Learning for Multi-View 3D Tracking in the Context of Particle Filters

Juergen Gall[1], Bodo Rosenhahn[1], Thomas Brox[2], and Hans-Peter Seidel[1]

[1] Max-Planck Institute for Computer Science
Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
{jgall, rosenhahn, hpseidel}@mpi-sb.mpg.de

[2] CVPR Group, Department of Computer Science, University of Bonn
Römerstr. 164, 53113 Bonn, Germany
brox@cs.uni-bonn.de

Abstract. In this paper we present an approach to use prior knowledge in the particle filter framework for 3D tracking, i.e. estimating the state parameters such as joint angles of a 3D object. The probability of the object's states, including correlations between the state parameters, is learned a priori from training samples. We introduce a framework that integrates this knowledge into the family of particle filters and particularly into the annealed particle filter scheme. Furthermore, we show that the annealed particle filter also works with a variational model for level set based image segmentation that does not rely on background subtraction and, hence, does not depend on a static background. In our experiments, we use a four camera set-up for tracking the lower part of a human body by a kinematic model with 18 degrees of freedom. We demonstrate the increased accuracy due to the prior knowledge and the robustness of our approach to image distortions. Finally, we compare the results of our multi-view tracking system quantitatively to the outcome of an industrial marker based tracking system.

1 Introduction

Model-based 3D tracking means to estimate the pose of a 3D object where the pose is determined by a value in a state space E. In the case of an articulated model of a human body, the pose is completely described by a 3D rigid body motion that has 6 degrees of freedom and the joint angles, which are 12 in this paper. This yields a high-dimensional state space that makes the tracking process difficult. Particle filters [1], however, can deal with high dimensions. A basic particle filter termed condensation has been used for contour tracking [2]. However, this algorithm lacks performance for 3D tracking. A heuristic that is based on these filters and that was successfully used for multi-view 3D tracking is the annealed particle filter (APF) [3]. In contrast to conventional particle filters, this method does not estimate the posterior distribution. Instead it performs a stochastic search for the global maximum of a weighting function. The two main drawbacks of the APF as applied in [3] are the simplified, unconstrained

kinematic model that results in a large number of particles needed for tracking and the assumption of a static background. The present paper addresses the first one by considering correlations between the state parameters as a soft constraint where the correlations are learned a priori. Using a level set based segmentation instead of background subtraction lifts the second assumption.

The idea to improve the model for 3D tracking by integrating prior knowledge is not new. In [4], training data acquired with a commercial motion capture system was used to learn a dynamical motion model (e.g. walking). This stabilizes the tracking as long as the assumptions are valid, but otherwise it is misleading and results in tracking failures. Hence, a large motion database is needed to learn more complicated motion models [5]. Hard constraints were also introduced for the 3D model such as anatomical joint angle limits and prevention of self-intersections [6]. This reduces the state space, but it does not consider the probability of different poses. In [7], it was suggested to learn a Gaussian mixture in a state space with reduced dimension, whereas the work in [8] captures the training data by a nonparametric Parzen density estimator. Our approach embarks on this latter strategy.

In previous works, a variational model for level set based image segmentation incorporating color and texture [9] has already been successfully used for pose estimation [10]. It is not based on background subtraction and, thus, does not necessarily need a static background. We combine this method with the APF to make the algorithm more flexible for applications.

The paper is organized as follows. We begin with a brief outline of the fundamental techniques, namely the APF and the variational model for image segmentation. Afterwards, in Section 3, we present the probabilistic model and motivate its choice. Furthermore, a full integration into a Bayesian framework is derived. Section 4 combines the prior knowledge with the methods from Section 2 and applies it to multi-view 3D tracking. The effect of the learned prior is demonstrated in Section 4. For our experiments we use a four camera set-up for tracking the lower part of a human body. Our articulated model consists of 18 degrees of freedom, and we will report on the robustness in the presence of occlusions and noise. Finally, we compare the results of our multi-view tracking system with a marker based tracking system. This provides a quantitative error measure. The paper ends with a brief summary.

