Exploiting Structural Hierarchy in Articulated Objects towards Robust Motion Capture

C. Canton-Ferrer, J.R. Casas, and M.Pardàs

Technical University of Catalonia, Barcelona, Spain,
{ccanton,josep,montse}@gps.tsc.upc.es

Abstract. This paper presents a general analysis framework towards exploiting the underlying hierarchical and scalable structure of an articulated object for pose estimation and tracking. The Scalable Human Body Model (SHBM) is presented as a set of human body models ordered following a hierarchy criteria. The concept of annealing is applied to derive a generic particle filtering scheme able to perform a sequential filtering over the models contained in the SHBM leading to a *structural annealing* process. This scheme is applied to perform human motion capture in a multi-camera environment. Finally, the effectiveness of the proposed system is addressed by comparing its performance with the standard and annealed particle filtering approaches over an annotated database.

1 Introduction

Automatic capture and analysis of human motion is a highly active research area due both to the number of potential applications and its inherent complexity. This research area contains a number of hard and often ill-posed problems such as inferring the pose and motion of a highly articulated and self-occluding non-rigid 3D object from a set of images. Applications of motion analysis range from gesture recognition or gait analysis to medical applications and human-computer interfaces.

Recovering the pose of an articulated structure such as the human body involves estimating highly dimensional and multi-modal statistic distributions. Monte Carlo based techniques [1] have been thoroughly applied due to its ability to perform this task with an affordable computational complexity. Particle filtering [5] has been the seminal idea to develop specific systems aiming at recovering human body pose such as the annealed particle filter [3], the hierarchical sampling [8] or the partitioned sampling [6] among others. A main characteristic of these approaches is a human body model that is selected beforehand and fitted to the input data. This paper presents a general analysis framework that exploits the underlying hierarchical and scalable structure of an articulated object by using a scalable human body model together with an annealed particle filtering strategy. A sequential fitting is performed over a set of human body models with increasing level of detail by applying the concept of *structural annealing*. Indeed, some of the aforementioned tracking schemes may be considered as particular cases of our general framework.

The proposed scheme is applied to recover and track human body pose in a multi-camera scenario. However, instead of performing our measures on each input image, our system first generates a 3D voxel-based representation of the person, and then performs the matching of the kinematic models directly in this 3D space. Finally, the efficiency of the proposed system is addressed by analyzing a set of sequences from the HumanEva-I database [4] and comparing the results with the standard and annealed particle filtering approaches.

2 Monte Carlo based tracking

The evolution of a physical articulated structure can be better captured with model-based tracking techniques. The articulated structure can be fully described by a state vector $\mathcal{X} \in \mathbb{R}^D$ that we wish to estimate. From a Bayesian perspective, the articulated motion estimation and tracking problem is to recursively estimate a certain degree of belief in the state vector $\mathcal{X}_t$ at time t, given the data $\mathcal{Z}_{1:t}$ up to time t. Thus, it is required to calculate the *pdf* $p(\mathcal{X}_t|\mathcal{Z}_{1:t})$.

Particle Filtering (PF) [1] algorithms are sequential Monte Carlo methods based on point mass (or "particle") representations of probability densities. These techniques are employed to tackle estimation and tracking problems where the variables involved do not hold Gaussianity uncertainty models and linear dynamics. PF expresses the belief about the system at time t by approximating the posterior distribution $p(x_t|\mathcal{Z}_{1:t})$, $x_t \in \mathcal{X}$, and representing it by a *weighted particle set* $\{(x, \pi)_j\}_t$, $1 \leq j \leq N_p$. In this paper, a Sample Importance Re-sampling (SIR) based strategy is adopted to drive particles along time.

PF is an appropriate technique to deal with problems where the posterior distribution is multimodal. To maintain a fair representation of $p(x_t|\mathcal{Z}_{1:t})$, a certain number of particles are required in order to find its global maxima instead of a local one. It has been proved in [6] that the amount of particles required by a standard PF algorithm [5] to achieve a successful tracking follows an exponential law with the number of dimensions. Articulated motion tracking typically employs state spaces with dimension $D \sim 25$ thus normal PF turns out to be computationally unfeasible.

