Recovery of the trajectories of multiple moving objects in an image sequence with a PMHT approach

Marc Gelgon, Patrick Bouthemy, Jean-Pierre Le Cadre

Submitted to *Image and Vision Computing*
First revised version, June 2004

Recovery of the trajectories of multiple moving objects in an image sequence with a PMHT approach

Marc Gelgon[a,1], Patrick Bouthemy[b] and Jean-Pierre Le Cadre[c]

[a]LINA / Ecole Polytechnique de l'Université de Nantes
rue C.Pauc 44306 Nantes cedex, France

[b]IRISA/INRIA [c]IRISA/CNRS
Campus universitaire de Beaulieu
35042 Rennes cedex, France

e-mail : Marc.Gelgon@polytech.univ-nantes.fr, bouthemy@irisa.fr, lecadre@irisa.fr
Tel : (33) 2.40.68.32.57 Fax : (33) 2.40.68.32.32

Abstract

This paper is concerned with the tracking of multiple moving objects in an image sequence and the reconstruction of the entire trajectories of these objects all over the sequence. More specifically, we address the joint issue of trajectory estimation and measurement-to-trajectory associations, which is the key problem in that context due to the occurrence of object occlusions or crossings. An original and efficient scheme is proposed, that adapts the Probabilistic Multiple Hypothesis Tracking (PMHT) technique to the case of tracking of regions in video, for which geometry and motion models can be introduced. Moreover, reliable partial associations can be obtained as an initialization. Data association and trajectory estimation are conducted within a probabilistic framework. The latter relies on Kalman filtering, while the former is solved with an EM algorithm for which a suitable initial configuration can be defined. The proposed tracking method is validated by experiments carried out on real image sequences depicting complex situations.

Keywords

Multiple object tracking, trajectory reconstruction, data association, EM algorithm, PMHT.

[1]Corresponding author. The work was carried out while the author was with IRISA.

1 Problem statement

This paper is concerned with the tracking of multiple moving objects in an image sequence and the reconstruction of the entire trajectories of these objects all over the sequence. More specifically, we address the joint issue of trajectory estimation and measurement-to-trajectory associations. This is the key problem in that context due to the occurrence of object occlusions or crossings.

In video content analysis, whether for interpretation, indexing or coding, trajectories of objects - manipulated as regions in images - are of much importance. For instance for surveillance purposes, trajectories of mobile objects are generally of key interest. It may occur, however, events (temporary misdetection, occlusions, crossings) from which important ambiguities in the association of successive measurements to a track can arise.

We specify the addressed problem by describing hereunder the input data to the algorithm designed in this paper. We are provided with a batch of motion segmentation maps using an approach presented in [20], of which Fig. 1 shows an example. This technique supplies a motion-based partition of images, in which the motion region homogeneity criterion is expressed by a 2D parametric motion model. Motion estimation is supplied by a multiresolution, robust estimator and the segmentation problem is expressed and solved as the statistical estimation of a pixel label map, within a Markov Random Field framework. The set of measurements (at each time instant), includes:

- the 2D spatial supports of the extracted moving regions ;
- the estimates of motion of these regions, i.e. the 2D parametric motion models estimated between the current frame and the next one associated to these regions;
- the regions labels, i.e., their numbers (symbolic information).

The motion segmentation algorithm employed has the property that if the same region (object) is continuously extracted in successive frames, the region label is maintained. This provides a short-term temporal link which we will assume reliable (e.g., as shown in Fig. 1, identity of the two labels is relevant over images b_0 to b_6). However, since an object may temporarily be static or totally occluded, there may be lacks of detections that break that temporal link. This introduces the concept of partial trajectory. When the region reappears and is segmented again, it then bears a new label, provided by the motion segmentation algorithm (as illustrated in Fig. 1 from images b_{24}). Our focus is on determining and associating partial trajectories of regions and jointly estimating the complete trajectories of these regions, while dealing with occlusion or crossing situations.

Besides, the silhouette of the extracted region is often affected by perturbations compared to the true projection of the object in the image. Moving shadows may enlarge the expected support, while partial occlusion may cause some pixels to miss. For instance, in the sequence displayed in Fig. 1a, the total occlusion (images 14 to 23) is preceded and followed by partial occlusions of the two moving elements. As illustrated in Fig. 1*b*, this has an obvious effect on the supplied motion segmentation maps.

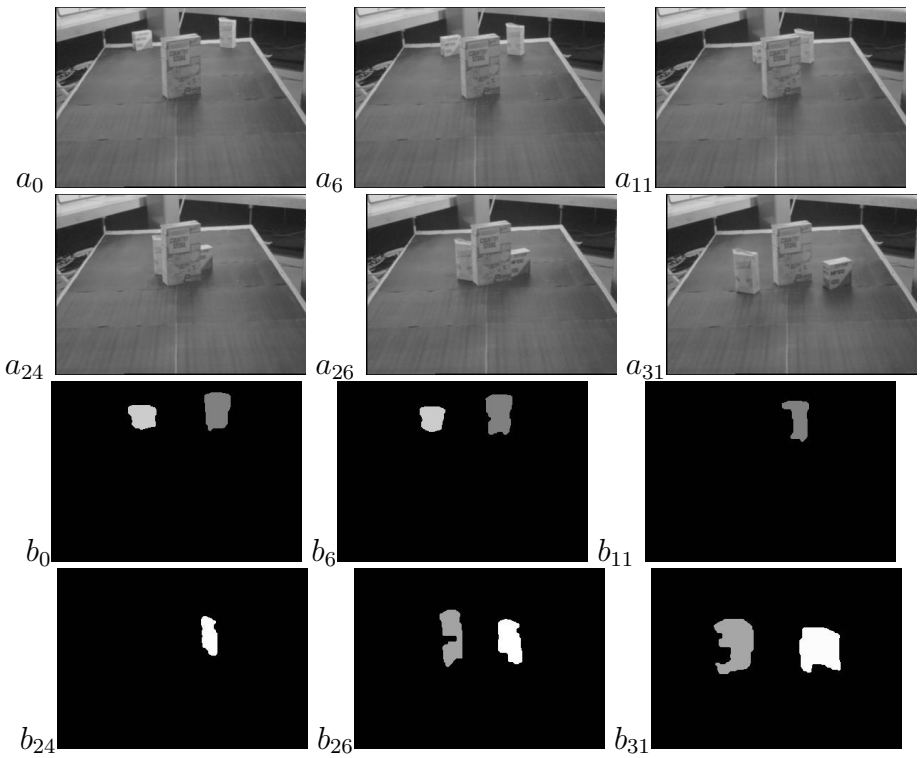

Figure 1: Original images (a) and resulting motion segmentation maps (b) at time t=0, t=6, t=11, t=24, t=26 and t=31. In this lab sequence, two moving boxes cross behind a third (static) one.

The desired output of the algorithm is two-fold :

- the correct association of the segmented regions over time, i.e., grouping of partial tracks;
- the complete trajectory of all the moving objects over the entire processed sequence, i.e. an estimated position of the object projections at each time instant (including at those when no measurement was initially available).

A core difficulty is that these two problems are tightly intricate. We briefly review below existing approaches for tracking, focusing on the issue of temporal data association.

2 State-of-the-art

Important research efforts in computer vision have been devoted to tracking objects in image sequence. In the case of region tracking, techniques based on active contours [2] or level-sets [21] have been employed, difficulties related to initialization and changes in topology being better handled by the latter approaches. It is insightful to distinguish between techniques that use a prediction and adjustment mechanism to track the image primitives, hence establishing a natural link between successive measurements and estimating model-based trajectories [14, 16, 20, 29], from those that determine merely

correspondence between primitives, and thus need to address an explicit data association problem (e.g., [18]).

