pmatrix} c_{ij} & d_{ij} \\ d^*_{ij} & e_{ij} \end{pmatrix}$, such that:

$$\theta_i(\cdot) = \alpha_i + \sum_{j \neq i} \beta_{ij} B(z_j) \ ,$$

or in a more explicit way:

$$\theta_{i,1}(\cdot) = a_i + \sum_{j \neq i} \left[c_{ij} \delta^*(z_j) - d_{ij} z_j^2 \right] \ , \ \theta_{i,2}(\cdot) = b_i + \sum_{j \neq i} \left[d^*_{ij} \delta^*(z_j) - e_{ij} z_j^2 \right] \ .$$

It can further be shown that the joint density of (z_i) is proportional to $\exp(-H)$ where the global energy H associated to the *mixed-state Gaussian auto-model* can be written as follows (with $\mathcal{Z} = (z_1, \ldots, z_{|S|})$):

$$H(\mathcal{Z}) = -[\sum_{i \in S} \left[a_i \delta^*(z_i) - b_i z_i^2 \right] + \sum_{\{i,j\}} (\delta^*(z_i), -z_i^2) \beta_{ij} (\delta^*(z_j), -z_j^2)^T], \qquad (12)$$

provided the parameters of H verify:

(i) for any $\{i, j\}$, $e_{ij} \leq 0$;
(ii) for any i and any part $A \subset S \backslash \{i\}$, $b_i + \sum_{j \in A} d^*_{ij} > 0$, (in particular, $b_i > 0$ for any i).

We now specify the mixed-state Gaussian auto-model for the 4-nearest neighbour system. The binary clique formed by two neighboring sites i and j will be denoted $i \sim j$. We will further assume that the model is spatially homogeneous, i.e., the model parameters are independent of the site i, but it can be anisotropic (different parameters can be associated to the horizontal and vertical directions). From the development above, there are a vector $\alpha = (a, b)$ and two 2×2 matrices $\beta_k = \begin{pmatrix} c_k & d_k \\ d^*_k & e_k \end{pmatrix}$, $k = 1, 2$, such that:

$$\forall i, j, \quad \alpha_i = \alpha \ , \qquad \beta_{ij} = \beta_1 \quad \text{if } \ j = i \pm (1, 0) \ , \quad \beta_{ij} = \beta_2 \quad \text{if } \ j = i \pm (0, 1) \ .$$

This model has ten parameters which have to satisfy the following conditions:

$$\begin{cases} b > 0, & e_1 \leq 0, & e_2 \leq 0 \ , \\ b + 2d^*_1 > 0, & b + 2d^*_2 > 0, & b + 2d^*_1 + 2d^*_2 > 0 \ . \end{cases}$$

The parameters of the conditional laws $\mu_i(z_i|\cdot)$ are given by:

$$\theta_{i,1}(\cdot) = a + \sum_{j = i \pm (1,0)} \left[c_1 \delta^*(z_j) - d_1 z_j^2 \right] + \sum_{j = i \pm (0,1)} \left[c_2 \delta^*(z_j) - d_2 z_j^2 \right] \ ,$$
$$\theta_{i,2}(\cdot) = b + \sum_{j = i \pm (1,0)} \left[d^*_1 \delta^*(z_j) - e_1 z_j^2 \right] + \sum_{j = i \pm (0,1)} \left[d^*_2 \delta^*(z_j) - e_2 z_j^2 \right] \ .$$

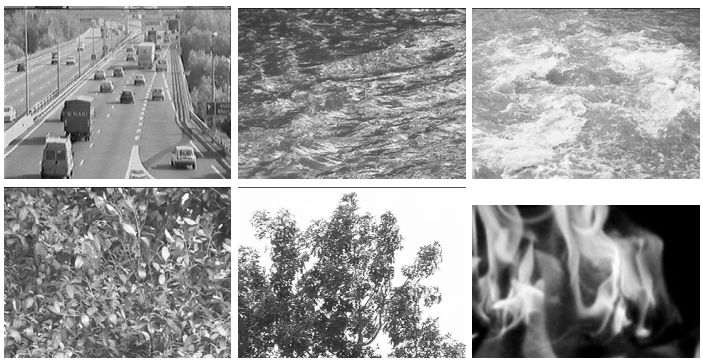

Fig. 1. One image of the considered video sequences (top row: *Highway, Water1, Water2* sequences; bottom row: *Leaves, Tree, Fire* sequences).

	a	b	c	d	d^*	e
Highway	-4.91	0.35	2.09	-3.45	0	0
Water1	-4.93	0.06	1.63	-3.19	0.01	0
Water2	-5.14	0.23	1.96	-3.49	0	0
Leaves	-4.86	0.91	2.11	-4.77	0	0
Tree	-4.66	1.89	2.16	-5.45	0	0
Fire	-7.03	0.08	2.5	-3.03	0	0

Table 1. Estimated parameters of the mixed-state Gaussian (isotropic) auto-model for the dynamic video contents of Fig.1.