There exist several possible strategies to reduce the complexity of the problem based on refinements and variations of the seminal PF idea. MacCormick et al. [6] presented partitioned sampling as a highly efficient solution to this problem. However, this technique imposes a linear hierarchy of sampling which may not be related to the true body structure assuming certain statistical independence among state variables. Hierarchical sampling presented by Mitchelson et al. [8] tackles the dimension problem by exploiting the human body structure and hierarchically explore the state space. Finally, annealed PF presented by Deutscher et al. [3] is one of the most general solutions to the problem of dimensionality. This technique employs a simulated annealing strategy to concentrate the particles around the peaks of the likelihood function.

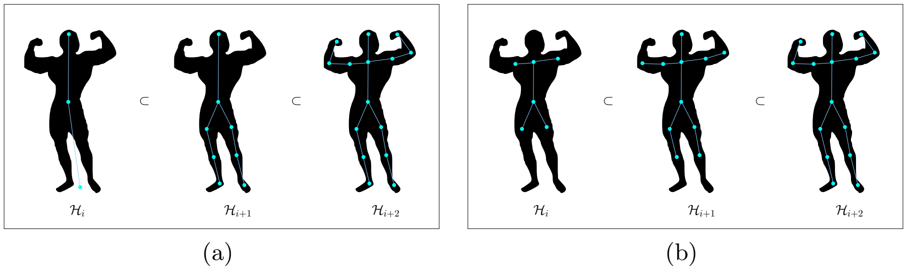


Fig. 1. Examples of inclusive Scalable Human Body Models. In (a), the refinement hierarchy model and, in (b), the construction model.

3 Scalable Human Body Model

A Human Body Model (HBM) $\mathcal{H}$ is employed to define a meaningful relation among the parameters contained in $\mathcal{X}$. This HBM mimics the structure of the skeleton representing it by a chain of rigid bodies (links) interconnected to one other by joints. The number of independent parameters will define the *degrees of freedom* (DoF) associated to a given HBM. HBMs employed in the literature range from simple configurations involving few DoF [2] to highly detailed models with up to 25 DoF [3]. A Scalable Human Body Model (SHBM) can be defined as a set of HBM:

$$\mathcal{M} = \{\mathcal{H}_0, \cdots, \mathcal{H}_i, \cdots, \mathcal{H}_{M-1}\}\,, \tag{1}$$

where the sub-index denotes an order within $\mathcal{M}$. To achieve scalability, a hierarchy among the elements of $\mathcal{M}$ must be defined. A criteria that grants hierarchy to the elements $\mathcal{H}_i$ in $\mathcal{M}$ is the inclusion condition:

$$\mathcal{H}_i \subset \mathcal{H}_j, \qquad i < j, \tag{2}$$

where the inclusion operation can be understood in terms of the detail or information provided by each model. This information measure is a design parameter and can be defined, for instance, as the number of joints/links, DoF, etc. Two examples of the inclusion operation are the refinement and constructive model. In the first one, depicted in Fig.1a, a model in a higher hierarchy level refines the one in the lower level by adding new limbs to it. In the constructive model, depicted in Fig.1b, segments are progressively added to all limbs until reaching the most detailed HBM.

4 Hierarchical Structure based Annealed Particle Filtering (HS-APF)

4.1 Theory

This papers presents a general analysis framework towards exploiting the underlying hierarchical structure of an articulated object for pose estimation and