Data association refers to the task of identifying, for each measurement, from which physical source (moving object, in our computer vision context) it arises. Potential association ambiguities and difficulties naturally appear when a scene contains several such physical elements. A similar issue is also encountered in general unsupervised classification tasks, but *data association* is the coined term when facing specific issues pertaining sequential data processing.

Explicit handling of the data association problem has received much attention, for a long time in the context of radar and sonar [7], more recently in computer vision. In the latter field, it has been applied to corners [4], segments [32], and regions [16, 22]. Trajectory estimation and data association problems are known to be two tightly interwoven problems. Indeed, the association between observations and objects depends on the estimated trajectories, which in turn should be computed from the whole set of measurements corresponding to a single physical element. The point is that this intricate issue is an NP combinatorial one.

A survey of data association techniques may be found in [3]. The measurement-to-trajectory model assignment can be hard, as in Multiple Hypothesis Tracking (MHT) algorithms [1, 24, 25]. Overall, MHT techniques consist in enumerating possible assignments and evaluating the pertinence of the trajectories formed, while introducing criteria to prune the assignment hypothesis tree, which otherwise would exponentially grow. Another classical tool for trajectory estimation/data association is the Joint Probabilistic Data Association Filter (JPDAF) [1], used for instance in [22] for region tracking. It is rooted in the Probabilistic Data Association Filter (PDAF) which, in e.g. Kalman filtering, updates the states using a combination of several competing measurements. The JPDAF is an enhanced version which, when there exists several such tracking processes, enforces some mutual exclusion in associations to prevent several trackers from fitting the same data. However, the JPDAF is rather a track updating technique.

In this paper, we propose an original approach relying on the Probabilistic Multiple Hypothesis Tracking technique (PMHT), which offers an attractive alternative to these classical techniques. Initially proposed in [28], a collection of works pertaining to the PMHT technique, and presenting variations thereof, may be found in [27]. They have been primarily explored in the radar and sonar domains. The statistical PMHT method consists in performing a MAP (Maximum A Posteriori) estimation of the models using Kalman filtering in the case of linear measurements and the EM algorithm for assigning, in a probabilistic manner, measurements to trajectory models. A key point is that doing so, it avoids the NP-hard combinatorial issue, in particular inherent in MHT techniques. We refer the reader to [8, 27, 28] for in-depth coverage.

In [17], the authors propose a recursive scheme closely related to PMHT in which the association variables form a Markov random field. The method we have designed remains, as in [28], with a batch approach, and a preliminary version was described in [9]. In [10], a modification was introduced to the PMHT, with a similar viewpoint to ours, so as to exploit the prior knowledge given by the existence of partial tracks, by constraining certain sets of measurements to be assigned to a single track.

A major aspect of target tracking with trajectory reconstruction is the modelling,

of the state temporal evolution and of the relation between state and measurements. In many naval surveillance scenarios, piecewise linear trajectories are assumed, while airborne applications usually require more flexible manoeuvering models. A classical solution is to employ Kalman filtering with dynamic and measurement models that are fixed in their form and parametrization [16]. We shall also take this approach. Recently, Hue et al. [12] have proposed a promising improvement on PMHT on this latter aspect, by introducing particle filtering (also known as Condensation or bootstrap filter [13]) which, compared to the abovementioned model, makes weaker assumptions on the form of the dynamic and observation processes. Flexibility in the dynamic process modelling has also recently been introduced in [31].

Applications of PMHT can so far be found in radar and sonar [8] and high-energy particle physics [26]. Still, to our knowledge, point-wise measurements are generally considered. Important contributions of the present work consist, besides demonstrating the effectiveness of PMHT for a common computer vision problem, in proposing the following adaptations :

- spatial extent (2D region support) and velocity information are properly incorporated into the PMHT scheme,
- a dedicated and efficient initialization is provided.

The remainder of the paper first presents the manner in which we model the problem, fitting in the PMHT framework (Section 3). We then recall how this category of problems may be solved using the Expectation-Maximization algorithm (Section 4). Section 5 presents the extension of the PMHT approach we have designed to handle tracking in video (in particular, initialization of the EM algorithm). Section 6 provides experimental results, and in section 7 we draw some concluding remarks.

3 Modelling of the problem

A *measurement* in our problem is a set of elements describing a segmented region at a given image instant, as listed in Section 1. They will be more formally defined hereafter. We shall call *partial track* a set of successive measurements linked over time by identity of the label attached to their corresponding regions. The goal is to recover *entire tracks* over the whole image sequence, each entire track being issued from the set of measurements corresponding to the same single physical moving object. To each partial track is associated a *2D trajectory model* of the mobile element, to be estimated from the measurements.

Let us denote $\mathcal{Z}$ the set of observed measurements $Z(t)$ in the batch $[t = 0, \ldots, t = T]$ corresponding to the processed image sequence. At each time instant t, $Z(t)$ is composed of a set of s_t measurements $z_j(t)$. They will be instanciated hereafter. We have :

$$\mathcal{Z} = \Big[Z(1), \ldots, Z(T) \Big] \tag{1}$$

$$Z(t) = \Big\{ z_1(t), \ldots, z_{s_t}(t) \Big\} \tag{2}$$

We assume that measurements originate from $\mathcal{M}$ moving objects in the scene. As $\mathcal{M}$ is unknown (and to be determined), the algorithm works throughout considering M trajectory models, where M is the number of partial tracks ($M > \mathcal{M}$). In a second stage, $\mathcal{M}$ will be determined by identifying redundant trajectory models among the M ones.

Each of the M trajectory models is described by a time-dependent state vector, and an evolution model of this state vector. Let us denote $x_m(t)$ the state vector of trajectory model m at time t. We also define the set $X(t)$ of state vectors at a given time t and their set $\mathcal{X}$ over the batch as follows :

$$\mathcal{X} = \Big[X(1), \ldots, X(T)\Big] \tag{3}$$

$$X(t) = \Big\{x_1(t), \ldots, x_M(t)\Big\} \tag{4}$$

Each region is represented by two elements :

- a geometric (polygonal) model of its contour. The polygonal approximation employs the technique described in [30];
- its kinematics, described by a 2D affine inter-frame motion model. Let us recall that a 2D affine motion model is defined as follows :

$$\omega_\theta(p) = [a_1 + a_2 x + a_3 y, a_4 + a_5 x + a_6 y]^T \tag{5}$$

  where $p(x, y)$ is an image point, $\theta = [a_1, a_2, a_3, a_4, a_5, a_6]^T$ and $\omega_\theta(p)$ is the velocity vector given by the considered motion model at point p.

The state vector $x_m(t)$ and the measurement vector $z_j(t)$ are hence made up of two components:

$$x_m(t) = \Big[\ \mathcal{G}_m(t)\ \ ,\ \ \Theta_m(t)\ \Big]^T \qquad m = 1, \ldots, M \tag{6}$$

$$z_j(t) = \Big[\ \tilde{\mathcal{G}}_j(t)\ \ ,\ \ \tilde{\Theta}_j(t)\ \Big]^T \qquad j = 1, \ldots, s_t \tag{7}$$

where

- $\mathcal{G}_m(t) = \{P_1^m(t), \ldots, P_{n(t)}^m(t)\}$ and $\Theta_m(t) = [a_1^m(t)\ ,\ \ldots\ ,\ a_6^m(t)]^T$ are respectively the geometric (i.e., the $n(t)$ vertices of the polygonal shape representing the region) and kinematic component of the state vector (i.e. the six parameters of the affine motion model);
- $\tilde{\mathcal{G}}_j(t) = \{\tilde{P}_j^1(t), \ldots, \tilde{P}_j^{\tilde{n}(t)}(t)\}$ is an ordered set of $\tilde{n}(t)$ vertices resulting from the polygonal approximation of the segmented region at time instant t;
- $\tilde{\Theta}_j(t) = [\tilde{a}_j^1(t)\ ,\ \ldots\ ,\ \tilde{a}_j^6(t)]^T$ is the estimated parameter vector of the affine motion model, obtained with the multiresolution robust estimation method described in [19].

