Data Fusion within a modified Annealed Particle Filter dedicated to Human Motion Capture

Mathias Fontmarty, Frédéric Lerasle and Patrick Danès

***Abstract*— This paper presents a new algorithm for human motion three-dimensional tracking based on a stereo camera system embedded on a mobile robot. The approach mixes advantages of the well-known ICONDENSATION and Annealed particle filters into a more reliable "I-Annealed" particle filter based tracker. Data fusion is also studied to show that a wide variety of visual cues must be used so that the system can adapt to various backgrounds. A complete implementation of the proposed tracker is described as well as some results on indoor sequences. Finally, evolutions and future work are discussed.**

I. Introduction and framework

A major challenge of Robotics is undoubtedly the personal robot, with the perspective for such an autonomous mobile platform to serve humans in their daily life. Embedding human motion capture (HMC) systems thanks to conventional cameras mounted on a robot would give it the ability to (i) act in a socially and human aware way, (ii) communicate with humans thanks to a natural and rich means.

Besides, 3D tracking from a mobile platform is a very challenging task, which imposes several requirements. First, the embedded sensors are positioned close to each other and so cover a narrow field of view comparatively to multi-ocular systems. As the robot's evolution takes place within a wide variety of environmental conditions, background modeling techniques [5], [15], [16] are precluded and the tracker gets inevitably faced with ambiguous data. Moreover, frequent occurrences of mutual occlusions between limbs require automatic (re-)initialization procedures. Clearly, several hypotheses must be handled simultaneously, and a robust integration of multiple visual cues is necessary. Finally, onboard processing power is limited and care must be taken to design computationally efficient algorithms.

Like many researchers in the Vision community, we aim at investigating markerless human motion capture systems based on vision techniques. Most of the existing approaches have concentrated on 3D articulated models of the tracked human limbs in order to make the problem more tractable (see a survey in [11]). They essentially differ in the sensor setup and the associated data processing so that two main classes can be exhibited, namely 3D reconstruction and appearance based approaches. The former ones try to fit the articulated model on the 3D-point cloud issued from a 3D-sensor system, *e.g.* a stereo head [3], [18] or a Swiss Ranger [9]. On the other hand, the appearance-based approaches infer the model configuration from its projection in monocular [10], [14], [16] or multi-ocular [5], [15] image sequences. These last strategies enable the derivation of abundant appearance information from the image contents, yet they may misestimate the motion-in-depth. This is so when using a single camera, or even a short-baseline stereo camera setup such as these widely used in mobile robotics.

Our challenge is to design a human motion capture system which copes with the above robotics requirements. Our observation model is based on a robust and probabilistically motivated integration of multiple cues. Thus, fusing 3D and 2D (image-based) information from the video stream of a stereo head—with cameras positioned at a distance of 30 centimeters—should enable to benefit both from reconstruction and appearance-based approaches.

Regarding the estimation process, Monte Carlo simulation methods, also known as particle filters [6] have proved well suited to our context. Indeed, they make no restrictive assumption on the probability distributions entailed in the characterization of the problem, and permit an easy fusion of diverse kinds of measurements. The main drawback for conventional particle filters remains the number of required particles which increases exponentially with the state-space dimensionality. Search space decomposition techniques [5], [2] undoubtedly enable to tackle this problem, yet global view strategies are often favored to determine the correct configuration. The Annealed Particle Filtering (APF), pioneered by Deutscher *et al.* in [4]—though with a much simplified observation model in a non-robotics context—is another way to address this difficulty.

In this paper, the APF is improved and extended in two ways. First, this framework together with the aforementioned data fusion principle can decrease the effective search space, through a multiple cue based likelihood function with gradually narrowing peaks. The second line of investigation concerns automatic (re-)initialization. When traditional PF algorithms loose track—as is always the case in cluttered scenes—the dimensionality of the state space makes any recovery difficult. Consequently, re-initialization is not straightforward. Feature detectors within the ICONDENSATION framework [8] do address this problem, but as far as we know this strategy has been exclusively devoted to 2D tracking. If one could detect some body parts then inverse kinematics could be used to solve for the model 3D pose (re-)initialization. We thus propose a modified APF (termed IAPF) which incorporates the prominent properties of an ICONDENSATION algorithm.