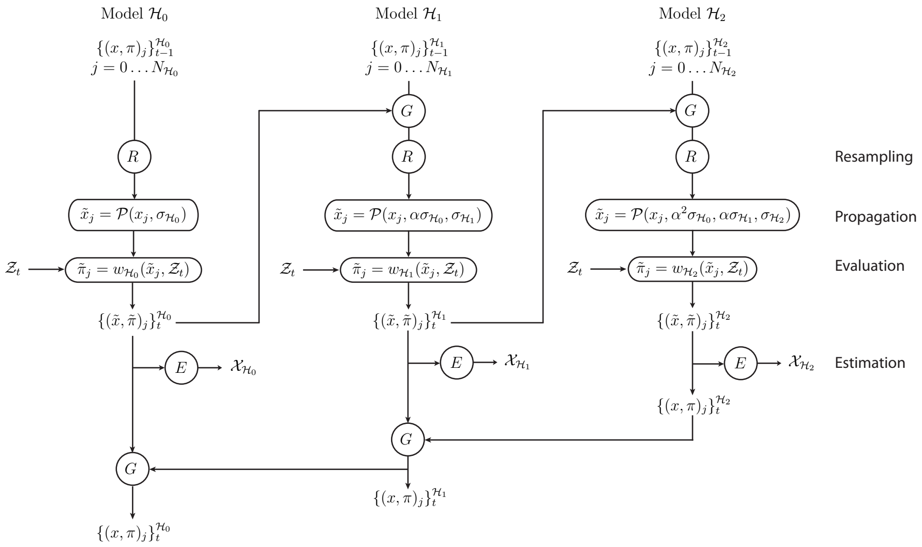

Model $\mathcal{H}_0$
$\{(x,\pi)_j\}_{t-1}^{\mathcal{H}_0}$
$j = 0 \dots N_{\mathcal{H}_0}$
Model $\mathcal{H}_1$
$\{(x,\pi)_j\}_{t-1}^{\mathcal{H}_1}$
$j = 0 \dots N_{\mathcal{H}_1}$
Model $\mathcal{H}_2$
$\{(x,\pi)_j\}_{t-1}^{\mathcal{H}_2}$
$j = 0 \dots N_{\mathcal{H}_2}$
G
R
Resampling
$\tilde{x}_j = \mathcal{P}(x_j, \sigma_{\mathcal{H}_0})$
$\tilde{x}_j = \mathcal{P}(x_j, \alpha\sigma_{\mathcal{H}_0}, \sigma_{\mathcal{H}_1})$
$\tilde{x}_j = \mathcal{P}(x_j, \alpha^2\sigma_{\mathcal{H}_0}, \alpha\sigma_{\mathcal{H}_1}, \sigma_{\mathcal{H}_2})$
Propagation
$\mathcal{Z}_t$
$\tilde{\pi}_j = w_{\mathcal{H}_0}(\tilde{x}_j, \mathcal{Z}_t)$
$\tilde{\pi}_j = w_{\mathcal{H}_1}(\tilde{x}_j, \mathcal{Z}_t)$
$\tilde{\pi}_j = w_{\mathcal{H}_2}(\tilde{x}_j, \mathcal{Z}_t)$
Evaluation
$\{(\tilde{x},\tilde{\pi})_j\}_t^{\mathcal{H}_0}$
$\{(\tilde{x},\tilde{\pi})_j\}_t^{\mathcal{H}_1}$
$\{(\tilde{x},\tilde{\pi})_j\}_t^{\mathcal{H}_2}$
E
$\mathcal{X}_{\mathcal{H}_0}$
$\mathcal{X}_{\mathcal{H}_1}$
$\mathcal{X}_{\mathcal{H}_2}$
Estimation
$\{(x,\pi)_j\}_t^{\mathcal{H}_2}$
$\{(x,\pi)_j\}_t^{\mathcal{H}_1}$
$\{(x,\pi)_j\}_t^{\mathcal{H}_0}$

Fig. 2. Hierarchical Structure Annealed Particle Filter scheme for $M = 3$ models in the SHBM.

tracking. Assuming that a SHBM $\mathcal{M}$ with a given hierarchy has been defined, a sequential fitting process is conducted over the several HBM $\mathcal{H}_i \in \mathcal{M}$. In order to carry out this task, we borrow the idea of annealing [3] where the particles are placed around the peaks of the likelihood function by means of a recursive search over a set of decreasingly smoothed versions of this function. Our proposal is to use the set of progressively refined HBMs constained in $\mathcal{M}$ instead of a set of smoothed versions of the likelihood function. This process mimics the annealing idea of the coarse-to-fine analysis of the likelihood function thus leading to a *structural annealing* process.