We assume that the temporal evolution of each component of the state vector $x_m(t)$ can be appropriately represented by a first order model, with additive Gaussian white noise. Besides, we consider that the measurements are corrupted by an additive Gaussian white noise, which covariance matrix is denoted R_m.

Kinematic component

The parameters of the motion model Θ_m are considered decorrelated and are estimated independently. A classical first order evolution model is selected for these parameters. It is expressed by relation (8) for any r^{th} parameter $(r = 1, \ldots, 6)$:

$$\begin{bmatrix} a_r^m(t+1) \\ \dot{a}_r^m(t+1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_r^m(t) \\ \dot{a}_r^m(t) \end{bmatrix} + \begin{bmatrix} \epsilon_{1,r}^m(t) \\ \epsilon_{2,r}^m(t) \end{bmatrix} \tag{8}$$

where $[\epsilon_{1,r}^m, \epsilon_{2,r}^m]^T$ is a Gaussian random vector, which covariance matrix Q_ϵ is expressed as :

$$Q_\epsilon \;=\; \sigma_\epsilon^2 \; \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \tag{9}$$

The measurement equation is defined by stating that an additive Gaussian measurement noise $\eta_m^r(t)$ of variance σ_η^2 affects each motion parameter :

$$\tilde{a}_r^m(t) = a_r^m(t) + \eta_r^m(t) \qquad (r = 1, \ldots, 6) \tag{10}$$

Considering we have no prior knowledge on the kinematics of the moving object, no training set, and that no reliable estimation of the measurement uncertainty is available, σ_ϵ^2 and σ_η^2 are empirically user-set parameters.

Geometric component

The geometric model is formed by the set of vertices of the polygon approximating the region boundary. The temporal evolution of each of these vertices is designed by involving the affine motion model $\widehat{\Theta}_m(t)$ estimated on the region m and filtered over time. We have, for any vertex :

$$P_q^m(t), q = 1, \ldots, n(t) \;:\; P_q^m(t+1) = P_q^m(t) + \omega_{\hat{\theta_m}(t)}(P_q^m(t)) \tag{11}$$

If we denote $P^m(t) = [u_q^m(t), v_q^m(t)]^T$ the temporal evolution model for the geometric component is specified by :

$$\begin{bmatrix} u_q^m(t+1) \\ v_q^m(t+1) \end{bmatrix} = \begin{bmatrix} a_0^m(t) \\ a_1^m(t) \end{bmatrix} + \begin{bmatrix} 1 + a_2^m(t) & a_3^m(t) \\ a_4^m(t) & 1 + a_5^m(t) \end{bmatrix} \begin{bmatrix} u_q^m(t) \\ v_q^m(t) \end{bmatrix} + \begin{bmatrix} \zeta_{q,1}^m(t) \\ \zeta_{q,2}^m(t) \end{bmatrix} \tag{12}$$

where the $\zeta_{q,1}^m(t)$ and $\zeta_{q,2}^m(t)$ are drawn from Gaussian distributions, which covariance matrix Q_ζ is expressed as :

$$Q_\zeta \;=\; \sigma_\zeta^2 \; \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{13}$$

The relation between the geometric model and the geometric measurements is also straightforwardly derived by assuming an additive Gaussian noise :

$$\begin{bmatrix} \tilde{u}_q^m(t) \\ \tilde{v}_q^m(t) \end{bmatrix} = \begin{bmatrix} u_q^m(t) \\ v_q^m(t) \end{bmatrix} + \begin{bmatrix} \beta_1^m(t) \\ \beta_2^m(t) \end{bmatrix} \tag{14}$$

where measurement noises $\beta_1^m(t)$ and $\beta_2^m(t)$ are assumed to be Gaussian random vectors of variance σ_β^2. Again, σ_ζ^2 and σ_β^2 are set empirically.

We now define notations related to the data association issue. We call K the set of assignements of measurements to trajectory models, which can be decomposed over time and measurements as follows :

$$K = \Big[K(1), \ldots, K(T)\Big] \tag{15}$$

$$K(t) = \Big\{k_1(t), \ldots, k_{s_t}(t)\Big\} \tag{16}$$

Each assignement variable $k_j(t) \quad (j = s, \ldots, s_t)$ takes values in $[1, \ldots, M]$, thereby indicating to which trajectory model the measurement j is assigned at time instant t.

Let us also introduce Π, the probability of trajectory models, which can also be decomposed over time as follows :

$$\Pi = \Big[\Pi(1), \ldots, \Pi(T)\Big] \tag{17}$$

$$\Pi(t) = \Big\{\pi_1(t), \ldots, \pi_M(t)\Big\} \tag{18}$$

Given a measurement at time t, $\pi_m(t)$ represents the probability that a measurement originates from model m, regardless of which measurement it may be. While K contains binary assignment random variables, the sets $\mathcal{X}$ and Π contain continuous random variables. Classical multi-track extraction methods (JPDAF, MHT) are based on the two following assumptions:

- the assumption that a measurement is associated to one and one trajectory model only, from which the following constraint on assignment variables is inferred :

$$\sum_{m=1}^{M} p\Big(k_j(t) = m\Big) \;=\; \sum_{m=1}^{M} \pi_m(t) = 1 \tag{19}$$

- the assumption that at most one measurement can originate from a moving object at a time. This implies a dependence of assignment variables.

In contrast, the approach we adopt, namely PMHT, relies only on the first of these two assumptions. Consequently, we assume independence of the assignment variables, which allows the factorization of the joint probability of $K(t)$ as described by :

$$p(K(t)) \;=\; \prod_{j=1}^{s_t} p(k_j(t)) \tag{20}$$

It is this very formulation which avoids enumeration of measurement-to-track association hypotheses.

4 Main theoretical aspects of PMHT

4.1 Joint estimation formulation and posterior probability

We recall in this section the main theoretical aspects of PMHT that are used in our method. The search for optimal assignments and states being two interlocking issues, Streit [28] proposed to include the data association problem in the estimation problem; more precisely, to consider the assignment variables as random variables to be estimated along with the state variables. Let us define $\Phi = (\mathcal{X}, \Pi)$. The $\{\pi_m\}_{m=1,\dots,M}$ represent the laws of the discrete variables $k_j(t)$, and estimating Φ according to the *Maximum A Posteriori (MAP)* criterion amounts to a joint estimation of assignments and states. The *a posteriori* distribution can be expressed by :

$$\begin{aligned} p(\Phi|\mathcal{Z}) \propto & p\Big(\mathcal{Z} \mid \mathcal{X}, \Pi\Big) p(\mathcal{X}, \Pi) \\ \propto & \underbrace{\prod_{t=1}^{T} p\Big(Z(t) \mid X(t), \Pi(t)\Big)}_{\text{measurement likelihood}} \qquad \underbrace{p\Big(X(1)\Big) \prod_{t=2}^{T} p\Big(X(t) \mid X(t-1)\Big)}_{\text{prior state evolution}} \end{aligned} \tag{21}$$

Our goal is to find an estimate of Φ which maximizes the posterior probability (21).

