Evaluation of Hierarchical Sampling Strategies in 3D Human Pose Estimation

Jan Bandouch[1], Florian Engstler[2], Michael Beetz[1]
[1]Intelligent Autonomous Systems Group, Department of Informatics,
[2]Ergonomics Department, Faculty of Mechanical Engineering,
Technische Universität München, Munich, Germany
bandouch@cs.tum.edu, engstler@lfe.mw.tum.de, beetz@cs.tum.edu

Abstract

A common approach to the problem of 3D human pose estimation from video is to recursively estimate the most likely pose via particle filtering. However, standard particle filtering methods fail the task due to the high dimensionality of the 3D articulated human pose space.

In this paper we present a thorough evaluation of two variants of particle filtering, namely *Annealed Particle Filtering* and *Partitioned Sampling*, that have been proposed to make the problem feasible by exploiting the hierarchical structures inside the pose space. We evaluate both methods in the context of markerless model-based 3D motion capture using silhouette shapes from multiple cameras. For that we created a simulation from ground truth sequences of human motions, which enables us to focus our evaluation on the sampling capabilities of the approaches, i.e. on how efficient particles are spread towards the modes of the distribution. We show the behavior with respect to the number of cameras used, the number of particles used, as well as the dimensionality of the search space. Especially the performance when using more complex human models ($\sim$ 40 DOF and above) that are able to capture human movements with higher precision compared to previous approaches is of interest in this work.

In summary, we show that both methods have complementary strengths, and propose a combined method that is able to perform the tracking task with higher robustness despite reduced computational effort.

1 Introduction

Human Pose estimation is a highly active research topic in Computer Vision. Among the possible areas of application are robotics, surveillance scenarios, human computer interaction, computer graphics and animation, as well as ergonomic industrial design and motion analysis in high performance sports. In the context of higher level action recognition, knowledge about the human motion is often considered a prerequisite for the recognition and understanding of human actions and intentions. While commercial systems for motion capture mostly rely on the recognition of easily detectable markers (e.g. infrared), research is mainly focusing on markerless motion capture systems.

In this paper we investigate different strategies for recursive estimation of human poses from video sequences in a Bayesian framework, or in other words for searching

BMVC 2008 doi:10.5244/C.22.92

the sequence of optimal poses given an initial estimate. Our work can be classified in the context of markerless human pose estimation using a 3D articulated human model. As human motion is non-linear in both its dynamics and the available observation models, particle filters [8, 2] are a reasonable choice for performing the tracking task. However, standard particle filtering is unsuitable for pose estimation as the number of particles needed grows exponentially with the number of dimensions. It is therefore advisable to exploit the hierarchical structure of the articulated pose space to guide particle spread into the most relevant areas, i.e. towards the modes of the distribution. Deutscher and Reid [7] proposed the *Annealed Particle Filter* (APF) to escape the local minima inherent in the high dimensional human pose space. Another particle filtering variant known as *Partitioned Sampling* (PS) has been proposed by MacCormick and Isard [11]. Both methods have been reported to achieve good results when standard particle filtering fails.

We present a detailed experimental evaluation of APF and PS in the context of 3D human pose estimation, and make the following contributions: • First, we provide a detailed comparison of APF and PS on a simulation sequence from ground truth data and point out strengths and weaknesses of both approaches, focusing on their sampling capabilities. We investigate the influence of the number of cameras used for tracking as well as the number of particles used. We show performance depending on the dimensionality of the pose space by distinguishing between upper-body and full-body motions. • Second, we propose an improved weighting function for silhouette-based correlation as compared to the commonly used SSD-based weighting functions. We show how this function can be used to control the survival rate, an important factor for the success of both approaches. • Third, we show that APF and PS have complementary strengths and propose a combination of both that is able to perform the tracking task with higher accuracy despite reduced computational effort.

The remainder of this paper is organised as follows. We briefly talk about related work in the next section. In section 3 we give an introduction to particle filtering in general and the two variants evaluated in this paper. Section 4 describes how we do 3D human pose estimation, focusing mainly on the 3D anthropometric model we use for tracking, but also on our choice of motion and observation model. Section 5 describes the experiments we conducted, including results and discussions of both approaches, and a description of the proposed combined approach. We finish in section 6 with our conclusions.

