IDIAP RESEARCH REPORT

Motion likelihood and proposal modeling in Model-Based Stochastic Tracking

Jean-Marc Odobez * Daniel Gatica-Perez *

IDIAP–RR 04-61

November, 2004

submitted for publication

* IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-61

Motion likelihood and proposal modeling in Model-Based Stochastic Tracking

Jean-Marc Odobez   Daniel Gatica-Perez

November, 2004

submitted for publication

Abstract. Particle filtering is now established as one of the most popular methods for visual tracking. Within this framework, there are two important considerations. The first one refers to the generic assumption that the observations are temporally independent given the sequence of object states. The second consideration, often made in the literature, uses the transition prior as proposal distribution. Thus, the current observations are not taken into account, requesting the noise process of this prior to be large enough to handle abrupt trajectory changes. As a result, many particles are either wasted in low likelihood regions of the state space, resulting in low sampling efficiency, or more importantly, propagated to distractor regions of the image, resulting in tracking failures. In this paper, we propose to handle both considerations using motion. We first argue that in general observations are conditionally correlated, and propose a new model to account for this correlation allowing for the natural introduction of implicit and/or explicit motion measurements in the likelihood term. Secondly, explicit motion measurements are used to drive the sampling process towards the most likely regions of the state space. Overall, the proposed model allows to handle abrupt motion changes and to filter out visual distractors when tracking objects with generic models based on shape or color distribution. Experimental results obtained on head tracking, using several sequences with moving camera involving large dynamics, and compared against the CONDENSATION algorithm, have demonstrated superior tracking performance of our approach.

1 Introduction

Visual tracking is an important problem in computer vision, with applications in teleconferencing, visual surveillance, gesture recognition, and vision based interfaces [BI98]. Although tracking has been intensively studied in the literature, it still represents a challenging task in adverse situations, due to the presence of ambiguities (e.g. when tracking an object in a cluttered scene or when tracking multiple instances of the same object class), the noise in image measurements (e.g. lighting chnages), and the variability of the object class (e.g. pose variations).

In the pursuit of robust tracking, Sequential Monte Carlo methods [AMGC01, DdFG01, BI98] have shown to be a successful approach. In this temporal Bayesian framework, the probability of an object configuration given the observations is represented by a set of weighted random samples, called particles. This representation allows in principle to simultaneously maintain multiple hypotheses in the presence of ambiguities, unlike algorithms that keep only one configuration state [CRM00], which are therefore sensitive to single failures in the presence of ambiguities or fast or erratic motion.

In this paper, we address two important issues related to tracking with a particle filter. The first issue refers to the specific form of the observation likelihood, that relies on the conditional independence of observations given the state sequence. The second one refers to the choice of an appropriate proposal distribution, which, tunlike the prior dynamical model, should take into account the new observations. To handle these issues, we propose a new particle filter tracking method based on visual motion. Our method relies on a new graphical model allowing for the natural introduction of implicit or explicit motion information in the likelihood term, and on the exploitation of explicit motion measurements in the proposal distribution. A longer description of the above issues, our approach, and their benefits, is given in the following paragraphs.

The definition of the observation likelihood distribution is perhaps the most important element in visual tracking with a particle filter. This distribution allows for the evaluation of the likelihood of the current observation given the current object state, and relies on the specific object representation. The object representation corresponds to all the information that characterizes the object like the target position, geometry, appearance, color, etc. Parametrized shapes like splines [BI98] or ellipses [WH01], and color distributions [RMG98, CRM00, PHVG02, WH01], are often used as target representation. One drawback of these generic representations is that they can be quite unspecific, which augments the chances of ambiguities. One way to improve the robustness of a tracker consists of combining low-level measurements such as shape and color [WH01].

The generic conditional form of the likelihood term relies on a standard hypothesis in probabilistic visual tracking, namely the independence of observations given the state sequence [AMF04, BJ98a, BI98, IB89, RC01, TB01, WH01]. In this paper, we argue that this assumption can be inaccurate in the case of visual tracking. As a remedy, we propose a new model that assumes that the current observation depends on the current and previous object configurations as well as on the past observation. We show that under this more general assumption, the obtained particle filtering algorithm has similar equations than the algorithm based on the standard hypothesis. To our knowledge, this has not been shown before, and so it represents the first contribution of this article. The new assumption can thus be used to naturally introduce implicit or explicit motion information in the observation likelihood term. The introduction of such data correlation between successive images will turn generic trackers like shape or color histogram trackers into more specific ones.

Another important distribution to define when designing a particle filter is the proposal distribution, that is, the function that predicts the new state hypotheses where the observation likelihood will be evaluated. In general, an optimal choice [AMGC01, DGA00] consists of drawing samples from the more likely regions taking into account both the dynamical model, which characterizes the prior on the state sequence, and the new observations. However, simulating from the optimal law is often difficult when using standard likelihood terms. Thus, a common assumption in particle filtering consists in using the dynamics as proposal distribution. With this assumption, the variance of the noise process in the dynamical model implicitly defines a search range for the new hypotheses. This assumption raises difficulties in modeling dynamics since this term should fulfill two contradictory objectives. On one hand, as prior distribution, dynamics should be tight to avoid the tracker being confused by distractors in the vicinity of the true object configuration, a situation that is likely to happen for unspecific object representations such as generic shapes or color distributions. On the other hand, as proposal distribution, dynamics should be broad enough to cope with abrupt motion changes. Furthermore, this proposal distribution does not take into account the most recent observations. Many particles drawn from it will probably have a low likelihood, which results in low sampling efficiency. Overall, such a particle filter is likely to be distracted by background clutter. To address these issues, we propose to use explicit motion measures in the proposal function. One benefit of this approach will be to increase the sampling efficiency by handling unexpected motion, allowing for a reduced noise variance in the prediction process. Combined with the new observation likelihood term, using our proposal distribution will reduce the sensitivity of the particle filter algorithm to the different noise variances setting in the proposal and prior since, when using larger values, potential distractors should be filtered out by the introduced correlation and visual motion measurements. Finally, our proposal allows to implement the intuitive idea according to which the likely configurations with respect to an object model are evolving in conformity with the visual motion.

The rest of the paper is organized as follows. In the next Section, we discuss the state-of-the-art and relate it to our work. For sake of completeness, in Section 3, we describe the standard particle filter algorithm. Our approach is motivated in Section 4, while Section 5 describes the specific parts of our model in details. Experiments and results are reported in Section 6. Section 7 concludes the article with some discussion and future work.

2 Related work

In this article, the first contribution refers to the introduction of a new graphical model for particle filtering. This model allows for the modeling of temporal dependencies between observations. In practice, it lead us to naturally introduce motion observation within the data likelihood.

The use of motion for tracking is not a new idea. Motion-based trackers, essentially deterministic, integrate two-frame motion estimates over time. However, without any object model, it is almost impossible to avoid some drift after a few seconds of tracking. For long term tracking, the use of appearence-based models such as templates [AMF04, SJ01, TSK01] lead to more robust results. However, a template representation do not allow for large changes of appearence over time.
To handle appearance changes, an often difficult template adaptation step is needed [JFEM03, NWvdB01], or more complex global appearence models are used (e.g. eigen-spaces [BJ98b] or examplars [RMD01, TB01]), which poses the problem of learning these models, either off-line [BJ98a, TB01] or on-line [RMD01]. For instance, in [JFEM03], a generative model relying on the past frame template, a long term template, and a non-Gaussian noise component is proposed. Adaptation is performed through the estimation of the optimal state parameters -comprising the spatial 2D localization and the long-term template-, via an EM algorithm that identifies the stable regions of the template as a byproduct. A similar approach is taken in [NWvdB01], where the gray level of each template pixel is updated using a Kalman filter, and the adaptation is blocked whenever the innovation is too large. In these two cases, although partial and total occlusion can be handled, nothing prevents the tracker from long term drifts. This drift happens when the 2D visual motion does not match perfectly the real state evolution. This corresponds to the problematic case, reported in [JFEM03], of a turning head remaining at the same place in the image; in [NWvdB01], tracked objects (mainly high resolution faces and people) undergo very little pose changes. Another interesting approach towards adaptation using motion is proposed in [VPGB02] where, in a particle filter framework, a color model is adapted on-line. Assuming a static camera, a motion detection module is employed to select the instants more suitable for adaptation, which leads to good results.

In the present article, however, the method we propose is not template-based, i.e. no *reference* appearence template is employed or adapted (see discussion at the end of subsection 5.3.2). The implementation of our model aims at evaluating, either explicitly or implicitly, the similarity between the visual motion estimated from low-level information and the motion field induced by the state change. Our approach is thus different from the above ones, and more similar to the methods proposed in [SB01, SBF00]. In particular, the work in [SB01] addresses the difficult problem of people tracking using articulated models, and their use of the motion measures implicitly corresponds to the graphical model we propose here.

In the introduction, we raised the problems linked to the choice of the dynamical model as proposal. In the literature, several approaches have been proposed to address these issues. For instance, when available, auxiliary information generated from color [IB89, VPGB02, PVB04], motion detection [PVB04], or audio in the case of speaker tracking [GPLMO03, PVB04], can be used to draw samples from. The proposal distribution is then expressed as a mixture of the prior and components of the likelihood distribution. An important advantage of this approach is to allow for automatic (re)initialization. However, one drawback of this approach is that, since these additional samples are not related to the previous samples, the evaluation of the transition prior term for one new sample involves all past samples, which

can become very costly [IB89, GPLMO03]. [PVB04] avoids this problem by defining the prior as a mixture of distributions that includes a uniform law component, and by relying on distinctive and discriminative likelihoods, allowing for reinitialization using the standard particle filter equations. Another auxiliary particle filter proposed in [PS99] avoids this problem. The idea is to use the likelihood of a first set of predicted samples at time $k+1$ to resample the seed samples at time k, and to then apply the standard prediction and evaluation steps on these seed samples. The feedback from the new data acts by increasing or decreasing the number of descendents of a sample according to its "predicted" likelihood. Such a scheme, however, works well only if the variance of the transition prior is small, which is usually not the case in visual tracking.
As an alternative, the work in [RC01] proposed to use the unscented particle filter to generate importance densities. Although attractive, it is still likely to fail in the presence of abrupt motion changes, and the method needs to convert likelihood evaluations (e.g. of shape) into state space measurements (e.g. translation, scale). This would be difficult with color distribution likelihoods and for some state parameters. In [RC01], only a translation state is considered. In [AMF04, AM04], all the equations of the filter are conditioned with respect to the images. This allows for the use of the inter-frame motion estimates as dynamical model instead of an auto-regressive model to improve the state prediction. Moreover, in their application (point tracking), thanks to the use of a linear observation model, the optimal proposal function can be employed. However, as in [RC01], measures in state space are needed, and only translations are thus considered. Although their utilization of explicit motion measures is similar to what we propose here, it was introduced in a different way (through the dynamics rather than the likelihood), and was in practice restricted to translation.