The overall operation of the proposed scheme is to filter the initial distribution associated to the simplest HBM $\mathcal{H}_0$ and then combine the resulting particle set with the initial particle set of the following model, $\mathcal{H}_1$. This process is performed for all the models in the SHBM until reaching the last one. Information contained by the particle set of the last model is back-propagated to the models with lower hierarchy rank thus refining their associated particle sets and closing the information filtering loop. The scheme of the proposed technique is depicted in Fig.2 for $M = 3$

Given a SHBM $\mathcal{M}$ containing M HBMs $\mathcal{H}_i$, a set of $N_{\mathcal{H}_i}$ particles, $\{(x,\pi)_j\}_t^{\mathcal{H}_i}$, associated to every $\mathcal{H}_i$ is defined at a given time t. It must be noted that, due to the hierarchy established in the SHBM, a mapping between the defining parameters of two consecutive HBMs can be always derived. Typically, this can be achieved by a linear or direct mapping between the involved variables. A fitness function $w_{\mathcal{H}_i}(x, \mathcal{Z}_t)$ measuring the likelihood between a particle and the incoming data $\mathcal{Z}_t$ is also constructed.

When a new measurement $\mathcal{Z}_t$ is available, an structural annealing iteration is performed. The hierarchical structure based annealed particle filtering can be summarized as follows:

- Starting from model $\mathcal{H}_0$, its associated particle set $\{(x, \pi)_j\}_{t-1}^{\mathcal{H}_i}$ is resampled with replacement. Then the filtered state $\{(\tilde{x}, \tilde{\pi})_j\}_t^{\mathcal{H}_i}$ is constructed by applying a propagation model $\mathcal{P}(\cdot, \cdot)$ and a weighting function $w_{\mathcal{H}_0}(\cdot, \cdot)$ to every particle as:

$$\tilde{x}_{j,t} = \mathcal{P}(x_{j,t}, \sigma_{\mathcal{H}_0}) = x_{j,t} + \mathbf{N}, \tag{3}$$

$$\tilde{\pi}_{j,t} = w_{\mathcal{H}_0}(\tilde{x}_{j,t}, \mathcal{Z}_t), \tag{4}$$

  where $\mathbf{N}$ is a multivariate Gaussian noise with mean $\mathbf{0}$ and a covariance matrix $\Sigma = \text{diag}\{\sigma_{\mathcal{H}_0}\}$. Weights are normalized such that $\sum_j \tilde{\pi}_j = 1$. At this point, the output estimation of this model $\mathcal{X}_{\mathcal{H}_0,t}$ can be computed by applying

$$\mathcal{X}_{\mathcal{H}_0,t} = \sum_{j=1}^{N_{\mathcal{H}_0}} \tilde{\pi}_{j,t}\tilde{x}_{j,t}. \tag{5}$$

- For the following HBMs, $i > 0$, the filtered particle set of the previous model in the hierarchy, $\{(\tilde{x}, \tilde{\pi})_j\}_t^{\mathcal{H}_{i-1}}$, is combined through the operator G with the particle set associated to model $\mathcal{H}_i$, $\{(x, \pi)_j\}_{t-1}^{\mathcal{H}_i}$. State space variables associated to $\mathcal{H}_i$ contain information from model $\mathcal{H}_{i-1}$ due to the imposed hierarchy relation. Since these variables have been already filtered, this updated information can be transferred to particles of model $\mathcal{H}_i$ in order to generate an improved initial particle set. Operator G has been inspired in the genetic algorithms theory and performs a crossover operation combining the common state variables of the two HBMs. Particles with a high weight in HBM $\mathcal{H}_{i-1}$ are combined with particles with a high weight in HBM $\mathcal{H}_i$. Common variables in $\mathcal{H}_i$ particles are replaced by the already filtered variables in $\mathcal{H}_{i-1}$ thus generating a new particle set that contains some information from the previous layer. However, it is also allowed some combination between particles with high weights from $\mathcal{H}_{i-1}$ with particles with low weights in $\mathcal{H}_i$ and viceversa. In this way, some variability is introduced thus being more robust to rapid motion and sudden pose changes.
  Then, the filtered state $\{(\tilde{x}, \tilde{\pi})_j\}_t^{\mathcal{H}_i}$ is constructed as:

$$\tilde{x}_{j,t} = \mathcal{P}(x_{j,t}, \alpha^i\sigma_{\mathcal{H}_0}, \alpha^{i-1}\sigma_{\mathcal{H}_1}, \ldots, \sigma_{\mathcal{H}_i}) = x_{j,t} + \mathbf{N}, \tag{6}$$

