Visual human motion capture from a mobile robot

Mathias Fontmarty †‡, Frédéric Lerasle †‡ and Patrick Danès †‡

Abstract— **In this paper, we present a visual human motion capture (HMC) system embedded on a mobile robot. The estimation of the 3D location and configuration of our 14 DOF model of the upper human body is performed with a particle filter. We use a stereoscopic camera to derive appearance based visual cues together with 3D measurements computed from a sparse 3D reconstruction. A comparative study is carried out in order to achieve the best association between measurement cues and filtering strategies in our robotics context. The system performs in real-time in various indoor environments.**

I. INTRODUCTION

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. In such a context, human motion capture (HMC) systems are of great interest as they enable the robot to localize the human and to analyze his/her configuration. HMC constitutes the groundwork of many activities such as interpretation of tasks, poses or motions, coordinated object manipulation, imitation learning. Embedding HMC systems thanks to conventional cameras mounted on a robot would give it the ability to act in a socially and human aware way, and enable the user to communicate 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 [1], [2], [3] 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, on board processing power is limited and care must be taken to design computationally efficient algorithms.

All these problems result in very few fully integrated 3D HMC systems on interactive robots. Consequently, our challenge is to design an HMC system which copes with the above robotics requirements. To tackle this problem, various methods have been proposed in the literature.

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 [4]). They essentially differ in the sensor setup and the associated data processing so that two main classes can be exhibited. On the one hand approaches based on 3D reconstruction aims at fitting the articulated model on the 3D point cloud issued from a 3D sensor system, *e.g.* a stereo head [5] or a Swiss Ranger [6]. On the other hand, the appearance-based approaches infer the model pose and configuration from its projection in monocular [7], [3] or multi-ocular [1], [2] 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 approach aims at mixing advantages of both appearance-based and 3D reconstruction methods.

Regarding the estimation process, Monte Carlo simulation methods, also known as particle filters (PF) [8] 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 PFs remains the number of required particles which increases exponentially with the state-space dimensionality. To tackle this problem, various search space decomposition techniques have been proposed together with more efficient sampling methods [1], [9].

In front of all these various approaches and in order to setup an embedded HMC system, we propose a quantitative evaluation of many state-of-the-arts methods so as to find out which combination of filtering strategy and measurement cues performs best.

The paper is organized as follows. Section II presents our robot and the system architecture. Then, sections III and IV respectively present the compared filtering strategies and the measurements cues. Associated evaluations together with some screenshots of the final system are presented in section V. Last, section VI summarizes our contribution and puts forward some future extensions.

II. SYSTEM ARCHITECTURE

We have set up the experiments on our JIDO robot. This MP-L655 platform from Neobotix is a mobile manipulator robot designed to interact with human beings. It embeds many sensors among which laser scanners and two stereo camera banks. In this paper, we use stereo cameras mounted on a pan-tilt unit at the top of its mast in order to exploit a larger field of view. A classical interaction context is presented on Fig. 1. JIDO is fitted with the Genom software architecture thoroughly presented in [10].

Our tracking system is implemented in a module named "HMC" fully integrated in the existing architecture. Its inner function is detailed in Fig. 2. First, it reads the stereo images

†CNRS; LAAS; 7 avenue du Colonel ROCHE, F-31077, Toulouse, France
‡Université de Toulouse; UPS, INSA, INP, ISAE; LAAS; F-31077, Toulouse, France
`firstname.lastname@laas.fr`

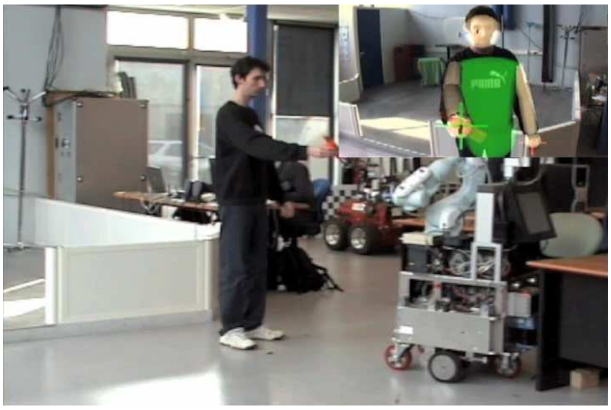

Fig. 1. A typical situation of Human-Robot interaction. The upper right part of the image represents the robot perception and the superimposed avatar is the estimated configuration.