3 Particle filtering

There exist at least two ways of introducing particle filters. The first one is through Sequential Importance Sampling (SIS) [AMGC01, DdFG01], and the second one is based on factored sampling [GCK91] applied to the filtering distribution [BI98]. While both approaches lead to the same algorithm with the standard assumptions, it is interesting to notice that the two methods do not lend themselves to the same extensions. In this paper, we follow the SIS approach, as it allows for the proposed extension.

Particle filtering is a technique for implementing a recursive Bayesian filter by Monte Carlo simulations. The key idea is to represent the required *posterior* probability density function (pdf) $\mathrm{p}(c_{0:k}|z_{1:k})$ of the state sequence $c_{0:k} = \{c_l, l = 0, \ldots, k\}$ up to time k conditionally to the observation sequence $z_{1:k} = \{z_l, l = 1, \ldots, k\}$, by a set of weighted samples $\{c^i_{0:k}, w^i_k\}_{i=1}^{N_s}$. Each sample (or particle) $c^i_{0:k}$ represents a potential trajectory of the state sequence, and w^i_k denotes its likelihood estimated from the sequence of observations up to time k. The weights are normalized ($\sum_i w^i_k = 1$) in order to obtain a discrete approximation of the true posterior :

$$\mathrm{p}(c_{0:k}|z_{1:k}) \approx \sum_{i=1}^{N_s} w^i_k \delta(c_{0:k} - c^i_{0:k}) \,. \tag{1}$$

Such a representation then allows to compute the expectation of any function f with respect to this distribution using a weighted sum :

$$\int f(c_{0:k})\mathrm{p}(c_{0:k}|z_{1:k})\mathrm{d}c_{0:k} \approx \sum_{i=1}^{N_s} w_k^i f(c_{0:k}^i) \tag{2}$$

In particular, the mean of the hidden state sequence can be computed from the first order moment (i.e. by using $f(x) = x$). Since sampling directly form the posterior is usually impossible, the weights are chosen using the principle of Importance Sampling (IS). This consists in simulating the samples from an importance (a.k.a proposal) function, and then introducing a correction factor (the weight) to account for the discrepancy between the proposal and the true posterior. More precisely, denoting by $q(c_{0:k}|z_{1:k})$ the importance density, the proper weights in (1) are given by :

$$w_k^i \propto \frac{\mathrm{p}(c_{0:k}^i|z_{1:k})}{q(c_{0:k}^i|z_{1:k})}\,. \tag{3}$$

The goal of the particle filtering algorithm is the recursive propagation of the samples and estimation of the associated weights as each measurement is received sequentially. Applying Bayes' rule, we obtain the following recursive equation for the posterior :

$$\mathrm{p}(c_{0:k}|z_{1:k}) = \frac{\mathrm{p}(z_k|c_{0:k}, z_{1:k-1})\mathrm{p}(c_k|c_{0:k-1}, z_{1:k-1})}{\mathrm{p}(z_k|z_{1:k-1})} \times \mathrm{p}(c_{0:k-1}|z_{1:k-1}) \tag{4}$$

Assuming a factorized form for the proposal (i.e. $q(c_{0:k}|z_{1:k}) = q(c_k|c_{0:k-1}, z_{1:k})q(c_{0:k-1}|z_{1:k-1})$) we obtain the following recursive update equation [AMGC01, DdFG01]:

$$w_k^i \propto w_{k-1}^i\ \frac{\mathrm{p}(z_k|c_{0:k}^i, z_{1:k-1})\mathrm{p}(c_k^i|c_{0:k-1}^i, z_{1:k-1})}{q(c_k^i|c_{0:k-1}^i, z_{1:k})}\,. \tag{5}$$

In order to simplify this general expression, conditional dependencies between variables are usually modeled according to the graphical model of Figure 1a, which corresponds to the following assumptions :

H1 : The observations $\{z_k\}$, given the sequence of states, are independent. This leads to $\mathrm{p}(z_{1:k}|c_{0:k}) = \prod_{i=1}^{k} \mathrm{p}(z_k|c_k)$, which requires the definition of the data-likelihood $\mathrm{p}(z_k|c_k)$. In Eq. 5, this assumption translates in $\mathrm{p}(z_k|c_{0:k}, z_{1:k-1}) = \mathrm{p}(z_k|c_k)$.

H2 : The state sequence $c_{0:k}$ follows a first-order Markov chain model. In Eq. 5, this means that $\mathrm{p}(c_k|c_{0:k-1}, z_{1:k-1}) = \mathrm{p}(c_k|c_{k-1})$.

We then obtain the simplified weight update equation :

$$w_k^i \propto w_{k-1}^i\ \frac{\mathrm{p}(z_k|c_k^i)\mathrm{p}(c_k^i|c_{k-1}^i)}{q(c_k^i|c_{0:k-1}^i, z_{1:k})} \quad (\text{ and } \sum_i w_k^i = 1\ )\,. \tag{6}$$

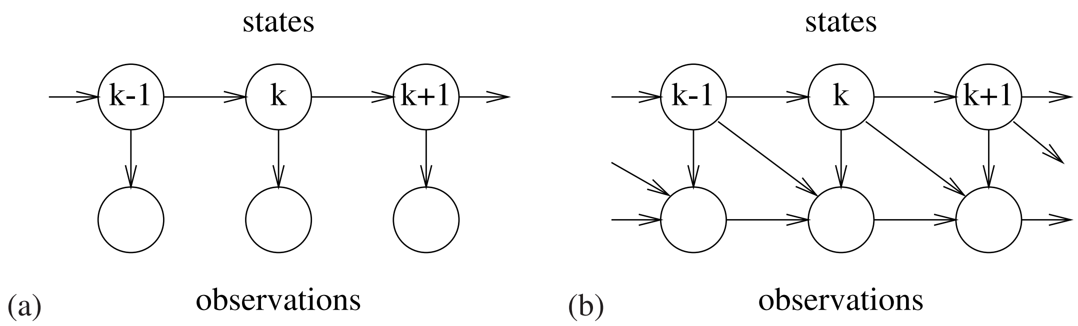


Figure 1: Graphical models for tracking. (a) standard model (b) proposed model.

1. Initialization
   - for $i = 1, \ldots, N_s$, sample $c_0^i \sim \mathrm{p}(c_0)$ and set $k = 1$.
2. Diffusion/propagation :
   - for $i = 1, \ldots, N_s$, sample $\tilde{c}_k^i \sim q(c_k^i | c_{0:k-1}^i, z_{1:k})$.
3. Weight updating
   - for $i = 1, \ldots, N_s$, evaluate the weight w_k^i with Equation (5)
4. Selection resample with replacement N_s particles
   - $\{c_k^j, \frac{1}{N_s}\} \leftarrow \mathrm{resample}(\{\tilde{c}_k^i, w_k^i\})$
   - set $k = k + 1$ and goto step 2.

Figure 2: The generic particle filter algorithm.

The set $\{c_{0:k}^i, w_k^i\}_{i=1}^{N_s}$ is then approximately distributed according to $\mathrm{p}(c_{0:k}|z_{0:k})$.

It is known that importance sampling is inefficient in high-dimensional spaces [DGA00], which is the case of the state space $c_{0:k}$ as k increases. In practice, this leads to the continuous increase of the weight variance, concentrating the mass of the weights onto a few particles only. To solve this problem, it is necessary to apply an additional resampling step, whose effect is to eliminate the particles with low importance weights and to multiply particles having high weights. Several resampling schemes exist [DGA00]. In our implementation, we used the one described in [AMGC01], and perform a systematic resampling. We finally obtain the particle filter displayed in Fig. 2.

The efficiency of a particle filter algorithm relies on the definition of a good proposal distribution. A temporally local strategy consists of choosing the importance function that minimizes the weight variance of the new samples at time k conditionally to trajectories $c^i_{1:k-1}$ and observations $z_{1:k}$. It can be shown [DGA00] that this optimal function is given by

$$q(c^i_k|c^i_{0:k-1}, z_{1:k}) = q(c_k|c^i_{k-1}, z_k) = \mathrm{p}(c_k|c^i_{k-1}, z_k) \propto \mathrm{p}(z_k|c_k)\mathrm{p}(c_k|c^i_{k-1})\,, \tag{7}$$

which leads to the following weight update equation :

$$w^i_k \propto w^i_{k-1}\ \mathrm{p}(z_k|c^i_{k-1})\,. \tag{8}$$

In practice, sampling from $\mathrm{p}(c_k|c^i_{k-1}, z_k)$ and evaluating $\mathrm{p}(z_k|c^i_{k-1})$ are only achievable in particular cases, involving for instance Gaussian noise and linear observation models [AMGC01, DGA00, AM04]. As an alternative, a choice often made consists of selecting the *prior* as importance function. In that case, we have :

$$w^i_k \propto w^i_{k-1}\ \mathrm{p}(z_k|c^i_k)\,. \tag{9}$$

Although this model is intuitive and simple to implement, this choice, which does not take into account the current observations, has several drawbacks, especially with high-dimensional vector spaces or narrow likelihood models.
Finally, notice that while the weighted set $\{c^i_{0:k}, w^i_k\}^{N_s}_{i=1}$ allows for the representation of the posterior pdf $\mathrm{p}(c_{0:k}|z_{0:k})$, the set $\{c^i_k, w^i_k\}^{N_s}_{i=1}$, that can be obtained from it, is also a representative sample of the filtering distribution $\mathrm{p}(c_k|z_{0:k})$, thanks to simple marginalization.

