Published in Int. Conference on Pattern Recognition (ICPR), Cambridge, aug. 2004.

Embedding motion in model-based stochastic tracking

J.M. Odobez and D. Gatica-Perez*
IDIAP Research Institute, Martigny, Switzerland
{odobez,gatica}@idiap.ch

Abstract

Particle filtering (PF) is now established as one of the most popular methods for visual tracking. Within this framework, two assumptions are generally made. The first is that the data are temporally independent given the sequence of object states, and the second one is the use of the transition prior as proposal distribution. In this paper, we argue that the first assumption does not strictly hold and that the second can be improved. We propose to handle both modeling issues using motion. Explicit motion measurements are used to drive the sampling process towards the new interesting regions of the image, while implicit motion measurements are introduced in the likelihood evaluation to model the data correlation term. The proposed model allows to handle abrupt motion changes and to filter out visual distractors when tracking objects with generic models based on shape representations. Experimental results compared against the CONDENSATION algorithm have demonstrated superior tracking performance.

1. Introduction

Visual tracking is an important problem in computer vision. Although intensively studied in the literature, tracking is still a challenging task in adverse situations. In the pursuit of robust tracking, particle filtering [4, 2] has shown to be a successful approach. In this temporal Bayesian framework, the posterior distribution is represented by a set of weighted random samples allowing to maintain multiple-hypotheses in the presence of ambiguities, unlike algorithms that keep only one configuration state [3], which can be therefore sensitive to single failure.

Visual tracking with a particle filter requires the definition of two main elements : a data likelihood and a dynamical model. The first term evaluates the likelihood of an observation given the object state. Parameterized shapes [2, 16] and color distributions [10, 3, 8, 16] are often used as target representation. One drawback of these generic representations is that they are quite unspecific which augment the chances of ambiguities. Combining low-level measurements such as shape and color [16], or using appearence-based models such as templates [13, 14] can render the target more discriminative. The latter representations, however, do not allow for large changes of appearence, unless more complex global models are used [1, 15].

The dynamical model characterizes the prior on the state sequence. A common assumption in particle filtering approaches is to use the dynamics as proposal distribution (the function that predicts the new state hypotheses) raising difficulties in the modeling since this term should fulfill two contradictory objectives. On one hand, as prior, dynamics should be tight enough to avoid the tracker being confused by distractors in the vicinity of the true object configuration, a common situation with unspecific object representations. On the other hand, as proposal distribution, it should be broad enough to cope with abrupt motion changes.

Besides, the prior distribution does not take into account the most recent observations. Thus, particles drawn from it will probably have a low likelihood, which results in a low efficiency of the sampling mechanism. Different approaches have been proposed to address these issues. For instance, auxiliary information, if available, can be used to draw samples from [5]. [9] proposed another auxiliary particle filter, whose idea is to increase or decrease (through resampling) the number of descendents of a sample according to a"predicted" likelihood estimated on the new data. This method works well only if the variance of the transition prior is small, which is usually not the case in visual tracking. [12] proposed to use the unscented particle filter to generate importance densities. Although attractive, the method needs to convert likelihood evaluations (e.g. color) into state space measurements (e.g. translation, scale). In [12], only a translation state is considered.

In this paper we propose a new PF tracking method based on visual motion with two main novelties. First, we argue that a standard hypothesis, namely the independence of observations given the state sequence [1, 2, 5, 12, 15, 16], is inaccurate in the case of visual tracking. In this view, we propose a model that assumes that the current observations depend on the current and previous object configurations as well as on the past observations. This model can be exploited to introduce an implicit object motion likelihood in the data term. Second, we exploit explicit motion measurements in the proposal distribution and in the likelihood term. The benefits of this new model are two-fold. On one hand, it increases the sampling efficiency by handling unexpected motion, allowing for a reduced noise variance in the propagation process. On the other hand, the introduction of data-correlation between successive images will turn generic trackers like shape trackers into more specific ones without resorting to complex appearence based models. As a consequence, it reduces the sensitivity of the algorithm to the difference noise variances setting in the proposal and prior since, when using larger values, potential distractors should be filtered out by the introduced correlation and motion measurements.

In Section 2, we motivate our approach and describe the proposed model. Section 3 presents the results and some concluding discussions.