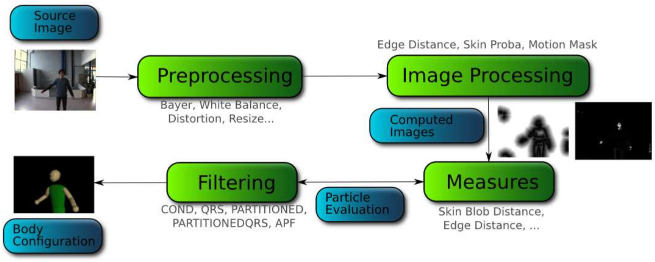
Source Image
Preprocessing
Bayer, White Balance, Distortion, Resize...
Edge Distance, Skin Proba, Motion Mask
Image Processing
Computed Images
Measures
Skin Blob Distance, Edge Distance, ...
Particle Evaluation
Filtering
COND, QRS, PARTITIONED, PARTITIONEDQRS, APF
Body Configuration

Fig. 2. The HMC module implementation.

from the camera. Some classical enhancement routines are applied in the preprocessing module: white balance (in order to lower illumination changes), distortion correction, image crop and resize, . . . Then, the processing module is in charge of extracting relevant information from the images using more or less classical visual cues (edge detection, motion flow, . . .). The filtering module finally estimates the 3D human pose thanks to a PF scheme, taking into account some prior dynamics and the extracted visual cues which are evaluated through the measurement module.

As we aim at proposing a markerless embedded HMC system, we focus on the tracking of the upper human body. Indeed, interactions rely mainly on hand and head movements. Thus, our model includes the torso, the head and the arms. We suppose that the head is fixed w.r.t. the torso because estimating head orientation needs very precise cues and the adopted image resolution is not sufficient to do so. The model is based on a kinematic tree consisting of 5 body segments and 14 DOF (6 for global localization and orientation, 3 for each shoulder and 1 for each elbow). It is fleshed out using truncated cones with fixed dimensions. These geometric primitives are easily handled and hidden part removal can be obtained in closed form. As our body model must be able to suit various subjects, member size are fixed to the human average. We assume Gaussian random walk prior dynamics.

The challenge here is to combine measurement cues and filtering strategies in the best way to provide a reliable tracking system under real-time constraints. To this end, we compare various combination of visual cues and filtering algorithms, which are presented below.

III. PARTICLE FILTER FRAMEWORK

A. Basics

In a stochastic Bayesian filtering approach to motion capture, the 3D template situation and configuration parameters to be estimated are first incorporated in a state vector $\mathbf{x}_k$, whose (given) initial probability density function (pdf) and prior dynamics write as $p_0(\mathbf{x}_0)$ and $p(\mathbf{x}_k|\mathbf{x}_{k-1})$. At any time k, the available visual data, symbolized by $\mathbf{z}_k$, is related to $\mathbf{x}_k$ by the observation density $p(\mathbf{z}_k|\mathbf{x}_k)$. Due to the high number of degrees of freedom (DOF) of the underlying articulated 3D model and to the difficulty to assess its projection onto the current images, the posterior pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ to be estimated is multimodal, defined in an high-dimensional state space, and unavailable in closed-form. A point-mass (or particle) approximation $p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)})$, $\sum_{i=1}^{N} w_k^{(i)} = 1$ is then recursively propagated along time through sequential Monte Carlo estimation methods [8], [11]. An approximation of the minimum mean-square error estimate (MMSEE) $\mathbb{E}(\mathbf{x}_k|\mathbf{z}_{1:k})$ follows.

The celebrated “Sampling Importance Resampling” (SIR) algorithm [8] operates in three major steps. First, the particles $\mathbf{x}_k^{(i)}$, $i = 1..N$ are drawn from an importance function $q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)$, selected to adaptively explore relevant areas of the state space. Then, the weights $w_k^{(i)}$ are updated to ensure the consistency of the point-mass approximation of the posterior pdf taking into account the observation density. Last, when the approximation tends to degenerate, a resampling stage is inserted. Importantly, the SIR framework encompasses the CONDENSATION [11] as well as importance sampling from the images. We present below some less well-known algorithms than the classical CONDENSATION or APF [1].

B. PARTITIONED particle filter

Contrarily to a common belief, the computation time of a particle filter for general problems, though linear in the number of particles, is exponential in the system order for a fixed dimension-free error [12]. To lower this complexity, many algorithms have been proposed. When the system dynamics comes as the sequence of M partial evolutions $p_m(\mathbf{x}_k^m|\mathbf{x}_k^{m-1})$ of the state vector $\mathbf{x}_k^m$ at step m and when intermediate likelihoods $l_m(\mathbf{x}_k^m|\mathbf{z}_k), m = 1..M$, can be assessed after applying each partial dynamics, PARTITIONED schemes apply [9].