4 Our approach

In this Section, we propose a new method that embeds motion in the particle filter. This is first obtained by incorporating motion information into the measurement process. This can be achieved by modifying the traditional graphical model represented in Fig. 1a, by making the current observation dependent not only on the current object configuration but also on the object configuration and observation at the previous instant (see Fig. 1b). Secondly, we propose to use explicit motion measurements in order to obtain a better proposal distribution. In the following Subsections, we motivate our approach by pointing out the limitations of the basic particle filter.

4.1 Revisiting the hypotheses in particle filtering

The filter described in Fig. 2 is based on the standard probabilistic model for tracking displayed in Fig. 1a and corresponding to hypotheses H1 and H2 of the previous section.

In visual tracking, hypothesis H1 of conditional independence of temporal measurements given the states may not be very accurate. Keeping only two time instants for simplicity, the assumption implies that for all state sequences $c_{k-1:k}$ and data sequences $z_{k-1:k}$,

$$\mathrm{p}(z_k, z_{k-1}|c_k, c_{k-1}) = \mathrm{p}(z_k|c_k, c_{k-1})\mathrm{p}(z_{k-1}|c_k, c_{k-1})\,.$$

This is a very strong assumption: in practice, there exist some state sequences $c_{k-1:k}$ of interest (e.g. the "true" or "target" state sequences, or state sequences close to the mean state sequence) for which the data are correlated, and hence, for which the standard assumption does not hold. This can be illustrated as follows.
In most tracking algorithms, the state space includes the parameters of a geometric transformation $\mathcal{T}$. Then, the measurements consist of implicitly or explicitly extracting some part of the image by :

$$\tilde{z}_{c_k}(\mathbf{r}) = z_k(\mathcal{T}_{c_k}\mathbf{r}) \qquad \forall \mathbf{r} \in R\,, \tag{10}$$

where $\tilde{z}_{c_k} = z_k|c_k$, $\mathbf{r}$ denotes a pixel 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 the pixel $\mathbf{r}$. The data likelihood is then usually computed from this local patch : $\mathrm{p}(z_k|c_k) = \mathrm{p}(\tilde{z}_{c_k})$. However, if c_{k-1} and c_k correspond to two consecutive states of a given object, it is reasonable to assume that :

$$\tilde{z}_{c_k}(\mathbf{r}) = \tilde{z}_{c_{k-1}}(\mathbf{r}) + \eta(\mathbf{r}) \qquad \forall \mathbf{r} \in R\,. \tag{11}$$

where $\eta(\mathbf{r})$ are prediction noise random variables, assumed to be symmetric with zero mean. This point is illustrated in Figure 3. Equation (11) is at the core of all motion estimation and compensation algorithms like MPEG and is indeed a valid hypothesis [Tek95]. More formally, if we consider the patch $\tilde{z}_\bullet$ as a vector of i.i.d components, we can compute the normalized cross-correlation (NCC) between two data vectors $\tilde{z}_{c_{k-1}}$ and $\tilde{z}_{c_k}$, for state couples $c_{k-1:k}$ of interest, to study their dependencies. The NCC of two patches $\tilde{z}_1$ and $\tilde{z}_2$ is given by :

$$\mathrm{NCC}(\tilde{z}_1, \tilde{z}_2) = \frac{\sum_{\mathbf{r}\in R} (\tilde{z}_1(\mathbf{r}) - \bar{\tilde{z}}_1) \cdot (\tilde{z}_2(\mathbf{r}) - \bar{\tilde{z}}_2)}{\sqrt{\mathrm{Var}(\tilde{z}_1)}\sqrt{\mathrm{Var}(\tilde{z}_2)}}\,, \tag{12}$$

where $\bar{\tilde{z}}_1$ represents the mean of $\tilde{z}_1$.
To perform experiments, we defined, as ground truth (GT) object sequences, ellipses manually fitted to the head of persons in two sequences of 300 images each. Next, we considered the state couple $(c_{k-1}, c_k) = (c^{gt}_{k-1}, c^{gt}_k + \vec{\delta})$, where c^{gt} denotes a GT object image position, and $\vec{\delta}$ corresponds to an offset around the GT state. Furthermore, the dimensions of the ellipse at time $k-1$ are used to define the ellipse at time k.
The dependency between measurements is illustrated in Fig. 4a and 4b, where the average NCC is ploted against the amplitude of $\vec{\delta}$, measured either in number of pixels, or in percentage of object size, where object size is defined as the average between the two ellipse's axis lengths. In the training data, object size ranges between 30 and 80 pixels, and there are between 600 and 12000 measurements per δ value. As can be seen, when the offset displacement reaches 50% of object size, correlation becomes close to 0. When the displacement is greater than 100%, the NCC should be 0 in average, as there is no more overlap between the two measurement vectors. Fig. 4c and 4d illustrates this further by displaying the histogram of the NCC for different values of δ. Again, while the histograms are peaked around 1 for small values for δ, it gradually moves towards a symmetric histogram centered at 0 with the increase of δ.

This issue bears similarities with the work on Bayesian correlation [SBIM99]. In such work, the dependence/independence of measurements (in this case, the output of a set of

Figure 3: Images at time t and $t + 3$. The two local patches corresponding to the head and extracted from the two images are strongly correlated.

filters) at different spatial positions, given the object state, was studied. It was shown that independence was achieved as long as the supports of the filters were distant enough. For foreground object modeling, however, the obtained measurement distributions were not specific enough. The work in [SBR00] further showed that the independence still holds *conditioned* on the availability of some form of object template to predict the filter output. In tracking terms, the patch extracted in the previous frame from the state at time $k-1$ plays the role of the conditioning template, as shown by Eq. (11), and the independence result of [SBR00] states that the noise variables $\eta(\mathbf{r})$ and $\eta(\mathbf{r}')$ are independent when $|\mathbf{r} - \mathbf{r}'|$ is large enough.

The above analysis illustrates that the independence of the data given the sequence of states is not a true assumption in general. More precisely :

$$\mathrm{p}(z_k|z_{1:k-1}, c_{1:k}) \neq \mathrm{p}(z_k|c_k)\,, \tag{13}$$

which means that we can not reduce the left hand side to the right one as usually done with the standard derivation of the particle filter equations. A more accurate model for visual tracking is thus represented by the graphical model of Fig. 1b.

The new model can be easily incorporated in the particle filter framework. First, note that all computation leading to Eq. 5 in Section 3 are general and do not depend on assumptions H1 and H2. Starting from there, replacing H1 by the new model gives :

$$\mathrm{p}(z_k|z_{1:k-1}, c_{1:k}) = \mathrm{p}(z_k|z_{k-1}, c_k, c_{k-1}^i)\,. \tag{14}$$

If we keep H2, it is easy to see that the new weight update equation is given by :

$$w_k^i \propto w_{k-1}^i \, \frac{\mathrm{p}(z_k|z_{k-1}, c_k^i, c_{k-1}^i)\mathrm{p}(c_k^i|c_{k-1}^i)}{q(c_k^i|c_{0:k-1}^i, z_{1:k})} \tag{15}$$

in replacement of equation (6).

4.2 Proposal distribution and dynamical model

According to our new graphical model, and following the same arguments as in [AMGC01, DGA00], we can show that the optimal proposal distribution and the corresponding update

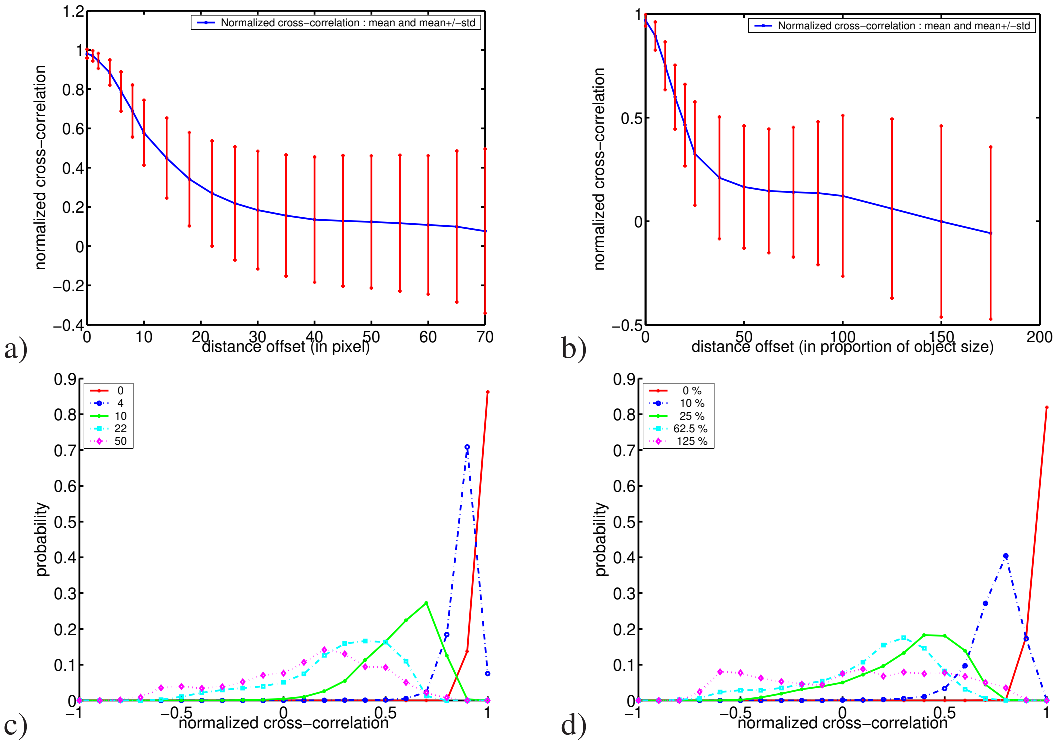


Figure 4: (a) and (b) Average of the NCC coefficient for state couples at varying distance from the ground truth state values. In (a), the distance is measured in pixels, while in (b) it is measured in proportion of object size. (c) and (d) Empirical distribution of the NCC coefficients for different displacement distance, measured either in pixel (c), or in proportion of object size (d).

rule are given by :

$$\begin{aligned} q(c_k|c^i_{k-1}, z_{1:k}) &= \mathrm{p}(c_k|z_k, z_{k-1}, c^i_{k-1}) \propto \mathrm{p}(z_k|c_k, z_{k-1}, c^i_{k-1})\mathrm{p}(c_k|c^i_{k-1}) \quad \text{and} \\ w^i_k &\propto w^i_{k-1}\mathrm{p}(z_k|c^i_{k-1}, z_{k-1})\,. \end{aligned}$$

As their homologous Equations (7) and (8), these equations are difficult to be used in practice.