We have applied this new motion model to different types of dynamic video contents (Fig.1). We have used the pseudo-likelihood criterion to estimate the auto-model parameters (with a gradient descent algorithm), and the computed values are given in Table 1.

Mixed-state auto-models can of course involve different kinds of mixed states. It is obviously not limited to one discrete value only, but any finite number K of discrete values $\{\xi_1, \xi_2, \ldots, \xi_K\}$, and pure symbols can be considered too. Any type of continuous domains included in $\mathbb{R}^n$ can also be considered. The mixed-state modeling framework introduced in this section should not be confused with the models previously developed either for motion segmentation (discrete labels l and continuous motion parameters Θ) or discontinuous optical flow computation (dense velocity field $\mathbf{w}$ and binary line process ζ), since the latter involve two different sets of variables defined on different set of sites. Here, we are dealing with one set of sites and one set of random variables x with mixed-state values. Several important issues need to be investigated such as the estimation of the mixed-state auto-model parameters, the handling of ML or MAP criteria, or the model selection issue. We also plan to exploit these models in different motion recognition tasks. It could be also interesting to revisit classical motion analysis issues such as motion detection, motion segmentation or discontinuous optical flow computation within that framework.

6 Motion grouping and detection of moving objects

6.1 Problem setting

Motion grouping is generally understood as the handling and analysis of multiple moving entities taken as a whole [44]. They may be either disconnected while sharing similar motion (such as flying birds, falling snow) or connected to form an articulated system. Here, we rather intend to revisit basic motion issues by considering perceptual grouping principles. More specifically, we aim at applying the Helmholtz principle ([12]) to motion detection, that is to compute an automatic criterion which ensures that a given region is not still (joint work with Thomas Veit and Frédéric Cao [43]).

Variational motion detection methods (in the sense of separating independent moving objects from background, [24,30,36,39]) paradoxically do not solve the problem of the detection itself: precisely, they enable to locate moving objects at each instant, assuming that one or several moving objects are present. Usually, deciding the presence of independent motion is achieved by hypothesis testing: a model of stillness is tested against a model of change and a decision is taken, for instance, by considering the likelihood ratio of both hypotheses ([1,30,24]). Nonetheless, this does not completely solve the decision threshold issue. A system that triggers off many false alarms cannot be efficient. This means that one should be able to explicitly formulate an automatic detection criterion and control the false alarm rate, which in turn can provide with a well-founded confidence measure.

The Helmholtz principle is a general perception law. It was recently applied to image feature detection in [12]. The Helmholtz principle states that an event is perceptible, that is to say significant, if its number of occurrences in a random situation is very small. According to this principle, significant events represent large deviations from randomness. Let us summarize the principle as follows. Entities to be detected are the conjunction of several local observations. We define a background model by assuming that all local observations are independent. By using this *a contrario* assumption, we can compute the probability that a given configuration occurs. More precisely, we call number of false alarms of a configuration, its expected number of occurrences in the background model. We say that an event is ϵ-meaningful if its number of false alarms is smaller than ϵ. ϵ can usually be set to 1 and the method considered as parameter-free.

Let us assume that there is no motion. Then, changes between two images of a sequence are due to noise and possible slight -not significant- changes of the images. We make the hypothesis that this noise (whatever its origin) is uncorrelated. We then consider this as the background model (in the statistical meaning), where no motion detection should occur. Let us assume now that an object is moving. The values of the image changes will increase. But what matters more is that this increase will certainly be very spatially (and temporally) correlated. Thus, if we use this background model, we can compute the probability that the change values increase (even slightly) in a compact region of the image. Because of the whiteness assumption in the background model, this probability is easy to compute. It will be very small in a region with a coherent change, leading to an *a contrario* detection. Here, spatial coherence is exploited in two ways: 1) changes associated with a moving object are supposed to be spatially correlated, 2) the detection criterion must be evaluated over a given spatial region. However, it is not

part of the designed model. We do not have a model for a moving region; we only use the background model and prove that a moving object does not conform to the *a contrario* noise model. In other words, the corresponding observation cannot result from a random situation. This approach is valid for sequences acquired by a static camera. It is straightforward to extend it to the case of a mobile camera, if we assume that we can compute and cancel the camera motion.

