TOWARDS REAL-TIME MARKERLESS HUMAN MOTION CAPTURE FROM AMBIANCE CAMERAS USING AN HYBRID PARTICLE FILTER

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

† CNRS ; LAAS ; 7, avenue du Colonel Roche, F-31077 Toulouse, France
‡ Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS-CNRS ; F-31077 Toulouse, France

ABSTRACT

In this article we present a two ambiance video camera system dedicated to markerless human motion capture. We introduce a new particle filter algorithm which entails an importance function enabling auto-(re)initialisation, and takes account of the global curvature of the likelihood so as to guide the search along poorly observable directions of the state space. The system robustness is improved by fusing different visual cues. Performances are nearly real time.

***Index Terms*—** visual tracking, human motion, particle filtering, data fusion

1. INTRODUCTION

The applications of Human Motion Capture are many, *e.g.* movie industry, video games, sport movement analysis, or, in our context, communication with instrumented environments or robots. Our purpose is to perform real time markerless human body tracking in indoor environments with a standard fire-wire camera system. The system does not need to be as precise as a commercial motion capture system, but must be robust to experimental conditions and able to recover from tracking failures. This raises several problems: (1) if we want to be as generic as possible, we cannot make any hypothesis on human appearance; (2) human models often present more than 20 degrees of freedom (DOF), which implies a huge state space, hard to explore in real time even with recent techniques; (3) likelihood functions show a lot of local minima which can "trap" optimization algorithms; (4) auto-(re)initialization capabilities are needed.

Much work has been done to propose efficient schemes suited to such high dimension spaces. The partitioned sampling technique [1] splits the state vector into sub-vectors which evolution and likelihood depend only on the sub-vectors of lower indices. The annealed particle filter of [2] entails simulated annealing arguments so as to iteratively explore the state space by sharpening the particle cloud towards likely areas. Covariance scaled sampling [3] estimates the particle cloud mode covariances at each time step in order to sample along the most uncertain directions; then it drives each sample towards local minima of the negative log-likelihood by local optimization. Partitioned annealed particle filter [4] incorporates this covariance sampling technique in the classical annealed particle filter.

Our filtering strategy is based on the latest one to which we add an importance sampling stage. This involves being able to infer a coarse 3D configuration of the human body model from the images. We also introduce Quasi Monte Carlo (QMC) sampling which has been proved superior in terms of convergence than classical Monte Carlo sampling [5].

Section 2 presents our approach of the human tracking problem and describes our filtering strategy. The measures it exploits are explained in section 3. Our experimental system configuration is described in section 4 together with some results. Then we conclude our work and propose some future research axes in section 5.

2. FILTERING SCHEME

As expected, we formulate the tracking problem as the Bayesian state estimation of a Markovian stochastic process. The human body configuration parameters to be estimated at time k constitute the hidden state vector x_k of a system which delivers the measurements —images— z_k. This system is characterized by its state dynamics $p(x_k|x_{k-1})$ —here a random walk— and the conditional probability $p(z_k|x_k)$ of the output given the state vector, which also defines the likelihood of the state w.r.t. the measurement.

In high dimension state space problems, the key to efficiency is the ability of focusing samples in regions of high likelihood. This is why an importance function is needed to guide the particle sampling.

The algorithm we propose is rooted in the Annealed Particle Filter of Deutscher *et al.* [2] and the Covariance Scaled Sampling of Sminchisescu *et al.* [6] on which we add an importance sampling stage to enable auto-(re)initialization. It is presented in table 1. The main idea is to split the processing step in two stages. The first one introduces importance sampling as in the ICONDENSATION [7]. The second one sharpens the exploration of the state space using covariance sampling. Furthermore, we introduce QMC sampling which can produce low-discrepancy samples. This avoids the "gaps and clusters" which are likely to occur when sampling in high

This work was partially conducted within the EU Projects COGNIRON ("The Cognitive Robot Companion" - www.cogniron.org) and Comm- Rob ("Advanced Behavior and High-Level Multimodal Communication with and among Robots" - www.commrob.eu) under Contracts FP6-IST-002020 and FP6-IST-045441.

dimension spaces.