A possibility then consists of using the dynamical model (i.e. the prior) as the proposal. This suffers from the generic drawbacks mentioned in the introduction, and in visual tracking, from the unspecificity of some state changes, which often plays in favor of the use of simple dynamical models (e.g. constant speed models). Also, the low temporal sampling rate and the presence of fast and unexpected motions, due either to camera or object movements, render the noise parameter estimation problem difficult.

An alternative, that we adopt in this paper, consists of using as proposal a mixture model built from the *prior* and observation likelihood distributions [PVB04]. In our case, the likelihood term $\mathrm{p}(z_k|c_k, z_{k-1}, c^i_{k-1})$ comprises an object-related term and one motion term (see paragraph 5.3). In this article, we will construct a proposal distribution from the latter. Moreover, as motivated by the rest of this section, this term happens to be more adapted to model state changes than dynamics relying only on state values.

The relevance of using a visual motion-based proposal rather than the dynamics is illustrated by the following experiments. Consider as state c the horizontal position of the head of the foreground person in the sequence displayed in Fig. 6, which has been hand held recorded and features a person moving around in an office, and denote by c^{gt} the GT value obtained from a manual annotation of the head position in 200 images. Furthermore, let us denote by ξ_k the state prediction error, whose expression is given by

$$\xi_k = c_k^{gt} - \hat{c}_k \, , \tag{16}$$

where $\hat{c}_k$ denotes the state prediction, computed by two methods. The first one uses a simple AR model :

$$\hat{c_k} = c_{k-1}^{gt} + \dot{c}_{k-1} \text{ with } \dot{c}_{k-1} = c_{k-1}^{gt} - c_{k-2}^{gt} \, , \tag{17}$$

where $\dot{c}$ denotes the state derivative and models the evolution of the state. In the second method, $\hat{c_k}$ is computed by exploiting the inter-frame motion to predict the new state value :

$$\hat{c}_k = c_{k-1}^{gt} + \dot{c}_{k-1}^{motion} \tag{18}$$

where $\dot{c}_{k-1}^{motion}$ is computed using the coefficients of an affine motion model robustly estimated on the region defined by c_{k-1}^{gt} (see Section 5.2).

Fig. 5a reports the prediction error obtained with the AR model. As can be seen, this prediction is noisy. The standard deviation of the prediction error, σ_ξ, is equal to 2.7. Furthermore, there are large peak errors (up to 30% of the head width)[1]. To cope with these peaks, the noise variance in the dynamics has to be overestimated to avoid particles near the ground truth to be too disfavored. Otherwise, only particles lying near the -erroneous- predicted states may survive the resampling step. However, a large noise variance has the effect of wasting many particles in low likelihood areas or spreading them on local distractors, which can ultimately lead to tracking failures. On the other hand, exploiting the inter-frame motion leads to a reduction of both the noise variance (σ_ξ=0.83) and the error peaks (Fig. 5b). There is another advantage of using image-based motion estimates. Let us first note that the previous state values (here c_{k-1}, c_{k-2}) used to predict the new state value $\hat{c}_k$ are affected by noise, due to measurement errors and uncertainty. Thus, in the standard AR approach, both the state c_{k-1} and state derivative $\dot{c}_{k-1}$ in Eq. 17 are affected by this noise, resulting in large errors (Fig. 5c). When using the inter-frame motion estimates, the estimation $\dot{c}_{k-1}^{motion}$ is almost not affected by noise (whose effect is to slightly modify the support region used to estimate the motion), as illustrated in Fig. 6, resulting again in a lower noise variance process (Fig. 5d).

Thus, despite needing more computation resources, inter-frame motion estimates are usually more precise than auto-regressive models to predict new state values; as a consequence, they are a better choice when designing a proposal function. This observation is supported

[1]Higher order models were also tested. Although they usually led to a variance reduction of the prediction error, they also increased the amplitude and duration of the error peaks.

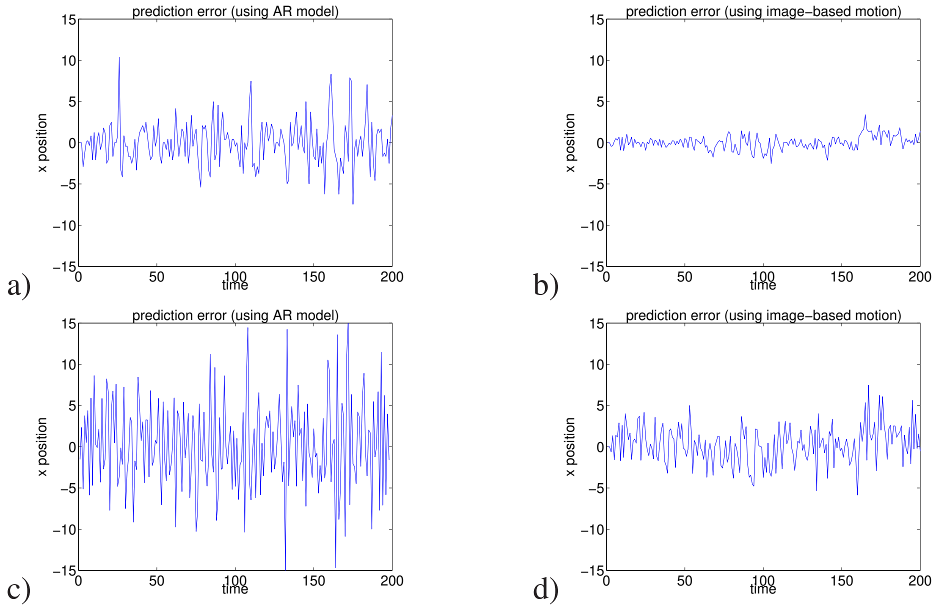


Figure 5: (a) Prediction error of the x position, when using an AR2 model . (b) Prediction error, but exploiting the inter-frame motion estimation. (c) resp. (d), same as (a) resp. (b) but now adding a random Gaussian noise (std=2 pixels) on the GT measurements used for prediction. With the AR model (Fig. c) both the previous state and state derivative estimates are affected by noise (σ_ξ=5.6), while with visual-motion (Fig. d) the noise mainly affects the previous measurement (σ_ξ=2.3).

by experiments on other state parameters -vertical position, scale-, and on other sequences. Finally, this observation can also be applied to a set of particles. If these are localized on modes of a distribution related to visual measurements, their prediction according to the visual motion will generally place them around the new modes associated with the current image.

5 The implemented model

The graphical model of Fig. 1b is generic. In this paper, our specific implementation will be based on the graphical model of Fig. 7, whose elements are described more precisely in the rest of this section.

5.1 Object representation and state space

We follow an image-based standard approach, where the object is represented by a region R subject to some valid geometric transformation, and is characterized either by a shape or by a color distribution. For geometric transformations, we have chosen a subspace of the

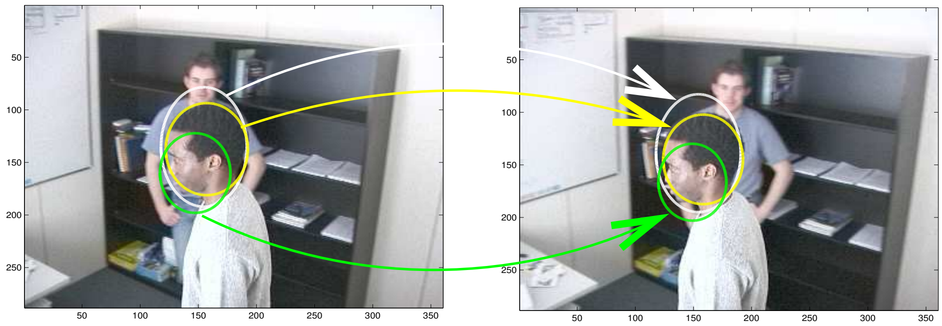


Figure 6: Example of motion estimates between two images from noisy states. The 3 ellipses correspond to different state values. Although the estimation support regions only cover part of the head and enclose textured background, the head motion estimate is still good.

