Particle Filtering Strategies for Visual Tracking dedicated to H/R Interaction

Ludovic Brèthes[†], Frédéric Lerasle[†‡], and Patrick Danès[†‡]

†LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse Cédex 4, FRANCE
‡Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cédex, FRANCE
{FirstName.Name }@laas.fr

Summary. This paper deals with visual tracking of people from a camera mounted on a mobile robot in a human, cluttered, environment. Various visual cues are described, relying on color, shape or motion, together with several particle filtering strategies taking into account all or part of the measurements. These strategies enable the combination/fusion of visual cues, both into an importance function from which the particles are sampled, and into a measurement model serving in the definition of weights. The paper describes some prominent visual-based interaction modalities for our tour-guide robot and checks which visual cues and filtering algorithms associations best fulfill their requirements. Extensions are finally discussed.

1 Introduction

The development of personal robots is a motivating challenge in robotics research. In this context, we have designed and implemented a new tour-guide mobile robot on the basis of an iRobot B21r platform (figure 1(a)). We have extended the standard equipment with one pan-tilt Sony camera EVI-D70, one digital camera mounted on a Directed Perception pan-tilt unit, one ELO touch-screen, a pair of loudspeakers, an optical fiber gyroscope and wireless Ethernet. Besides endowing the robot with robust and efficient basic navigation abilities in a dynamic environment, our efforts concern the design of onboard visual tracking functions in order to interpret the motion of visitors attending an exhibition. We have outlined three visual modalities (figure 1) our robot must basically deal with:

1. **the "search for interaction",** where the robot, static and left alone, visually tracks visitors thanks to the camera mounted on its helmet, in order to heckle them when they enter the exhibition;
2. **the "proximal interaction",** where a user can interact through the ELO touch-screen, for example to select the area he wants to visit; during this interaction, the robot remains static and must keep, thanks to the camera materializing its eye, the visual contact with the user;
3. **the "guidance mission",** where the robot drives the visitor to the selected area; during its mission, the robot must also maintain the interaction with the guided visitor.

(a) (b) (c) (d)

Fig. 1. The Rackham robot (a) and its three visual modalities: search for interaction (b), proximal interaction (c), guidance mission (d).

As the robot's evolution takes place into dynamic and cluttered environments, several hypotheses must be handled at each instant concerning the parameters to be estimated, and a robust integration of multiple visual cues must be developed. Particle filtering seems well-suited to this context. Indeed, it makes no restrictive assumption on the probability distributions entailed in the characterization of the problem, and enables an easy combination/fusion of diverse kinds of measurements. Nevertheless, it can be argued that data fusion using particle filtering schemes has been fairly seldom exploited in the robotics context, for it has often been confined to a restricted number of visual cues. Moreover, despite numerous particle filtering strategies have been described in the literature, it is still not clear which ones best fit the requirements of the three above visual modalities, so that a study comparing their efficiency must be carried out in this robotics context.

The paper is organized as follows. Section 2 briefly sums up the well-known particle filtering formalism, and describes some variants which enable data fusion for tracking. Then, section 3 specifies some visual measurements which rely on the color, shape or image motion of the observed target. Section 4 describes the three tracking setups which best fulfill the requirements for the aforementioned visual modalities. Last, section 5 summarizes our contribution and puts forward some future extensions.

2 Particle filtering algorithms for data fusion

2.1 A generic algorithm

Particle filters are sequential Monte Carlo simulation methods for the state vector estimation of any Markovian dynamic system subject to possibly non-Gaussian random inputs [1, 2]. Their aim is to recursively approximate the a posteriori probability density function (pdf) $p(x_k|z_{1:k})$ of the state vector x_k at time k conditioned on the set of measurements $z_{1:k} = z_1, \ldots, z_k$, through the linear point-mass combination

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)}), \ \sum_{i=1}^{N} w_k^{(i)} = 1, \tag{1}$$

which expresses the selection of a value – or "particle" – $x_k^{(i)}$ with probability – or "weight" – $w_k^{(i)}$, $i = 1, \ldots, N$. An approximation of the conditional expectation of any function of x_k then follows from (1).