Gauvrit and Le Cadre [8] have shown that, in the above expression, the measurement likelihood term can be expressed as the product of conditional likelihoods of measurements $z(t)$, which in turn are defined as a mixture density law, in which the parameters weighing the respective contributions of the elementary laws to the mixture are the prior probabilities of the trajectory models. This can be written as follows :

$$\prod_{t=1}^{T} p\Big(Z(t) \mid X(t), \Pi(t)\Big) \; = \tag{22}$$

$$= \prod_{t=1}^{T} \prod_{j=1}^{s_t} \sum_{m=1}^{M} p\Big(z_j(t) \mid x_m(t)\Big) \pi_m(t) \tag{23}$$

An essential point is that, thanks to the independence assumption between assignment variables, writing (22) as a product of mixture laws (23) is made possible. Direct maximization of (21) is however not feasible, since it is parameterized by the unknown weights $\pi_m(t)$.

Following the work by Redner and Walker [23], the EM algorithm [6] can be used to estimate the parameters of such a mixture density, through an iterative procedure. Let us assume that an initial estimate Φ^0 is available. At the $i + 1^{th}$ iteration of the algorithm, in a first step ("E-Expectation" step), an approximation of the *a posteriori* distribution is computed, via its expectation, from measurements and current estimates Φ^i of Φ. In a second step ("M-Maximization" step), a new estimate Φ^{i+1} is computed from the approximation that has just been determined. "E" and "M" steps are alternatively

iterated until (guaranteed [6]) convergence. An appropriate and efficient initialization of the recovery problem of multiple trajectories in an image sequence is specified in the next section.

4.2 Association between partial tracks and trajectory models

Spatial proximity or other criteria can supply a short-term temporal link between measurements but, due to the possible lack of detections, in case of occlusion or crossing for instance, this link is sometimes broken. Therefore, our association problem is not more the assignement of the measurements to the trajectory models at each time instant, but the association of available partial tracks to the trajectory models. To this respect, we adapt the method proposed by Giannopoulos et al. [10] for radar and sonar data, and summarize below the main results.

Let us denote $\mathcal{P}$ the set of M partial tracks and $K_l^{\mathcal{P}}$ the assignment of partial track $\mathcal{P}^l$. This assignment takes values in $[1, \ldots, M]$. $\mathcal{P}$ and the set $K^{\mathcal{P}}$ of assignments can be decomposed as follows :

$$\mathcal{P} = \left\{ \mathcal{P}_1, \ldots, \mathcal{P}_M \right\} \tag{24}$$

$$K^{\mathcal{P}} = \left\{ K_1^{\mathcal{P}}, \ldots, K_M^{\mathcal{P}} \right\} \tag{25}$$

To apply the EM algorithm, we need to derive the expectation of the logarithm of the *a posteriori* distribution of variables Φ given an estimate Φ_i. This can be expressed as follows, starting from (21) and (23) :

$$\begin{aligned} Q(\Phi \mid \Phi^i) \;=\; & \sum_{m=1}^{M} \sum_{\mathcal{P}_l \in \mathcal{P}} w_{\mathcal{P}_l,m}^{i+1}(t)\; ln[\pi_m(t)] \\ & + \sum_{m=1}^{M} \sum_{\mathcal{P}_l \in \mathcal{P}} \sum_{z_j \in \mathcal{P}_l} ln[p(z_j(t) \mid x_m(t))\,]\; w_{\mathcal{P}_l,m}^{i+1}(t) \\ & + \sum_{m=1}^{M} ln[p(x_m(1))] + \sum_{m=1}^{M} \sum_{t=2}^{T} ln[p(x_m(t) \mid x_m(t-1))] \end{aligned} \tag{26}$$

where $w_{\mathcal{P}_l,m}^{i+1}$ is a weighing factor corresponding to the probability of assigning partial track $\mathcal{P}_l$ to model m, and is defined by :

$$w_{\mathcal{P}_l,m}^{i+1} = \prod_{z_j \in \mathcal{P}_l} \Big(\frac{\pi_m^i p(z_j \mid x_m(t))}{\sum_{m=1}^{M} \pi_m^i p(z_j \mid x_m(t))} \Big) \tag{27}$$

The maximization of $Q(\Phi \mid \Phi^i)$ can be decomposed into two independent maximizations, first with respect to the parameters of the mixture, the $\pi_m(t)$'s, and second w.r.t. to the states (i.e. the trajectory models), the $x_m(t)$'s. Through these maximizations, one updates the estimate $\Phi^i = (\Pi^i, X^i)$ at iteration $i+1$ to get $\Phi^{i+1} = (\Pi^{i+1}, X^{i+1})$.

The first maximization problem has a simple analytic solution. For every t and m, we get :

$$\pi_m^{i+1}(t) = \frac{1}{s_t} \sum_{j=1}^{s_t} w_{j,m}^{i+1}(t) \tag{28}$$

The second problem consists of the state estimation :

$$\begin{aligned}(x_m(0), \ldots, x_m(T)) \in \\ \underset{X_m}{\text{argmax}} \Bigg\{ & \sum_{\mathcal{P}_l \in \mathcal{P}} \sum_{z_j \in \mathcal{P}_l} ln(\ p(z_j(t) \mid x_m(t))\ )\ w_{j,m}^{i+1}(t) \\ & + ln[p(x_m(1))] \\ & + \sum_{t=2}^{T} ln[p(x_m(t) \mid x_m(t-1))] \Bigg\}\end{aligned} \tag{29}$$

In the case of a Markovian process, it is more relevant to maximize the exponential of the expression included in relation (29), that is :

$$p(x_m(1)) \prod_{t=2}^{T} \Big\{ p(x_m(t) \mid x_m(t-1)) \prod_{j=1}^{s_t} p(z_j(t) \mid x_m(t))^{w_{j,m}^{i+1}(t)} \Big\} \tag{30}$$

Taking advantage of the Gaussian nature of the measurement noise, this expression can be simplified by introducing a fictitious "synthetic" measurement $\tilde{z}_m(t)$ and its covariance matrix $\tilde{R}_m$, defined below (relations (32) and (33)). $\mathcal{N}[\tilde{z}_m(t), x_m(t), \tilde{R}_m]$ denotes the Gaussian probability distribution of variable $\tilde{z}_m(t)$, parameterized by its mean $x_m(t)$ and covariance matrix $\tilde{R}_m$. At each instant t, we have :

$$\prod_{j=1}^{s_t} p(z_j(t) \mid x_m(t))^{w_{j,m}^{i+1}(t)} \propto \prod_{j=1}^{s_t} \mathcal{N}[z_j(t), x_m(t), (w_{j,m}^{i+1}(t))^{-1} R_m] \propto \mathcal{N}[\tilde{z}_m(t), x_m(t), \tilde{R}_m] \tag{31}$$

$$\text{with} \qquad \tilde{z}_m(t) \ = \ \frac{1}{s_t \pi_m^{i+1}(t)} \sum_{j=1}^{s_t} w_{j,m}^{i+1}(t) z_j(t) \tag{32}$$

$$\tilde{R} \ = \ \frac{R_m}{s_t \pi_m^{i+1}(t)} \tag{33}$$

This transform leads to the classical expression (34) of the *a posteriori* distribution of the state for a *single track* :

$$p(x_m(1)) \prod_{t=2}^{T} \Big\{ p(x_m(t) \mid x_m(t-1)) p(\tilde{z}_m(t) \mid x_m(t)) \Big\} \tag{34}$$

The practical resulting algorithm is particularly simple, since the optimal estimation of $\mathcal{X}$ amounts to M independent estimations using Kalman filtering with smoothing.

5 Initialization stage and tracking algorithm

Let us stress that, in general, the result of the EM algorithm is strongly dependent on the initialization provided for the parameters to be estimated. For our problem, this means that care should be taken to provide the best possible initial guesses for each trajectory model. It is the main purpose of this section to describe the solution we propose to this issue. We expose below how, by utilizing rich information about geometry and velocity of the regions, a meaningful and robust initialization can be elaborated, leading to an original and effective PMHT multiple-object tracking scheme.