2 Related Work

Several surveys give a good overview on recent work and taxonomies in human pose estimation [12, 13]. Wang and Rehg [15] have evaluated variants of particle filters for figure tracking with 22 DOF. However, they do not evaluate strategies that take advantage of the hierarchical structure of articulated poses. Balan et al. [3] have done experimental evaluations on 3D pose estimation using the APF, but no comparison to PS is performed. They also provide experiments on the choice of motion and observation models. Gall et al. [9] have presented a detailed mathematical derivation and analysis of the APF, but again lack comparison to other methods such as PS.

Apart from particle filtering, some pose estimation approaches use optimization to find the best pose [10]. Usually a good initial guess is needed to avoid getting trapped in local minima, so tracking should be done at high frame rates. One option to reduce the dimensionality of the pose space is to project it to a lower-dimensional manifold by learning the manifold for specific activities [14]. This works well for the specified motions, but also constrains the amount of detectable motions. Methods also differ in their used observation models. Several recent approaches use the visual hull estimated from multiple cameras for a precise fitting of 3D models [10, 1]. Especially Anguelov et al. [1] combine this with a highly accurate and deformable human model learned from a database of 3D laser scans (SCAPE). However, such highly realistic models are difficult to use with particle filtering due to the high computational complexity.

3 Particle Filtering

In a Bayesian framework, the problem of tracking human motion can be formulated as one of estimating the *posterior* probability density function (pdf) $p(x_t \,|\, y_{1:t})$ for the pose x_t at time t given a sequence of image observations $y_{1:t}$ up to time t. This pdf can be obtained recursively in a *prediction* and an *update* stage, given the *motion model* $p(x_t \,|\, x_{t-1})$ and the *observation model* $p(y_t \,|\, x_t)$ (see [8, 2] for a more detailed description).

In *Sampling Importance Resampling* (SIR) particle filtering (also known as *Condensation*), the pdf if approximated by a set of N weighted particles $\{x_t^{(i)}, \pi_t^{(i)}\}_{i=1}^{N}$ consisting of the state $x_t^{(i)}$ and associated normalized weights $\pi_t^{(i)}$. In each timestep, the following steps are performed: • During *Resampling*, a weighted particle set is transformed into a new set of unweighted particles by drawing particles with probability according to their weights. • In the *Prediction* step, particles are moved according to the motion model and dispersed to represent the growing uncertainty. • The *Update* step produces the new weighted set of particles representing the pdf by assigning weights according to the observation model. An estimate of the state x_t can be found by either selecting the particle with maximum weight or by calculating the weighted mean sample of all particles.

A drawback of SIR is that the number of particles needed to approximate the pdf (respectively for successful tracking) grows exponentially with increased dimensionality.

Annealed Particle Filtering:

Traditional SIR aims at an approximation of the pdf, whereas in high dimensional tracking (more than 10 DOF), the number of particles is clearly insufficient for such an approximation. Usually, the strategy becomes one of focusing particles around the modes of the pdf. To avoid getting stuck in local maxima, Deutscher and Reid [7] proposed the *Annealed Particle Filter* (APF) as a combination of particle filtering with simulated annealing.

In each timestep, a multi-layered search (starting from layer $m = M$ to layer 0) is conducted so that the sparse particle set is able to gradually move towards the global maximum without being distracted by local maxima. Each of these layers corresponds to standard SIR particle filtering, however, the weighting functions $\omega_m(Y,X) = \omega(Y,X)^{\beta_m}$ are heavily smoothed in the first layers. This enables relatively unconstrained particle motion and escape from local maxima. The smoothing of the weighting functions is achieved by a set of values $\beta_M < .. < \beta_1 < \beta_0$, with $\omega(Y,X)$ being the original weighting function. The effect of these values is similar to the annealing schedule in simulated annealing. The bigger β_m becomes, the more constrained will the particle movement be by the current weight. At the same time, the amount of diffusion added during the prediction step of each annealing layer is decreased, to tighten the particle spread around the promising areas. This is where Deutscher and Reid take the hierarchical configuration into account, by setting the diffusion covariance P_m proportional to the quality of the localization of each individual pose parameter (Adaptive Diffusion). The search becomes then focused in regions where the optimal parameter could not yet be determined.