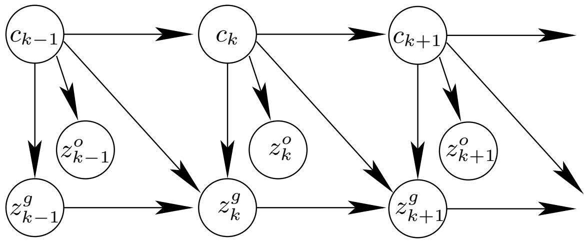


Figure 7: Specific graphical model for our implementation.

affine transformations comprising a translation **T**, a scaling factor s, and an aspect ratio e :

$$\mathcal{T}_\alpha \mathbf{r} = \begin{pmatrix} \mathbf{T}_x + x s_x \\ \mathbf{T}_y + y s_y \end{pmatrix} \quad , \tag{19}$$

where $\mathbf{r} = (x, y)$ denotes a point position in the reference frame, $\alpha = (\mathbf{T}, s, e)$, and :

$$s = \frac{s_x + s_y}{2}\,,\ e = \frac{s_x}{s_y}\,,\ s_x = \frac{2es}{1+e} \text{ and } s_y = \frac{2s}{1+e} \tag{20}$$

A state is then defined as $c_k = (\alpha_k, \alpha_{k-1})$.

5.2 Motion estimation

As mentioned in the previous Section, we use inter-frame motion estimates both as observations and to sample the new state values. More precisely, an affine displacement model $\vec{d}_\Theta$ parameterized by $\Theta = (a_i), i = 1..6$ is computed using a gradient-based robust estimation method described in [OB95][2]. $\vec{d}_\Theta$ is defined by:

$$\vec{d}_\Theta \mathbf{r} = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} . \tag{21}$$

[2]We use the code available at http://www.irisa.fr/vista

This method takes advantage of a multiresolution framework and an incremental scheme based on the Gauss-Newton method. It minimizes an M-estimator criterion to ensure the goal of robustness, as follows :

$$\widehat{\Theta}(c_{k-1}) = \underset{\Theta}{\operatorname{argmin}} \sum_{\mathbf{r} \in R(c_{k-1})} \rho\left(\mathrm{DFD}_{\Theta}(\mathbf{r})\right)$$

$$\textit{with} \quad \mathrm{DFD}_{\Theta}(\mathbf{r}) = z_k(\mathbf{r} + \vec{d}_{\Theta}\mathbf{r}) - z_{k-1}(\mathbf{r}), \tag{22}$$

where z_{k-1} and z_k are the images, and $\rho(\cdot)$ is a robust estimator (bounded for high values of its argument). Owing to the robustness of the estimator, an imprecise region definition $R(c_{k-1})$ due to a noisy state value does not sensibly affect the estimation (see Fig. 6). From these estimates, we can construct an estimate $\widehat{\dot{\alpha}}_{k-1}$ of the variation of the coefficients between the two instants. Assuming that the coordinates in Eq. 21 are expressed with respect to the object center (located at $\mathbf{T}$ in the image), we propose the following derivative estimates :

$$\left\{ \begin{array}{l} \dot{\mathbf{T}}_x = a_1 \\ \dot{\mathbf{T}}_y = a_4 \end{array} \right., \left\{ \begin{array}{l} \dot{s}_x = a_2 s_x \\ \dot{s}_y = a_6 s_y \end{array} \right. \text{ and } \left\{ \begin{array}{l} \dot{s} = \frac{s}{1+e}(a_2 e + a_6) \\ \dot{e} = e(a_2 - a_6) \end{array} \right. \tag{23}$$

The estimated predicted value is then given by $\widehat{\alpha}_k = \alpha_{k-1} + \widehat{\dot{\alpha}}_{k-1}$. Although not used in the reported experiments, the covariance matrix of the estimated parameters can also be computed. With model-based approaches involving more state parameters, this would be useful to account for uncertainty and underconstrained optimization.

5.3 Data likelihood modeling

To implement the new particle filter, we assume that the measurements z_k are of two types: object measurements z_k^o (i.e. edges or color), and patch gray level measurements z_k^g. Then, we consider the following data likelihood :

$$\begin{aligned} \mathrm{p}(z_k|z_{k-1}, c_k, c_{k-1}) &= \mathrm{p}(z_k^o, z_k^g|z_{k-1}^o, z_{k-1}^g, c_k, c_{k-1}) && (24)\\ &= \mathrm{p}(z_k^o|z_k^g, z_{k-1}^o, z_{k-1}^g, c_k, c_{k-1})\mathrm{p}(z_k^g|z_{k-1}^o, z_{k-1}^g, c_k, c_{k-1}) && (25)\\ &= \mathrm{p}(z_k^o|c_k)\mathrm{p}(z_k^g|z_{k-1}^g, c_k, c_{k-1}) && (26) \end{aligned}$$

where the last derivations exploit the properties of the graphical model of Fig. 7. Two assumptions were made to derive this model. The first one assumed that object observations are independent of patch observations given the state sequence measurements. This choice decouples the model of the dependency existing between two images, whose implicit goal is to ensure that the object trajectory follows the optical flow field implied by the sequence of images, from the shape or appearence object model. When the object is modeled by a shape, our assumption is valid since shape observations will mainly involve measurements on the border of the object, while the correlation term will apply to the regions inside the object. When a color representation is employed, the assumption is valid as well, as color measurements can usually be considered as being independent of gray-scale measurements. The second

assumption we made is that object measurements are uncorrelated over time. When considering shape measurements, the assumption is quite valid as the temporal auto-correlation function of contours is peaked. However, with the color representation [CRM00, PHVG02], the temporal independence assumption might not hold. Better models need to be searched for to handle this case.
We describe the specific observations models as follows.

5.3.1 Visual object measurement

For the experiments, we considered both contour models or color models.
Shape 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 a vector of candidate positions $\{\nu_m^l\}$ relative to a point lying on the contour ν_0^l. With some usual assumptions [BI98], the shape likelihood can be expressed as

$$p(z_k^o|c_k) \propto \prod_{l=1}^{L} \max\left(K, \exp(-\frac{\|\hat{\nu}_m^l - \nu_0^l\|^2}{2\sigma_s^2})\right), \tag{27}$$

where $\hat{\nu}_m^l$ is the nearest edge on l, and K is a constant used when no edges are detected. In all experiments, we used $L = 16$ search lines, the search range along each line was 10 pixels inside and outside the contour, σ_s was set to half the search range (i.e. 5), and $K = \exp^{-2}$ (value we obtain for the farthest edge detection).
Color model :
As color models we used color distributions represented by normalized histograms in the HSV space and gathered inside the candidate region $R(c_k)$ associated with the state c_k. To be robust to illumination effects, we only considered the HS values. Then, a normalized multidimensional histogram was computed, resulting in a vector $\mathbf{b}(c_k) = (\mathrm{b}^j(c_k))_{j=1..N}$, where $N = N_h \times N_s$ with N_h and N_s representing the number of bins along the hue and saturation dimensions respectively ($N_h = N_s = 8$), and where the index j corresponds to a couple (h, s) with h and s denoting hue and saturation bin numbers. At time k, the candidate color model $\mathbf{b}(c_k)$ is compared to a reference color model $\mathbf{b}_{ref}$. We use the histogram computed in the first frame as reference model. As a distance measure, we employed the Bhattacharyya distance measure [CRM00, PHVG02]:

$$D_{bhat}(\mathbf{b}(c_k), \mathbf{b}_{ref}) = \left(1 - \sum_{j=1}^{N} \sqrt{\mathrm{b}^j(c_k)\mathrm{b}^j_{ref}}\right)^{1/2} \tag{28}$$

and assumed that the probability distribution of the square of this distance for a given object follows an exponential law,

$$p(z_k^o|c_k) \propto \exp\{-\lambda_{bhat}\, D^2_{bhat}(\mathbf{b}_k(c_k), \mathbf{b}_{ref})\}\,. \tag{29}$$

We used a value of $\lambda_{bhat} = 20$ in all experiments.

5.3.2 Image correlation measurement

To model this term, we used two possibilities :

- The first one consists of extracting measures in the parameter space. Usually, this is achieved by thresholding and/or extracting local maxima of some interest function [AM04, PVB04]. In our case, this corresponds to the extraction of peaks of a correlation map, as done in [AM04] for translations. One advantage of such a method is to provide a well-behaved likelihood (i.e. involving only a few well identified modes). One drawback is that the extraction process can be time consumming.
- In the second approach, gray-level patches are directly compared after having warped them according to the state values (see Eq.(10)). The advantages of this method are to supply more “detailed” likelihoods that can be computed directly from the data.

In this paper, we employ both options, by assuming that observations are made of the measured parameters $\widehat{\alpha}_k$ obtained using the estimated motion, and of the local patches $\tilde{z}^g_{c_k}$. We model the correlation term in the following way :

$$\mathrm{p}(z^g_k|z^g_{k-1}, c_k, c_{k-1}) \propto \mathrm{p}_{c1}(\widehat{\alpha}_k, \alpha_k)\mathrm{p}_{c2}(\tilde{z}^g_{c_k}, \tilde{z}^g_{c_{k-1}}) \tag{30}$$

with :

$$\mathrm{p}_{c1}(\widehat{\alpha}_k, \alpha_k) = \mathcal{N}(\widehat{\alpha}_k; \alpha_k, \Lambda_{\xi_p}) \tag{31}$$

$$\mathrm{p}_{c2}(\tilde{z}^g_{c_k}, \tilde{z}^g_{c_{k-1}}) = \mathrm{Z}^{-1} \exp^{-\lambda_c {D_c}^2(\tilde{z}^g_{c_k}, \tilde{z}^g_{c_{k-1}})} \tag{32}$$

$$\mathrm{Z} = \int_{z',z''} \exp^{-\lambda_c {D_c}^2(z',z'')} \, \mathrm{d}z' \mathrm{d}z'' \tag{33}$$

where $\mathcal{N}(.; \mu, \Lambda)$ represents a Gaussian distribution with mean μ and covariance matrix Λ, D_c denotes a distance between two image patches, $\Lambda_{\xi_p} = diag(\sigma^2_{\xi_p,j})$ is the covariance of the measurements, Z is a normalization constant whose value can be computed from (33), where the integral runs over pairs of consecutive patches corresponding to the same tracked object extracted in training sequences [TB01]. In practice, however, we did not compute this value and assumed it to be constant for all object patches. The first probability term in Eq. 30 compares the measured parameters with the sampled values, and assumes a Gaussian noise process in parameter space. The second term introduces a non-Gaussian model, by comparing directly the patches defined by c_k and c_{k-1} using the similarity distance D_c. It has been derived by assuming that all patches are equally probable. Although the use of those two terms is somewhat redundant, it proved to be a good choice in practice and its purpose can be illustrated using Fig. 6. While all the three predicted configurations will be weighted equally according to p_{c1}, the second term p_{c2} will downweight the two predictions (green and white ellipses) 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 and used in the literature [SJ01, TB01, SB01]. The choice of the distance should take into account the followings considerations :

1. the distance should still model the underlying motion content, i.e. the distance should increase as the error in the predicted configuration grows;

2. the random nature of the prediction process in the SMC filtering will rarely produce configurations corresponding to exact matches. This is particularly true when using a small number of samples;

3. particles covering both background and object, each undergoing different motions, should have a low likelihood.

