Unscented Kalman Filtering for Articulated Human Tracking

Anders Boesen Lindbo Larsen, Søren Hauberg, and Kim Steenstrup Pedersen

Department of Computer Science
University of Copenhagen
{abll,hauberg,kimstp}@diku.dk
http://diku.dk/

Abstract. We present an articulated tracking system working with data from a single narrow baseline stereo camera. The use of stereo data allows for some depth disambiguation, a common issue in articulated tracking, which in turn yields likelihoods that are practically unimodal. While current state-of-the-art trackers utilize particle filters, our unimodal likelihood model allows us to use an unscented Kalman filter. This robust and efficient filter allows us to improve the quality of the tracker while using substantially fewer likelihood evaluations. The system is compared to one based on a particle filter with superior results. Tracking quality is measured by comparing with ground truth data from a marker-based motion capture system.

1 Introduction

Articulated human motion tracking is the process of estimating the human body configuration over time from a series of sensor inputs [1]. Motion tracking has a wide variety of uses ranging from computer gaming and film making to medical applications. Currently, the most accurate methods are based on physical markers attached to the human body that can be tracked in three dimensions using multiple calibrated cameras. These methods have serious drawbacks since they are cumbersome to set up and too intrusive to be used easily outside laboratory settings, e.g. in private homes. Therefore, an accurate markerless tracking method based solely on input from a camera is needed for a vast array of non-laboratory applications.

To alleviate this need, much research has gone into markerless articulated tracking. The most common solution is to use a nonlinear filter with a likelihood model based on monocular images. Due to the lack of depth information from such data, these likelihood models are inherently multimodal, which has forced researchers to perform the inference using very general techniques such as particle filters [2,3,4,5,6]. However, recent boosts in computational power has made consumer stereo cameras possible, see e.g. the Bumblebee[1] or the Microsoft

[1] http://www.ptgrey.com/products/bumblebee2/

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 228–237, 2011.

Kinect[2] camera. Using such cameras allows us to construct approximately unimodal likelihood models. This, in turn, allows us to perform the inference using the more constrained *Unscented Kalman Filter (UKF)* [7,8]. These constraints allow for a more robust estimation using fewer computational resources compared to a particle filter. Both these features are sorely needed in practical applications.

The objective of articulated tracking is to estimate joint angles of a skeleton model in each frame of an image sequence. The most common approach is to infer these joint angle from monocular images using a particle filter, see e.g. [2,3,4,5,6].

Due to the flexibility of the human body, the skeleton model needs to exhibit many degrees of freedom. Robust estimation of joint angles then requires many samples in the particle filter, rendering the approach computationally very demanding. A commonly used approach to deal with this problem is to reduce the degrees of freedom in the model by confining the set of legal joint angles to some (often nonlinear) subspace of the angle space. It seems that most researchers taking this route focus on simple low-dimensional motions, such as *walking* [9,10,11,12], *golf swings* [11,12], *tennis playing* [13] etc. This approach can be both robust and computationally efficient, but suffers from the main drawback that the resulting trackers only work with very specific motions.

The need for particle filters stems from the fact that the used likelihood models often are multimodal, making the posterior distribution of the joint angles multimodal as well. The multimodality of the likelihood comes from the use of monocular images that makes depth ambiguities an inherent part of the problem. Examples of such likelihoods include a combination of edge strength and horizontal flow [14], silhouettes extracted using background subtraction [5] and texture models for each limb [2]. One way of making the largest mode of the likelihood easier to locate is to use multiple calibrated cameras, as was done by Deutscher et al. [3]. The need for several calibrated cameras, however, makes the approach hard to use outside the laboratory. One compromise is to use a single pre-calibrated stereo camera as suggested by Hauberg et al. [6]. This is also the approach we will be taking as it will allow us to infer the joint angles using an unscented Kalman filter.

Unscented Kalman filters have seen little use in articulated tracking as the likelihood models have usually been multimodal which does not fit with the Gaussian assumptions of this filter. One notable exception is the work of Ziegler et al. [15] whose approach shares many similarities with ours. Using four stereo cameras placed at a 90° angle from each other, they are able to track a human upper body reliably using the UKF. This is also to be expected as data from the four stereo cameras should be sufficient to avoid any observational ambiguities. Another example of articulated tracking with the UKF is found in [16], where a hand is tracked. Here, the likelihood is based on edges in a monocular image, so there is little reason to believe that the likelihood actually is unimodal. Further-

[2] `http://www.xbox.com/kinect`

more, it seems that they only conduct experiments on image sequences of hands where the articulation of the fingers remains unchanged for the entire sequence.

In this work we make the following contributions.