Partitioned Sampling:
Partitioned sampling is an approach at hierarchical decomposition of the state space that has been introduced by MacCormick and Isard [11]. It can be seen as the statistical analogue to a hierarchical search, and is especially suited to cope with the high dimensionality of articulated objects. PS consists of a series of sequentially coupled SIR filters, so that each filter estimates parts of the state space independently. The prerequisites for using PS are fulfilled in the case of human pose estimation: The pose space can be partitioned as a Cartesian product of joint angles, the dynamics of joint angles do not influence the dynamics of hierarchically preceding joint angles, and the weighting function can be evaluated locally for each body part.

4 3D Human Pose Estimation

We use the digital human model *RAMSIS* for the pose estimation (see Bandouch et al. [4] for a detailed introduction). It is an industry-proven and far-developed model from the ergonomics community that we have optimized for use in motion tracking. The model consists of an inner and an outer model (Figure 1) capable of capturing different body types according to anthropometric considerations, i.e. the different appearance of a wide range of humans. The locations of the inner joints correspond precisely to the real human joint locations. Poses are parametrized via joint angles. *RAMSIS* is able to capture most of the movements humans can perform while retaining a correct outer appearance. Absolute motion limits as well as inter-frame motion limits are integrated and help to reduce the search space when tracking. We have simplified the original model from 65 DOF to 41 DOF for our evaluations by interpolating dependencies in the joints of the spine and disregarding hands and fingers. The triangulation of the outer model is a good compromise between accurate outer appearance and fast computations. An additional speed-up is provided by caching of body part transformations and surface meshes, which facilitates the use of the model for pose estimation using particle filters.

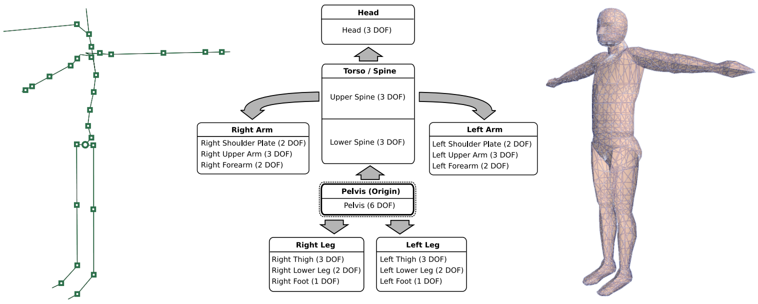

Figure 1: 3D anthropometric human model *RAMSIS* with inner model (left), simplified hierarchical structure used for tracking (center) and outer model (right).

Motion model:
As we want to be able to track unconstrained human motions, we do not use a specific motion model except for Gaussian distributed diffusion ($x_{t+1} = x_t + \mathcal{N}(0, \sigma^2)$). The amount of diffusion for each joint angle j is dependent on the image frames per second (fps) and has been estimated in markerbased motion capture experiments. For a sequence captured at 25 fps, standard deviations σ_j range from 0.5 deg for some joints in the spine

up to 38 deg for torsion of the forearms. In our experiments, we have limited σ_j to a maximum of 12.5 degrees, or else the tracking would become inaccurate.

Observation model:
We use the match between the projected outer model and the silhouettes observed in the video frames. Although no depth or luminance information is given, silhouettes provide rich and almost unambiguous information about a human pose, given enough cameras (see Section 5). Furthermore, they are easy to extract using standard background subtraction techniques, and they fulfil the requirement of being locally evaluable for each body part, as requested by PS. In contrast to [7, 3], we do not use the Mean Squared Error between the predicted foreground pixels and the observed silhouette mask. Doing so can result e.g. in arms being detected in front of the torso despite being apart, as the error measure is always low for limb predictions coinciding with the torso. The usual way to compensate for this effect is to additionally match model contours with detected image edges [7]. We propose the following error function as an alternative:

$$E^{(i)} = \sum_{x,y} I_e^{(i)}(x,y) \; ; \qquad I_e^{(i)} = I_p^{(i)} \; XOR \; I_s \; ; \qquad i = 0 \ldots N \; ; \tag{1}$$

$$e^{(i)} = \frac{E^{(i)} - \min_i(E^{(i)})}{\max_i(E^{(i)}) - \min_i(E^{(i)})} \tag{2}$$