2. Approach, motivation, and algorithm

Particle filtering (PF) implements a recursive Bayesian filter by Monte-Carlo simulations. Let $c_{0:k} = \{c_l, l = 0, \ldots, k\}$ (resp. $z_{1:k} = \{z_l, l = 1, \ldots, k\}$) represents the sequence of states (resp. of observations) up to time k. Furthermore, let $\{c^i_{0:k}, w^i_k\}_{i=1}^{N_s}$ denote a set of weighted samples that characterizes the posterior

* The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research on "Interactive Multimodal Information Management (IM2).

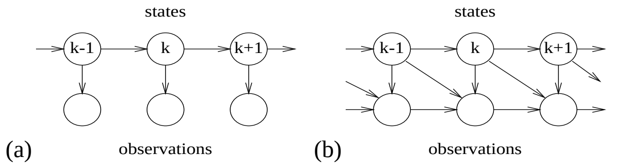

Figure 1. Graphical models for tracking. (a) standard model (b) proposed model.

probability density function (pdf) $\mathrm{p}(c_{0:k}|z_{0:k})$, where $\{c^i_{0:k}, i = 1, \ldots, N_s\}$ is a set of support points with associated weights w^i_k. The samples and weights can be chosen using the Sequential Importance Sampling (SIS) principle, which leads to the following recursive update equation [4]:

$$w^i_k \propto w^i_{k-1} \frac{\mathrm{p}(z_k|c^i_{0:k}, z_{1:k-1})\mathrm{p}(c^i_k|c_{0:k-1}, z_{1:k-1})}{q(c^i_k|c^i_{0:k-1}, z_{1:k})}, \quad (1)$$

$$= w^i_{k-1}\ \mathrm{p}(z_k|c^i_k) \quad (2)$$

where q is the proposal function from which the new samples c^i_k are drawn, and Eq. 2 derives from three common hypotheses :

H1 : observations $\{z_k\}$ are independent given the sequence of states. Hence, $\mathrm{p}(z_k|c_{0:k}, z_{1:k-1}) = \mathrm{p}(z_k|c_k)$;

H2 : the state sequence $c_{0:k}$ follows a first-order Markov chain model, characterized by $\mathrm{p}(c_k|c_{k-1})$;

H3 : the prior distribution $\mathrm{p}(c_{0:k})$ is employed as proposal. Hence, $q(c_k|c_{0:k-1}, z_{1:k}) = \mathrm{p}(c_k|c_{k-1})$.

To avoid sampling impoverishment, an additional resampling step is necessary [4]. Altogether, we obtain the standard PF :

1. Initialisation : $\forall i_{i\in 1:N_s}$, sample $c^i_0 \sim \mathrm{p}(c_0)$; set $k = 1$
2. IS step: $\forall i$ sample $\tilde{c}^i_k \sim q(c^i_k|c^i_{0:k-1}, z_{1:k})$; evaluate $\tilde{w}^i_k$ using (1) or (2).
3. Selection: Resample N_s particles $\{c^i_k, w^i_k = \frac{1}{N_s}\}$ from the sample set $\{\tilde{c}^i_k, \tilde{w}^i_k\}$; set $k = k + 1$; go to step 2.

2.1. Motivations

Conditional independence of data. In visual tracking, hypothesis H1 may not be very accurate. Usually, the configuration state includes the parameters of a geometric transformation $\mathcal{T}$. Then, the measurements consist of implicitly or explicitly extracting some part $\tilde{z}_{c_k}$ of the image by :

$$\tilde{z}_{c_k}(\mathbf{r}) = z_k(\mathcal{T}_{c_k}\mathbf{r}) \qquad \forall \mathbf{r} \in R, \quad (3)$$

where $\mathbf{r}$ denotes a position, R denotes a fixed reference region, and $\mathcal{T}_{c_k}\mathbf{r}$ represents the application of the transform $\mathcal{T}$ parameterized by c_k to $\mathbf{r}$. However, if c_{k-1} and c_k correspond to two consecutive states of a given object, it can easily be seen that $\tilde{z}_{c_k}$ and $\tilde{z}_{c_{k-1}}$ are strongly correlated (Fig. 2). Thus, the independence of the data given the sequence of states is not a strictly valid assumption. A better model is given by $\mathrm{p}(z_k|c_{0:k}, z_{1:k-1}) = \mathrm{p}(z_k|z_{k-1}, c^i_k, c^i_{k-1})$ (cf graphical model of Fig. 1b). which can be incorporated in the particle framework. For instance, keeping hypotheses H2 and H3, derivations lead to:

$$w^i_k \propto w^i_{k-1}\ \mathrm{p}(z_k|z_{k-1}, c^i_k, c^i_{k-1}) \quad (4)$$

in replacement of equation (2) (derivation details in [11]).