From a succession of sampling operations followed by resampling based on the intermediate likelihoods, the particle cloud can be successively refined towards areas of the state space in which the posterior is dense. The computational complexity then becomes linear in the number of partitions. It can be applied for hierarchical systems where the root has to be placed before the leaves, *i.e.* arm/fist/fingers, or torso/arms in our case.

TABLE I

PARTITIONED QRS: PARTITIONED PARTICLE FILTER EXPLOITING QMC TECHNIQUES.

```
$\{(\mathbf{x}_k^{(i)}, w_k^{(i)})\}_{i=1}^N = PARTITIONEDQRS(\{(\mathbf{x}_{k-1}^{(i)}, w_{k-1}^{(i)})\}_{i=1}^N, \mathbf{z}_k)$

IF $k = 0$, THEN Sample a uniform randomized Sobol QMC sequence $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(N)}$ then turn it into $\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(N)} \sim p_0(\mathbf{x}_0)$ - Set $w_0^{(i)} = \frac{1}{N}$. END IF
IF $k \geq 1$ THEN
    Set $\tau_0^{(i)} = w_{k-1}^{(i)}$ and $\mathbf{x}_k^{0,(i)} = \mathbf{x}_{k-1}^{(i)}$, $i = 1..N$
    FOR $m = 1..M$, DO
        Independently draw $s^{(1)}, \ldots, s^{(N)}$ into $1..N$ such that $P(s^{(i)} = j) = \tau_{m-1}^{(j)}$ - Set $C_j = card(\{i | s^{(i)} = j\})$
        FOR $j = 1..N$, DO
            Sample a uniform randomized Sobol QMC sequence $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(C_j)}$ then turn it into $\mathbf{x}_k^{m,(\sum_{l=1}^{j-1} C_l + 1)}, \ldots, \mathbf{x}_k^{m,(\sum_{l=1}^{j-1} C_l + C_j)}$ i.i.d. according to $p_m(\mathbf{x}_k^m | \mathbf{x}_k^{m-1,(j)})$
        END FOR
        Update the weights $\tau_m^{(i)} \propto l_m(\mathbf{z}_k | \mathbf{x}_k^{m,(i)})$, then normalize them so that $\sum_i \tau_m^{(i)} = 1$
    END FOR
    Set $w_k^{(i)} = \tau_M^{(i)}$ and $\mathbf{x}_k^{(i)} = \mathbf{x}_k^{M,(i)}$ for $i = 1..N$
    Approximate the MMSEE $\mathbb{E}(\mathbf{x}_k | \mathbf{z}_{1:k})$ by $\sum_{i=1}^N w_k^{(i)} \mathbf{x}_k^{(i)}$
END IF
```

C. *Quasi Monte Carlo filtering methods*

Pure random importance sampling leads to "gaps and clusters" in the particle support, especially in high-dimension spaces. An excessive Monte Carlo variation of the predictions can follow, making the filter unreliable or even leading to failures. Substituting the random particles by a deterministic or randomized low-discrepancy—or "Quasi Monte Carlo" (QMC)—sequence can lead to a better convergence rate w.r.t. the number of particles N, while lowering the root mean square (RMS) estimation error and leading to a variability reduction from 5% to 20% [13], [14].

Among the main issues on QMC filters are the difficulty to design low-discrepancy sequences in spite of the resampling steps, the exploitation of the current measurement in the definition of these sequences, and the possible trade-off between the reduction of the (quadratic) complexity and the mathematical soundness of the algorithms. A QMC counterpart of CONDENSATION, henceforth termed QRS (for Quasi Random Sampling), is proposed in [15]. We adapted this idea to the PARTITIONED filter proposed in [9]. The final algorithm is described Table I. The key idea here is to gather importance sampling and resampling stages. This enables the generation of low discrepancy sequences from a particle to be resampled, thus resulting in a more regular state space exploration.

IV. MEASUREMENT CUES

A. *Likelihood sub-functions*

Our likelihood function $p(\mathbf{z}_k|\mathbf{x}_k)$ involved in the filtering algorithm relies on the use of appearance-based cues and geometric features. Some more or less widely used atomic likelihood functions are described below.