For these purposes, we found out in practice that it was preferable not to use robust norms such as L1 saturated distance or a Haussdorf distance [TB01]. Additionally, we needed to avoid distances which might *a priori* favor patches with specific content. This is the case of the L2 distance, which corresponds to an additive Gaussian noise model in Eq.(11) and generally provides lower scores for tracked patches with large uniform areas[3]. Instead, we used a distance based on the normalized-cross correlation coefficient (Eq. (12)) defined as :

$$D_c(\tilde{z}_1, \tilde{z}_2) = 1 - \mathrm{NCC}(\tilde{z}_1, \tilde{z}_2) \tag{34}$$

Regarding the above equation, it is important to emphasize again that the method is not performing template matching, as in [SJ01]. No object template is learned off-line or defined at the begining of the sequence, and the tracker does not maintain a single template object representation at each instant of the sequence. Thus, the correlation term is not object specific (except through the definition of the reference region R). A particle placed on the background would thus receive a high weight if the predicted motion is in adequation with the background motion. Nevertheless, the methodology can be extended to be more object dependent, by using more object specific regions R and by allowing the region R to vary over time, as is done in articulated object tracking [SB01].

5.4 Dynamics definition

To model the prior, we use a standard second order AR model (Eq. 17) for each of the components of α. However, to account for outliers (i.e. unexpected and abrupt changes) 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

$$\mathrm{p}(c_k|c_{k-1}) = \prod_{j=1}^{4} \rho_c\left(\alpha_{k,j} - (2\alpha_{k-1,j} - \alpha_{k-2,j}), \sigma^2_{\xi_{d,j}}\right) \ . \tag{35}$$

where $\sigma^2_{\xi_{d,j}}$ denotes the dynamics noise variance of the j^{th} component. Moreover, as the correlation likelihood term is indeed more reliable than the prior to constraint trajectories, we set $\sigma_{\xi_{d,j}}$ to three times the $\sigma_{\xi_{p,j}}$ value.

[3] This issue is related to our assumption of equally probable patches. Given our likelihood model for joint tracked patches, Eq. (32), this assumption is only approximate.

5.5 Proposal distribution

As motivated in Section 4.2, the proposal distribution function relies on the estimated motion. More precisely, we define it as :

$$q(c_k|c_{0:k-1}, z_{1:k}) \propto \mathcal{N}(\alpha_k; \widehat{\alpha}_k, \Lambda_{\xi_p}) \tag{36}$$

which means that we sample new positions around the predicted state value.

6 Results

To illustrate the method, we considered three sequences involving head tracking. Results should be appreciated by looking directly at typical video results that can be found on our website [4]. To differentiate the different elements of the model, we considered three kinds of trackers :

- condensation tracker M1 : this tracker corresponds to the standard CONDENSATION algorithm [BI98], with the object likelihood p_o (Eq. 27 or 29) combined with the same AR model with Gaussian noise for the proposal and the prior.
- implicit correlation tracker M2 : it corresponds to CONDENSATION, with the addition of the implicit motion likelihood term in the likelihood evaluation (i.e now equal to $\mathrm{p}_o . \mathrm{p}_{c2}$). This method does not use explicit motion measurements.
- motion proposal tracker M3 : it is the full model. The samples are drawn from the motion proposal, Eq. 36, and the weight update is performed using Eq. 5. After simplification, the update equation becomes :

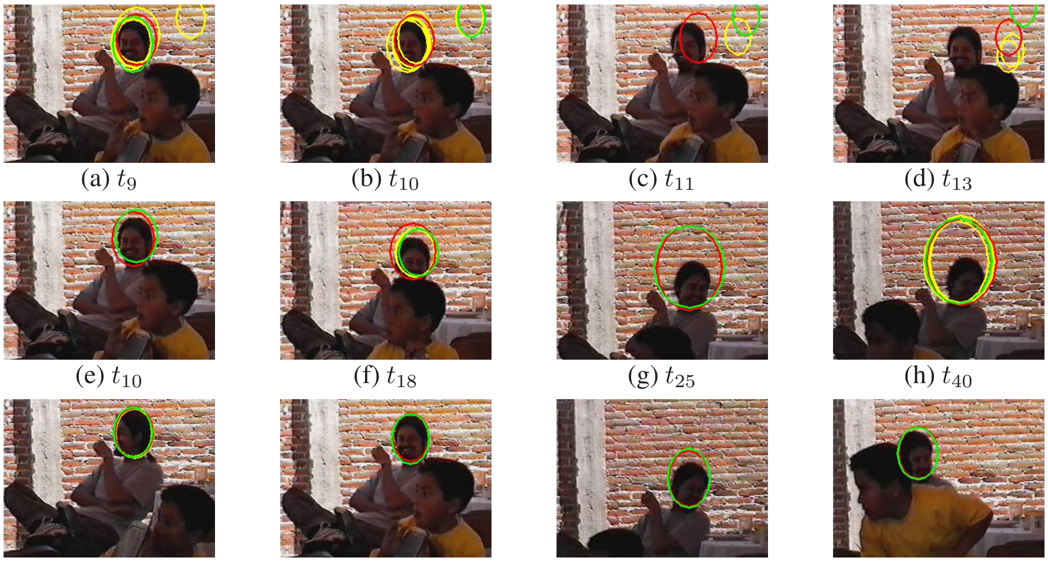
  $$w_k^i = w_{k-1}^i \ \mathrm{p}_o(z_k^o|c_k^i)\mathrm{p}_{c2}(\tilde{z}^g_{c_k^i}, \tilde{z}^g_{c_{k-1}^i})\mathrm{p}(c_k^i|c_{k-1}^i) \tag{37}$$

  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. We use $K = max(20, N_s/10)$ clusters.

For 200 particles, the shape-based M1 tracker runs in real time (on a 2.5GHz P IV machine), M2 at around 20 image/s, and M3 around 8 image/s. In experiments, all the common parameters are kept identical. Only the number of samples N_s, and the noise variance in the proposal distribution (and consequently in the dynamics) will be changed. The noise standard deviations will be denoted σ_{T} for the translation components (i.e. $\sigma_{\xi_p,1} = \sigma_{\xi_p,2} = \sigma_{\mathrm{T}}$) and σ_s for the scale (i.e. $\sigma_{\xi_p,3} = \sigma_s$). The aspect ratio component noise is kept fixed, with $\sigma_{\xi_p,4} = 0.01$.

The first sequence (Fig. 8 and 9), containing 64 images of size 240×320, illustrates qualitatively the benefit of the method in the presence of strong ambiguities. The sequence

[4] *www.idiap.ch/~odobez/IPpaper/EmbeddingMotion.html*



(a) t_9 (b) t_{10} (c) t_{11} (d) t_{13}

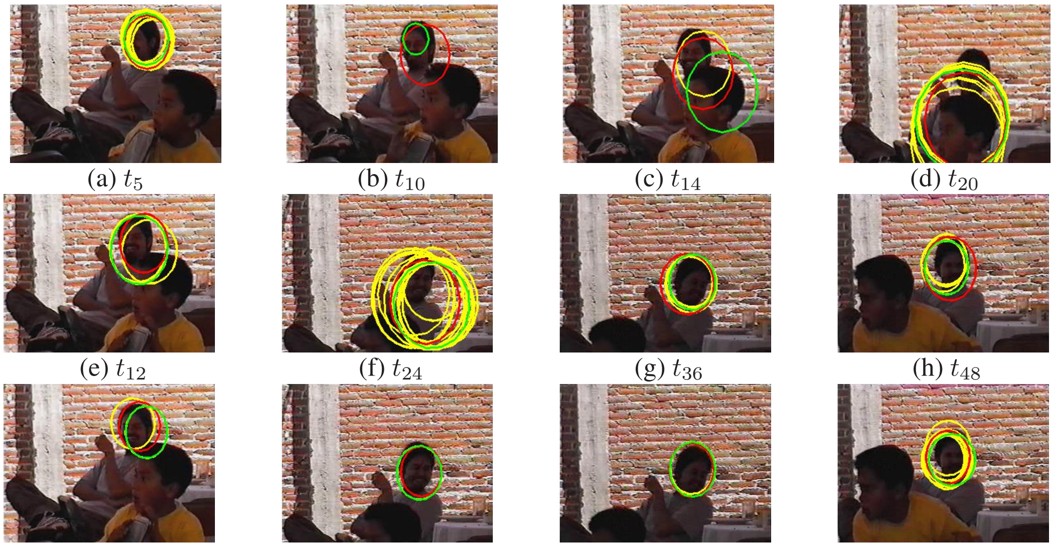
(e) t_{10} (f) t_{18} (g) t_{25} (h) t_{40}

(i) t_1 (j) t_{10} (k) t_{35} (l) t_{63}

Figure 8: Head tracking 1 : first row : shape-based tracker M1. Second row : shape-based tracker M2. Third row : shape-based tracker M3. In red, mean state; in green, mode state; in yellow, likely particles.