6.2 Designed method

The designed motion detection method is fully described in [43] (joint work with Thomas Veit and Frédéric Cao). Its main features only will be described hereafter. The case of a mobile camera is considered. The dominant image motion is represented by a 2D parametric model (affine or quadratic one) and is computed with the robust multi-resolution method [35]. First, a motion observation has to be defined at pixel level accounting for the adequacy to the estimated dominant motion. A first choice could be the Displaced Frame Difference $DFD_{\hat{\theta}_t}(p,t)$. However, in uniform regions (with very low image gradient) the DFD is always small regardless of the adequacy to the dominant motion. On the contrary, along highly contrasted edges, the DFD can be large even if the residual motion is low. A small error in the global motion estimation will be immediately enhanced. Therefore, the observation we use is the normal residual flow magnitude given by $|\nu_{res}(p,t)|$, $\nu_{res}(p,t)$ being defined in (4). A high value of this quantity indicates that the motion of the corresponding point differs from the estimated dominant motion, and is likely to be generated by a moving object in the scene (the points of the image where the spatial gradient is too small are ignored). In order to deal with occlusion, a three-image scheme on images $I(t-1)$, $I(t)$ and $I(t+1)$ is considered. Two dominant motions are estimated: a forward one from $I(t)$ to $I(t+1)$, leading to a set of parameters θ_t^{t+1}, and a backward one from $I(t)$ to $I(t-1)$, leading to θ_t^{t-1}. The resulting quantity considered is now:

$$C_t(p) = \min(\nu_{res}(p,t,\theta_t^{t+1}), \nu_{res}(p,t,\theta_t^{t-1}) \ . \tag{13}$$

To apply the Helmholtz principle, the above-defined motion detection variable $C_t(p)$ and a spatial segmentation are jointly exploited. Each region is tested for conformity with the estimated dominant image motion. The *a contrario* model is specified as follows: the value of C_t is distributed randomly according to its empirical distribution. Moreover, the value at each pixel is supposed to be independent of the values at all other pixels. The *a contrario* model is built upon the empirical inverse cumulative distribution function of the observations C_t:

$$F_t(\mu) = \frac{1}{A}\#\{p/C_t(p) \geq \mu\} \ , \tag{14}$$

where A is the surface of the image counted in pixels. Given a region R, the event of interest E is "for at least k points among the n points of the region, C_t assumes a value larger than μ". The probability of this event according to the *a contrario* model is:

$$B(k,n,F_t(\mu)) = \sum_{i=k}^{n} \binom{n}{i} F_t(\mu)^i (1-F_t(\mu))^{n-i} \ , \tag{15}$$

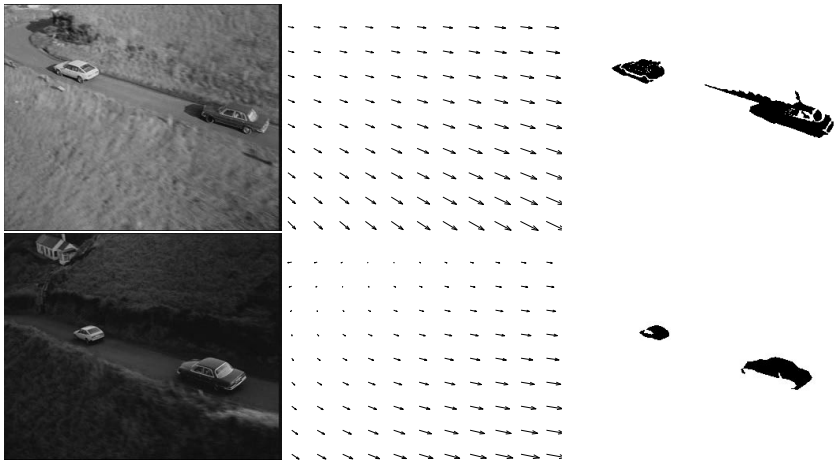

Fig. 2. Sequence "Road" (provided by INA). *Left column:* two original images of the sequence at distant time instants. The cars move leftward along the road. The camera is tracking the cars. *Middle column:* the estimated global dominant motion field is plotted (a 8-parameter quadratic model is used). *Right column:* the detection maps. Cars and associated cast shadows are detected. Detection extends slightly to parts surrounding the dark moving car. NFAs for the left image are about 10^{-10} and 10^{-30} for the white car and the dark car respectively. For the right image, NFAs are about 10^{-4} and 10^{-10} for the white car and the dark car respectively. As demonstrated in this example, detection is effective in quite different illumination conditions.