The generic particle filtering algorithm – or "Sampling Importance Resampling" (SIR) – is shown on Table 1. The particles $x_k^{(i)}$ evolve stochastically over time, being sampled from an *importance function* $q(x_k|x_{k-1}^{(i)}, z_k)$ which aims at adaptively exploring "relevant" areas of the state space. Their weights $w_k^{(i)}$ are updated accordingly, so as to guarantee the consistency of the approximation (1). In order to limit the degeneracy phenomenon, which says that whatever the sequential Monte Carlo simulation method, after few instants all but one particle weights tend to zero, step 8 inserts a resampling stage. There, the particles $x_k^{(j)}$ associated to high weights $w_k^{(j)}$ are duplicated while the others collapse, so that the sequence $\tilde{x}_k^{(1)}, \ldots, \tilde{x}_k^{(N)}$ is i.i.d. according to $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$. Note that this resampling stage should rather be fired only when the filter efficiency – related to the number of "useful" particles – goes beyond a predefined threshold [2].

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \mathrm{SIR}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$**,** Draw $x_0^{(i)} \sim p(x_0)$, set $w_0^{(i)} = \frac{1}{N}$, so that $\{x_0^{(i)}, w_0^{(i)}\}$ depicts $p(x_0)$ **END IF**
2: **IF** $k \geq 1$ **THEN** $\{$—$\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}$ being a particle description of $p(x_{k-1}|z_1^{k-1})$—$\}$
3: **FOR** $i = 1, \ldots, N$, **DO**
4: "Propagate" the particle $x_{k-1}^{(i)}$ by independently sampling $x_k^{(i)} \sim q(x_k|x_{k-1}^{(i)}, z_k)$
5: Update the weight $w_k^{(i)}$ according to the formula $w_k^{(i)} \propto w_{k-1}^{(i)} \dfrac{p(z_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}|x_{k-1}^{(i)}, z_k)}$,
prior to a normalization step so that $\sum_i w_k^{(i)} = 1$
6: **END FOR**
7: Compute the conditional mean of any function of x_k, *e.g.* the MMSE estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
8: At any time, or according to an "efficiency" criterion, resample $\{x_k^{(i)}, w_k^{(i)}\}$ according to $P(\tilde{x}_k^{(i)} = x_k^{(i)}) = w_k^{(i)}$, which leads to a an equivalent weighted particle set $\{\tilde{x}_k^{(i)}, \frac{1}{N}\}$ describing $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$; set $x_k^{(i)}$ and $w_k^{(i)}$ with $\tilde{x}_k^{(i)}$ and $\frac{1}{N}$
9: **END IF**

Table 1. Generic particle filtering algorithm (SIR)

2.2 Importance sampling from either dynamics or measurements: basic strategies

The CONDENSATION – for "Conditional Density Propagation" [3] – can be viewed as the instance the SIR algorithm in which the particles are drawn according to the system dynamics, viz. when $q(x_k|x_{k-1}^{(i)}, z_k) = p(x_k|x_{k-1}^{(i)})$. This endows CONDENSATION with a prediction-update structure, in that $\sum_{i=1}^N w_{k-1}^{(i)} \delta(x_k - x_k^{(i)})$ approximates the prior $p(x_k|z_{1:k-1})$. The weighting stage becomes $w_k^{(i)} \propto w_{k-1}^{(i)} p(z_k|x_k^{(i)})$.

In a visual tracking context, the original algorithm [3] defines the particles likelihoods from contour primitives, yet other visual cues have also been exploited [7].

Resampling by itself cannot efficiently limit the degeneracy phenomenon. In addition, it can lead to a loss of diversity in the state space exploration. The importance function must thus be defined with special care.