Here, $E^{(i)}$ is the absolute error between the observed silhouette mask I_s and the model projection $I_p^{(i)}$, calculated by applying a pixelwise XOR and counting all non-zero pixels. $e^{(i)}$ is the normalized error scaled between 0 (lowest particle error) and 1 (highest particle error). Scaling the error according to the minimal and maximal encountered particle errors is a nice way to influence the survival diagnostic $\mathscr{D}$. The survival diagnostic was introduced by MacCormick and Isard [11], and gives an estimate of the number of particles that will survive a resampling step. It is an important tool for controlling the particle spread in both APF and PS. In APF, it is directly related to the rate of annealing, and is controlled via β_m. Instead of exponentiating the weight $\tilde{\pi}^{(i)} = 1 - e^{(i)}$ by β_m, a nicer way to control survival is to use the function $\pi^{(i)} = 1 - \left(1 - \tilde{\pi}^{(i)^a}\right)^b$, where the parameters a and b smoothly influence the survival rate as shown in Figure 2.

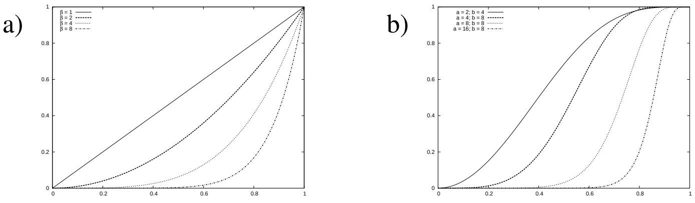

Figure 2: Weight scale functions for controlling the survival rate $\mathscr{D}$. a) π^β as in APF. b) $1 - (1 - \pi^a)^b$ for better control of the survival rate.

5 Experiments

For the experimental evaluation we have created a simulated sequence of silhouette shapes of our model as seen from virtual cameras (Figure 3). We use motion recovered from a real video sequence (1700 frames, 25 fps) captured by 3 cameras as the ground truth. The motion consists of walking in a circle and some movement on the spot, and was captured using the PS approach (with 10000 particles) as described in the last sections, followed by a smoothing step to remove motion jitter. The reason why we don't use any of the publicly available data sets is that we need the ground truth data to coincide precisely with the joint locations in our model, so that the silhouette shapes correspond exactly to the model projections for the ground truth. This way, the optimal pose will always coincide with the global maximum of the weighting function. By having this exact weighting function without any noise (and by refusing to use a specific motion model), we can concentrate the evaluation solely on the sampling capabilities of the approaches, i.e. on how good they are in finding the global maximum.

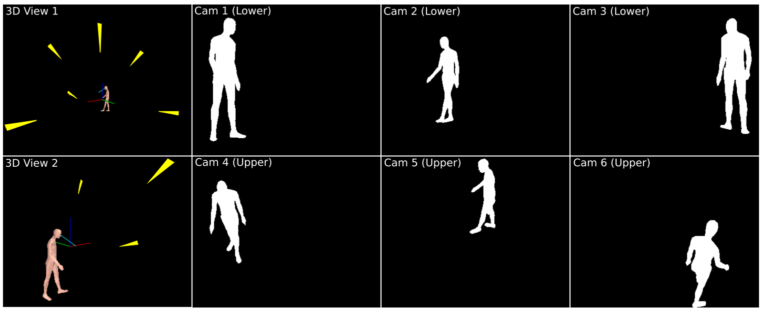

Figure 3: Simulated camera positions (first column) and generated model silhouettes.

We use the 3D joint locations of the inner model for comparison, as angular data might be ambiguous with respect to the silhouette appearance. Our error measure is the mean Euclidean distance error in 3D of all 28 joints in the inner model. In cases where the mean error is not able to discriminate between different approaches, we use the maximal Euclidean distance error in 3D of any of the 28 joints. This is a good indicator when localization fails partially, e.g. only an arm is localized incorrectly.