The paper is organized as follows. Section II first briefly summarizes the APF algorithm and describes the amended version to account for diverse kinds of measurements. Then section III specifies these 3D or image-based cues. Sec-

M. Fontmarty, F. Lerasle and P. Danès are with Université de Toulouse: LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse Cédex 4, FRANCE and Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cédex, FRANCE. firstname.name@laas.fr

tion IV describes our setup and associated evaluations. Last, section V summarizes our contribution and puts forward some future extensions.

II. Annealed Particle Filter for data fusion

A. The APF algorithm

Like all other particle filters, the Annealed Particle Filter (APF) algorithm is a Monte Carlo method dedicated to the recursive estimation of the state vector of a Markovian stochastic system. Its aim is to approximate the posterior probability $p(x_k|z_{1:k})$ of the state vector x_k at time k conditionally to the measurements $z_{1:k} = z_1, \dots, z_k$ by a point-mass distribution :

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

which represents the selection of a value—or particle—$x_k^{(i)}$ with probability—or weight—$w_k^{(i)}$, $i = 1, \dots, N$. The posterior conditional mean of any function of x_k, *e.g.* the Minimum Mean Square Error (MMSE) estimate $\mathrm{E}[x_k|z_{1:k}]$, immediately follows.

Consider a system of state x_k, whose dynamics and observation density can respectively be described by $p(x_k|x_{k-1})$ and $p(z_k|x_k)$. The basic APF scheme is presented in table I where the "$Strategy$" parameter must take the value "APF". The main idea is to split the classical CONDENSATION [1] main loop into L layers. Each stage $l \in \{1, \dots, L\}$ processes the set of particles computed by the previous level. It applies a "layer dynamics function" $p_l(x_{k,l}|x_{k,l-1})$ to samples $x_{k,l-1}^{(i)}$, then focusses the resulting particles in regions where a "layer likelihood function" $p_l(z_k|x_{k,l})$ presents high values. The main idea is to define these functions smartly enough in order to improve the results of a classical CONDENSATION in high dimension spaces. Deutscher *et al.* in [4] propose to select

$$\begin{aligned} p_l(x_{k,l}|x_{k,l-1}) &= [p(x_k|x_{k-1})^{\alpha_l}]_{x_k = x_{k,l}; x_{k-1} = x_{k,l-1}} \\ p_l(z_k|x_{k,l}) &= [p(z_k|x_k)^{\beta_l}]_{x_k = x_{k,l}} \end{aligned}$$

where $\alpha_l \in [1, +\infty)$ and $\beta_l \in [0, 1]$ are increasing sequences of parameters. Hence, as l grows, α_l increases, and the state space exploration becomes sharper. Meanwhile the coefficients β_l increase, and while the first layers use a very smoothed likelihood function (β_l is small), the last layers use a potentially peaked one (β_l tends towards 1).

Sampling from $p(x_k|x_{k-1})^{\alpha_l}$ may not be trivial. However, in our context, since we use a random walk dynamics $p(x_k|x_{k-1}) = \mathcal{N}(x_{k-1}, \Delta_k)$ with Δ_k diagonal, sampling from $p(x_k|x_{k-1})^{\alpha_l}$ is equivalent to sampling from $\mathcal{N}(x_{k-1}, \frac{1}{\alpha_l}\Delta_k)$.

B. The amended APF algorithm

The extended APF algorithm we propose is inspired from the ICONDENSATION [8]. The main idea is to explore the state space using an importance function $q(x_k|x_{k-1}, z_k)$ in place of the system dynamics $p(x_k|x_{k-1})$. As the initial particle cloud is sharpened within each layer of the APF, the importance sampling is introduced only in the first layer. The algorithm is presented in table I with the parameter "$Strategy$" taking the value "IAPF". The use of importance sampling enables self-initialization or reinitialization in case of target loss, which must be taken in account in our mobile robotics context. The importance function $q(x_k|x_{k-1}, z_k)$ we use in our amended APF involves both measurements and dynamics, such as the classical ICONDENSATION does.

We must notice that, as is the case for the APF, whose authors say "its only disadvantage is not being able to work in a robust Bayesian framework" [4], IAPF is not a mathematically sound Monte Carlo method.

III. Description of the cues

The importance function $q(.)$ generally involves discriminant but possibly intermittent visual cues[1] while measurement functions $p(z_k|x_k)$ involve cues which are persistent yet proner to ambiguity for cluttered scenes [12]. Fusing several cues confers robustness w.r.t temporary failures in some of the measurement processes. The next subsections describe our importance function followed by our multiple cues based measurement function.