In visual tracking, the modes of the likelihoods $p(z_k|x_k)$, though multiple, are generally pronounced. As $\mathrm{CONDENSATION}$ draws the particles $x_k^{(i)}$ from the system dynamics but blindly w.r.t. the measurement z_k, many of these can be assigned a low weight in step 5, thus significantly worsening the overall filter performance. An alternative – henceforth labelled "Measurement-based SIR" (MSIR) – may merely consist in sampling the particles at time k – or just some of their entries – according to an importance function $q(x_k|z_k)$ defined from the current image. The first MSIR strategy was ICONDENSATION [4], which guided the state space exploration by a color blobs detector. Other visual detection functionalities can be used as well, *e.g.* face detector (§3), or any other intermittent primitive which, despite its sporadicity, is very discriminant when present [7]: motion, sound, etc.

2.3 Advanced strategies

In an MSIR scheme, a particle $x_k^{(i)}$ whose entries are drawn from the current image may be inconsistent with its predecessor $x_{k-1}^{(i)}$ from the point of view of the state dynamics. Of course, the smaller the value $p(x_k^{(i)}|x_{k-1}^{(i)})$, the lesser the weight $w_k^{(i)}$. One solution to this problem, as proposed in the genuine ICONDENSATION algorithm, consists in sampling some of the particles w.r.t. the dynamics,

An interesting alternative is proposed in [8, Table 4]. Dynamic models of order greater than or equal to 2 are considered, in which the state vector reads as $x_k = (u'_k, v'_k, h'_k)'$, with $'$ the transpose operator. The subvector $(u'_k, v'_k)'$ – or "innovation part" – of x_k obeys a stochastic state equation on x_{k-1} while h_k – called "history part" – is a deterministic function $f(x_{k-1})$. It is assumed that the particles $({u_k^{(i)}}', {u_k^{(i)}}')'$ are sampled from an importance function such as $q(u_k, v_k|x_{k-1}^{(i)}, z_k) = \pi(u_k|z_k)p(v_k|u_k^{(i)}, x_{k-1}^{(i)})$ – *i.e.* the subparticles $u_k^{(i)}$ are positioned from the measurement only while the $v_k^{(i)}$'s are drawn by fusing the state dynamics with the knowledge of $u_k^{(i)}$ –, and that the pdf of the measurement conditioned on the state satisfies $p(z_k|x_k) = p(z_k|u_k, v_k)$. This context is particularly well-suited to visual tracking, for equivalent state-space representations of linear AR models entail the above decomposition of the state vector, and because the output equation does not involve its "history part".

The authors define procedures enabling the avoidance of any contradiction between $({u_k^{(i)}}', {u_k^{(i)}}')'$ and its past $x_{k-1}^{(i)}$. Their "Rao-Blackwellised Subspace Particle Filter with History Sampling" ($\mathrm{RBSSHSSIR}$) is summarized in Table 2. Its step 5 noticeably consists, for each subparticle $u_k^{(i)}$ positioned using z_k, in the resampling of a predecessor particle – and thus of the "history part" of $x_k^{(i)}$ – which is at the same time likely w.r.t. $u_k^{(i)}$ from the dynamics point of view and assigned with a significant weight. The $\mathrm{RBSSHSSIR}$ algorithm differs from ICONDENSATION precisely because of this stage, yet necessary lest the weighted particles $\{x_k^{(i)}, w_k^{(i)}\}$ may not be a consistent description of the posterior $p(x_k|z_{1:k})$. Last, though the demonstration

goes outside the scope of this paper, it can be shown that the algorithm also applies when the state process is of the first order, in which case it just suffices to suppress the entry $f(x_{k-1})$ from x_k.