Figure 2 includes an overview of the proposed scheme. Since the true number of moving objects, and consequently of trajectories to recover in the image sequence is unknown, we initially set it to M as stated in section 3, where M is the number of partial tracks found within the batch, i.e. in the processed image sequence. The PMHT algorithm requires initializing states and prior probabilities of trajectory models. For the latter, we initially set them in a uniform way, for every instant t and for every model m: $\pi_m^0(t) = 1/M$. Then, the objective is to determine the number of actual trajectories by grouping the partial tracks through the joint trajectory estimation process introduced in section 4.

We exploit the partial tracks to build the M initial trajectories (initial states). Each trajectory model is initially assigned the measurements forming a partial track. We then estimate independently the M models over the whole sequence. Figure 3 illustrates this operation in an example involving three models. A prediction-only estimation mode is used in the Kalman filtering step at time instants when measurements are not available (dashed polygons in fig. 3).

Handling of the geometric component

Tracking of the geometric models by Kalman filters cannot be directly applied by considering that the vertices of the polygonal approximation of the segmentation mask form the measurements of the geometric component. As illustrated in Fig. 4, since polygonal approximations are carried out independently over time, even slightly time-varying segmentation masks may generate significantly different sets of polygonal approximation vertices (regarding the location and the number of these vertices). To solve this issue and supply correct vertices $\tilde{P}_r^j$ for correspondence, we operate as follows (fig. 4) : (1) the predicted polygon and the extracted one are spatially registered with a translation, minimizing the inter-polygon distance defined in [5] with local gradient-descent; (2) for each vertex of the predicted geometric component, the nearest point on the polygon extracted from the image is chosen to be the corresponding measurement.

Let us point out that the prediction/update principle applied to the geometric component by Kalman filtering enables some (limited) degree of non-rigidity in the motion (in addition to the sequence of affine transforms). More precisely, the affine transform assumption is used for the prediction step (use of the global affine motion for all the vertices of a given region), but the adjustement step is carried out locally at each vertex, hence handling, to some extent, articulation and deformation.

1. First estimation of trajectory models $X^0 = (X_1^0, \ldots, X_M^0)$ from partial tracks.

2. Detection and elimination of measurements corresponding to occlusion /dis-occlusion phases, if any.

3. Re-estimation of trajectory models $X^0 = (X_1^0, \ldots, X_M^0)$, having discarded perturbed measurements in step 2. These models serve as the initialization for the EM algorithm.

4. Initialization of the mixture model parameters

$$\Pi^0 = (\pi_1^0, \ldots, \pi_M^0), i = 0.$$

5. For each partial track and for each trajectory model, compute :

- the probability of association of the partial track to the model :

$$w_{\mathcal{P}_l,m}^{i+1} = \prod_{j \in \mathcal{P}_l} \Big(\frac{\pi_m^i p(z_j \mid x_m(t))}{\sum_{m=1}^M \pi_m^i p(z_j \mid x_m(t))} \Big)$$

- the prior probability of each trajectory model :

$$\pi_m^{i+1}(t) = \frac{1}{m_t} \sum_{j=1}^{m_t} w_{j,m}^{i+1}(t)$$

6. For each trajectory model and each time instant, compute the *a posteriori* measurement and its covariance matrix :

$$\begin{aligned} \tilde{z}_m(t) &= \frac{1}{s_t \pi_m^{i+1}(t)} \sum_{j=1}^{s_t} w_{j,m}^{i+1}(t) z_j(t) \\ \tilde{R_m} &= \frac{R_m}{s_t \pi_m^{i+1}(t)} \end{aligned}$$

7. For each trajectory model, estimate the state at each time instant by Kalman filtering with smoothing, considering the *a posteriori* measurements and their covariance matrices computed in step 6.

8. Increment $i = i + 1$ and go to step 5 if the stopping criterion (35) is not met.

9. Decision on possible association of several partial tracks.

Figure 2: Overview of the proposed scheme.

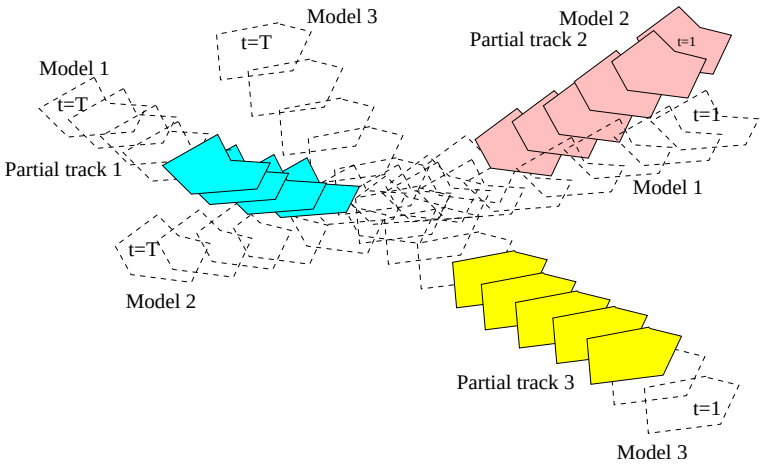

Figure 3: Building initial states, in the case of three partial trajectories (only the geometric component is shown here). Dashed lines represent temporal extensions, when a prediction-only mode is employed.

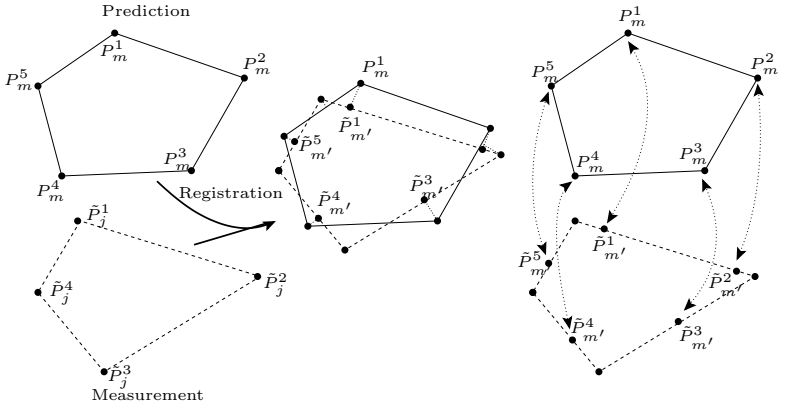

Figure 4: The two polygons, one corresponding to the prediction computed from the current region trajectory model and the other to the extracted region, are first registered using a translation. Then, for each vertex on the model, the closest point on the measurement polygon is considered, so as to attempt to obtain pairs of points that approximately correspond to the same physical point. For the sake of figure clarity, the predicted geometric model and the polygonal silhouette of the extracted region are drawn far apart.

Discarding perturbed measurements

We noticed that the reliability in the "prediction-only" mode of the state is strongly dependent on the accuracy of the last few measurements before the filter switches to this mode. Typically, these last few measurements can correspond to a progressive occlusion phase (Fig .1). Such an issue arises both for progressive appearance and disappearance of a region. The geometric component is particularly affected, since the extracted region and its measured silhouette reveal only the visible part of the object. Therefore, we decided to discard such "uncertain" measurements. We carry out detection of occlusion and disocclusion phases according to the criterion introduced in [15], since it has proved effective enough. In short, it consists in detecting unexpected strong temporal variation of the area of the tracked region support. We predict the area of this region from time t to time $t+1$, using the divergent component of the 2D motion field of the region (due to object motion towards or away from the camera, or camera motion). It can be straightforwardly computed from the 2D affine motion model (given by $\frac{1}{2}(a_2 + a_5)$) estimated over the considered region at time instant t. We then examine an "innovation" variable, which is the difference between area of the segmented region at time t, and its prediction. Temporal upward or downward jumps of this variable are then detected using Hinkley's test. Besides its simplicity, the interest of this test is two-fold. Since it is cumulative over time, it can detect (dis)occlusion phases with various speed with the same threshold. It also provides conveniently the time at which the (dis)occulsion phase starts (which is by construction a little earlier than the time at which it is detected). Once the (dis)occulsion phases have been identified, if any, the corresponding measurements are discarded, and the states of all models are re-estimated over the batch.