Proposal and dynamical model. Finding a good dynamical model is very difficult because of presence of fast and unexpected motions, due either to camera or object movments. To illustrate this, let us consider the following simple dynamical model :

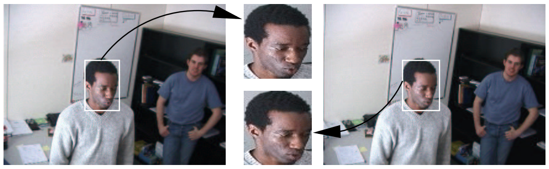

Figure 2. Images at time t and $t + 2$. The two local patches extracted from the two images are strongly correlated.

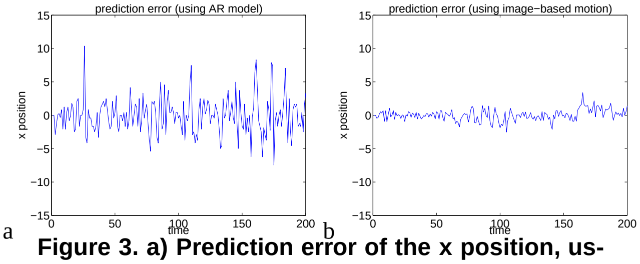

Figure 3. a) Prediction error of the x position, using an AR2 model. b) Prediction error, but exploiting the inter-frame motion estimation.

$$c_k = c_{k-1} + \dot{c}_{k-1} + w_k\ , \quad (5)$$

where $\dot{c}$ denotes the state derivative and models the evolution of the state. As state, consider the horizontal position of the head of the sequence in Fig. 6. Fig. 3a reports the prediction error w calculated using ground-truth data and obtained by estimating $\dot{c}$ with an auto-regressive model ($\dot{c}_k = c_k - c_{k-1}$). As can be seen, this prediction is noisy (σ_w=2.7) . Furthermore, there are large peak errors (up to 30% of the head width). To cope with these peaks, the noise variance in the dynamics, used as proposal distribution, has to be set to a large enough value, with the downside that many particles are wasted in low likelihood areas, or spread on local distractors that can ultimately lead to tracking failure. On the other hand, exploiting the inter-frame motion to estimate $\dot{c}$ and predict the new state value (see Section 2.2) can lead to a reduction of both the noise variance and of the error peaks (Fig. 3b, σ_w=0.83). Despite needing more computation resources, inter-frame motion estimates are usually more precise than auto-regressive models to predict new state values of geometric transformation parameters; as a consequence, they are a better choice when designing a proposal function. This observation is supported by experiments on other parameters -vertical position, scale- and on other sequences.

2.2. The proposed model

Object representation and state space. An object is represented by a region R subject to some valid geometric transformation, and is characterized by a shape. The chosen transformation comprises a translation $\mathbf{T}$, a scaling factor s, and an aspect ratio e. A state is defined as $c_k = (\alpha_k, \alpha_{k-1})$ where $\alpha = (\mathbf{T}, s, e)$.

Proposal distribution : We use inter-frame motion estimates to predict the new state values. More precisely, an affine displacement model d_Θ is computed using a gradient-based robust and multiresolution estimation method [7]. Owing to the robustness of the estimator, an imprecise region definition $R(c_k)$ due

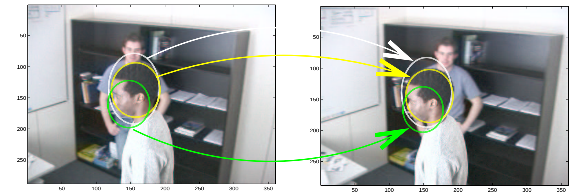

Figure 4. Example of motion estimates between two images from noisy states (the 3 ellipses). Though the estimation support regions may only cover part of the head and enclose textured background, the head motion estimate is still good.