1) Color histogram: As in [16], we associate color histograms to N_C specific regions of interest (ROI) on our body model. For a given state hypothesis $\mathbf{x}_k$, the likelihood is computed through

$$p(\mathbf{z}_k^C|\mathbf{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_C^2}\right), \ D = \frac{1}{N_C}\sum_{i=1}^{N_C} \mathcal{B}(\mathbf{c}_{\mathbf{x}_k,i}, \mathbf{c}_i),$$

where $\mathcal{B}$ denotes the Bhattacharyya distance, $\mathbf{c}_{\mathbf{x}_k,i}$ is the color histogram modeling the image appearance of the i^{th} ROI on the projected model under hypothesis $\mathbf{x}_k$, $\mathbf{c}_i$ is the reference histogram, learnt on the first frame, and σ_C the *a priori* standard deviation of our Gaussian measure model. We chose to set up ROIs on the middle of arms, forearms and torso for a better learning of the model appearance.

2) Edge distance: The shape related likelihood is classically computed using the sum of the squared distances between model points and the nearest image edges [11]. These N_p model points $\mathbf{p}_{\mathbf{x}_k,i}, i \in \{1, \ldots, N_p\}$ for a configuration $\mathbf{x}_k$ are chosen to be uniformly distributed along the model projected segments. Edges of the image are extracted with a Canny detector. The result can be filtered with a movement mask in order to focus on moving edges. In some implementations, the image can be converted into a Distance Transform image, noted I_{ED}, which is used to pick the distance value [17]. The likelihood of a state hypothesis $\mathbf{x}_k$ is given by

$$p(\mathbf{z}_k^{ED}|\mathbf{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{ED}^2}\right), \ D = \frac{1}{N_p}\sum_{i=1}^{N_p} I_{ED}(\mathbf{p}_{\mathbf{x}_k,i}),$$

where $I_{ED}(\mathbf{p}_{\mathbf{x}_k,i})$ is the associated value in the edge distance image,

3) Skin color distance: Hands of the subject play a great role in the interaction process. Moreover, their motions are often faster than these of the other body parts as they constitute the ends of our kinematic chains. Consequently, we set up an additional visual cue dedicated to their location. A skin color probability image I_S is computed thanks to the back-projection of an offline learnt skin color histogram. High probability zones correspond to the head and hands. As is the case for the previous cue, we can filter the probability map with a movement mask and build a skin distance image I_{SD} from the detected skin blobs. The associated likelihood of a state hypothesis $\mathbf{x}_k$ is then computed by

$$p(\mathbf{z}_k^{SD}|\mathbf{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{SD}^2}\right), D = \frac{1}{3}\sum_{i=1}^{3} I_{SD}(\mathbf{h}_{\mathbf{x}_k,i}),$$

where $\mathbf{h}_{\mathbf{x}_k,i}, i \in \{1, 2, 3\}$ are the 2D coordinates of hands and head after model projection.

4) 3D skin blob distance: To complete the above visual cues and to improve the motion-in-depth estimation, we add a geometric 3D information issued from a sparse 3D reconstruction. In the vein of [18], we compute the 3D positions $\widehat{\mathbf{H}}_j = (X_j, Y_j, Z_j)'$ of the person hands and head ($j \in \{1, 2, 3\}$) : skin color blobs are extracted from each skin probability image (if many small blobs are near enough they are merged), and they are matched in the stereo images according to criteria defined in [19] and triangulation is performed. We define :

$$p(\mathbf{z}_k^{3D}|\mathbf{x}_k) \propto \exp\left(-\frac{D^2}{2\sigma_{3D}^2}\right), D = \frac{1}{3}\sum_{i=1}^{3}||\mathbf{H}_{\mathbf{x}_k,i} - \widehat{\mathbf{H}}_{j_i}||_2,$$

where $||\mathbf{H}_{\mathbf{x}_k,i} - \widehat{\mathbf{H}}_{j_i}||_2$ is the Euclidean distance between the mass center $\widehat{\mathbf{H}}_{j_i}$ of blob j_i ($j_i \in \{1, \ldots, N_{Blob}\}$) and the 3D position $\mathbf{H}_{\mathbf{x}_k,i}$ of a hand or the head of the model under hypothesis $\mathbf{x}_k$. Link between i and j_i is done by a simple heuristics involving detected blob 3D position and a face detector [20]. In some cases, we cannot triangulate the 3D positions of hands and head (too low number of detected blobs, triangulation error too high, ...). Consequently, 3D information cannot be exploited, and the 2D skin color distance is then the only cue enabling hand localization.