A. Importance function

We sample α percent of the particles according to the dynamics, β percent according to the measure and the last ones according to the prior $p(x_0)$. The particles sampled from the measure are drawn from a Gaussian distribution centered on a configuration x_k^D computed from 3D positions of head and hands thanks to an analytical Inverse-Kinematics (IK) algorithm. These features are extracted by skin color blob segmentation, then matched in the image pair. 2D blob matching procedure is based upon criteria defined in [13, Chap. 4]. The centroids of the matched regions are finally triangulated using the parameters of the calibrated stereo setup. Thus, hands and head can be understood as three natural markers.

Sampling x according to $q(x_k|x_{k-1}, z_k)$ is then analogous to drawing $u \sim \mathcal{U}(0, 1)$ and then sampling:

- $x \sim p(x_k|x_{k-1})$ if $u < \alpha$
- $x \sim \mathcal{N}(x_k^D, \Delta_k)$ if $\alpha \leq u < \alpha + \beta$
- $x \sim \mathcal{N}(x_0, \Delta_k)$ if $u \geq \alpha + \beta$

where Δ_k is a covariance matrix. In our context, it is the same matrix as the system dynamics covariance.

B. Likelihood sub-functions

1) Edge distance: This likelihood requires the projection of the 3D model and the removal of its hidden parts. The shape related likelihood is classically computed using the sum of the squared distances between model points and the nearest image edges. These N_p measurement points $p_i, i \in \{1, \dots, N_p\}$ for a configuration x_k are chosen to be uniformly distributed along the model projected segments. In this implementation, the edge image is converted into a Distance Transform image, noted I_{DT}, which is used to pick the distance value [7]. This likelihood is given by

$$p(z_k^e|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_e^2}\right), \; D = \frac{1}{N_p}\sum_{i=1}^{N_p} I_{DT}(p_i),$$

[1] due to occlusions or mis-segmentation in our case.

```
$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^{N} = \mathrm{APF}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^{N}, z_k, Strategy)$
IF $k = 0$, Sample $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$. Set $\alpha_1, \ldots, \alpha_L \in [1, +\infty)$ the increasing dynamics function exponents
  and $\beta_1, \ldots, \beta_L \in [0, 1]$ the increasing likelihood functions exponents. END IF
IF $k \geq 1$ THEN {—$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^{N}$ describes a particle approximation of $p(x_{k-1}|z_{1:k-1})$—}
  Set $[\{x_{k,0}^{(i)}, w_{k,0}^{(i)}\}]_{i=1}^{N} = [\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^{N}$
  FOR $l = 1, \ldots, L$, DO
    IF $l == 1$ && $Strategy ==$ IAPF THEN
      FOR $i = 1, \ldots, N$, DO
        Propagate the particle $x_{k,l-1}^{(i)}$ using importance function $x_{k,l}^{(i)} \sim q(x_{k,l}|x_{k,l-1}^{(i)}, z_k)$
        Update the weight $w_{k,l-1}^{(i)}$ associated to $x_{k,l-1}^{(i)}$ by $w_{k,l}^{(i)} \propto \frac{p_l(z_k|x_{k,l}^{(i)})p_l(x_{k,l}^{(i)}|x_{k,l-1}^{(i)})}{q(x_{k,l}^{(i)}|x_{k,l-1}^{(i)}, z_k)}$ with $p_l(x_{k,l}^{(i)}|x_{k,l-1}^{(i)}) = p(x_{k,l}^{(i)}|x_{k,l-1}^{(i)})^{\alpha_l}$
                                                                                             and $p_l(z_k|x_{k,l}^{(i)}) = p(z_k|x_{k,l}^{(i)})^{\beta_l}$
      END FOR
      Normalize the weights $w_{k,l}^{(i)}$ so that $\sum_i w_{k,l}^{(i)} = 1$
      Resample the particle representation $[\{x_{k,l}^{(i)}, w_{k,l}^{(i)}\}]_{i=1}^{N}$
    ELSE
      FOR $i = 1, \ldots, N$, DO
        Propagate the particle $x_{k,l-1}^{(i)}$ using the dynamics function $x_{k,l}^{(i)} \sim p_l(x_{k,l}|x_{k,l-1}^{(i)})$ with $p_l(x_{k,l}|x_{k,l-1}^{(i)}) = p(x_{k,l}|x_{k,l-1}^{(i)})^{\alpha_l}$
        Update the weight $w_{k,l-1}^{(i)}$ associated to $x_{k,l-1}^{(i)}$ by $w_{k,l}^{(i)} \propto p_l(z_k|x_{k,l}^{(i)})$ with $p_l(z_k|x_{k,l}^{(i)}) = p(z_k|x_{k,l}^{(i)})^{\beta_l}$
      END FOR
      Normalize the weights $w_{k,l}^{(i)}$ so that $\sum_i w_{k,l}^{(i)} = 1$
      Resample the particle representation $[\{x_{k,l}^{(i)}, w_{k,l}^{(i)}\}]_{i=1}^{N}$
    END IF
  END FOR
  Set $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^{N} = [\{x_{k,L}^{(i)}, w_{k,L}^{(i)}\}]_{i=1}^{N}$
  Compute the posterior mean $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, from particle representation $\sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)})$ of $p(x_k|z_{1:k})$
END IF
```