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \mathrm{RBSSHSSIR}([\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$, Draw $x_0^{(i)} \sim p(x_0)$, set $w_0^{(i)} = \frac{1}{N}$, so that $\{x_0^{(i)}, w_0^{(i)}\}$ depicts $p(x_0)$ **END IF**
2: **IF** $k \geq 1$ **THEN** $\{$—$\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}$ being a particle description of $p(x_{k-1}|z_1^{k-1})$—$\}$
3: **FOR** $i = 1, \ldots, N$, **DO**
4: Draw $u_k^{(i)} \sim \pi(u_k|z_k)$
5: Sample in $(1, \ldots, N)$ the index $I_k^{(i)}$ of the predecessor particle of $u_k^{(i)}$ according to the weights $(w_{k-1}^{(1)} p(u_k^{(i)}|x_{k-1}^{(1)}), \ldots, w_{k-1}^{(N)} p(u_k^{(i)}|x_{k-1}^{(N)}))$
6: Draw $v_k^{(i)} \sim p(v_k|u_k^{(i)}, x_{k-1}^{I_k^{(i)}})$
7: Set $x_k^{(i)} = \left(u_k^{(i)\prime}, v_k^{(i)\prime}, f(x_{k-1}^{(I_k^{(i)})})'\right)'$
8: Update the weights, prior to their normalization, by setting $w_k^{(i)} \propto \dfrac{p(z_k|u_k^{(i)}) \sum_{l=1}^N w_{k-1}^{(l)} p(u_k^{(i)}|x_{k-1}^{(l)})}{\pi(u_k^{(i)}|z_k)}$
9: Compute the conditional mean of any function of x_k, *e.g.* the MMSE estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$
10: **END FOR**
11: **END IF**

Table 2. Rao-Blackwellised Subspace Particle Filter with History Sampling (RBSSHSSIR)

Last, it must be mentioned that the "optimal recursive strategy" [2] – in terms of filter efficiency – should define $q^*(x_k|x_{k-1}, z_k) \triangleq p(x_k|x_{k-1}, z_k)$ and $w_k^{*(i)} \propto w_{k-1}^{*(i)} p(z_k|x_{k-1}^{(i)})$ in the SIR algorithm Table 1. Except in very particular cases, such formulae can only be approximated in practice, e.g. through the Auxiliary Particle Filter (APF) [6]. Though this strategy has also been evaluated in the current visual tracking context, its details are not included for space reasons, all the more because it was shown to be superseded by the aforementioned schemes.

3 Importance and measurement functions

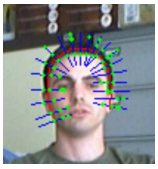

Fig. 2. Shape cue.

Importance sampling offers a mathematically principled way of directing search according to visual cues which are discriminant though possibly intermittent, e.g. motion. Such cues are logical candidates for detection modules and efficient proposal distributions. Besides, each sample weight is updated taking into account its likelihood w.r.t. the current image. This likelihood is computed by means of measurement functions, according to visual cues (e.g. color, shape) which must be persistent but may however be proner to ambiguity in cluttered scenes. In both importance sampling and weight update steps, combining or fusing multiple cues enables the tracker to better benefit from distinct information sources, and can decrease its sensitivity to temporary failures in some of the measurement processes. Measurement and importance functions are depicted in the next subsections.

3.1 Measurement functions

1. Shape cue: The use of shape-based cues requires that silhouette templates of human limbs have been learnt beforehand (figure 2). Each particle x is classically given an edge-based likelihood $p(z^S|x)$ that depends on the sum of the squared distances between N_p points uniformly distributed along the template corresponding to x and their nearest image edges [3], *i.e.*

$$p(z^S|x) \propto \exp\left(-\frac{D^2}{2\sigma_S^2}\right),\ D = \sum_{j=1}^{N_p} |x(j) - z(j)|, \tag{2}$$

where the similarity measure D involves each j-th template point $x(j)$ and associated closest edge $z(j)$ in the image, the standard deviation σ_S being determined *a priori*.

2. Color cue: Reference color models can be associated with the targeted ROIs. These models are defined either *a priori*, or on-line using some automatic detection modules. We denote the N_{bi}-bin normalized reference histogram model in channel $c \in \{R, G, B\}$ by $h^c_{ref} = (h^c_{1,ref}, \ldots, h^c_{N_{bi},ref})$. The color distribution $h^c_x = (h^c_{1,x}, \ldots, h^c_{N_{bi},x})$ of the region B_x corresponding to the state x is computed as

$$h^c_{j,x} = c_H \sum_{u \in B_x} \delta_j(b^c_u),\ j = 1, \ldots, N_{bi}, \tag{3}$$

where $b^c_u \in \{1, \ldots, N_{bi}\}$ denotes the histogram bin index associated with the intensity at pixel u in channel c of the color image, δ_a terms the Kronecker delta function at a, and c_H is a normalization factor. The color likelihood model must be defined so as to favor candidate color histograms h^c_x close to the reference histogram h^c_{ref}. The likelihood $p(z^C|x)$ has a form similar to (2), provided that D terms the Bhattacharyya distance $D(h^c_x, h^c_{ref})$ [7] between the two histograms h^c_x and h^c_{ref}.