to a noisy state value does not sensibly affect the estimation (see Fig. 4). Moreover, the algorithm delivers the covariance matrix of the affine parameters. From these estimates, we can easily construct an estimate $\widehat{\dot{\alpha}}_k$ of the variation of the coefficients between the two instant, with their variance $\widehat{\dot{\Lambda}}_k$. Denoting the predicted value $\widehat{\alpha}_{k+1} = \alpha_k + \widehat{\dot{\alpha}}_k$, and assuming the noise on the estimate $\widehat{\dot{\alpha}}_k$ independent of the noise process w in Eq. 5, we define the proposal distribution as :

$$q(c_{k+1}|c_{0:k}, z_{1:k+1}) \propto \mathcal{N}(\alpha_{k+1}; \widehat{\alpha}_{k+1}, \widehat{\Lambda}_{k+1}) \tag{6}$$

where $\mathcal{N}(.;\mu,\Lambda)$ represents a Gaussian distribution with mean μ and variance Λ, and $\widehat{\Lambda}_{k+1} = \widehat{\dot{\Lambda}}_k + \Lambda_{w_p}$, Λ_{w_p} being the variance of the process noise w_p.

Dynamics definition. We use a standard second order AR model for each of the components of α. However, to account for outliers and reduce the sensitivity of the prior in the tail, we model the noise process with a Cauchy distribution $\rho_c(x,\sigma^2) = \frac{\sigma}{\pi(x^2+\sigma^2)}$. This leads to :

$$\mathrm{p}(c_{k+1}|c_k) = \prod_{j=1}^{4} \rho_c\left(\alpha_{k+1,j} - (2\alpha_{k,j} - \alpha_{k-1,j}), \sigma_{w_d,j}^2)\right) \ . \tag{7}$$

where σ_{w_d} denotes the dynamics noise variance.

Data likelihood modeling. We modeled the data likelihood as :

$$\mathrm{p}(z_k|z_{k-1}, c_k, c_{k-1}) = \mathrm{p}_{sh}(z_k^s|c_k)\mathrm{p}_c(z_k^g|z_{k-1}^g, c_k, c_{k-1}) \ , \tag{8}$$

with $z_k = (z_k^s, z_k^g)$ and z_k^s (resp. z_k^g) denotes the shape (resp. the gray-level) measurements, and where p_c models the correlation between the two observations and p_{sh} is a shape likelihood. This choice decouples the model of the dependency between two images, whose implicit goal is to ensure that the object trajectory follows the optical flow field from the object shape model. We assumed that these two terms are independent [11].

• *Object shape observation model.* The observation model assumes that objects are embedded in clutter. Edge-based measurements are computed along L normal lines to a hypothesized contour, resulting for each line l in the nearest edge position $\{\hat{\nu}_m^l\}$ relative to a point lying on the contour ν_0^l. With some usual assumptions [2], the shape likelihood $\mathrm{p}_{sh}(z_k|c_k)$ can be expressed as

$$\mathrm{p}_{sh}(z_k|c_k) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{\|\hat{\nu}_m^l - \nu_0^l\|^2}{2\sigma^2})\right) , \tag{9}$$

where K is a constant used when no edges are detected.

• *Image correlation measurement.* We model this term as :

$$\mathrm{p}_c(z_k^g|z_{k-1}^g, c_k, c_{k-1}) \propto \mathrm{p}_{c1}(\widehat{\alpha}_k, \alpha_k)\mathrm{p}_{c2}(\tilde{z}_{c_k}^g, \tilde{z}_{c_{k-1}}^g) \tag{10}$$

with

$$\mathrm{p}_{c1}(\widehat{\alpha}_k, \alpha_k) \propto \mathcal{N}(\widehat{\alpha}_k; \alpha_k, \widehat{\Lambda}_k) \tag{11}$$

$$\mathrm{p}_{c2}(\tilde{z}_{c_k}^g, \tilde{z}_{c_{k-1}}^g) \propto \exp^{-\lambda_c \mathrm{d}_c(\tilde{z}_{c_k}^g, \tilde{z}_{c_{k-1}}^g)} \tag{12}$$

where d_c denotes a distance between two image patches. The first pdf compares the parameter values predicted using the estimated motion with the sampled values. This term assumes a Gaussian noise process in parameter space. This assumption, however, is only valid around the predicted value. To introduce a non-Gaussian modeling, we use a second term that compares directly the patches around c_k and c_{k-1}. Its purpose is illustrated using Fig. 4. While all the three predicted configurations will be weighted equally from p_{c1} (assuming their estimated variance are approximately the same), the second term p_{c2} will downweight the two predictions whose corresponding support region is covering part of the background which is undergoing a different motion than the head. The definition of p_{c2} requires the specification of a patch distance. Many such distances have been defined in the literature [13, 15]. We use the normalized-cross correlation coefficient defined as :

$$\mathrm{d}_c(\tilde{z}_1, \tilde{z}_2) = \frac{\sum_{\mathbf{r}\in R}\left(\tilde{z}_1(\mathbf{r}) - \bar{\tilde{z}}_1\right) \cdot \left(\tilde{z}_2(\mathbf{r}) - \bar{\tilde{z}}_2\right)}{\sqrt{\mathrm{Var}(\tilde{z}_1)}\sqrt{\mathrm{Var}(\tilde{z}_2)}} \tag{13}$$

where $\bar{\tilde{z}}_1$ represents the mean of $\tilde{z}_1$ [11].