TABLE I
APF FRAMEWORK.

where i indexes the N_p model points, $I_{DT}(p_i)$ is the associated value in the DT image, and σ_e the *a priori* standard deviation of our Gaussian measure model.

2) ROI color histograms distance: Clothes colors create a clear distinction between the limbs (feet, trunk of sleeves, …) of the observed person. So, the use of clothing patches of characteristic color distributions seems very promising. N_{ROI} reference color models are associated with these targeted ROIs. Then, the histogram distance is written

$$p(z_k^{ROI}|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_{ROI}^2}\right),$$

$$D = \frac{1}{N_{ROI}} \sum_{i=1}^{N_{ROI}} \left(D_B(h_{x_k,i}, h_{ref,i})\right)$$

where D_B is the Bhattacharyya distance used to compare the normalized histograms $(h_{ref,i}, h_{x_k,i})$. The histograms appearances, *i.e.* the histograms $h_{ref,i}$, are learned on the first image of the sequence.

3) 3D blob distance: In the vein of our importance function $q(.)$, this measure involves the 3D positions $\widehat{P}_j = (X_j, Y_j, Z_j)'$ of the person hands and head ($j \in \{1, 2, 3\}$) after triangulation. We define :

$$p(z_k^{3d}|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_{3d}^2}\right), D = \frac{1}{3}\sum_{i=1}^{3} D_E(P_{x_k,i}, \widehat{P}_{j_i}),$$

where $D_E(P_{x_k,i}, \widehat{P}_{j_i})$ is the Euclidean distance between the mass center $\widehat{P}_{j_i}$ of blob j_i ($j_i \in \{1, \ldots, N_{Blob}\}$) and $P_{x_k,i}$ the 3D position of a hand or the head of the model under hypothesis x_k. Link between i and j_i is done by a simple heuristic involving 3D position of detected blobs and a face detector [17].

4) Skin color distance: In some cases, we cannot triangulate the 3D positions of hands and head (not enough detected blobs, triangulation error too high, …). Consequently, 3D information cannot be exploited. Nevertheless, we can still use the skin color segmented image I_S and define a new color-based likelihood $p(z_k^s|x_k)$. For a given state x_k, the 2D coordinates $p_{x_k,i}, i \in \{1, \ldots, 3\}$ of hands and head after model projection are supposed to be in skin color high probability areas, so that one can define

$$p(z_k^s|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_s^2}\right), D = \frac{1}{3}\sum_{i=1}^{3}(1 - I_S(p_{x_k,i})).$$

5) Homogeneous color distance: This measure uses N_m disjoints sets $E_i, i \in \{1, \ldots, N_m\}$ of uniformly sampled points inside each of the N_m projected body members for a configuration x_k.