This algorithm can be seen as an evolution of the I-Annealed Particle Filter with only 2 stages [8]. Using such a simplified scheme provides the advantage of being freed from the use of α and β exponents which are very difficult to tune in practice. Recall that the goal here is not accuracy, but rather to be able to maintain a limited-size particle cloud which explores smartly enough the state space.

- $[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^{N}$ describes the state vector posterior distribution at time $k-1$
- Layer 1 : Classical importance sampling step
  - Sample each particle $x_k^{(i)}$ according to $q(x_k|x_{k-1}^{(i)}, z_k)$
  - Update each weight according to $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}|x_{k-1}^{(i)}, z_k)}$
  - Approximate the posterior mean μ_k and covariance Σ_k of the state x_k at time k from the particle cloud: $\mu_k = \sum_{i=1}^{N} w_k^{(i)} x_k^{(i)}$, $\Sigma_k = \sum_{i=1}^{N} w_k^{(i)} (x_k^{(i)} - \mu_k)(x_k^{(i)} - \mu_k)^T$
- Layer 2 : Sharpening the estimation *via* covariance sampling
  - Resample an outer distribution $\mathcal{N}(\mu_k, \lambda\Sigma_k)$ enclosing the mean- and covariance- matched Gaussian approximation $\mathcal{N}(\mu_k, \Sigma_k)$ of the posterior distribution at time k using QMC sampling: $x_k^{(i)} \underset{\text{QMC}}{\sim} \mathcal{N}(\mu_k, \lambda\Sigma_k)$
  - Update the weights according to $w_k^i = \frac{p(z_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{\mathcal{N}(x_k^{(i)}; \mu_k, \lambda\Sigma_k)}$
- $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^{N}$ describes the posterior distribution of the state vector at time k

---

Table 1. Particle filter algorithm

3. MEASURES

The importance function $q(x_k|x_{k-1}, z_k)$ generally involves discriminant but possibly intermittent visual cues — due to occlusions or mis-segmentation in our case — while measurement functions $p(z_k|x_k)$ involve cues which are persistent yet proner to ambiguity for cluttered scenes [9]. 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.

3.1. 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 a prior $p(x_0)$. The particles sampled from the measure are drawn from a multi-Gaussian distribution which modes are centered on configurations $x_k^{D(i)}, i \in 1..N_{3D}$ computed from 3D possible positions of head and hands thanks to an analytical Inverse-Kinematics (IK) algorithm. To obtain 3D position of head and hands, we apply a skin color segmentation on each image, then triangulate the detected 2D blobs to get 3D blobs. Thus, the hands and the 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 $u < \alpha$
- $x \sim \sum_{i=1}^{N_{3D}} \frac{1}{N_{3D}} \mathcal{N}(x_k^{D(i)}, \Delta_k)$ 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.

3.2. Likelihood function

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

$$p(z_k|x_k) = p(z_k^{s1}, z_k^{s2}|x_k) = p(z_k^{s1}|x_k)p(z_k^{s2}|x_k).$$

They are based on the use of a segmented silhouette image I_s obtained with background subtraction. Each measure is evaluated in both images.

3.2.1. Silhouette distance

This likelihood requires the projection of the 3D model. N_p points $p_i, i \in \{1, \ldots, N_p\}$ are sampled from the silhouette corresponding to the projection of the configuration x_k. The silhouette distance is then defined by

$$p(z_k^{s1}|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_{s1}^2}\right),\ D = \frac{1}{N_p}\sum_{i=1}^{N_p}(1 - I_s(p_i)),$$

where i indexes the N_p model points, $I_s(p_i)$ is the associated value in the segmented image, and σ_s the *a priori* standard deviation of our Gaussian measure model.

3.2.2. Dual Silhouette distance

This measure is the dual of the one above. We sample N_s points p_j from the segmented image I_s. For each configuration x_k the dual silhouette distance is defined by

$$p(z_k^{s2}|x_k) \propto \exp\left(-\frac{D^2}{2\sigma_{s2}^2}\right),\ D = \frac{1}{N_s}\sum_{j=1}^{N_s}(1 - f(p_j, x_k)),$$

where $f(p_j, x_k) = 1$ if the point p_j is in the silhouette corresponding to the projection of x_k, 0 otherwise.