2 Previous Work

2.1 Annealed Particle Filter

The APF does not approximate a distribution, usually the posterior distribution, like other particle filters [11]. Instead it performs a stochastic search for the global minimum of an "energy" function $V \geq 0$ by using n particles that are random variables in the state space. In accordance with simulated annealing [12], the weighting function is a Boltzmann-Gibbs measure that is defined in terms of V

and an inverse "temperature" $\beta > 0$ by

$$g(x)^{\beta}\,\lambda(dx) := \frac{1}{Z}\exp\left(-\beta V(x)\right)\,\lambda(dx), \tag{1}$$

where λ is the Lebesgue measure and $Z := \int \exp(-\beta V)\,d\lambda$. These measures have the property that the probability mass concentrates at the global minimum of V as $\beta \to \infty$. For avoiding that the particles are misguided by a local minimum, an annealing scheme $0 < \beta_M < \ldots < \beta_0$ is used. It provokes that the particles are weighted by smoothed versions of the weighting function with β_0 where the influence of the local minima is first reduced but then increases gradually as depicted in Figure 1. After the particles are initialized in accordance with an

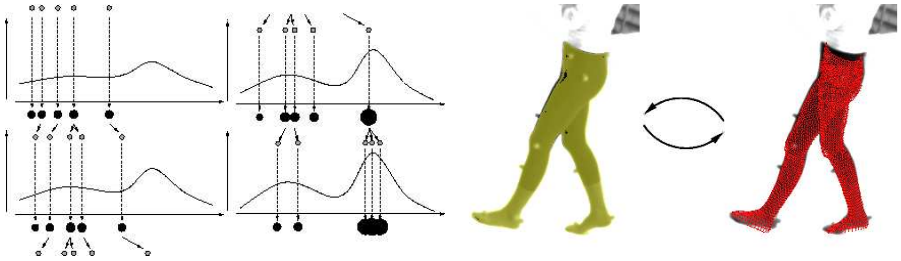

Fig. 1. Left: Illustration of the annealing effect with three runs. After weighting the particles (*black circles*), the particles are resampled and diffused (*gray circles*). Due to annealing, the particles migrate towards the global maximum without getting stuck in the local maximum. **Right:** The pose estimate (*right*) is obtained by weighting the particles according to the segmentation result (*left*). In return the pose result is used as shape prior for the segmentation of the next frame.

initial distribution, the APF with M annealing runs consists of a prediction step and an update step:

Prediction: Sample $\tilde{x}^{(i)}_{t+1,M}$ from $p(x_{t+1}|x^{(i)}_{t,0})\,\lambda(dx_{t+1})$
Update: For m from M to 0:

- Calculate weight $\pi^{(i)} = g(\tilde{x}^{(i)}_{t+1,m})^{\beta_m}$ and normalize weights to $\sum_i \pi^{(i)} = 1$.
- Generate $x^{(i)}_{t+1,m}$ by resampling with replacement, where $\tilde{x}^{(j)}_{t+1,m}$ is selected with probability $\pi^{(j)}$.
- Diffuse particles.

In the last run $m = 0$, the pose at time $t+1$ is estimated by $\widehat{x}_{t+1} = \sum_i \pi^i\,\tilde{x}^{(i)}_{t+1,0}$, and the particles are not diffused.

2.2 Variational Model for Segmentation

Level set based segmentation for r views splits the image domain Ω^i of each view into two regions Ω^i_1 and Ω^i_2 by level set functions $\Phi^i : \Omega^i \to \mathbb{R}$, such

that $\Phi^i(x) > 0$ if $x \in \Omega_1^i$ and $\Phi^i(x) < 0$ if $x \in \Omega_2^i$. The contour of an object is thus represented by the zero-level line. The approach described in [13] uses a variational model that integrates the contour of a prior pose $\Phi_0^i(\widehat{x}\,)$ for each view i. It minimizes the energy functional $E(\widehat{x}, \Phi^1, \ldots, \Phi^r) = \sum_{i=1}^{r} E(\widehat{x}, \Phi^i)$ where

$$\begin{aligned} E(\widehat{x}, \Phi^i) = &- \int_{\Omega^i} H(\Phi^i) \ln p_1^i + (1 - H(\Phi^i)) \ln p_2^i \, dx \\ &+ \nu \int_{\Omega^i} \left|\nabla H(\Phi^i)\right| \, dx + \lambda \int_{\Omega^i} \left(\Phi^i - \Phi_0^i(\widehat{x}\,)\right)^2 \, dx \end{aligned} \tag{2}$$

and H is a regularized version of the step function.