We suppose the tracked person wears a cloth with a homogeneous color on each limb. We then use the following measure, with $\sigma_{E_i,c}$ the standard deviation of the color distribution on channel $c \in \{R, G, B\}$ associated to point set E_i of member i :

$$p(z_k^m|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_m^2}\right),$$

$$D = \frac{1}{N_m}\sum_{i=1}^{N_m}\left(\frac{1}{3}\sum_{c \in \{R,G,B\}} \sigma_{E_i,c}\right).$$

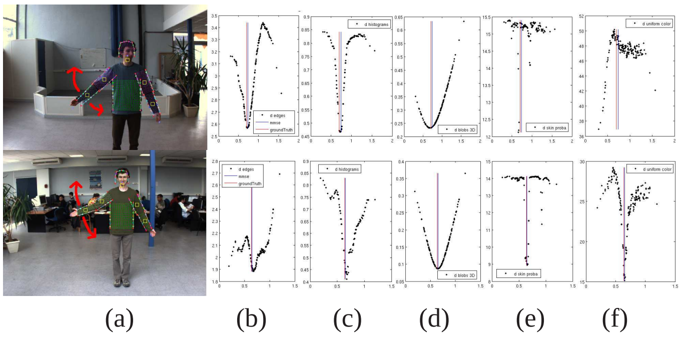

Fig. 1. Distance evolution regarding the position of a 1-DOF arm on clear and cluttered backgrounds. Figures (b), (c), (d), (e) and (f) respectively show distances relative to edges, ROI histograms, 3D blobs, skin color and uniform color on limbs. Red and blue lines respectively represent the ground truth (set manually) and the filter MMSE.

C. Cues study and discussion

The above measurements are assumed mutually independent conditioned on the state, so that the global measurement function factorizes as :

$$\begin{aligned} p(z_k|x_k) &= p(z_k^e, z_k^{ROI}, z_k^{3d}, z_k^s, z_k^m|x_k) \\ &= p(z_k^e|x_k)p(z_k^{ROI}|x_k)p(z_k^{3d}|x_k)p(z_k^s|x_k)p(z_k^m|x_k) \end{aligned}$$

Fig. 1 plots the distances obtained by sweeping a subspace of the configuration space formed by the orientation of the model right arm involving moderate or heavy background clutter. These plots bring out that appearance based measures are less discriminant than the ones involving 3D information (Fig. 1 (d)). In cluttered background, edge distance measure is not sufficiently discriminant as multiple *minima* are present (b) while color histograms on ROIs are quite robust to background clutter (c), but still very sensitive to illumination changes. The skin probability measure (e) is extremely sharp but shows some false positives as soon as spurious skin color like regions are detected. Color uniformity (f) performs well in a cluttered background if the tracked person wears a single color shirt, but very poorly otherwise.

Moreover, we can notice that among all of the cues presented above, no assumption has been made on the fact that the robot must not move, that is, trackers can perform well even if the scene background changes due to a robot motion (as long as this displacement is consistent with the system dynamics).

IV. System setup

A. Implementation and associated architecture

We use a human model based on truncated cones, even if boxes are prefered for 3D visualization as they provide a better way to see rotations. It uses 14 degrees of freedom : 6 for global localization, 3 rotations for each shoulder and 1 for each elbow. The head is supposed to be rigidly linked with the torso.

The software architecture is presented on fig. 2. We have chosen to set up 5 main modules :

- The image preprocessing module is in charge of a few standard tasks in order to obtain a good base image to work on. First, the module converts the raw data acquired from two mono-CCD stereo fire-wire cameras to RGB images. Then a white balance is applied, and the image is conveniently resized.
- The image processing stage computes various images to be used by the measure module. The first of them is the distance image I_{DT}, obtained by applying a distance transform on an edge image computed by a Canny edge detector. The module also computes a skin color image I_S, by back-projecting an off-line learned skin region histogram. These ones are used to triangulate the 3D position of head and hands. A face detector is also implemented among various other functionalities.
- The filtering module implements some classical particle filter schemes, among which a CONDENSATION, an ICONDENSATION, an APF, and our amended version of the latest : IAPF. The filters based on an importance function use data from the IK algorithm to draw samples from the measure. This stage provides the output from the global tracking system, *i.e.* the estimated 3D configuration of the person.
- The measure module is in charge of particle likelihood evaluations. For each particle of the chosen filtering strategy, it computes the distances presented in III-B, using some of the images computed by the image processing module.
- The IK algorithm computes a 3D configuration of the human body model from head and hand detections provided by the image processing module when these ones are available. The computed configuration is the nearest one to a given rest position.