To overcome the ROIs appearance changes in the video stream, the target reference model is updated from the estimates in each frame through a first-order filtering process, so that the farther a frame in the past, the least its contribution [7]. Moreover, in order to avoid the tracker to be distracted by color-like clutters, one can make the likelihood $p(z^C|x)$ depict the similarity of several color patches related to the particle x w.r.t. convenient reference values. In other words, it may be worth splitting the ROI into subregions, e.g. the face and clothes of a person, each with its own reference color model.

3. Motion cue: In our context, it is highly possible that the targeted subject be moving, at least intermittently. To cope with background clutter, we thus favor the moving edges (if any) by combining motion and shape cues into the definition of the likelihood of particle x. Given $\overrightarrow{f}(z(j))$ the optical flow vector for pixel $z(j)$, the similarity distance D in (2) is then replaced by

$$D = \sum_{j=1}^{N_p} |x(j) - z(j)| + \rho\gamma(z(j)), \tag{4}$$

where $\gamma(z(j)) = 0$ (resp. 1) if $\overrightarrow{f}(z(j)) \neq 0$ (resp. if $\overrightarrow{f}(z(j)) = 0$) and $\rho > 0$ terms a penalty.

4. Multi-cues fusion: The above measurements are assumed mutually independent conditioned on the state, *i.e.* weak correlation exists between the color, motion and shape of the tracked object. Given M measurement sources $(z^1, \ldots, z^M)$, the global measurement function thus factorizes as

$$p(z|x) = \prod_{m=1}^{M} p(z^m|x). \tag{5}$$

3.2 Importance functions

1. Shape cue: We use the face detector introduced by Viola *et al.* [9]. It is based on a boosted cascade of Haar-like features to measure relative darkness between eyes and nose/cheek or nose bridge. Let B be the number of detected faces and $\mathbf{b}_n^{'S}$ the centroid coordinate of each such region. An importance function $q(.)$ at location $\mathbf{x} = (u_k, v_k)$ follows, as the Gaussian mixture

$$q(\mathbf{x}|z^S) = \sum_{n=1}^{B} \delta_n^S . \mathcal{N}(\mathbf{b}_n^S, \Sigma_B^S), \tag{6}$$

where $\mathbf{b}_n^S = \mathbf{b}_n^{'S} + \bar{X}_B^S$. The vector $\bar{X}_B^S$ and matrix Σ_B^S, which respectively term the mean and covariance of the offset from the ROI position to the centroid of the associated contour describing a face, are learnt off-line.

2. Color cue: Human skin colors have a specific distribution in color space. Training images from database [5] are used to construct a reference color histogram model in (R, G, B). Blobs detection is performed by subsampling the input image prior to grouping the classified skin-like pixels. The importance function $q(x|z^C)$ on the detected blobs is defined by a Gaussian mixture similar to (6).

3. Motion cue: For a static camera, a basic method consists in computing the luminance absolute difference image from successive frames. To detect regions of significant motion activity, we define the reference motion histogram h_{ref}^M as $h_{j,ref}^M = \frac{1}{N_{bi}}, \; j = 1, \ldots, N_{bi}$ (see [7] for details). We evaluate here the Bhattacharyya distance $D(h_x^M, h_{ref}^M)$ on a subset of locations obtained by subsampling the image. These locations are taken as the nodes of a regular grid. Locations that satisfy $D^2(h_x^M, h_{ref}^M) > \tau$ are selected. In the vein of (6), the importance function $q(\mathbf{x}|z^M)$ is a mixture centered on the detected locations of high motion activity.