3. Results

To illustrate the method, we consider two sequences involving head tracking. Three configurations of the tracker are considered. The first model (M1) is CONDENSATION [2], which corresponds to the shape likelihood combined with the same AR model with Gaussian noise for the proposal and the prior. The second model (M2) corresponds to CONDENSATION, with the addition of the implicit motion likelihood term in the likelihood evaluation (i.e now equal to $\mathrm{p}_{sh}\cdot\mathrm{p}_{c2}$). This method does not use explicit motion measurements. The third model (M3) is the full model. For this model, the motion estimation is not performed for all particles since it is robust to variations of the support region. At each time, the particles are clustered into K clusters. The motion is estimated using the mean of each cluster and exploited for all the particles of the cluster. Currently we use max(20,N_s/10) clusters. For 200 particles, the M1 tracker runs in real time (on a 2.5GHz machine), M2 at 20 image/s, and M3 at around 4 image/s.

The first example is a 12 s sequence of 330 frames (Fig. 5) extracted from a hand-held home video. Table 1 reports the tracking performance of the three trackers for different dynamics and sampling rates (all other parameters are left unchanged). A tracking failure is considered when the tracker looses the head and locks on another part of the image. As can be seen, while CONDENSATION performs quite well for tuned dynamics (D1), it breaks down rapidly, even for slight increases of dynamics variances (D2 to D4). Fig. 5 illustrates a typical failure due to the small size of the head at the begining of the sequence, the low contrast And the clutter. On the other hand, the implicit tracker M2 performs well under almost all circumstances, showing its robustness against clutter, partial measurements (around time t_{250} and partial occlusion (end of the sequence). Only when the number of samples is low (100 in S2) does the tracker fail. These failures are occuring at different parts of the sequence. Finally, in all experiments, the M3 tracker produces a correct tracking rate equal to 98%, even with a small number of samples, up to the partial occlusion. At this part of the sequence, as the occlusion reaches 50% of the tracked head,

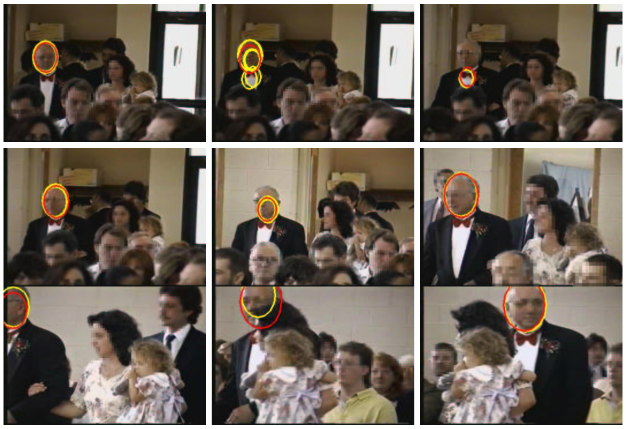

Figure 5. top row : CONDENSATION at time t_1, t_8 and t_{15} (N_s=500). center and bottom : M2 tracker (N_s=200) at time t_{20}, t_{60}, t_{165}, t_{250}, t_{295}, t_{305}. In red, mean shape. In yellow, highly likely particles.