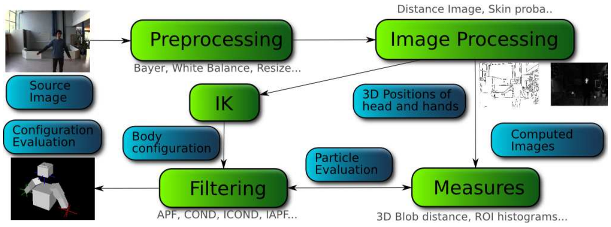
Distance Image, Skin proba..
Preprocessing
Image Processing
Bayer, White Balance, Resize...
Source Image
IK
3D Positions of head and hands
Configuration Evaluation
Body configuration
Computed Images
Particle Evaluation
Filtering
Measures
APF, COND, ICOND, IAPF...
3D Blob distance, ROI histograms...

Fig. 2. Software architecture of the tracking module.

B. Evaluations and discussion

Our IAPF-based tracker was tested on a number of challenging sequences, acquired from the robot, of human movement including temporary occlusions, jumps in the dynamic, heavy or moderate cluttered backgrounds. Refer to the URL `www.laas/∼mfontmar` for more videos and images. The current processing rate ranges from $1\ Hz$ to $4\ Hz$ on a $1.8\ GHz$ Pentium IV Centrino personal computer. We saw that by using both our importance and multiple cues measurement functions, we could reduce the size of the search space. This reduction of search space allows to limit drastically the required number of particles within $[200; 1000]$. The IAPF and APF we set up use 3 layers.

1) Accuracy: Accuracy in our context is very difficult to evaluate since we are not able to have access to a precise ground truth. That's why we first tested our tracker performance on synthetic data. Results are presented on figure 4. Since algorithms take place in a stochastic framework,

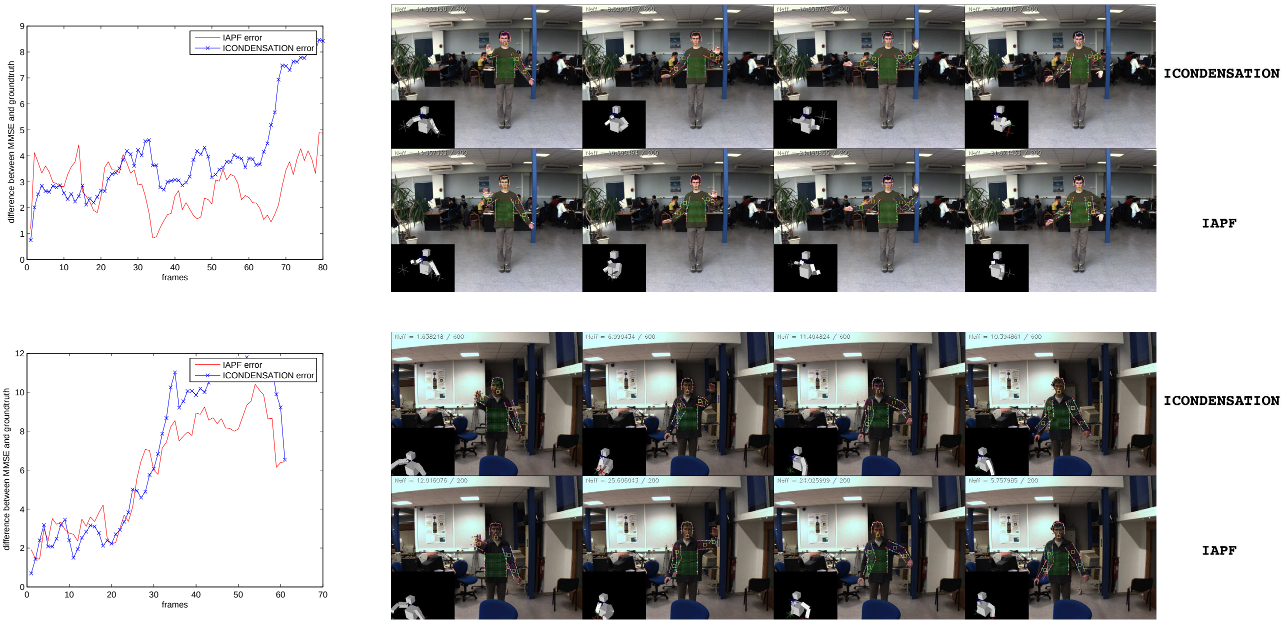

Fig. 3. ICONDENSATION and IAPF runs on two different sequences (up and down). Left plots show the error (Euclidean square distance) between estimates of each filter and the ground truth. For each plot, right sequences show the ICONDENSATION and IAPF runs.