i.e., the tail of a binomial law of parameters k, n, and $F_t(\mu)$. Now, the question of how to choose the threshold μ arises. One way to solve this problem is to consider a set of thresholds μ_i, $i \in \{1, \ldots, N_\mu\}$ reasonably sampled. In practice, we take μ_i such that $F_t(\mu_i) = \frac{i}{1+N_\mu}$, i.e., the probabilities $P(C_t \geq \mu_i)$ are uniformly sampled in $[0, 1]$. Now, we can define the number of false alarms (NFA) with respect to motion for a region R containing n points. For $1 \leq i \leq N_\mu$, we denote by k_i the number of points at which C_t has a value larger than μ_i. The NFA of a region R with respect to motion is defined as follows:

$$NFA_m(R) = N_r \cdot N_\mu \cdot \min_{1 \leq i \leq N_\mu} B(k_i, n, F_t(\mu_i)), \tag{16}$$

where N_r is the number of regions. We say that R has an ϵ-meaningful motion if $NFA_m(R) \leq \epsilon$. A result is reported in Fig.2. More details can be found in [43].

On-going work is concerned with extending this motion detection scheme to include temporal integration. Besides, we plan to investigate this kind of approach to address other motion analysis issues such as region matching and tracking.

Acknowledgements: I would like to thank Patrick Pérez and Jian-Feng Yao for their comments on this paper, and the contributors (quoted in the text) of the described works.

References

1. T. Aach, A. Kaup. Bayesian algorithms for change detection in image sequences using Markov random fields. Signal Processing: Image Communication, 7(2):147-160, 1995.
2. L. Alvarez, J. Weickert, J. Sánchez. Reliable estimation of dense optical flow fields with large displacements. International Journal of Computer Vision, 39(1):41-56, 2000.
3. J. Besag. Spatial interactions and the statistical analysis of lattice systems. Journal Royal Statistical Society, B, 148:1-36, 1974.
4. J. Bigün, G.H. Granlund, J. Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(8) :775-790, August 1991.
5. M.J. Black, P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding, 63(1):75-104, 1996.
6. M.J. Black, A.D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(10):972-986, October 1996.
7. P. Bouthemy, E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. Int. Journal of Computer Vision, 10(2):157-182, April 1993.
8. I. Cohen, I. Herlin. Non uniform multiresolution method for optical flow and phase portrait models: Environmental application. IJCV, 33(1):29-49, September 1999.
9. T. Corpetti, E. Mémin, P. Pérez. Dense estimation of fluid flows. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(3):365-380, March 2002.
10. D. Cremers, T. Kohlberger, C. Schnörr. Nonlinear shape statistics in Mumford-Shah based segmentation. 7th European Conference on Computer Vision, ECCV'2002, Copenhagen, Vol. LNCS 2351, Springer Verlag, 2002.
11. D. Cremers, S. Soatto. Variational space-time motion segmentation. Proc. 9th IEEE Int. Conf. on Computer Vision, ICCV'2003, Nice, October 2003.
12. A. Desolneux, L. Moisan, J.-M. Morel. A grouping principle and four applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(4):508-513, April 2003.
13. R. Fablet, P. Bouthemy. Non-parametric scene activity analysis for statistical retrieval with partial query. Journal of Mathematical Imaging and Vision, 14(3):257-270, May 2001.
14. R. Fablet, P. Bouthemy, P. Pérez. Non-parametric motion characterization using causal probabilistic models for video indexing and retrieval. IEEE Trans. on Image Processing, 11(4):393-407, April 2002.
15. R. Fablet, P. Bouthemy. Motion recognition using non parametric image motion models estimated from temporal and multiscale cooccurrence statistics. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(12):1619-1624, December 2003.
16. D.J. Fleet, A.D. Jepson, Computation of component image velocity from local phase information. International Journal of Computer Vision, 5(1):77-104, 1990.
17. M. Gelgon, P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image partition. Pattern Recognition, 33(4):725-745, April 2000.
18. X. Guyon, C. Hardouin, J.-F. Yao. Markovian auto-models with mixed states. Preprint, 2004.
19. H.W. Haussecker, D.J. Fleet. Estimating optical flow with physical models of brightness variation. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(6):661-673, 2001.
20. F. Heitz, P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. IEEE Trans. on PAMI, 15(12):1217-1232, December 1993.
21. F. Heitz, P. Pérez, P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. CVGIP : Image Understanding, 59(1):125-134, January 1994.
22. B.K.P. Horn, B.G. Schunck. Determing optical flow. Art. Intelligence, 17:185-203, 1981.