Iteration and convergence of the EM algorithm

From these initial state estimates and prior model probabilities, the two steps of the EM algorithm are iterated : computation of the measurement-to-model assignment probabilities given the current states, derivation of prior probabilities of models and of the "synthetic" measurements $\tilde{z}_m(t)$, estimation of the states over the batch. Convergence is considered obtained when the following condition is met:

$$\max_{j,m,t} \mid w^i_{j,m}(t) - w^{i-1}_{j,m}(t) \mid < \delta_w \tag{35}$$

The parameter δ_w is typically set to 0.001.

The key parameters of the algorithm that the user should set are the process and measurement noises. Automatic learning of appropriate values from image sequences are beyond the scope of this paper, notably because their setting should exploit application-dependent knowledge, or extensive training data.

Convergence of the EM algorithm leads to an optimal (in the sense defined of relation (21)), stable, assignment of measurements to trajectory models. A policy to recover the *full tracks*, in other words to associate *partial tracks*, can be defined on the basis of the values obtained for these assignments $w^{i+1}_{\mathcal{P}_l,m}$. In practical experiments, we observe that a clear convergence of $w^{i+1}_{\mathcal{P}_l,m}$'s to 1 or 0 occurs in most cases, respectively if two partial

tracks should intuitively clearly be associated or not. Simple thresholding below e.g. 10e-3 or above 1-10e-3 easily identifies such situation. On the other side, typical ambiguous cases include :

- two partial tracks which trajectories are not clearly the continuation of one another, but might be (this may occur in the presence of temporary occlusions) ;
- two partial tracks overlapping in time, that both are in plausible continuity of a third partial track, that occurs earlier or later.

In the first case, weights take intermediate values between 0 and 1. In the second case, the weights associating the third partial track to the two trajectory models arising initially from the two plausible matching partial tracks are typically close to 0.5, since these weights should sum to 1. Existence of such configurations may be identified.

A practical rule, in the context of region tracking, is suggested by our experiments. In [15], two trajectory models are to be grouped if, over a sufficient time interval, they are consistent both in position and velocity. In contrast, we suggest to only demand consistency in position, and leave more flexibility on the evolution of the kinematics during occlusion phases. Besides, the influence of kinematics remains via the state equation (12). Moreover, we globally handle the determination of multiple trajectories, whereas in [15], the problem is stated by considering each trajectory individually.

More generally, the probabilistic nature of the results provided by our technique opens interesting perspectives for variations in the decision-taking phase. The present paper proposes a technique for *inferring* the association probabilities. From there, one may introduce some cost associated to each type of error, depending on the application, and apply various decision strategies (Bayesian, minimax,...) to conclude. Finally, formalisms that penalize overall complexity in explaining the scene may be introduced to supply automatically an interpretation of the scene, by trading trajectory continuity for global scene simplicity.

6 Experimental results

We report experimental results for two real image sequences involving complex situations. The first one is the "Breakfast" sequence, acquired in our lab and which was already described in Section 1 (Fig.1). The scene comprises four partial tracks : two per object, as each object undergoes temporary total occlusion. Then, four trajectory models are initially created and estimated. At convergence, finally two global trajectories are retained and estimated. For this sequence, initial and final estimated trajectory models are respectively plotted on Fig. 5a and 5b, with measurements. It can be noticed that, at convergence of our algorithm, the four partial tracks are correctly grouped in two pairs, despite the relatively complex crossing situation. Only the gravity centers of the geometric models are indicated for clarity sake.

Fig. 6a and 6b respectively show the computed geometric measurements, and the estimated geometric models at convergence, superimposed over the first image of the

sequence. The algorithm supplies relevant geometric models, including the whole silhouette of the regions at instants when partial or total occlusions take place. Convergence is obtained in about 20 iterations for this sequence.

As an example, a result for the kinematic model is provided in Fig. 7, for the translational parameter a_1 of the motion model. Measurements and estimated values of a_1 are plotted for two trajectory models corresponding to two partial tracks in the "Breakfast" sequence, that should be associated. They are provided at initialization (Fig 7a,b) and at convergence (Fig 7c,d) of the EM algorithm. The (conservative) prediction-only mode employed for estimating the kinematic model when no measurement is available consists in keeping the last filtered value available constant. The need for this switching of evolution model arises from the following observation : the last few measurements before switching to prediction-only mode (e.g. corresponding to a occlusion) are not reliable enough to allow long-term in prediction-only mode based on a higher-order evolution model on motion parameters, so this simpler model is only employed in this context. As the two partial tracks are correctly associated at convergence, it appears that the state estimation corresponds to Kalman smoothing.

The second sequence depicts an outdoor scene. The "Van" sequence is a crossroads scene (a few images of the sequence are displayed in Fig 8a), in which the white vehicule (partial track 2) crosses (behind) a van (partial track 1), and reappears on its left (partial track 3). Fig 8b shows the corresponding motion-based segmentation maps. The dark car closely following the van is not differenciated by the motion-segmentation scheme from the van it is following, as their motions are very similar. Due to the short-term linkage provided by the motion segmentation algorithm, three partial tracks and associated object trajectory models are generated for the sequence, two of which actually correspond to the same white vehicule. Values of the kinematic measurements and estimated motion models, exemplified by a_1, are provided in Fig. $8c_1$ and $8c_2$ respectively at initialization and at convergence of the EM algorithm. It can be observed that model 2 fits partial track 3, while model 3 mismatches partial track 2. As explained in the previous section, we state that a one-direction fit suffices to associate the two partial tracks at hand. The evolution of the association weights $w_{\mathcal{P}_l,m}$ over iterations is supplied, for trajectory models 2 and 3 with partial track 3, in Fig $8c_3$. Hence, our tracking method was able to correctly decide that there were only two relevant different entities (i.e., $\mathcal{M} = 2$), and to accurately recover the corresponding two entire trajectories, despite the first partial, then total occlusion, and the crossing situation.

The running time of the technique on a 60-image batch is about 2 seconds (C++ implementation) for the data association part, which is the contribution of this paper. The processing time required by prior motion segmentation from the image sequence is about an order of magnitude higher.

The MHT technique is based on the NP-complete enumeration of association hypotheses, usually requiring application of pruning techniques to the hypotheses tree. In the PMHT technique, computational complexity only grows moderately with the number of partial tracks. The examples considered here only involve a few regions and computational cost should be low both for MHT and PMHT. In general, however, PMHT posseses three advantages for the region-tracking problem:

- The more computationally-expensive features are added to the regions (e.g. the geometric features, included in this paper; color distribution, as a valuable extension), the greater the computational advantage of PMHT over hypothesis enumeration. Besides, introduction of pruning/gating techniques for MHT would require ad-hoc tuning for each feature.

- The context chosen was that of a availability of a short-term link between regions. In situations where this link does not exist, the combinatorial issue is strong even for sequences such as the ones presented in the paper.

- Besides combinatorial issue, there is an intrinsic advantage in probabilistic modelling of the associations, in that it takes naturally into account uncertainties on measurements and models, and also provides confidence evaluation as an output and hence enabling various decision-taking policies.

7 Conclusion

We have presented an original and efficient method for tracking multiple objects in an image sequence. It involves the association of partial tracks of regions, while jointly estimating the trajectories of these regions. We have introduced the modelling of geometric and kinematic components of regions in the PMHT framework. From an adequate model initialization scheme, an iterative EM procedure leads to a stable configuration of trajectory models from which associations can be inferred and entire trajectories of the physical moving objects recovered. The proposed tracking method has been validated by experiments on real image sequences involving complex events such as partial occlusion, total (temporary) occlusion and crossing.