Assuming all the above likelihoods are mutually independent conditioned on the state $\mathbf{x}_k$, the unified likelihood factorizes as :

$$p(\mathbf{z}_k|\mathbf{x}_k) = p(\mathbf{z}_k^C|\mathbf{x}_k)p(\mathbf{z}_k^{ED}|\mathbf{x}_k)p(\mathbf{z}_k^{SD}|\mathbf{x}_k)p(\mathbf{z}_k^{3D}|\mathbf{x}_k).$$

In the next section, we present the evaluations of these measurement cues together with various filtering strategies.

V. ROBOTICS EXPERIMENTS

Evaluations are performed in order to check which HMC strategy, regarding visual cues and PF algorithms, best fulfill our robotics requirements.

A. Quantitative evaluations

30 runs of the tracker are performed on 4 different sequences including various movement types (arm waving, object manipulation, fitness...). Results of our system are compared to a groundtruth acquired thanks to a commercial HMC system from Motion Analysis [21]. We study the RMS (root mean square) error, which is the distance between the 3D average joint position of the estimated model and the true 3D joint positions given by the commercial HMC.

Concerning the visual cues comparison, our basis system focuses on the classical moving edges only ($\mathbf{z}_k^{ED}$) through a CONDENSATION with 1000 particles. The tracking is not satisfactory with an average RMS error greater than $40\ cm$ per joint. It appears that adding 2D skin color blob detection ($\mathbf{z}_k^{SD}$) greatly improves the estimation RMS error (until $24\ cm$ per joint). The constraints imposed on the end of the kinematic chains enable a more robust tracking of the hands. This speaks in favor of the combination of global and local attributes.

Moreover, the use of sparse 3D reconstruction to localize the hands and the head ($\mathbf{z}_k^{3D}$) lowers the estimation RMS error of $10\ cm$, mainly on the Z axis. Thus, mixing 2D and 3D cues enable a better behavior of the tracker. Concerning the color histograms ($\mathbf{z}_k^C$), they do not improve results if used with these other cues. This is due to the very peaky nature of the likelihood profile and the limited number of particles. Moreover, they are very heavy to compute and slow down the processing. The final RMS error per joint is about $12\ cm$, with $5\ cm$ on X and Y axis and $10\ cm$ on Z axis.

With this optimal measurement set, filtering strategies CONDENSATION, QRS PARTITIONED, PARTITIONED QRS and APF [1] have been compared. PARTITIONED strategies use 2 partitions: one for the torso localization and orientation (6 DOF) and one for the arm orientation (8 DOF). The APF filter is tuned following the hints in [1]. Finally, all strategies have been normalized with respect to the number of likelihood evaluations which is the most time-consuming part. We use Gaussian dynamics with standard deviation of $5\ cm$ for translations and $0.2\ rad$ for rotations.

Contrarily to our evaluations on visual cues, the filtering strategy does not seem to have a great impact on the system behavior. Fig. 3 shows the average RMS joint error and the estimate variance, computed as the average variance of the estimated 3D joint positions over the 30 runs of the tracker. PARTITIONED strategies seem to provide slightly better results in terms of RMS error and estimator dispersion. We can also notice that the famous APF performs even worst than the classical CONDENSATION. This can be explained by our robotics context where sensors do not provide enough information (especially in depth) to build well-conditioned likelihood functions. To support this, we have tested our algorithms in a classical context of multi-camera HMC, and we joined the results of the literature [22]. This means that in our ill-conditioned multi-modal robotic context, measurement cues are of greater importance than the filtering approach, what has – to our knowledge – never been reported in the literature. In addition, it seems that RMS estimation error does not lower very much for a number of particle $N > 500$. Thus, using an higher number of particles would only slow down the system by adding more computation without improving the average error. In our real-time robotics context, it is essential to take into account this key point.

B. Qualitative evaluations

Our embedded tracking system have been tested in various contexts including clear and cluttered background, with 3 different subjects. However, even if our robot is a mobile platform, we assume that it does not move when it is interacting with the subject. So, on a given sequence, cameras are fixed. We also suppose that there is only one interacting person in front of the cameras. However, even if we use a motion mask for skin blob and edge detection, the subject can still be tracked even if he stops moving for a few seconds, because the filter Gaussian dynamics is a random walk. Consequently to the above evaluation, we have implemented a PARTITIONED QRS filter with $\mathbf{z}^{ED}, \mathbf{z}^{SD}$ and $\mathbf{z}^{3D}$ cues. Our system currently performs at $5\ Hz$ on 320×240 pixel frames with 400 particles.