Number of cameras needed:
Our first experiment evaluates the influence of the number of cameras used when using only silhouette shapes (Figure 4a). We tested tracking (PS with 5000 particles) using between 1 and 6 cameras. Camera placement is visualized in Figure 3. We placed 3 cameras parallel to the ground plane, and 3 cameras as if they were hanging from the upper corners of a room, a setup favourable in many situations. Furthermore, the cameras clearly differ in their viewpoint, and no two cameras are opposite of each other, as that would result in mirrored silhouettes without additional information gain. Obviously, tracking using only one camera results in high errors, as occlusions of the limbs can not be disambiguated. Two cameras still seem to be insufficient for accurate tracking, however from three cameras upwards there is no qualitative improvement any more. There is also no difference whether the 3 lower or the 3 upper cameras are used. In our further experiments, we thus use the 3 lower cameras for tracking. An important aspect of this experiment is that the quality of tracking can not be improved by providing more cameras, as 3 cameras seem to be enough to dissolve all ambiguities. The minimal achievable error is then only determined by the capabilities of the hierarchical sampling strategies. Our results approve a similar experiment conducted by Balan et al. [3].

Upper Body Motions (21 DOF):
To test the behavior at around 20 DOF, we have additionally created a simulated sequence

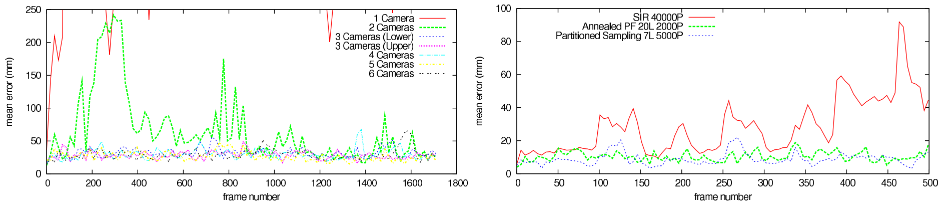

Figure 4: Mean Euclidean distance errors of all joints. Left: for camera count experiment. Right: for upper body tracking experiment.

of 500 frames for upper body motions only. The sequence consists of throwing darts and features subtle motions in the spine and the shoulder plates as well as fast motions of the arms. We have tried to compare each method by the amount of necessary image evaluations performed, e.g. APF with 20 layers and 2000 particles needs 40000 evaluations (even more when using adaptive diffusion) and PS with 5000 particles and 7 partition layers needs 35000 particles. The total running time is hard to generalize due to implementation details. Figure 4b shows that both APF and PS perform very well with a mean error below 20mm. In contrast, standard SIR (with 40000 particles) performs much worse. The same experiment conducted with a fifth of the particles still showed successful tracking with only slightly increased mean errors for both APF and PS.

Full Body Motions (41 DOF):
Accurate motion capture (e.g. in ergonomic applications) requires models with higher complexity than the often used ∼30 DOF models. In our full body tracking experiments, we evaluated the behavior of APF and PS with respect to such challenging models. Note that most of the DOFs in our model are critical degrees in the spine and shoulders. These are harder to estimate than e.g. hands, that can be localized quite independently.

We started by evaluating different versions of APF, both with respect to the number of particles/layers as well as choice of the annealing schedule. Furthermore, we investigated the influence of adaptive diffusion as proposed by Deutscher et al. [6]. We cannot provide the full evaluations here, but our results indicate that at least 20 layers are necessary for somewhat successful tracking. For the annealing schedule, we tried to estimate the parameters that worked best for our problem. Adaptive diffusion seems to be clearly beneficial for tracking, as it provides a soft partitioning of the search space. However, the estimation of the degree of localization for specific body parts is difficult, as a mapping from body parts to parts of the observed silhouettes is missing. We circumvent this problem using a heuristic based on the development of the mean correlation error over APF layers. Figure 5a shows the comparison of different APF methods.

We have compared the most successful APF method found (adaptive diffusion, 20 layers, 2000 particles) with PS using both 1000 and 5000 particles with 11 partition layers. While the mean error does not seem to favour either method, the maximum error shows that PS is more robust, even when using only a fraction of image evaluations (Figure 5b). High maximum errors indicate partial tracking loss, e.g. when a single arm is falsely estimated. We have also investigated different partitioning strategies for PS, i.e. breadth first partitioning (descending all hierarchies alternately) and depth first partitioning (descending each hierarchy to the bottom before moving to the next one). However, the order of partitioning did not have an influence on the quality of the tracking in our experiments.