4. Multi-cues mixture: The importance function $q(.)$ can be extended to consider the output from M detection modules, *i.e.*

$$q(\mathbf{x}|z) = \frac{1}{M} \sum_{j=1}^{M} q(\mathbf{x}|z^j). \tag{7}$$

4 Trackers for our "Tour-Guide Robot" modalities

For our three visual modalities, the aim is to fit the *template* relative to the tracked visitor all along the video stream, through the estimation of its image coordinates (u, v), its scale factor s, as well as, if the template is shape-based, its orientation θ. All these parameters are accounted for in the state vector x_k related to the k-th frame. With regard to the dynamics model $p(x_k|x_{k-1})$, the image motions of observed people are difficult to characterize over time. This weak knowledge is thus formalized by defining the state vector as $x_k = [u_k, v_k, s_k, \theta_k]'$ and assuming that its entries evolve according to mutually independent random walk models, viz. $p(x_k|x_{k-1}) = \mathcal{N}(x_k|x_{k-1}, \Sigma)$, where $\mathcal{N}(.|\mu, \Sigma)$ is a Gaussian distribution with mean μ and covariance $\Sigma = \mathrm{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2, \sigma_\theta^2)$.

A preliminary evaluation enables the selection of the most meaningful visual cues associations in terms of discriminative power, robustness to artefacts (*e.g.* clutter or illumination changes) and time consumption, be these cues involved in the importance or measurement functions. As a result, dedicated visual cues are selected for each modality.

The filtering strategies depicted in § 2 are then evaluated in order to check which ones best fulfill the requirements of the considered H/R interaction modalities. For the sake of comparisons, importance functions rely on dynamics or measurements alone (and are respectively noted DIF for "Dynamics-based Importance Function" and MIF for "Measurement-based Importance Function"), or combine both (and are termed DMIF for "Dynamics and Measurement-based Importance Function"). Further, each modality is evaluated on a database of sequences acquired from the robot in a wide range of typical conditions: cluttered environments, appearance changes or sporadic disappearance of the targeted subject, jumps in his dynamics, etc. For each sequence, the mean estimation error with respect to "ground truth", together with the mean failure ratio ($\%$ of target loss), are computed from several filter runs. The associated figure plots are not shown here for space reasons but they can be found at `www.laas.fr/~lbrethes/KE2_2k5`. These results motivate our choices depicted hereafter for the three visual modalities. The processing sampling rate of all these modalities ranges from $20Hz$ to $50Hz$ on a $3GHz$ Pentium IV for a particles number ranging from 100 to 200.

4.1 Tracker dedicated to the search for interaction

Fig. 3. The template.

Regarding this modality, color and motion ROIs, as shown in figure 3, are fused into the particles likelihood (5). The importance function involves the motion detector, yielding $q(x_k|z_k^M)$.

The two filters $\mathrm{MSIR/RBSSHSSIR}$ are well-suited to this modality. The hierarchical scheme contitutes an alternative, along the lines of Perez *et al.* in [7]. However, though its intermediate sampling step enables the particles cloud to remain more focused on the target, which results in a tracking error decrease, its failure ratio is significantly pronounced on sequences showing occultations of the target. As robustness is prefered to precision

for our application, we finally opt for the RBSSHSSIR algorithm using a Dynamics and Measurement-based Importance Function. Figure 4 shows a tracking run for such a scenario.

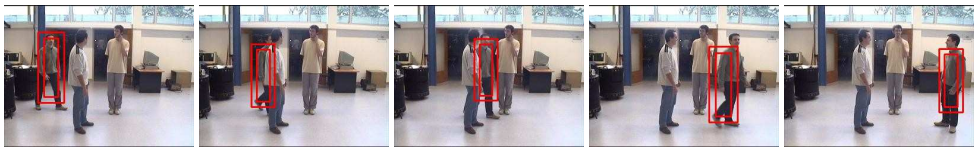

Fig. 4. A scenario involving persistent occlusions due to persons. Tracker based on a DMIF into the RBSSHSSIR algorithm.