Fig. 4 presents a complex sequence with a cluttered background and a subject whose motion varies in depth and who goes out of the field of view of the camera. Only the images from one camera are shown. Despite the somewhat high RMS error on the Z axis, we can see that the tracking is visually satisfying. Furthermore, our tracker “initializes” automatically after a few frames due to our hybrid measurement cues (appearance based and geometric) which improve the state space exploration: this enables the

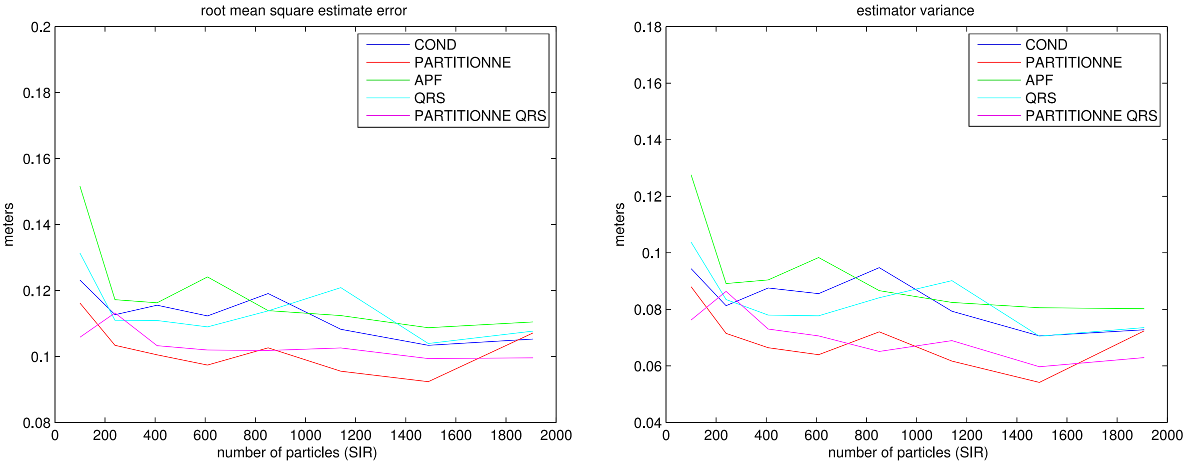

Fig. 3. RMS estimation error (left) and estimator dispersion (right) for different filtering strategies.

target to be locked even if the particle cloud is very far from it. When the subject grabs an object to give it to the robot, the tracker successfully localizes his hands. Currently, our system is efficient from $1\ m$ to $\sim 4\ m$.

In the sequence presented on Fig. 5, a different subject performs a reading activity in front of the robot. The automatic initialization is still efficient and the whole tracking is successful, despite a partial occlusion of the torso behind the table.

All these screenshots have been taken from the robot while the tracker was running in real-time. All videos can be seen at the URL `www.laas.fr/∼mfontmar`.

VI. CONCLUSIONS AND FUTURE WORKS

We proposed a visual HMC system embedded on a mobile robot. In order to deal with the robotics drastic constraints, we proposed a comparison of various measurement cues and particle filtering strategies. It appears that the selection of visual cues greatly affects the tracking behavior while the filtering strategy seems less important in our ill conditioned robotics context. The fusion of fast computed classical appearance based visual cues and 3D geometric information derived from a sparse 3D reconstruction of the scene as well as the combination of local and global cues are of great contribution. These hybrid measurements finally enable a satisfying tracking of the upper human body model in real time in various indoor environments, showing that visual based Human-Robot interaction is within reach.

Some improvements of our system could entail the definition of less classical visual cues to enable a more robust tracking, especially for elbow localization. One could also implement them on a GPU as this would significantly reduce computation time, and thus, improve the reactivity of our system. To go further in mixing appearance based and geometric cues, we could exploit 3D sensors such as the Swiss Ranger. Furthermore, in order to provide a rich means of communication between the human and the robot, this work will be coupled with motion/activity recognition techniques.

ACKNOWLEDGEMENTS

This work was partially conducted within the french ANR project AMORCES.

REFERENCES

[1] J. Deutscher and I. Reid, “Articulated body motion capture by stochastic search,” *International Journal of Computer Vision (IJCV’05)*, vol. 21, no. 3, pp. 185–205, 2005.