features a highly textured background producing very noisy shape measurements, camera and head motion, change of appearence of the head, and partial occlusion. Whatever the number of particles or the noise variance in the dynamical model, the shape-based tracker M1 alone is unable to perform a correct tracking after time t_{12}. In contrast, tracker M2 is able to do the tracking correctly on a large majority of runs when using small dynamics $((\sigma_{\mathbf{T}}, \sigma_s) = (1, 0.005))$. However, with an increase of the noise variance, it fails (see second row of Fig. 8) : the observations are clearly multimodal, and the head motion is only occasionaly different from the background, which makes it especially hard for the correlation term to keep configurations enclosing only the head. Using tracker M3, however, leads to correct tracking, even with large noise values. There might be two reasons for this. The first one consists of the use of the correlation likelihood measure in parameter space. The second one is due to its ability to better maintain multimodality[5]. Consider a mode that is momentarily represented by only a few particles. With a "blind" proposal, these particles are spread with few chances to hit the object likelihood mode, decreasing their probability of survival in the next selection step. On the other hand, with the motion proposal, these chances are increased. Finally, for small dynamics, the color-based tracker M1 usually succeeds, but it fails with standard dynamics (e.g. dynamics used in [PVB04]), as shown in the first row of Fig. 9). This is due to the presence of the brick color and more importantly, the face of the boy. Exploiting correlation leads to successful tracking, but with a lower precision when using M2 (see images 9(e) to 9(h)), than with M3 (images 9(i) to 9(l)).

[5] In [VDP03], it has been shown on simulated experiments that even when the true density is a two Gaussian mixture model with the same mixture weight for each Gaussian, and with the appropriate likelihood model, the



(a) t_5 (b) t_{10} (c) t_{14} (d) t_{20}

(e) t_{12} (f) t_{24} (g) t_{36} (h) t_{48}

(i) t_{12} (j) t_{24} (k) t_{36} (l) t_{48}

Figure 9: Head tracking sequence 1 : color-based model. First row : M1, second row : M2, last row : M3. All experiments (including those of Fig. 8) with N_s=250 and $(\sigma_{\mathbf{T}}, \sigma_s) = (5, 0.01)$. In red, mean state; in green, mode state; in yellow, likely particles.

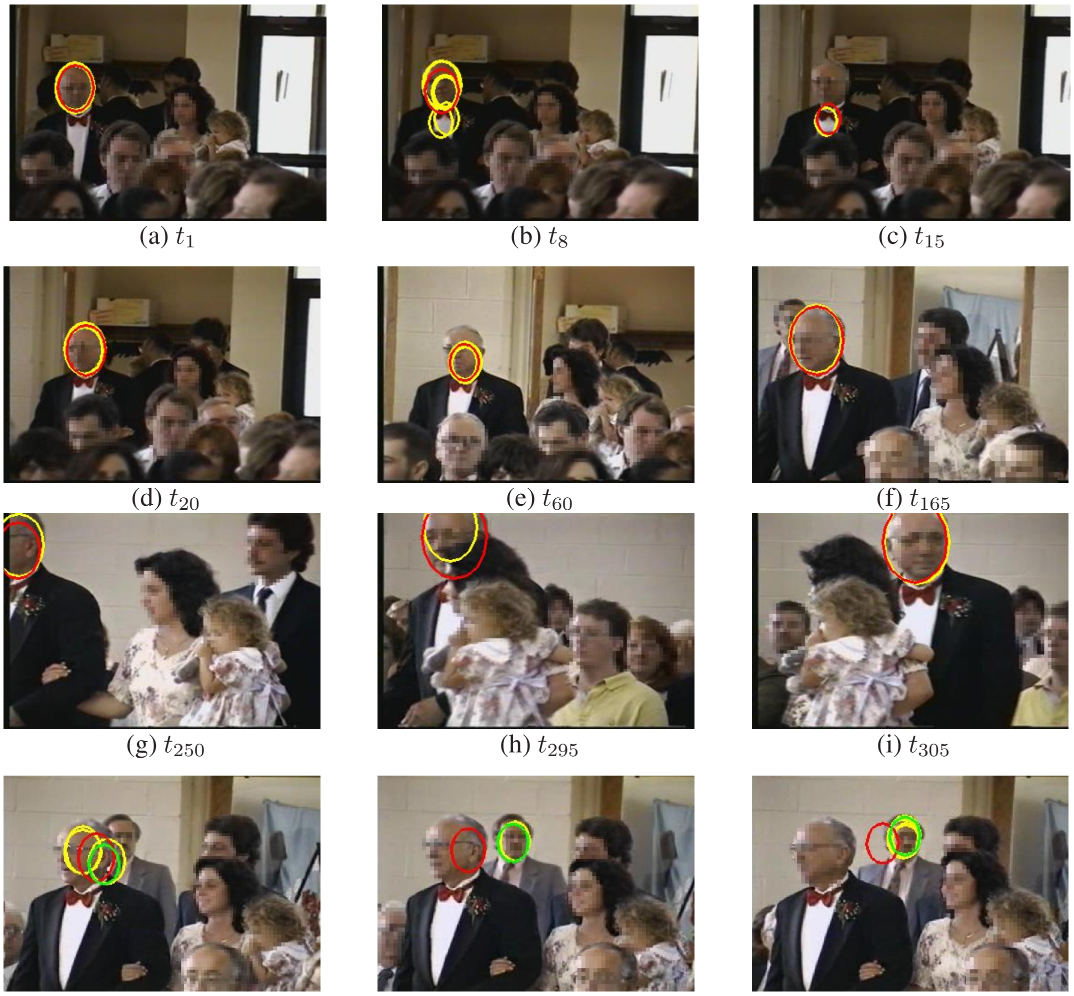
The second sequence is a 330 frame sequence (Fig. 10) extracted from a hand-held home video. Figure 11 reports the tracking performance of the three trackers for different dynamics and number of particle. At each frame, the resulting tracked region R_t (obtained from the mean state value) is considered as successful if the recall and precision are both higher than 25%, where these rates are defined by :

$$R_{\cap,t} = R_{gt,t} \cap R_t \, , r_{prec} = \frac{|R_{\cap,t}|}{|R_t|}, \, r_{rec} = \frac{|R_{\cap,t}|}{|R_{gt,t}|} \tag{38}$$

where $R_{gt,t}$ is the ground truth region, and $| \cdot |$ denotes the set cardinality operator.

A tracking failure is considered As can be seen, while the shape-based tracker M1 performs quite well for tuned dynamics (parameter set D1), it breaks down rapidly, even for slight increases of dynamics variances (parameters D2 to D4). Fig. 10 illustrates a typical failure due to the small size of the head at the begining of the sequence, the low contrast at the left of the head, and the clutter. On the other hand, the shape-based 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 (see Fig. 11(g)) does the tracker fail. These failures are occuring at different parts of the sequence. Finally, in all experiments, the shape-based tracker M3 produces a correct tracking rate. When looking at the color-based tracker M1, we can see that it performs much better than its shape equivalent (compare Fig. 11(d) and 11(a)). However, due to the presence of a person in the background, it fails around 25% of the time with standard noise values. Incorporating the motion leads to perfect tracking, though with a very small

standard particle filter loses rapidly one of the modes.



(a) t_1 (b) t_8 (c) t_{15}

(d) t_{20} (e) t_{60} (f) t_{165}

(g) t_{250} (h) t_{295} (i) t_{305}

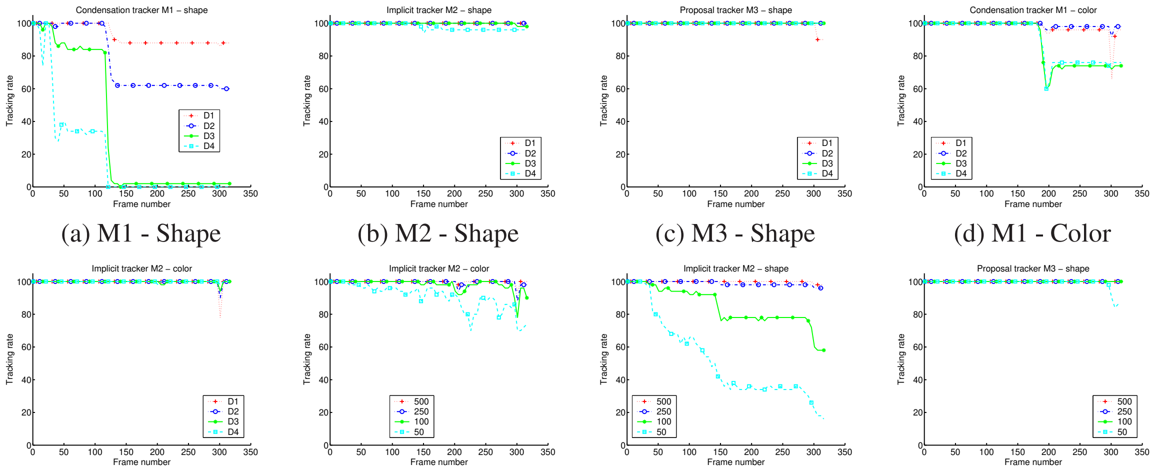
(j) t_{180} (k) t_{185} (l) t_{190}

Figure 10: Head tracking sequence 2 (N_s=500) : top row : shape-based tracker M1. Second and third rows : shape-based tracker M2. Last row : color-based tracker M1. In red, mean state. In green, mode state. In yellow, highly likely particles.

number of samples (N_s=50, see 11(f)), the M2 tracker sometimes fails while the full model is always successful.