Minimizing the first term corresponds to maximizing the a-posteriori probability of all pixel assignments given the probability densities p_1^i and p_2^i of Ω_1^i and Ω_2^i, respectively. These densities are modeled by local Gaussian densities. The second term minimizes the length of the contour and smoothes the resulting contour. The last one penalizes the discrepancy to the shape prior. The relative influence of the three terms is controlled by the constant weighting parameters $\nu \geq 0$ and $\lambda \geq 0$. The interaction between segmentation with shape prior and the APF is illustrated in Figure 1. It has been shown that this method is robust in the case of a non-static background and that it is also able to deal with clutter, shadows, reflections, and noise [13].

3 Prior Knowledge in the Bayesian Framework

In the Bayesian framework, the particles are first predicted according to the transition density $p(x_{t+1}|x_t)$ and then updated by the likelihood $p(y_{t+1}|x_{t+1})$, where y_t is the observation at time t. The transition density, denoted by p_{pred}, is often modeled as zero-mean Gaussian since an accurate model is not available. This weak model does not include prior knowledge in an appropriate way. Since a precise model of the dynamics is not available for many applications, we combine the simple dynamical model p_{pred} with the probability density of the resulting pose p_{pose} that leads to a new transition density

$$p(x_{t+1}|x_t) := \frac{1}{Z(x_t)} \, p_{pred}(x_{t+1}|x_t) \, p_{pose}(x_{t+1}), \tag{3}$$

where $Z(x_t) := \int p_{pred}(x_{t+1}|x_t) \, p_{pose}(x_{t+1}) \, dx_{t+1}$. As it is often expensive to sample from the corresponding distribution, we show that it is possible to integrate p_{pose} in the update step. Following the basic notations of [1, p. 6], we obtain

$$p_\star(x_{t+1}|y_0, \ldots, y_t) := \int \frac{1}{Z(x_t)} \, p_{pred}(x_{t+1}|x_t) \, p(x_t|y_0, \ldots, y_t) \, dx_t, \tag{4}$$

$$p(x_{t+1}|y_0, \ldots, y_{t+1}) = \frac{p(y_{t+1}|x_{t+1}) \, p_{pose}(x_{t+1}) \, p_\star(x_{t+1}|y_0, \ldots, y_t)}{\int p(y_{t+1}|x_{t+1}) \, p_{pose}(x_{t+1}) \, p_\star(x_{t+1}|y_0, \ldots, y_t) \, dx_{t+1}} \tag{5}$$

where Equation (4) describes the prediction step and Equation (5) the update step. It is obvious that $p_\star$ is a density function, but not a probability density function, satisfying $p(x_{t+1}|y_0,\ldots,y_t) = p_{pose}(x_{t+1})\, p_\star(x_{t+1}|y_0,\ldots,y_t)$. Note that sampling from the distribution $p_{pred}(x_{t+1}|x_t)/Z(x_t)\,\lambda(dx_{t+1})$ is equivalent to sample from $p_{pred}(x_{t+1}|x_t)\,\lambda(dx_{t+1})$ for a given x_t. Hence, the prediction step of the particle filter remains unchanged, while the particles are weighted by the product $p(y_{t+1}|x_{t+1})\, p_{pose}(x_{t+1})$ instead of the likelihood during updating.

Only in rare cases we are able to give an analytical expression for p_{pose}. Instead, we suggest to learn the probability of the various poses from a finite set of training samples. For a nonparametric estimate of the density we use a Parzen-Rosenblatt estimator [14]

$$p_{pose}(x) = \frac{1}{(2\,\pi\,\sigma^2)^{d/2}\,N} \sum_{i=1}^{N} \exp\left(-\frac{d(x,x_i)^2}{2\,\sigma^2}\right) \tag{6}$$

to deal with the complexity and the non-Gaussian behavior of the distribution, where N denotes the number of training samples and the function d is a distance measure in E. This estimate depends on the window size σ that is necessary to be chosen in an appropriate way. While a small value of σ forces the particles to stick to the training data, a greater value of σ approximates the density smoother. In order to cope with this, we chose σ as the maximum second nearest neighbor distance between all training samples, i.e. the two neighbors of a sample are at least within a standard deviation. Other values for the window size are discussed in detail in [15].