[2] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, “Tracking loose-limbed people,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR’04)*, Washington, DC, USA, 2004, pp. 421–428.

[3] C. Sminchisescu and B. Triggs, “Estimating articulated human motion with covariance scaled sampling,” *International Journal on Robotic Research (IJRR’03)*, vol. 6, no. 22, pp. 371–393, May 2003.

[4] T. Moeslund, A. Hilton, and V. Krüger, “A survey of advanced vision-based human motion capture and analysis,” *Computer Vision and Image Understanding (CVIU’06)*, vol. 104, pp. 174–192, Dec. 2006.

[5] J. Ziegler, K. Nickel, and R. Stiefehagen, “Tracking of the articulated upper body on multi-view stereo image sequences,” in *International Conference on Computer Vision and Pattern Recognition (CVPR’06)*, New York, USA, 2006, pp. 774–781.

[6] S. Knoop, S. Vacek, and R. Dillman, “Sensor fusion for 3D human body tracking with an articulated 3D body model,” in *International Conference on Robotics and Automation (ICRA’06)*, Orlando (USA), May 2006, pp. 1686–1691.

[7] P. Menezes, F. Lerasle, and J. Dias, “Visual tracking modalities for a companion robot,” in *International Conference on Intelligent Robots and Systems (IROS’06)*, Beijing, China, 2006.

[8] A. Doucet, N. De Freitas, and N. J. Gordon, *Sequential Monte Carlo Methods in Practice*, ser. Series Statistics For Engineering and Information Science. New York: Springer-Verlag, 2001.

[9] J. MacCormick and M. Isard, “Partitioned sampling, articulated objects, and interface-quality hand tracking,” in *European Conference on Computer Vision (ECCV’00)*, Dublin, Ireland, 2000, pp. 3–19.

[10] R. Alami, R. Chatila, S. Fleury, and F. Ingrand, “An architecture for autonomy,” *International Journal of Robotics Research (IJRR’98)*, vol. 17, no. 4, pp. 315–337, 1998.

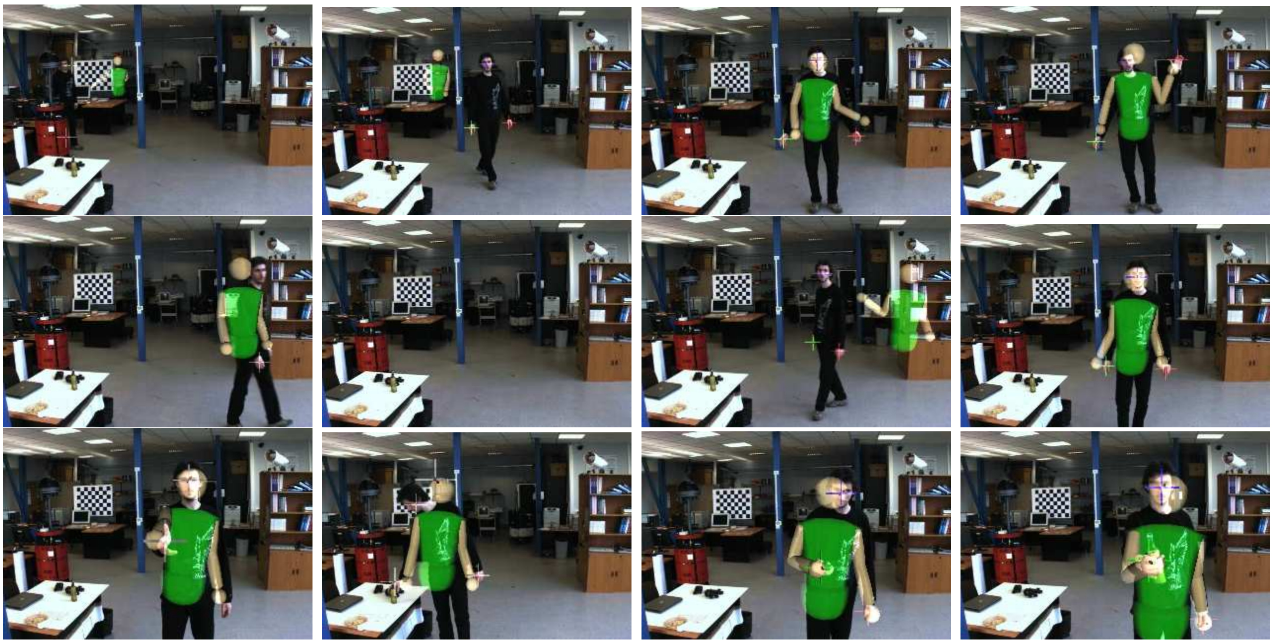

Fig. 4. Tracking in a cluttered background with various moves. The filter enables automatic (re-)initialization.