We sometimes also use other likelihood functions (based on a skin color distance image, on motion or on collision computation) but they are not described here by lack of space.

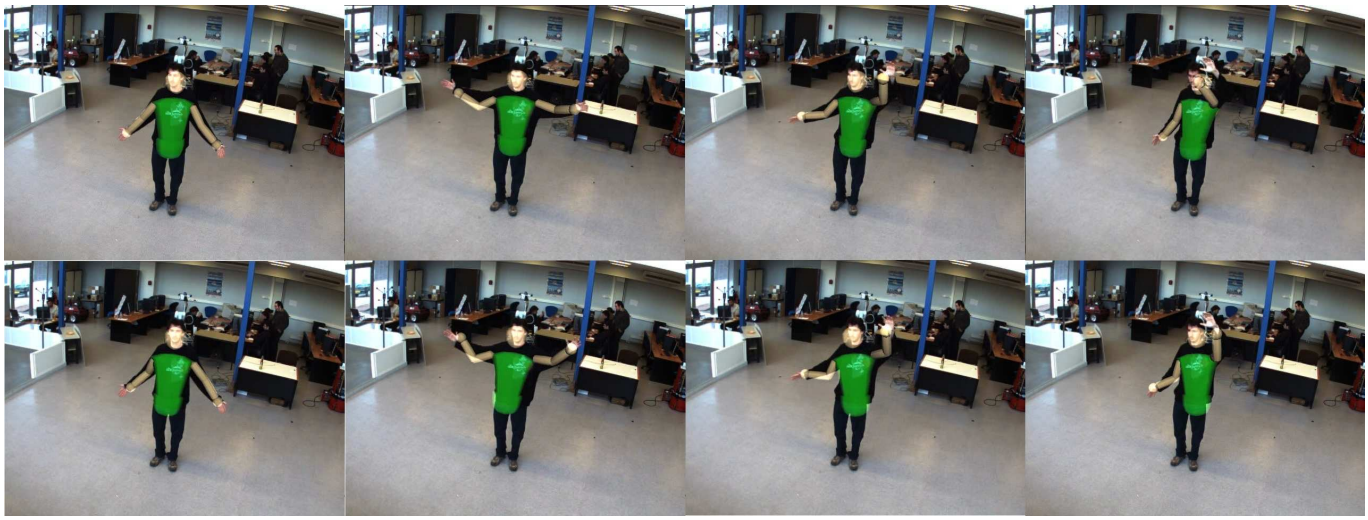

Fig. 1. Comparison between our tracker (bottom) and an annealed particle filter (top).

Name	Description	Value
(W, H)	Image size	$(640, 480)$
N	Number of particles	400
$(\sigma_{s1}, \sigma_{s2})$	Likelihood parameters	$(0.2, 0.4)$
λ	Covariance sampling factor	from 2 to 5
$p(x_k \vert x_{k-1})$	Dynamics	$\mathcal{N}(x_{k-1}, \Delta_k)$
Δ_k	Gaussian dynamics covariance	diag$(\delta_1..\delta_{N_{DOF}})$
δ_i	Scalar variance of i^{th} DOF	0.07 if translation 0.1 if rotation

Table 2. Parameter values used in the trackers

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

The system uses two fire-wire ambiance mono CCD color cameras. We acquire 1024×768 bayerised images. They are first converted to RGB, then downsampled to 640×480, prior to performing a white balance. We use a 22-DOF template built with truncated cones, even if the visualization of the configuration is done using a meshed model. On some video sequences, we can reduce the complexity of this model by taking into account only the 14 DOF of the upper body part.

The filtering strategies we compare involve a number of particles so that they all process one frame within the same time. The aim is to assess the tracking properties for a given computation power.