$$\tilde{\pi}_{j,t} = w_{\mathcal{H}_i}(\tilde{x}_{j,t}, \mathcal{Z}_t), \tag{7}$$

  where $\mathbf{N}$ is a multivariate Gaussian noise with mean $\mathbf{0}$ and a covariance matrix $\Sigma = \text{diag}\{\alpha^i\sigma_{\mathcal{H}_0}, \alpha^{i-1}\sigma_{\mathcal{H}_1}, \ldots, \sigma_{\mathcal{H}_i}\}$ with $\alpha < 1$. This propagation function assigns a higher drift to the newly added variables of model $\mathcal{H}_i$ while assigning a lower drift to those that have been more recently filtered

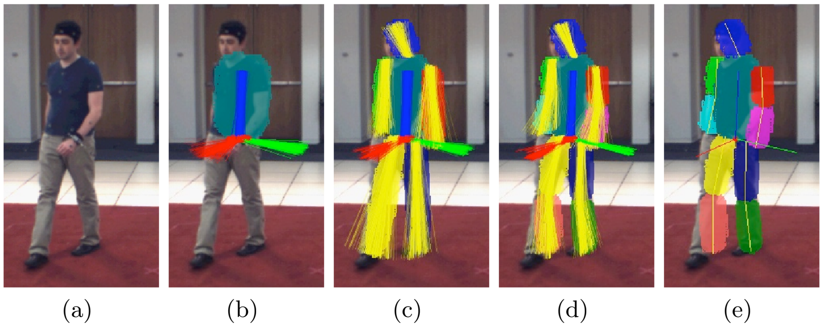


Fig. 3. Example of the execution of the HS-APF using a SHBM based on a refinement hierarchy. In (a), the original image is depicted and in (b)-(d) the sequential fitting of the models contained in the SBHM is shown (where the superimposed yellow lines are the particles associated to this HBM). In (e), the final pose estimation.

in the previous models. At this point, the output estimation of this model $\mathcal{X}_{\mathcal{H}_i,t}$ can be computed.

- Once reaching the highest hierarchy level, that is the most detailed HBM, the information contained in the particle set $\{(\tilde{x},\tilde{\pi})_j\}_t^{\mathcal{H}_M}$ is back-propagated to the other models in the hierarchy by means of the aforementioned crossover operator G. In this way, the particle sets will be refined thus closing the filtering loop.

An example of the execution of this scheme is depicted in Fig.3.

4.2 Implementation

For a given frame in the video sequence, a set of N images are obtained from the N cameras (see a sample in Fig.4a). Each camera is modeled using a pinhole camera model based on perspective projection with camera calibration information available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [9] as shown in Fig.4b. Redundancy among cameras is exploited by means of a Shape-from-Silhouette (SfS) technique [7]. This process generates a discrete occupancy representation of the 3D space (voxels). A voxel is labelled as foreground or background by checking the spatial consistency of its projection on the N segmented silhouettes (see a sample in Fig.4c). These data will be the input information fed to our HS-APF scheme, that is $\mathcal{Z}_t$. However, this 3D reconstruction is corrupted by spurious voxels, holes, etc. introduced due to wrong segmentation and camera calibration inaccuracies.

Every particle defines an instance of the pose of a given HBM $\mathcal{H}_i$. In order to relate this pose with the input 3D data, this model is fleshed out with super-ellipsoids associated to every limb part (see an example in Fig.3e). Let us denote

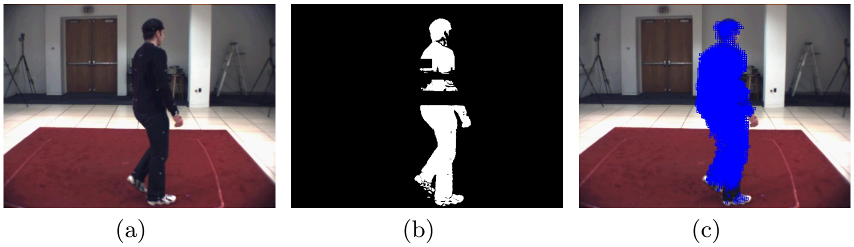


Fig. 4. Input data generation example. In (a), a sample of the original images. In (b), foreground segmentation of the input images employed by the SfS algorithm. In (c), example of the binary 3D voxel reconstruction.