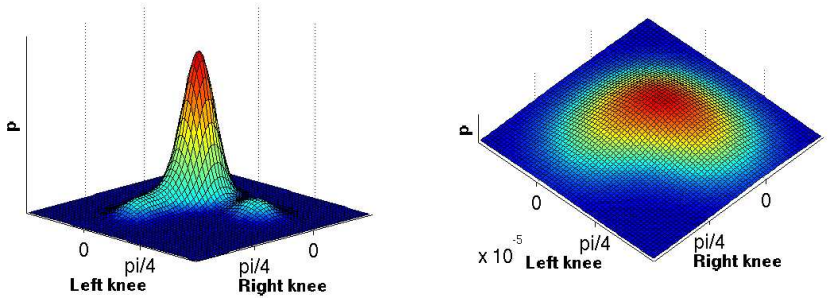

Fig. 2. The Parzen estimate subject to the angles of the knee joints. **Left:** Using the Euclidean distance leads to a domination of the knee joints. The density rapidly declines to zero as the values differ from the data. **Right:** The influence of the knees is reduced by the weighted Euclidean distance.

We have not yet specified the norm for evaluating the distance between a training sample x_i and a value x in the d-dimensional state space E for Equation (6). The commonly used Euclidean distance weights all dimensions of the

state space uniformly. This means in the context of human motion estimation that a discrepancy of the knee contributes to the measured distance in the same matter as a discrepancy of the ankle. As illustrated in Figure 2, this involves a dominated measure by joints with a relatively large anatomical range as the knee in comparison to joints with a small range as the ankle. Therefore, we propose using a weighted Euclidean distance measure that incorporates the variance of the various joints, i.e.

$$d(x, x_i) := \sqrt{\sum_{k=1}^{d} \frac{((x)_k - (x_i)_k)^2}{\rho_k}}, \qquad \rho_k := \frac{\sum_{i=1}^{N} \left((x_i)_k - \overline{(x)_k}\right)^2}{N-1} \tag{7}$$

where $\overline{(x)_k}$ denotes the arithmetic mean of the samples in the kth dimension. This distance is generally applied in image analysis [16] and is equivalent to a Mahalanobis distance in the case that the covariance matrix is diagonal. A full covariance matrix significantly increases the computation in high dimensional spaces.

Additionally, the prior knowledge is suitable for setting the covariance matrix of the zero-mean Gaussian density p_{pred}. One approach is to estimate the variance of the differences between succeeding samples. However, this has the drawback that training data from a quite large range of dynamics are needed a priori. In the case where the sample data only include walking sequences, the prediction is not accurate for tracking a running person. Thus setting the variances proportional to ρ_k is generally applicable and better than adjusting the parameters manually. We remark finally that not all parameters of a pose can be learned. For example, it does not make sense to learn the position of an object. Therefore, the density is usually estimated in a slightly lower dimensional space than the state space.

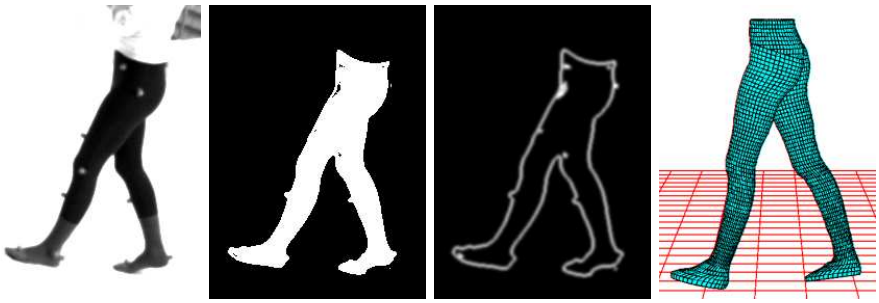

Fig. 3. Feature extraction by level set segmentation. **From left to right:** (a) Original image. (b) Extracted silhouette. (c) The smoothed contour is slightly deformed by the markers needed for the marker based system. (d) 3D model with 18 DOF used for tracking.

4 Application to Multi-View 3D Tracking

4.1 Feature extraction

For weighting the particles during the update step of the APF, features from an image y_t have to be extracted. In previous works, only low-level features assuming a static background as foreground silhouette, edges, or motion boundaries [3, 6] were considered. In our work, the level set based image segmentation from Section 2.2 with the experimentally determined parameter $\nu = 4$ is applied using the estimated pose $\widehat{x}_{t-1}$ from the previous time step. The resulting level set describes the silhouette and the contour of the observed object. We remark that the extraction of this image feature is not independent of the estimate anymore. This yields a weighting function that depends not only on the current image and the particle itself, but also on the whole set of particles defining the estimate. Even though particle filters already provide an interaction between the particles due to the normalization of the weights, it holds the danger that a segmentation error leads to an estimate error and vice-versa. However, the influence of the estimate on the segmentation can be regulated by the parameter λ. Our experiments, where we set $\lambda = 0.04$, show indeed that a proper value for this parameter avoids this problem.