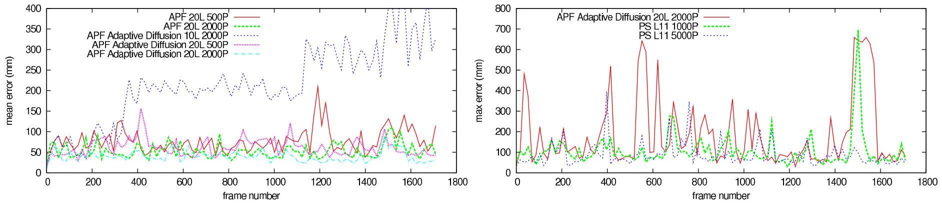

Figure 5: Left: Mean Euclidean distance errors of several APF strategies. Strategies using Adaptive Diffusion perform best. Right: Max Euclidean distance errors for APF and PS methods, high peaks indicate partial tracking loss. PS is more stable than APF.

Combining Partitioned Sampling and Annealing:

We are able to draw the following conclusions from our experiments:

• APF works fine between ∼10 and 30 DOF, but performance decreases at higher dimensions. This is unsurprising, as there is no partitioning of the search space (see Figure 7a). During each layer, pose parameters are estimated that are dependent on hierarchically preceding parameters, that in turn have not yet been estimated. In theory, even when using a million particles, only about 1.4 particles per dimension are available in a 40 DOF state space using combinatorics. Therefore, the amount of annealing layers must be very high for successful tracking. Adaptive diffusion [6] is a way to improve APF by coupling search dynamics to the covariances of the local estimates, thus creating a soft partitioning. This improves the tracking capabilities of APF significantly, but is not sufficient to cope with the exponential grow of the search space when tracking 40+ DOF.

• PS outperforms APF at higher dimensions as it is not affected by the exponential grow of the search space due to its hard partitioning (see Figure 7b). However, PS is forced to provide a good localization at early stages, or a good solution will not be found. This is easier to achieve when using models with accurate outer appearance (as in our experiments). Still it is difficult to localize the torso (usually the origin of the hierarchy) with good precision without knowledge of the protruding limbs. In our example, the full torso including shoulders has 16 DOF, which is problematic for the standard SIR filters used in each partition. We therefore need to partition the torso in a lower and an upper part, which further complicates local evaluation.

Given these insights, we propose the following combination of APF and PS. As APF outperforms traditional SIR, we propose to use APF filters in each partition of PS. This way, a larger initial partition (the torso) can be estimated accurately at once, reducing the risk of pursuing wrong initial estimates due to a bad localization of the torso (see Figure 7c). We have run the following experiment to prove the validity of our proposition. We use an APF with 10 layers and 1000 particles in the first partition of our APF+PS approach, that consists of the torso plus thighs and the head. This partition has 19 DOF, and can be estimated with high accuracy from silhouette shapes. The following limb partitions are then estimated either using standard SIR or APF with only 2-3 layers. Figure 6 shows both the mean and the max errors for APF, PS and the combined approach. Although the combined approach runs with fewer particles and image evaluations, it outperforms both APF and PS and shows a higher robustness. According to a dependent t-test, APF+PS performs better than PS at 99.9% confidence level, and PS performs better than APF at 99.9% confidence level.

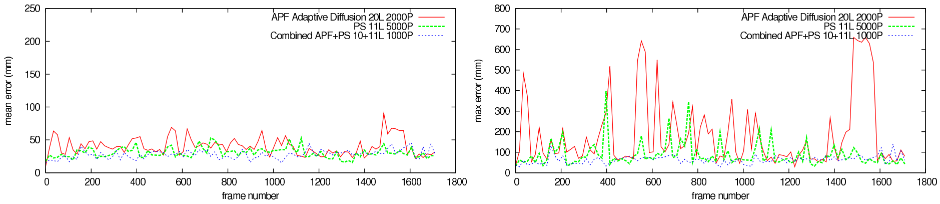

Figure 6: Comparison of APF, PS and the combined approach. Left: Mean error. Right: Max error. The combined approach runs at twice the speed and is more accurate.