this 3D HBM representation of the particle state as $\mathcal{D}_{x_j}$. The weighting function $w_{\mathcal{H}_i}(x_j, \mathcal{Z}_t)$ relating the state of a particle x_j with the input data $\mathcal{Z}_t$ is defined as:

$$w_{\mathcal{H}_i}(x_j, \mathcal{Z}_t) = \exp\left\{ -\left(1 - \frac{\#\left(\mathcal{D}_{x_j} \cap \mathcal{Z}_t\right)}{\#\mathcal{D}_{x_j}} \right) \right\}, \tag{8}$$

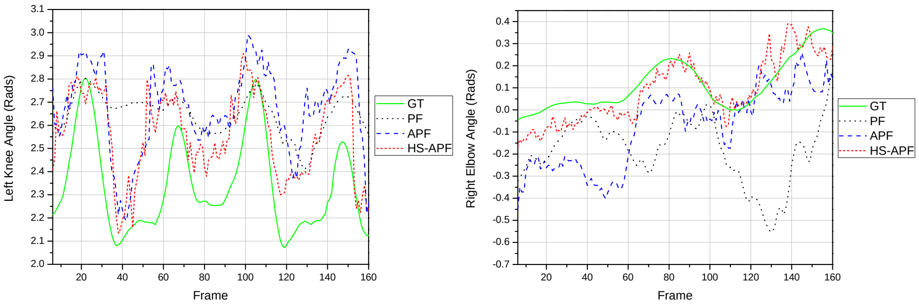
where $\#(\cdot)$ indicates the cardinality of the set, that is the number of foreground voxels in enclosed volume. This likelihood function may be seen as a measure of the overlap between the $\mathcal{D}_{x_j}$ and $\mathcal{Z}_t$.

5 Evaluation and Results

In order to prove the effectiveness of the proposed pose estimation and tracking scheme, a series of experiments have been conducted over a part of the HumanEva-I database [4], thus allowing fair comparison with other algorithms. The original data contained approximately 2000 frames at 25 fps recorded with 3 color and 4 greyscale calibrated cameras at a resolution of 640x480 pixels (however, only the 4 cameras were used to generate the 3D reconstruction of the scene). The 3D position of the most relevant joints in the body is provided in this database captured by means of a professional MOCAP system. This information will allow computing quantitative metrics in order to compare the body pose estimation with respect to the groundtruth.

Several metrics are employed to quantify the performance of the employed algorithm. HumanEva-I project proposes two point-based metrics based on the error measured at the position of the joints, namely the mean of the error μ and its associated standard deviation σ. Since the most natural way to encode a pose is by using the angles associated to every joint, we also provide two angle-based metrics: the mean of the angular error μ_θ and its associated standard deviation σ_θ.

The proposed system was compared to the Standard PF (SPF) [5] and the Annealed PF (APF) [3] approaches. The HS-APF scheme employed a SHBM



(a) (b)

Fig. 5. Evolution of the estimation of the compared tracking systems for the elbow and the knee angles during a walking cycle.

based on a refinement hierarchy as depicted in Fig.1a. In order to provide a fair comparison, the input data and the initial variance parameters employed by the propagation function were the same for all the filters. These variance parameters were set to be the half of the maximum expected movement associated to each joint. The overall number of particles employed by every filter was the same: $N_{\mathrm{SPF}} = 1000$, $N_{\mathrm{APF}} = 250$ (with 4 annealing layers) and $N_{\mathcal{H}_i} = \{250, 500, 250\}$.

Quantitative results of these experiments have been reported in Table 1 and Fig.5 showing the effectiveness of the HS-APF in comparison with the PF and APF approaches. Metrics μ and σ quantize the error when estimating the position of the body joints and there is a relative 33% error reduction from the SPF to the HS-APF, and a 19.5% reduction from APF to HS-APF. Metrics μ_θ and σ_θ quantize the error in terms of angles, being perhaps a more informative measure. In this case, there is a relative 51.5% angular error reduction from SPF to HS-APF and a 33% reduction from APF to HS-APF. Typically, when the state space has a high dimensionality, the number of particles required by the SPF is very high thus not operating accurately with 1000 particles. This problem is efficiently addressed by APF and a noticeable improvement is achieved. Finally, HS-APF exploits the underlying structure of the articulated model thus achieving a better performance. A visual example is depicted in Fig.6. In this example, PF scheme is unable to properly estimate the pose with only 1000 particles while APF does not recover the pose of some limbs (in this case a leg). Finally, HS-APF can retrieve the correct pose taking advantage of the scalable human body model.