The last sequence (Fig. 12) better illustrates the benefit of using the motion proposal approach. This 72s sequence acquired at 12 frame/s is specially difficult because of the occurence of several head turns[6](which prevents us from using the color trackers), and abrupt motion changes (translations, zooms in and out), and importantly, due to the absence of head contours as the head moves near (frames 160 to 200) or in front of the bookshelves (frames 620 to the end). Because of these factors, the shape-based tracker M1 fails due to a local ambiguity with the whiteboard frame (around frame 65), or because of camera jitter (frame 246) (cf Fig.13(a)). The M2 tracker works better, handling correctly the jitter situation when

[6]Head turns are difficult cases for the new method, as in the extreme case, the motion inside the head region indicates a right (or left) movement while the head outline remains static.



(a) M1 - Shape (b) M2 - Shape (c) M3 - Shape (d) M1 - Color

(e) M2 - Color (f) M2 - Color (g) M2 - Shape (h) M3 - Shape

Figure 11: Head tracking sequence 2 : successful tracking rate (in %, computed over 50 trials with different random seeds). Experiments (a) to (e) : parameter sets D1 to D4 correspond to N_s=500, with dynamics $(\sigma_{\mathbf{T}}, \sigma_s)$: D1 (2,0.01), D2 (3,0.01), D3 (5,0.01) D4 (8,0.02). Experiments (f) to (h), different number of particles are tested (500/250/100/50) using the D3 (5,0.01) noise values.

the dynamic noise is large enough, but fails when the head moves in front of the bookshelves, due to the temporally lack of head contours, combined with background clutter. In contrast, all these issues are resolved by the M3 tracker, which better capture the state variations, and allows a successful track of the head until the end of the sequence under almost all conditions (Fig. 13(c) and 13(d)).

7 Conclusion

We presented a methodology to embed data-driven motion measurements into particle filters. This was first achieved by proposing a new graphical model that accounts for the temporal correlation existing between successive images of the same object. We show that this new model can be easily handled by the particle filter framework. The new introduced observation likelihood term can be exploited to model the visual motion using either implicit or explicit measurements. Secondly, explicit motion estimates were exploited to predict more precisely the new state values. This data-driven approach allows for designing 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 or color-based object models, and to reduce the sensitivity to the different parameters of the prior model.

The conducted experiments have demonstrated the benefit of exploiting the proposed scheme. However, this should not obliterate the fact that the tracking performance depends on the choice of a good and robust object model. This was also illustrated in the

(a) t_1 (b) t_{65} (c) t_{170}

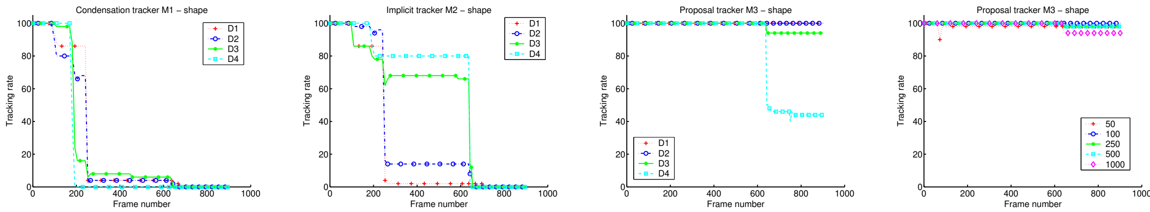
(d) t_{446} (e) t_{640} (f) t_{660}

(g) t_{710} (h) t_{745} (i) t_{800}

Figure 12: Head tracking sequence 3. Tracker with motion proposal (N_s=500). In red, mean shape; in green, mode shape; in yellow, likely particles.

reported experiments. The color tracker, when its use is appropriate, performs better than its shape equivalent. Thus, when dealing on a specific object tracker, like head tracker for instance, building more precise or adaptive object likelihood may further improve the proposed method. This can be achieved by developing better probability density functions to model the likelihood of observations of different nature, or measured at different spatial or temporal positions.

Finally, we have showed that the exploitation of explicit motion measurements in the proposal improved the tracking efficiency. The described approach is general. For instance, it can be used to track deformable objects, by exploiting the integration of motion measurements along the shape curve, as described in [KHP96]. However, in this case, the usefulness and the robustness of the low-level motion measurements to model the temporal variation of fine scale parameters need to be demonstrated. The use of an hybrid scheme, in which one part of the state parameters (e.g. translation, scale, rotation,...) are sampled from a data driven motion proposal, while the other part is drawn from a standard AR model, might be more appropriate.



(a) M1 - Shape (b) M2 - Shape (c) M3 - Shape (d) M3 - Shape

Figure 13: Head tracking sequence 3 : successful tracking rate (in %, computed over 50 trials with different random seeds). Experiments 13(a) to 13(c) : parameter sets D1 to D4 correspond to N_s=1000, with dynamics $(\sigma_{\mathbf{T}}, \sigma_s)$: D1 (2,0.01), D2 (3,0.01), D3 (5,0.01) D4 (8,0.02). In experiments 13(d), different number of particles are tested using the D3 (5,0.01) noise values.

Acknowledgment

The authors would like to thank the Swiss National Fund for Scientific Research that supported this work through the IM2 MUCATAR project.

References

[AM04] E. Arnaud and E. Mémin. Optimal importance sampling for tracking in image sequences:application to point tracking. In *Proc. of* 8^{th} *European Conf. Computer Vision*, Prague, Czech Republic, May 2004.

[AMF04] E. Arnaud, E. Mémin, and B. Cernushi Frias. Filtrage conditionnel pour la trajectographie dans des séquences d'images - application au suivi de points. In *14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, Toulouse, France, January 2004.

[AMGC01] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian. *IEEE Trans. Image Process.*, pages 100–107, 2001.

[BI98] Andrew Blake and Michael Isard. *Active Contours*. Springer, 1998.

[BJ98a] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In H. Burkhardt and B. Neumann, editors, *Proc. of* 5^{th} *European Conf. Computer Vision, Lecture Notes in Compter Science, vol. 1406*, pages 909–924, Freiburg, Germany, 1998. Springer-Verlag.

[BJ98b] M.J. Black and A.D. Jepson. Eigentracking : robust matching and tracking of articulated objects using a view based representation. *Int. J. Computer Vision*, 26(1):63–84, 1998.

[CRM00] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 142–151, 2000.

[DdFG01] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[DGA00] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[GCK91] U. Grenander, Y. Chow, and D.M. Keenan. *HANDS. A Pattern Theoretical Study of Biological Shapes*. Springer-Verlag, New-York, 1991.

[GPLMO03] Daniel Gatica-Perez, Guillaume Lathoud, Iain McCowan, and Jean-Marc Odobez. A Mixed-State I-Particle Filter for Multi-Camera Speaker Tracking. In *IEEE Int. Conf. on Computer Vision Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, 2003.

[IB89] M. Isard and A. Blake. Icondensation : Unifying low-level and high-level tracking in a stochastic framework. In *Proc. of* 5^{th} *European Conf. Computer Vision, Lecture Notes in Compter Science, vol. 1406*, pages 893–908, 1889.

[JFEM03] A.D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust on-line appearance models for visual tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(10):661–673, October 2003.

[KHP96] C. Kervrann, F. Heitz, and P. Pérez. Statistical model-based estimation and tracking of non-rigid motion. In *Proc. 13th Int. Conf. Pattern Recognition*, pages 244–248, Vienna, Austria, August 1996.

[NWvdB01] H. T. Nguyen, M. Worring, and R. van den Boomgaard. Occlusion robust adaptive template tracking. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, pages 678–683, Vancouver, July 2001.

[OB95] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.

[PHVG02] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proc. of* 7^{th} *European Conf. Computer Vision, Lecture Notes in Compter Science, vol. 2350*, pages 661–675, Copenhaguen,, June 2002.

[PS99] Michael K. Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Jl of the American Statistical Association*, 94(446):590–599, 1999.

[PVB04] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.

[RC01] Y. Rui and Y. Chen. Better proposal distribution: object tracking using unscented particle filter. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 786–793, December 2001.

[RMD01] A. Rahimi, L.P. Morency, and T. Darrell. Reducing drift in parametric motion tracking. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, pages 315–322, Vancouver, July 2001.

[RMG98] Y. Raja, S. McKenna, and S. Gong. Colour model selection and adaptation in dynamic scenes. In *Proc. of* 5^{th} *European Conf. Computer Vision, Lecture Notes in Compter Science, vol. 1406*, pages 460–474, 1998.

[SB01] H. Sidenbladh and M.J. Black. Learning image statistics for bayesian tracking. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, volume 2, pages 709–716, Vancouver, Canada, July 2001.

[SBF00] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Proc. European Conf. Computer Vision*, volume 2, pages 702–718, Dublin, Ireland, June 2000.

[SBIM99] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by bayesian correlation. In *Proc.* 7^{th} *IEEE Int. Conf. Computer Vision*, pages 1068–1075, 1999.

[SBR00] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modeling for object localisation. In *Proc. European Conf. Computer Vision*, pages 307–323, 2000.

[SJ01] J. Sullivan and Rittscher J. Guiding random particles by deterministic search. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, pages 323–330, Vancouver, July 2001.

[TB01] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, Vancouver, July 2001.

[Tek95] A.M. Tekalp. *Digital video processing*. Signal Processing series. Prentice Hall, 1995.

[TSK01] Hai Tao, Harpreet S. Sawhney, and Rakesh Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(1):75–89, 2001.

[VDP03] J. Vermaak, A. Doucet, and P. Pérez. Maintaining multi-modality through mixture tracking. In *Proc.* 9^{th} *IEEE Int. Conf. Computer Vision*, Nice, France, June 2003.

[VPGB02] J. Vermaak, P. Pérez, M. Gangnet, and A. Blake. Towards improved observation models for visual tracking: Selective adaptation. In *Proc. of* 7^{th} *European Conf. Computer Vision, Lecture Notes in Compter Science, vol. 2350*, pages 645–660, Copenhague, Danemark, 2002.

[WH01] Y. Wu and T. Huang. A co-inference approach for robust visual tracking. In *Proc.* 8^{th} *IEEE Int. Conf. Computer Vision*, Vancouver, July 2001.