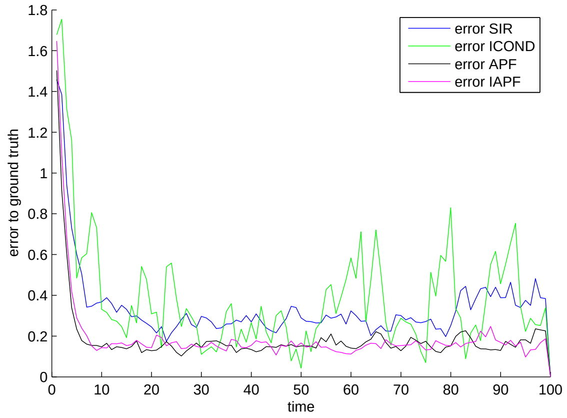

Fig. 4. Error between ground truth and estimates of different filtering strategies (SIR, ICONDENSATION, APF, IAPF)

Name	**Description**	**Value**
(W, H)	Image size	$(640, 480)$
N	Number of particles	600
N_{layers}	APF layer number	3
$(\alpha_1, \alpha_2, \alpha_3)$	Dynamics exponents	$(1, 4, 25)$
$(\beta_1, \beta_2, \beta_3)$	Likelihood exponents	$(0.1, 0.4, 1)$
(α, β)	$q(.)$ parameters	$(0.8, 0.2)$
$(\sigma_e, \sigma_{ROI}, \sigma_{3D},$ $\sigma_s, \sigma_m)$	Likelihood parameters	(10, 0.3, 0.15, 0.02, 10)
$p(x_k \mid x_{k-1})$	Dynamics	$\mathcal{N}(x_{k-1}, \Delta_k)$

TABLE II

PARAMETERS VALUES USED IN THE TRACKERS

we launched 15 runs of different filtering strategies, among which the IAPF one, for a linear Gaussian stochastic system using random walk dynamics. We then plot the mean error to ground truth for a state space of dimension 10. We can see that our IAPF strategy performs as well as the APF, and better than well-known ICONDENSATION and SIR. We noticed an average error reduction of 70 % for the IAPF with respect to ICONDENSATION.

In our visual tracking context, it seems that our IAPF gives at least similar results to ICONDENSATION, and, in some tricky cases, better ones. Once again, 15 runs of each filter have been performed on the same data to evaluate the mean behavior of each tracker. The fig. 3 shows the distance between a ground truth built "by hand" and the estimates provided by ICONDENSATION and IAPF trackers on a few sequences representative of the mean behavior of the filters. We noticed the error to ground truth is lowered from 0 to 15 degrees in rotations.

Nevertheless, those latest data must be handled carefully. For more reliable results, one should use a commercial HMC system, what we are not able to do at the moment.

2) Robustness: The main advantage of our IAPF over APF in our high dimension state tracking context is the possibility to initialize or re-initialize automatically — and so aid recovery from transient tracking failures —, which frees the tracker from the classical "by hand" initialization, as a detection of head and hands is enough to induce a 3D configuration of the model. We can even add more constraints to be sure that reference histograms are acquired correctly, *e.g.* the person must face the camera and have straight arms. A short sequence showing the IAPF-based tracker (re-)initialization is presented on fig 5. In these sequence, no prior draw is used.

3) Computation time: In all tests presented in this paper, ICONDENSATION has been run using N particles, while IAPF and APF has been run using N_L layers and N/N_L particles. This choice is justified by the limited CPU

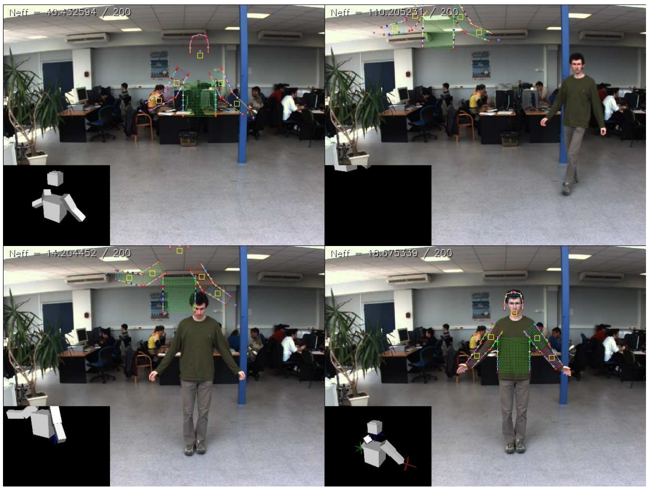

Fig. 5. From left to right and from top to down : the tracker is initialized with a default configuration which does not make sense as nobody is present. The tracker diverges. Head and hands detections enable to build a basic configuration near from the real one to initialize the particle cloud : the initialization succeeds.

resources we have : by choosing this ratio between particles and layers, all strategies use the same time to perform one iteration. Hence we can compare results for a given CPU resource. This also means that for same accuracy results, one should use less particles with IAPF strategy than with ICONDENSATION.