6 Conclusions and Future Work

This paper presents a general framework to address estimation and tracking problems where a scalable hierarchy can be defined within the analysis model. Exploiting this hierarchy allows the system to deal with noisy input data thus

	μ	σ	μ_θ	σ_θ
SPF	172.87	28.43	14.75	07.86
APF	143.16	23.01	10.79	05.77
HS-APF	115.21	20.32	07.21	03.14

Table 1. Quantitative results for the walking action of the subjects $S2$ and $S3$ of the HumanEva-I dataset. Results are shown in millimeters and degrees.

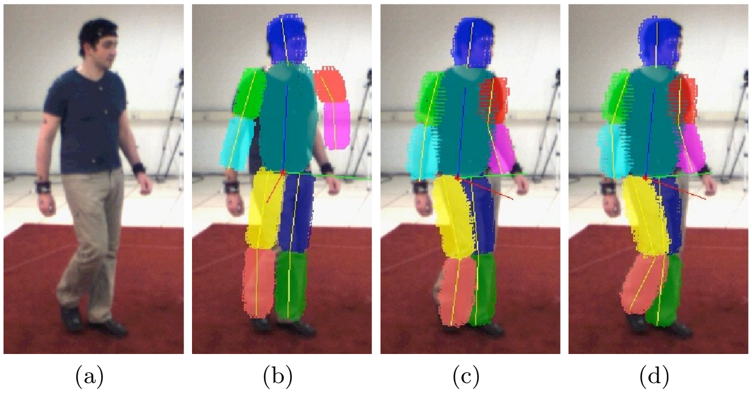


Fig. 6. Example of pose estimation with several filtering schemes. Legend: (a) original image, (b) PF, (c) APF, (d) HS-APF.

providing a robust solution. Human motion capture is one of such cases and the proposed scheme proved effective to estimate and track pose. Quantitative results comparing the Hierarchical Structure based Annealed Particle Filtering with the Standard Particle Filter and the Annealed Particle Filter showed the effectiveness of our approach.

Future research involves defining new hierarchy relations within the analysis models and a further validation of this system with larger databases including unconstrained motion and more than one subject in the scene. Including surface and color information will allow constructing more discriminative likelihood functions leading to a lower number of particles required by the HS-APF scheme. Applications in other signal processing fields such as audio processing are under study.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. on Signal Processing, vol. 50:2, pp. 174–188, 2002.
2. Canton-Ferrer, C., Casas, J.R., Pardàs, M.: Human Model and Motion Based 3D Action Recognition in Multiple View Scenarios. In Proc. European Signal Processing Conf., 2006.

3. Deutscher, J., Reid, I., “Articulated body motion capture by stochastic search”. In Int. Journal of Computer Vision, vol.61(2), pp. 185–205, 2005.
4. HumanEva - Synchronized video and motion capture dataset for evaluation of articulated human motion, http://vision.cs.brown.edu/humaneva
5. Isard, M., Blake, A., “CONDENSATION–Conditional density propagation for visual tracking”. In Int. Journal of Computer Vision, vol.29(1), pp. 5–28, 1998.
6. MacCormick, J., Isard, M., “Partitioned sampling, articulated objects and interface-quality hand tracking”. In Proc. European Conference on Computer Vision, vol.2, pp. 3–19, 2000.
7. Mikič, I., “Human body model acquisition and tracking using multi-camera voxel Data”. PhD Thesis, University of California, San Diego, 2003.
8. Mitchelson, J., Hilton., A., “Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling”. In Proc. British Machine Vision Conference, 2003.
9. Stauffer, C., Grimson, W., “Adaptive background mixture models for real-time tracking”. In Proc. IEEE Int. on Computer Vision and Pattern Recognition, pp.252–259, 1999.