4.2 Weighting Function

The error between a particle and the observed image y is calculated pixel-wise similar to [3]. Each particle $x \in E$ determines a pose of our 3D model. The projected surface of the model into the image plane gives a set of silhouette points $S_i^S(x)$ and a set of contour points $S_i^C(x)$ for each view $i = 1, \ldots, r$, where a set contains all pixels $p \in \mathbb{R}^2$ of the silhouette and the contour, respectively. The silhouette S_i^y of the observed object is obtained from the level set function Φ^i, where $S_i^y(p) = 1$ if $\Phi^i(p) > 0$ and $S_i^y(p) = 0$, otherwise. The contour C_i^y is just the boundary of the silhouette smoothed by a Gaussian filter and normalized between 0 and 1, cf. Figure 3. Then the error functions are defined by

$$err_L(x, y, i) := \frac{1}{\left|S_i^L(x)\right|} \sum_{p \in S_i^L(x)} (1 - L_i^y(p))^2. \tag{8}$$

for $L \in \{S, C\}$. Following Section 3, we integrate the learned prior knowledge in form of the probability density p_{pose}. Altogether the energy function of the weighting function (1) can be written as

$$V(x, y) := \sum_{i=1}^{r} (err_S(x, y, i) + err_C(x, y, i)) - \eta \ln(p_{pose}(x)), \tag{9}$$

where the parameter $\eta \geq 0$ controls the influence of the prior knowledge. It is obvious that $V \geq 0$ and $g(x, y)^{\beta_m} \lambda(dx)$ is thus a Boltzmann-Gibbs measure. Furthermore, the constant term $(2\,\pi\,\sigma^2)^{d/2}$ of p_{pose} can be omitted since it is

canceled out when normalizing the weights. Note that the prior knowledge is embedded in accordance with the Bayesian framework by multiplying the old weighting function with $(p_{pose})^{\eta}$. Our method performs well with $\eta \in [0.06, 0.1]$ as we demonstrate below.

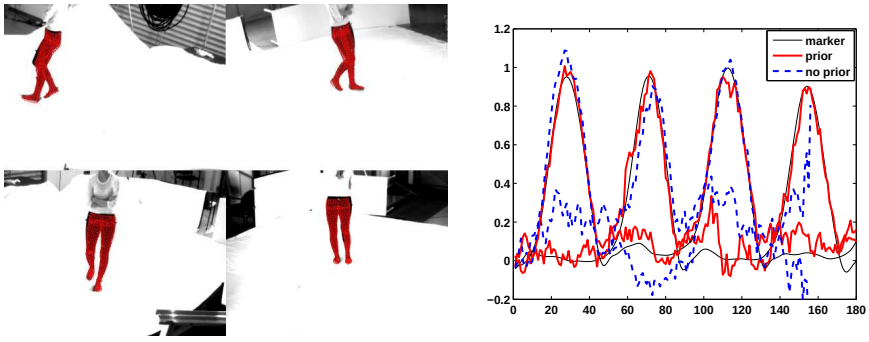

Fig. 4. Left: Results for a walking sequence captured by four cameras. **Right:** The joint angles of the right and left knee. *Solid (thin):* Marker based system. *Solid (thick):* Prior with weighted distance. *Dashed:* Without prior (Tracking fails).

5 Experiments

In our experiments we track the lower part of a human body using four calibrated and synchronized cameras. The sequences are simultaneously captured by a commercial marker based system[3] allowing a quantitative error analysis. The black leg suit and the attached retroflective markers are required by the marker based system, see Figure 3.

The training data used for learning p_{pose} consists of 480 samples obtained from walking sequences of the same person. The data was captured by the commercial system before recording the test sequences. The parameters of the APF are set during the experiments as follows: 10 annealing runs are applied with $\beta_m = 8\,(1-1.6^{m-11})$ and 250 particles. The resampling step includes a crossover operator [3], and the particles are diffused according to a zero-mean Gaussian distribution with covariance matrix determined by $0.1\,\rho_k$, see (7). The initial distribution is the Dirac measure of the initial pose. Our implementation took several minutes for processing 4 images of one frame.