4.2 Tracker dedicated to the proximal interaction

The selected tracking strategy for proximal interaction combines shape and motion cues (figure 5, eq. (4)) into a CONDENSATION algorithm. Indeed, evaluations show that Dynamics-based Importance Functions lead to a better precision together with a low failure ratio, so that detection modules are not necessary in this easiest context. Moreover, among the filtering strategies, the CONDENSATION enjoys the least time consumption.

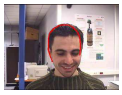

Fig. 5. The template.

4.3 Tracker dedicated to the guidance mission

Regarding this modality, shape and color cues are also fused into the particles likelihoods (5). Considering multiple patches of color distribution (figure 6) along with an update of the reference histograms h^c_{ref}, enables the tracker to keep focusing on the guided visitor even if several persons enter the camera field of view.

The importance functions of the MSIR and RBSSHSSIR strategies combine the outputs from color blob and face detectors along eq. (7). These associations lead to lower false negatives for a given detection rate. Experiments on sequences including cluttered background and/or appearance changes prove that fusing measurements clearly improves the discriminative power while MSIR and RBSSHSSIR strategies are shown to perform as well as CONDENSATION. Experiments on the sequences subset including additional sporadic disappearances (due to occlusions or to the limits of the camera field of view) highlight the efficiency of MSIR/RBSSHSSIR strategies in term of failure ratio. Figure 7 shows a tracking run in this context. In fact, these two strategies have the ability to recover from such artefacts because some particles are drawn from the visual detectors in the proposal. Finally, the RBSSHSSIR filter leads to a slightly better precision than MSIR. The nice property of the RBSSHSSIR strategy is to associate more efficiently particles sampled from the proposal and their plausible predecessors thanks to resampling stages.

Fig. 6. The template.

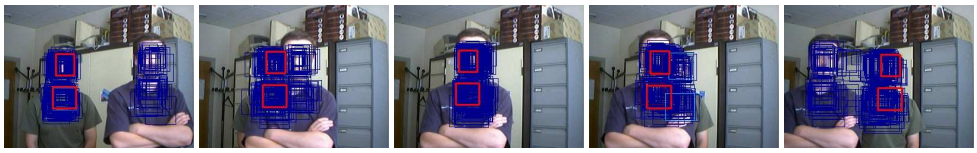

Fig. 7. Tracking scenario involving occlusions with RBSSHSSIR and DMIF. The blue (resp. red) rectangles depict all particles (resp. the MMSE estimate).

5 Conclusion

In this paper we introduced mechanisms for data fusion within particle filtering to develop trackers combining/fusing color, motion and shape cues in a novel way. The most persistent of them were used in the particles weighting stage. The others, logically intermittent, act in detection and initialization modules. Dedicated particle filtering strategies have been evaluated in order to check which trackers regarding visual cues and algorithms associations best fulfill the requirements of considered robotics scenarii dedicated to H/R interaction. The multi-cues associations proved to be more robust than any of the cues individually. Finally, we have integrated these trackers on our robot to highlight the relevance of our visual modalities.

In a near future, we plan to fuse other information such as sound cues and adapt our tracker to be able to track multiple persons simultaneously. Further evaluations will consider trackers not exclusively limited to particle filtering.

References

1. A. Doucet, N. De Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Series Statistics For Engineering and Information Science. Springer-Verlag, New York, 2001.
2. A. Doucet, S. J. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.
3. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. Comput. Vision*, 29(1):5–28, 1998.
4. M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *ECCV '98: Proceedings of the 5th European Conference On Computer Vision-Volume I*, pages 893–908, London, UK, 1998. Springer-Verlag.
5. M.. Jones and J. Rehg. Color detection. Technical report, Compaq Cambridge Research Lab, 11 1998.
6. Michael K. Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.
7. P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.
8. P. Torma and C. Szepesvári. Sequential importance sampling for visual tracking reconsidered. In *AI and Statistics*, pages 198–205, 2003.
9. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Int. Conf. On Computer Vision and Pattern Recognition*, 2001.