The practical interest of the proposed method is several fold. The understanding of the sequence content is improved and a rich description of the content is provided: region motions and trajectories with the whole silhouette of objects are estimated over the whole sequence, including when measurements are either not available, or not reliable. A possible major improvement on the performance of the scheme could be obtained by adding intensity or color related descriptors to the measurements, and modelling their temporal evolution, as for instance described in [11].

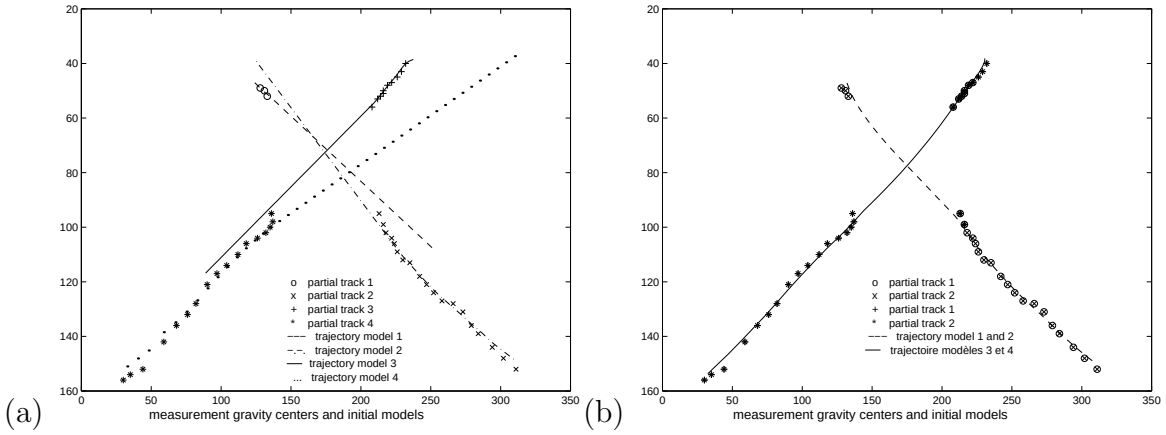

Figure 5: "Breakfast" sequence : measurements and four initially estimated partial trajectories (a) and the two finally estimated global trajectories at convergence (b). Only the gravity centers of the geometric models are displayed.

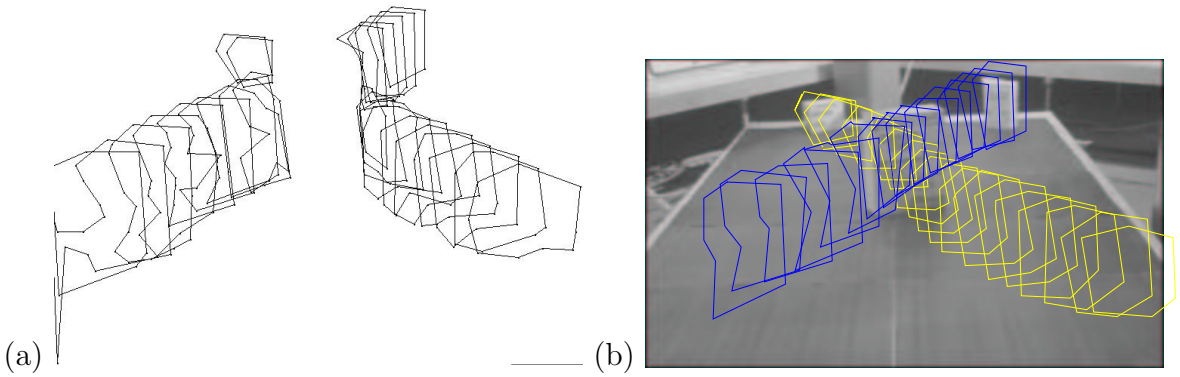

Figure 6: "Breakfast" sequence : measured polygonal silhouettes (a), estimated geometric models at convergence, superimposed on the original image at $t = 0$ (b). For the sake of clarity, only one out of two geometric models (in time) are represented.

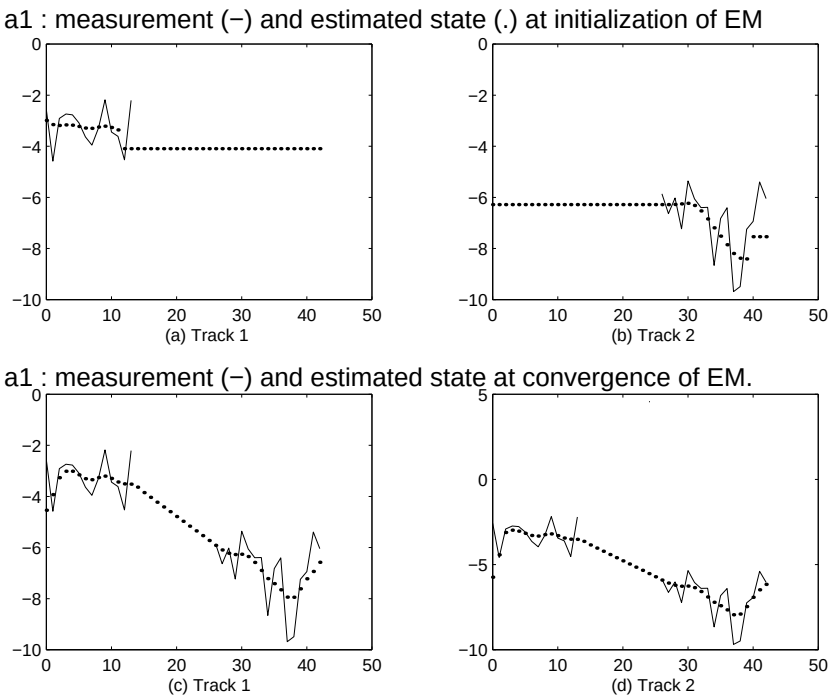

Figure 7: "Breakfast" sequence : estimated (filtered) values (dotted line) of parameter a_1 (kinematic component) for two of the four trajectory models, plotted at initialization (a,b) and at convergence (c,d) of the EM algorithm.