Figure 4 visualizes results of a walking sequence that is not contained in the training data. For the sake of comparison, the results of the APF without using prior knowledge at all are also visualized in Figure 5. The estimated angles of the left and the right knee are shown in the diagram in Figure 4 where the values acquired from the marker based system provide a ground truth with an

[3] We used the Motion Analysis system with 8 Falcon cameras

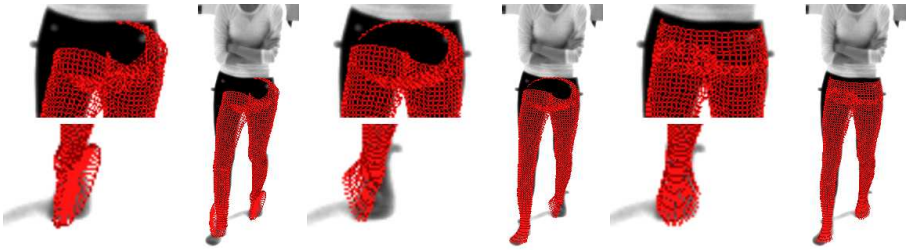

Fig. 5. Visual comparison of results. **From left to right:** (*a*) Without prior. (*b*) Without weighted distance. (*c*) With weighted distance.

accuracy of about 3 degrees. It allows to analyze the quantitative error of our method in contrast to previous works, e.g. [3], where visual comparisons indicate roughly the accuracy of the pose estimates. The root mean square (RMS) error for both knees is 6.2 degrees (red line). While tracking with 100 particles failed, our method also succeeded using 150 and 200 particles with RMS errors 15.3 and 8.8 degrees, respectively.

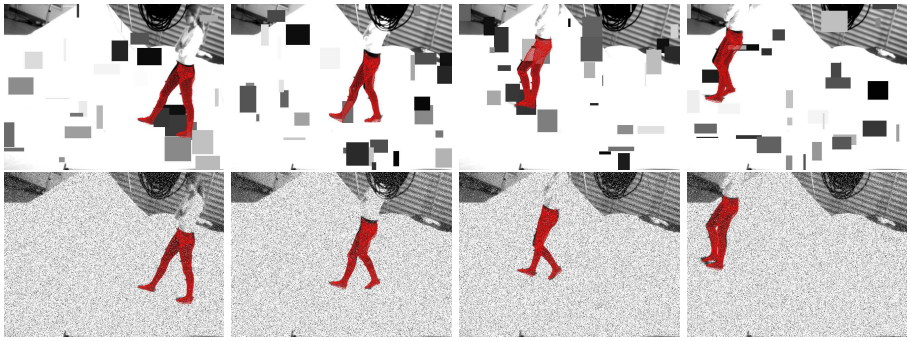

Fig. 6. Results for distorted sequences (4 of 181 frames). Only one camera view is shown. **Top:** Occlusions by 30 random rectangles. **Bottom:** 25% pixel noise.

Figure 6 shows the robustness in the presence of noise and occlusions. Each frame has been independently distorted by 25% pixel noise and by occluding rectangles of random size, position and gray value. The legs are tracked over the whole sequence with RMS errors 8.2 and 9.0 degrees, respectively. Finally, we applied the method to a sequence with scissor jumps, see Figure 7. This demonstrates that our approach is not restricted to the motion patterns that were used for training as it is when learning the patterns instead of the poses. However, the 7th image also highlights the limitations of the prior. Since our training data are walking sequences, the probability that both knees are bended is almost zero, cf. Figure 2. Therefore a more probable pose is selected with less

bended knees. It yields a higher hip of the 3D model than in the image. Overall, the RMS error is 8.4 degrees. A similar error can be observed for the feet since they are more bended for jumping as for walking. Nevertheless, the result is much better than without using any prior.