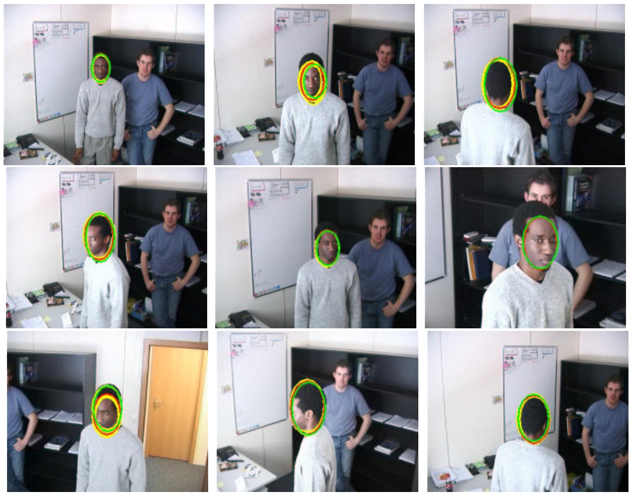

Figure 6. Tracker with motion proposal (N_s=1000) at time t_2, t_{40}, t_{85}, t_{100}, t_{130}, t_{145}, t_{170}, t_{195}, and t_{210} . In red, mean shape; in green, mode shape; in yellow, likely particles.

the motion estimation sometimes lock onto the woman's head motion, leading to the reported tracker failures.

The second sequence (Fig. 6) illustrates more clearly the benefit of using the motion proposal. This 24s sequence acquired at 12 frame/s is specially difficult because of the occurence of several head turns and abrupt motion changes (translations, zooms) the large variations of scale, and importantly, the absence of head contours as the head moves in front of the bookshelves. Because of these, CONDENSATION is again lost very quickly. On the other hand, the M2 tracker successfuly tracks the head at the beginning, but usually gets lost when the person moves in front of the bookshelves (around frames t_{130}-t_{145}), due to the lack of contour measurements coupled with a large zooming effect. This latter problem is resolved by the motion proposal, which better capture the state variations, and allows a successful track of the head until the end of the sequence (time t_{340}). More results can be found in [11].

Tracker	D1	D2	D3	D4	S1	S2
N_s	500				200	100
$\sigma_{\mathbf{T}}$	2	3	5	8	5	
σ_s	0.01			0.02	0.01	
CONDENS.	88	36	2	0	0	0
M2 (Implicit)	100	98	100	94	90	50
M3 (see text)	70	82	92	90	96	80

Table 1. Successful tracking rate (in %, out of 50 trials with different seeds). $\sigma_{\mathbf{T}}$ (resp. σ_s) denotes the dynamics and proposal noise standard deviation of the $\mathbf{T}$ (resp. s) state components.

4. Conclusion

We presented a methodology to embed data-driven motion into particle filters. This was first achieved by introducing a likelihood term that models the temporal correlation existing between successive images of the same object. Secondly, a data-driven approach based on explicit motion estimates is used to design better proposals that take into account the new image. Altogether, the algorithm allows to better handle unexpected and fast motion changes, to remove tracking ambiguities that arise when using generic shape-based object models, and to reduce the sensitivity to the different parameters of the prior model. The method is general and could also be applied to the tracking of deformable objects [6].

References

[1] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *ECCV*, pp 909–924, 1998.

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

[3] D. Comaniciu, V. Ramesh, P. Meer. Real-time tracking of non-rigid objects using mean shift. In *CVPR*, 2000.

[4] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[5] M. Isard and A. Blake. ICONDENSATION : Unifying low-level and high-level tracking in a stochastic framework In *5th ECCV*, pp 893-908, 1998.

[6] C. Kervrann, F. Heitz and P. Pérez Statistical model-based estimation and tracking of non-rigid motion In *ICPR*, 1996.

[7] J.-M. Odobez and P. Bouthemy Robust multiresolution estimation of parametric motion models In *Jl of Visual Com. and Image Representation*, vol 6, num. 4, pp 348-365, 1995.

[8] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *ECCV* , pp 661–675, June 2002.

[9] M.K. Pitt and N. Shephard. Filtering via Simulation: Auxiliary Particle Filters In *Journal of the American Statistical Association*, pp 590–599, vol. 94, num. 446, 1999.

[10] Y. Raja, S. McKenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. In *ECCV*, 1998.

[11] J-M. Odobez and D. Gatica-Perez Embedding Motion in Model-Based Stochastic Tracking IDIAP RR-03-72, 2003.

[12] Y. Rui and Y. Chen. Better proposal distribution: object tracking using unscented particle filter In *CVPR*, 2001.

[13] J. Sullivan and J. Rittscher Guiding random particles by deterministic search. In *ICCV*, pp 323–330, 2001.

[14] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE PAMI*, 24(1):75–89, 2001.

[15] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc.* 8^{th} *ICCV*, 2001.

[16] Y. Wu and T. Huang. A co-inference approach for robust visual tracking. In *Proc.* 8^{th} *ICCV*, 2001.