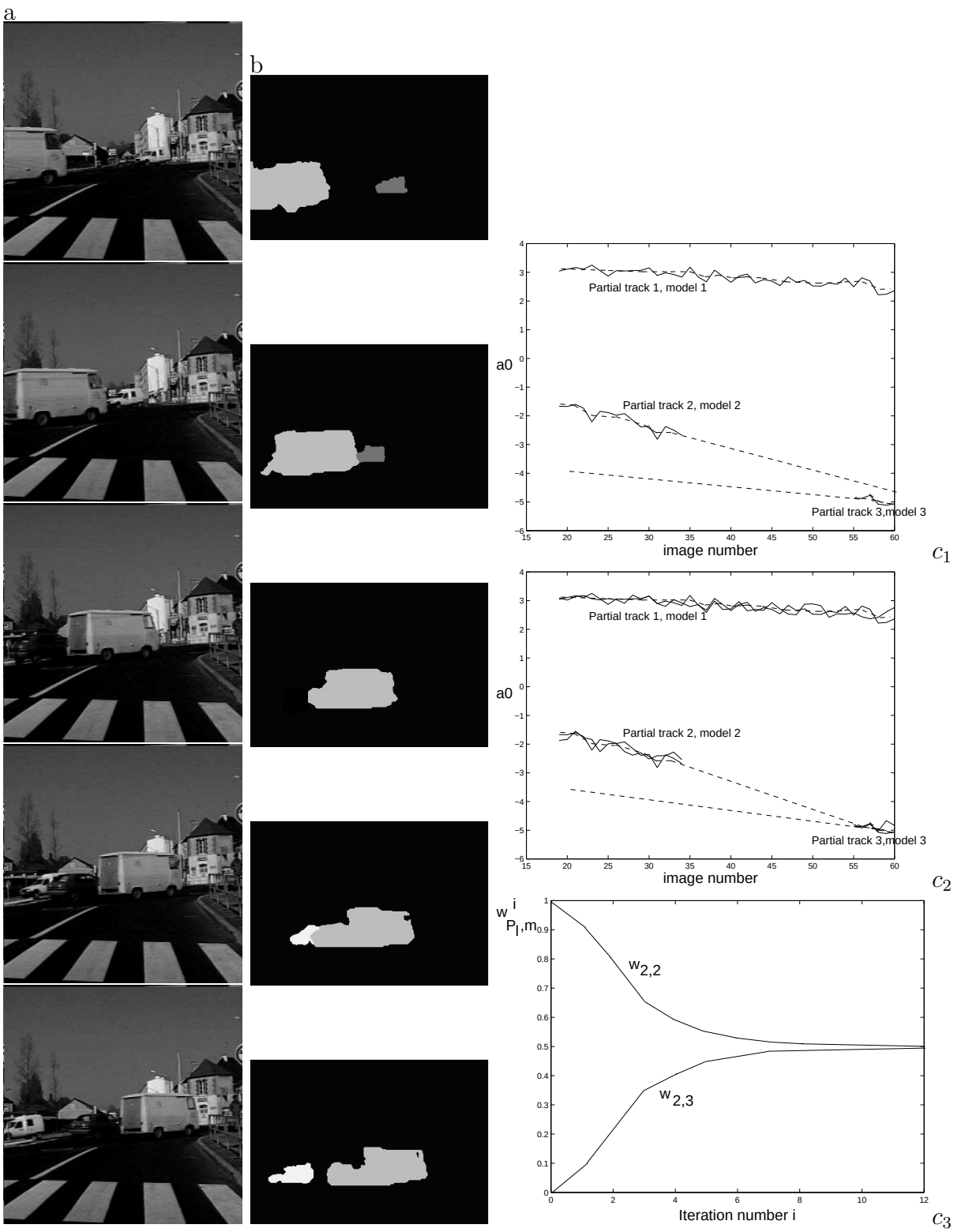

Figure 8: Column (a): images from the "Van" sequence, at time instants $t = 19, 31, 47, 55, 61$. Column (b): obtained motion segmentation maps for these images. Column (c): evolution over the sequence of the affine motion parameter a_1 for the three models and three partial tracks, at initialization (c_1) and at convergence of the EM algorithm (c_2), evolution over the iterations of association weights $w^i_{\mathcal{P}_l,m}$, for $l = 2$, $m = 2$ and $m = 3$.

References

[1] Y. Bar-Shalom and X.R. Li. *Estimation and Tracking : Principles, Techniques and Software*. Artech House, Boston, 1993.

[2] A. Blake and M. Isard. *Active contours*. Springer, 1998.

[3] I.J. Cox. A review of statistical data association techniques for motion correspondance. *Int. Journal of Computer Vision*, 10(1):53–66, 1993.

[4] I.J. Cox and S.L. Hingorani. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(2):138–150, February 1996.

[5] P. Cox, H. Maître, M. Minoux, and C. Ribeiro. Optimal matching of convex polygons. *Pattern Recognition Letters*, (9):327–334, June 1989.

[6] A.P. Dempster, N.M Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Ser. B*, 39:1–38, 1977.

[7] T.E. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Research*, pages 173–184, July 1983.

[8] H. Gauvrit, C. Jauffret, and J.P. Le Cadre. A formulation of multitarget tracking as an incomplete data problem. *IEEE Trans. on Aerospace and Electronic Systems*, 33(4):1242–1257, October 1997.

[9] M. Gelgon, P. Bouthemy, and J-P. Le Cadre. Associating and estimating trajectories of multiple moving regions with a probabilistic multi-hypothesis tracking approach. In *Proceedings of Int. Symposium of Physics in Image Processing*, pages 80–83, Paris, January 1999.

[10] E. Giannopoulos, R. Streit, and P. Swaszek. Multi-target track segment bearing-only association and ranging. In *31rst Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, November 1997.

[11] R. Hammoud and R. Mohr. Mixture densities for video objects recognition. In *International Conference on Pattern Recognition (ICPR'2000)*, pages 71–75, Barcelona, Spain, September 2000.

[12] C. Hue, J-P. Le Cadre, and P. Pérez. Tracking multiple objects with particle filtering. *IEEE Trans. on Aerospace and Electronic Systems*, 38(3):791–812, July 2002.

[13] M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 1(29):5–28, 1998.

[14] F. Marques and C. Molina. Object tracking for content-based functionalities. In *SPIE Visual Communication and Image Processing (VCIP-97)*, volume 3024, pages 190–198, San Jose, 1997.

[15] F. Meyer and P. Bouthemy. Exploiting the temporal coherence of motion for linking partial spatio-temporal trajectories. In *Proc of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 746–747, New-York, June 1993.

[16] F. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long image sequences. *CVGIP : Image Understanding*, 60(2):119–140, September 1994.

[17] K.J. Molnar and J.W Modestino. Application of the EM algorithm for the multitarget/multisensor tracking problem. *Signal Processing*, 46(1):115–128, January 1998.

[18] F. Moscheni, S. Bhattacharjee, and M. Kunt. Spatiotemporal segmentation based on region merging. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(9):897–915, September 1998.

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

[20] J.M. Odobez and P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. *Signal Processing*, 66(3):143–156, May 1998.

[21] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:266–280, March 2000.

[22] C. Rasmussen and G.D. Hager. Joint probabilistic techniques for tracking multi-part objects. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pages 18–26, Santa-Barbara, June 1998.

[23] R.A. Redner and H.F Walker. Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics - SIAM Review*, 26(2):195–239, 1984.

[24] D.B Reid. An algorithm for tracking multiple targets. *IEEE Trans. on Automatic Control*, 24(6):843–854, December 1979.

[25] M. Ringer and J. Lasenby. Multiple hypothesis tracking for automatic optical motion capture. In *Proc. of European Conference on Computer Vision (ECCV'2002)*, pages 524–536, Copenhaguen, Denmark., May 2002.

[26] A. Strandlie and J. Zerubia. Particle tracking with iterated Kalman filters and smoothers: the PMHT algorithm. *Computer Physics Communications*, 123(1-3):77–86, 1999.

[27] R.L. Streit. *Studies in Probabilistic Multi-Hypothesis Tracking and Related Topics*, volume SES 98-01. Naval Underwater Warfare Center Division, February 1998.

[28] R.L. Streit and T.E. Luginbuhl. A probabilistic multi-hypothesis tracking algorithm without enumeration and pruning. In *Proc. of the 6th Joint Service Data Fusion Symposium*, pages 1015–1024. Laurel, June 1993.

[29] J-P. Tarel, S-S. Ieng, and P. Charbonnier. Using robust estimation algorithms for tracking explicit curves. In *Proc. of European Conference on Computer Vision (ECCV'2002)*, pages 492–507, Copenhaguen, May 2002.

[30] K. Wall and P.E. Danielsson. A fast sequential method for polygonal approximation of digitized curves. *Computer Vision, Graphics, and Image Processing*, (28):220–227, 1984.

[31] M.A. Zaveri, U.B. Desai, and S.N. Merchant. Pmht based multiple point targets tracking using multiple models in infrared image sequence. In *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS'03)*, pages 73–79, Miami, USA, July 2003.

[32] Z. Zhang and O Faugeras. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. *Int. Journal of Computer Vision*, 7(3):211–241, 1992.