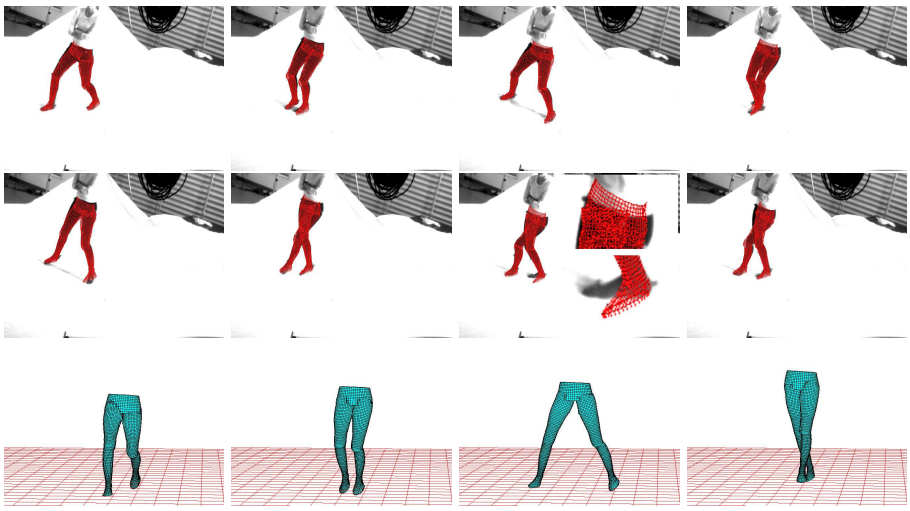

Fig. 7. Rows 1-2: Results for a sequence with scissor jumps (8 of 141 frames). **Row 3:** The 3D models for the 4 poses on the left hand side of rows 1 and 2 are shown from a different viewpoint.

6 Summary

We have presented a method that integrates a-priori knowledge about the distribution of pose configurations into the general model of particle filters as well as into the special APF scheme. Thereby, the prior ensures that particles representing a familiar pose are favored. Since only single pose configurations and not whole motion patterns are learned, a relatively small set of training samples is sufficient for capturing a variety of movements. Our experiments provide a quantitative error analysis that clearly demonstrates the increased accuracy of the APF due to the incorporated prior knowledge. Moreover, we have shown that our approach combined with a variational model for level set based image segmentation is able to deal with distorted images, a case where common techniques that rely on background substraction fail. Since we were restricted to use artificial distortions by the marker-based system, further work will be done to evaluate the system in real examples like crowded and outdoor scenes. Work on acquiring training data from motion databases and handling occlusions by clothes is also in progress.

References

1. Doucet, A., de Freitas, N., Gordon, N., eds.: Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science. Springer, New York (2001)
2. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. of Computer Vision **29** (1998) 5–28
3. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. Int. J. of Computer Vision **61** (2005) 185–205
4. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In: European Conf. on Computer Vision. Volume 2. (2000) 702–718
5. Sidenbladh, H., Black, M., Sigal, L.: Implicit probabilistic models of human motion for synthesis and tracking. In: European Conf. on Computer Vision. Volume 1. (2002) 784–800
6. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. Int. J. of Robotics Research **22** (2003) 371–391
7. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearily embedded visual inference. In: Int. Conf. on Machine Learning. (2004)
8. Brox, T., Rosenhahn, B., Kersting, U., Cremers, D.: Nonparametric density estimation for human pose tracking. In: Pattern Recognition (DAGM). LNCS, Springer (2006) To appear.
9. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Unsupervised segmentation incorporating colour, texture, and motion. In Petkov, N., Westenberg, M.A., eds.: Computer Analysis of Images and Patterns. Volume 2756 of LNCS., Springer (2003) 353–360
10. Rosenhahn, B., Brox, T., Smith, D., Gurney, J., Klette, R.: A system for markerless human motion estimation. Künstliche Intelligenz **1** (2006) 45–51
11. Crisan, D., Doucet, A.: A survey of convergence results on particle filtering methods for practitioners. IEEE Transaction on Signal Processing **50** (2002) 736–746
12. Kirkpatrick, S., Gelatt, C., Vecchi, M.: Optimization by simulated annealing. Science **220** (1983) 671–680
13. Brox, T., Rosenhahn, B., Weickert, J.: Three-dimensional shape knowledge for joint image segmentation and pose estimation. In Kropatsch, W., Sablatnig, R., Hanbury, A., eds.: Pattern Recognition (DAGM). Volume 3663 of LNCS., Springer (2005) 109–116
14. Parzen, E.: On estimation of a probability density function and mode. Annals of Mathematical Statistics **33** (1962) 1065–1076
15. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
16. Mukundan, R., Ramakrishnan, K.: Moment Functions in Image Analysis: Theory and Application. World Scientific Publishing (1998)