4.2. Results

4.2.1. Efficiency

By introducing a covariance scaled QMC sampling along the directions of lowest observability, less particles are needed to perform an efficient tracking as they are focused in pertinent areas of the state space. This also reduces the variance of the estimate along consecutive trials. Figure 1 shows a comparison between our filtering strategy and the annealed particle filter on the same video sequence. We can see that our algorithm provides estimates as good as the annealed particle filter and —in some tricky cases— slightly better ones (one can notice the last sequence image where the right hand is confused with the head).

4.2.2. Robustness

The main advantage of our particle filter in our high dimension state space tracking context is its ability to initialize or re-initialize automatically — and so aid recovery from transient tracking failures —. This makes the classical manual initialization unnecessary, as a detection of head and hands is enough to induce a 3D configuration of the model. A short sequence showing the tracker (re-)initialization is presented on fig 2. In these sequence, no prior draw is used.

4.2.3. Material considerations

Our algorithm has a complexity $\mathcal{O}(N)$ where N is the number of particles, what is absolutely necessary in a real time context. In its actual form, it runs on 640×480 pixel images at about $1\ Hz$. Most of the time consumption — about $200\ ms$ — is spent in image processing and preprocessing (conversion from Bayer to RGB, skin probability computations, blob segmentation, ...), and we hope to optimize this step. Furthermore, we could speed up the tracking by reducing the image size. A few parameter values used for the main sequences presented here are shown in table 2. Most of them have been set up by experimenting different values, sometimes guided by simple heuristics. More videos are visible at the URL `www.laas.fr/~mfontmar`.

Fig. 2. From left to right and from top to down : the tracker is initialized with a default configuration which does not make sense with regard to the real state. Head and hands detections enable to build a basic configuration near from the real one to initialize the particle cloud : the initialization succeeds.

5. CONCLUSION

We have set up a system performing human body tracking from two ambiance cameras. A new algorithm has been proposed, mixing advantages of the well-known Annealed and Covariance Scaled Sampling strategies, introducing an importance sampling stage, and enabling auto-re(initialization). This algorithm is built on a QMC low-discrepancy sampling of the state space. As it is built in a particle filter framework, data fusion from heterogeneous sensors is easy and theoretically sound. The tracking system is independent of the human clothes and runs close to real-time.

While our results are promising, a detailed comparison between our filter and those of the literature would be interesting in a context where accuracy is important. For more precision, we plan to extend our filtering strategy by using Gaussian mixture to approximate the posterior which could be more justified than a simple Gaussian (even if this is a key step to pass through). However such an approach raises the problem of the algorithm complexity which is always critical in real-time systems. We will also try to incorporate the use of the IK in the 2^{nd} step of the filter to sharpen, once again, the state space exploration.

Future work could also be focused on mixing data from 3D time-of-flight sensors [10] with classical color camera images. This could provide more stable localisation information as blob triangulation is not always reliable. Efforts can still be done to increase the frame rate of the tracking, by exploiting more efficiently the measurements — building a 3D voxel model of the scene for instance.

6. REFERENCES

[1] 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.

[2] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *International Conference on Computer Vision and Pattern Recognition (CVPR'00)*, Hilton Head Island, South Carolina, USA, 2000, vol. 2, pp. 126–133.

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

[4] J. Deutscher, A. Davison, and I. Reid, "Automatic partitioning of high dimensional search spaces associated with articulated body motion capture," in *International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Kauaii Marriott, Hawaii, USA, 2001, pp. 669–676.

[5] 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.

[6] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3d body tracking," in *Conference on Pattern Vision Recognition, (CVPR'01)*, Kauaii Marriott, Hawaii, USA, Dec. 2001, pp. 447–454.

[7] M. Isard and A. Blake, "CONDENSATION – conditional density propagation for visual tracking," *International Journal on Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[8] M. Fontmarty, F. Lerasle, and P. Danès, "Data fusion within a modified annealed particle filter dedicated to human motion capture," in *International Conference on Intelligent Robots and Systems*, San Diego, CA, USA, Nov. 2007.

[9] 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.

[10] S. Knoop, S. Vacek, and R. Dillman, "Sensor fusion for 3D human body tracking with an articulated3D body model," in *International Conference on Robotics and Automation (ICRA'06)*, Orlando (USA), May 2006, pp. 1686–1691.