4) Material considerations: In its actual form, our tracker performs on 640×480 pixel images. Most of the time consumption — about $200\ ms$ — is spent in image processing and preprocessing (conversion from Bayer to RGB, skin probability computations, distance image, ...), and we hope to optimize this step. Furthermore, we could improve the tracker frequency by reducing the image size. A few parameter values used for the main sequences presented here are shown in table II. Most of them have been set up by experimenting different values, sometimes guided by simple heuristics.

V. Conclusion

This paper presents a fully automatic approach for tracking the upper human body parts in 3D. Two lines of investigations have been pursued. First, we endow the APF with the nice properties of ICONDENSATION scheme. This amended APF is shown to efficiently track articulated body motion and automatically recover from transient target loss or occlusion. In addition, it is more accurate than ICONDENSATION scheme for the same computation time. Second, data fusion principle is shown to improve the tracker versatility and robustness to clutter. Combined with today's powerful off-the-shelf PCs, such quasi real-time HMC approach devoted to mobile robot nearly becomes a reality and would have a great number of robotics applications.

While our results are promising, further investigations regarding implementation and evaluation are postponed to future research. Our strategy is based on simple detectors dedicated to specific features. Hands are segmented by color while head detection is based on Adaboost [17] which proves faster and more robust than color segmentation. Such machine learning methods will be extended to other limbs [15] in order to provide additional "initialization" cues. To our belief, the difference with conventional particle filters would be more significant if we had better part detectors. Next, our observation model will be enriched with sparse stereo-correlation data. Further evaluations will be also performed using a motion capture testbed that provides more accurate "ground truth" from a commercial HMC system, such as VICON, that will be synchronized with the video streams.

References

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Trans. on Signal Processing*, 2(50):174–188, 2002.

[2] J. Mac Cormick and M. Isard. Partitioned sampling, articulated objects, and interface quality hand tracking. In *European Conf. on Computer Vision (ECCV'00)*, pages 3–19, 2000.

[3] Q. Delamarre and O. Faugeras. 3D articulated models and multi-view tracking with physical forces. *Computer Vision and Image Understanding (CVIU'01)*, 81:328–357, 2001.

[4] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'00)*, pages 126–133, 2000.

[5] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, pages 669–676, 2001.

[6] 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.

[7] J. Giebel, D. M. Gavrila, and C. Schnorr. A bayesian framework for multi-cue 3D object. In *European Conf. on Computer Vision (ECCV'04)*, Pragues, 2004.

[8] M. Isard and A. Blake. I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *European Conf. on Computer Vision (ECCV'98)*, pages 893–908, 1998.

[9] S. Knoop, S. Vacek, and R. Dillman. Sensor fusion for 3D human body tracking with an articulated3D body model. In *Int. Conf. on Robotics and Automation (ICRA'06)*, pages 1686–1691, Orlando (USA), May 2006.

[10] P. Menezes, F. Lerasle, and J. Dias. Visual tracking modalities for a companion robot. In *Int. Conf. on Intelligent Robots and Systems (IROS'06)*, Beijing, 2006.

[11] T. Moeslund and E. Granum. A survey on computer vision-based human motion capture. *Computer Vision and Image Understanding (CVIU'01)*, 81:231–268, 2001.

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

[13] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, 1996.

[14] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *European Conf. on Computer Vision (ECCV'00)*, pages 702–718, 2000.

[15] L. Sigal, S. Bhatia, S. Roth, M.J. Black, and M. Isard. Tracking loose-limbed people. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, 2004.

[16] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *Int. Journal on Robotic Research (IJRR'03)*, 6(22):371–393, 2003.

[17] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, 2001.

[18] J. Ziegler, K. Nickel, and R. Stiefehagen. Tracking of the articulated upper body on multi-view stereo image sequences. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, 2006.