23. M. Irani, B. Rousso, S. Peleg. Computing occluding and transparent motion. International Journal of Computer Vision, 12(1):5-16, 1994.
24. J. Konrad. Motion detection and estimation. in Handbook of Image and Video Processing, A.C. Bovik ed., Academic Press, 2000.
25. P. Kornprobst, R. Deriche, G. Aubert Image sequence analysis via partial differential equations. Journal of Mathematical Imaging and Vision, 11(1):5-26, 1999.
26. A.-R. Mansouri and J. Konrad. Multiple motion segmentation with level sets. IEEE Trans. Image Processing, vol. 12, pp. 201-220, February 2003.
27. J.-L. Marroquin, E.A. Santana, S. Botello. Hidden Markov measure field models for image segmentation. IEEE Trans. on PAMI, 25(11):1380-1387, November 2003.
28. E. Mémin, P. Pérez. Optical flow estimation and object-based segmentation with robust techniques. IEEE Trans. on Image Processing, 7(5):703-719, May 1998.
29. E. Mémin, P. Pérez. Hierarchical estimation and segmentation of dense motion fields. Int. Journal of Computer Vision, 46(2):129-155, February 2002.
30. A. Mitiche, P. Bouthemy. Computation and analysis of image motion: A synopsis of current problems and methods. International Journal of Computer Vision, 19(1):29-55, 1996.
31. D.W. Murray, B.F. Buxton. Scene segmentation from visual motion using global optimization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 9(2):220-228, 1987.
32. H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. Artificial Intelligence, 33:299-324, 1987.
33. H.-H. Nagel, A. Gehrke. Spatiotemporally adaptive estimation and segmentation of OF-fields. 5th Eur. Conf. on Comp. Vis., ECCV'98, Freiburg, Vol. LNCS 1407, Springer, 1998.
34. M. Nicolescu, G. Medioni. Layered 4D representation and voting for grouping from motion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(4):492-501, April 2003.
35. J-M. Odobez, P. Bouthemy. Robust multiresolution estimation of parametric motion models. Journal of Visual Communication and Image Representation, 6(4):348–365, December 1995.
36. J.M. Odobez, P. Bouthemy. Separation of moving regions from background in an image sequence acquired with a mobile camera. In Video Data Compression for Multimedia Computing, H.H. Li, S. Sun, H. Derin (eds.), Chap. 8, pp. 283-311, Kluwer, 1997.
37. J-M. Odobez, P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. Signal Processing, 6(2):143-155, 1998.
38. C. Papin, P. Bouthemy, E. Mémin, G. Rochard. Tracking and characterization of highly deformable cloud structures. 6th European Conference on Computer Vision, ECCV'2000, Dublin, Vol. LNCS 1843, Springer Verlag, 2000.
39. N. Paragios, R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. IEEE Trans. on PAMI, 22(3):266-280, March 2000.
40. G. Piriou, P. Bouthemy, J-F. Yao. Extraction of semantic dynamic content from videos with probabilistic motion models. 8th Eur. Conf. on Comp. Vis., ECCV'04, Prague, May 2004.
41. P. Smith, T. Drummond, R. Cipolla. Layered motion segmentation and depth ordering by tracking edges. IEEE Trans. on PAMI, 26(4):479-494, April 2004.
42. C. Stiller, J. Konrad. Estimating motion in image sequences (A tutorial on modeling and computation of 2D motion). IEEE Signal Processing Magazine, vol. 16, pp. 70-91, July 1999.
43. T. Veit, F. Cao, P. Bouthemy. Probabilistic parameter-free motion detection. IEEE Conf. Computer Vision and Pattern Recognition, CVPR'04, Washington DC, June 2004.
44. Y. Wang, S.-C. Zhu. Modeling textured motion: Particle, wave and sketch. IEEE Int. Conf. on Computer Vision, ICCV'03, Nice, October 2003.
45. J. Weickert, A. Bruhn, N. Papenberg, T. Brox. Variational optic flow computation: From continuous models to algorithms. International Workshop on Computer Vision and Image Analysis (ed. L. Alvarez), IWCVIA'03, Las Palmas de Gran Canaria, December 2003.