6 Conclusion

We have presented detailed experimental evaluations of two common particle filter variants in the context of high dimensional 3D human pose estimation. Our results indicate that both PS and especially APF experience difficulties when dealing with human models considerably more complex than 30 DOF. We have proposed a combined approach that incorporates the complementary strengths of both methods to create a highly robust sampling strategy. Future directions are to further investigate this approach and to find the best balance between APF and PS. We also want to extend the comparison to evaluate the performance of particle filtering compared to optimization methods. *Smart particle filtering* [5] is an interesting combination of both that is worth investigating. Results of our work and accompanying real tracking videos can be found on our webpage `http://memoman.cs.tum.edu`. Although our tracking results from only silhouette shapes are quite good (we were able to continuously track a sequence of 9000 frames with only occasional errors in the upper limbs and head), a more informed observation model incorporating e.g. local appearance is recommended to further improve robustness.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005.

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

[3] A. O. Balan, L. Sigal, and M. J. Black. A quantitative evaluation of video-based 3d person tracking. In *ICCCN '05*, pages 349–356, 2005.

[4] J. Bandouch, F. Engstler, and M. Beetz. Accurate human motion capture using an ergonomics-based anthropometric human model. In *Proc. of the 5th International Conference on Articulated Motion and Deformable Objects (AMDO)*, 2008.

[5] M. Bray, E. Koller-Meier, and L. Van Gool. Smart particle filtering for high-dimensional tracking. *Computer Vision and Image Understanding (CVIU)*, 106(1):116–129, 2007.

[6] J. Deutscher, A. J. Davison, and I. D. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. *CVPR '01*, pages 669–676, 2001.

[7] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV)*, 61(2):185–205, 2005.

[8] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

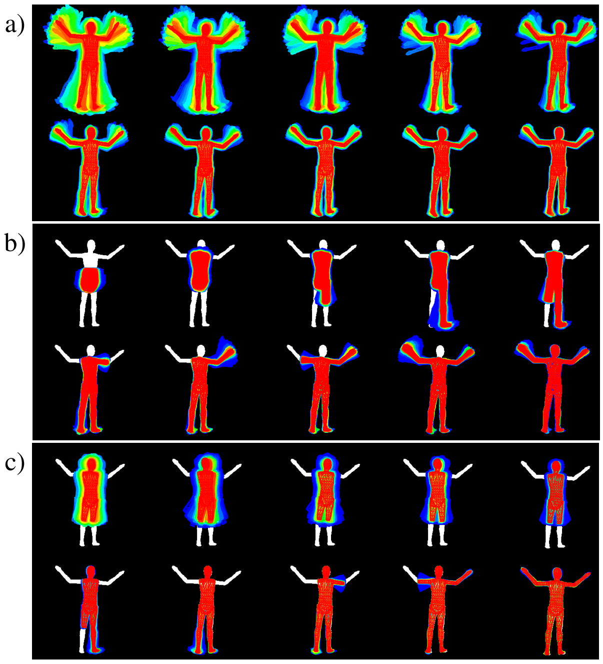

Figure 7: Particle development in different sampling strategies (from real data). Higher color temperature corresponds to higher weights; a) Iterations of APF with 20 layers (every second layer); b) Iterations of PS with 11 hierarchies (without the 6th); c) Iterations during combined APF and PS (first row shows the annealing partition of the extended torso, second row shows partitioning of the limbs).

[9] J. Gall, J. Potthoff, C. Schnörr, B. Rosenhahn, and H.-P. Seidel. Interacting and annealing particle filters: Mathematics and a recipe for applications. *Journal of Mathematical Imaging and Vision*, 28(1):1–18, 2007.

[10] R. Kehl and L. Van Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding (CVIU)*, 104(2):190–209, 2006.

[11] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV '00*, pages 3–19, 2000.

[12] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2):90–126, 2006.

[13] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):4–18, 2007.

[14] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR '06*, pages 238–245, 2006.

[15] P. Wang and J. M. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *CVPR '06*, pages 790–797, 2006.