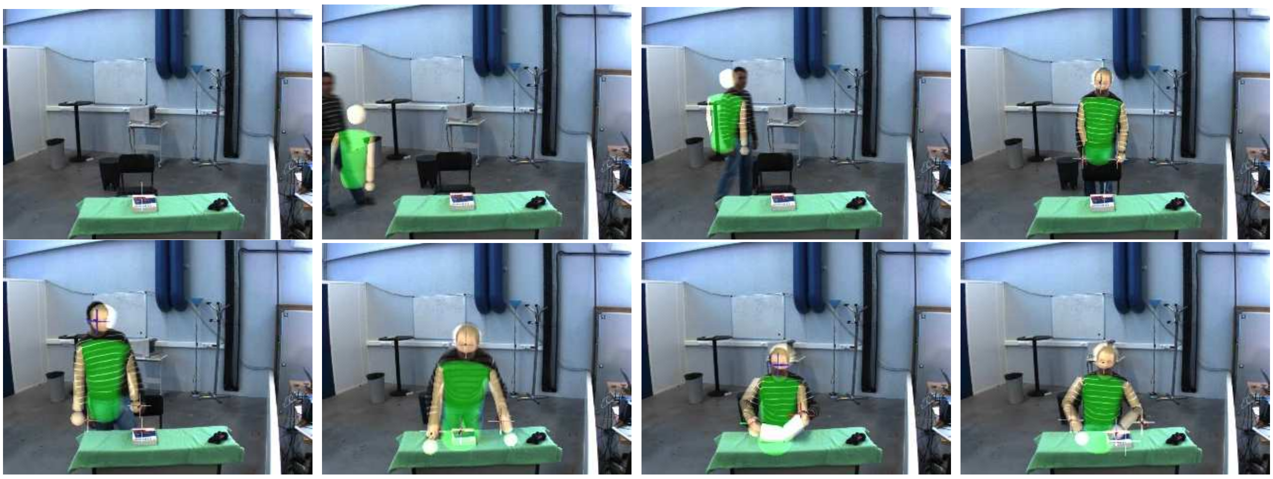

Fig. 5. Tracking on a different subject who performs a reading activity.

[11] M. Isard and A. Blake, "CONDENSATION – Conditional density propagation for visual tracking," *International Journal on Computer Vision (IJCV'98)*, vol. 29, no. 1, pp. 5–28, 1998. [Online]. Available: citeseer.ist.psu.edu/isard98condensation.html

[12] F. Daum and J. Huang, "Mysterious computational complexity of particle filters," in *Signal and Data Processing of Small Targets*, ser. Proceedings of SPIE, vol. 4728, Bellingham, MA, USA, Aug. 2003.

[13] V. Philomin, R. Duraiswami, and L. S. Davis, "Quasi-random sampling for CONDENSATION," in *European Conference on Computer Vision (ECCV'00)*, Dublin, Ireland, 2000, pp. 134–149. [Online]. Available: citeseer.ist.psu.edu/455759.html

[14] D. Ormoneit, C. Lemieux, and D. Fleet, "Lattice particle filters," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI'01)*, San Francisco, CA, USA, 2001, pp. 395–402.

[15] D. Guo and X. Wang, "Quasi-Monte Carlo filtering in nonlinear dynamic systems," *IEEE transactions on signal processing*, vol. 54, no. 6, pp. 2087–2098, June 2006.

[16] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proceedings of IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

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

[18] P. Azad, A. Ude, T. Asfour, G. Cheng, and R. Dillmann, "Image-based markerless 3D human motion capture using multiple cues," in *International Workshop on Vision Based Human-Robot Interaction*, Palermo, Italy, Mar. 2006.

[19] C. Schmid, "Appariement d'images par invariants locaux de niveaux de gris," Ph.D. dissertation, Institut National Polytechnique de Grenoble, 1996.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Kauaii Marriott, Hawaii, USA, 2001.

[21] "http://www.motionanalysis.com — Motion Analysis Corporation."

[22] A. Balan, L. Sigal, and M. Black, "A quantitative evaluation of video-based 3D person tracking," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05)*, Washington, USA, October 2005, pp. 349–356.