- We show that unimodality of the human pose distribution can be assumed when working with data from a single, narrow-baseline stereo camera.
- We apply the unscented Kalman filter for articulated tracking and achieve superior results in terms of accuracy and realism of body movements. Furthermore, UKF gives us the benefit of requiring significantly fewer likelihood evaluations resulting in a lower computational complexity.

This paper is organized as follows. In the next section we describe the general nonlinear filtering framework and two possible implementations: the UKF and the particle filter. This is then specialized to articulated tracking in Sec. 3. Results are presented in Sec. 4 and the paper is concluded in Sec. 5.

2 Nonlinear Filtering

The articulated tracking of human motion can be formulated as a nonlinear estimation problem modelled by the two difference equations

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}, \boldsymbol{v}_{t-1}) \tag{1}$$

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t, \boldsymbol{n}_t) \tag{2}$$

where $\boldsymbol{x_t} \in \mathbb{R}^{n_x}$ denotes the state of the system at time t and $\boldsymbol{y}_t \in \mathbb{R}^{n_y}$ the observation. With our motion tracking, the system state corresponds to the pose of a human body while the observation is a stereo image of the human. The function f models the transition between system states over time while h relates the hidden state space to the observation space. Both f and h are deterministic. $\boldsymbol{v}_t$ and $\boldsymbol{n}_t$ are random variables representing process noise and measurement noise respectively.

2.1 The Unscented Kalman Filter

Below follows a very brief introduction to the UKF, we refer to [7,8] for a thorough presentation.

The UKF provides a sequential estimation of the posterior density $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ where $\boldsymbol{y}_{1:t} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t\}$. This is achieved by updating the posterior density recursively. In each time step, UKF selects a set of $2n_x + 1$ sample points $\boldsymbol{\mathcal{X}}_i$, $i = 0, 1, \ldots, 2n_x$ that completely captures the mean and covariance of the state distribution $p(\boldsymbol{x})$. These sample points (called *sigma points*) are then updated according to the state prediction function f and propagated through the observational model h into observation space. In observation space, their deviation from the observation is measured by the likelihood model $p(\boldsymbol{y}|\boldsymbol{\mathcal{X}}_i)$. From the likelihood of all sigma points, the Kalman gain $\boldsymbol{K}$ is updated. $\boldsymbol{K}$ is then used to update the state estimate $\boldsymbol{x}$ as well as the state distribution $p(\boldsymbol{x})$.

2.2 The Particle Filter

The particle filter works by generating a set of n weighted random sample points $\boldsymbol{\mathcal{X}}_i$, $i = 1, 2, \ldots, n$ from the prior distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. Like with the sigma points of UKF, each of these sample points are projected into observation space where their likelihood $p(\boldsymbol{y}|\boldsymbol{\mathcal{X}}_i)$ is quantified and weights assigned accordingly. The new pose estimate $\boldsymbol{x}_t$ becomes the mean of $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. A more comprehensive description of the particle filter is presented in [17].

3 Filtering for Articulated Tracking

3.1 The State Model

The articulated human body model is built from a kinematic skeleton consisting of rigid bones connected by joints with up to three degrees of freedom depending on the joint type (e.g. an elbow joint has one degree of freedom and a shoulder joint has three). This approach is common within articulated tracking [1,3]. The set of joint angles of the kinematic skeleton constitutes our state model vector $\boldsymbol{x}$. In this work we limit our tracker to consider only a human body from the hip and up as depicted in Fig. 1. Furthermore, we assume that the human is standing still and only moving the upper body parts. Notice however, that it is trivial to extend the model to include full body motion.

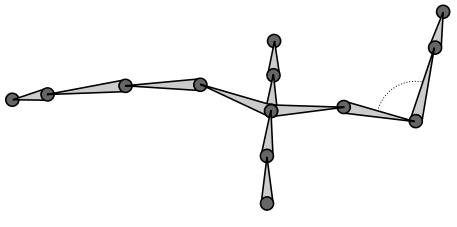


Fig. 1. The kinematic skeleton of the upper human body that we wish to track

As joints of the human skeleton cannot move freely due to physical constraints, we enforce similar angular constraints on our model. More specifically, we limit each angle to some interval $[l, u]$ where l and u denote the lower and upper bound. These box constraints are applied to both the sigma points and the samples in the particle filter to ensure that the prediction does not consider illegal joint angles. However, we do not handle self-intersections between body parts.

We initialize the first state $\boldsymbol{x}_0$ manually so that it matches the actual state as close as possible. We also provide a probability density estimate $p(\boldsymbol{x}_0)$ of the initial state. The state is propagated in time by adding zero mean Gaussian noise to each joint angle independently, i.e.

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}) \tag{3}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix. In our experience, it is not worthwhile to perform prediction of the state transition between frames as the changes are too small. Therefore, we perform no actual state transition between frames by letting function f from Eq. 1 be the identity function. The above model has, among others, been applied by Sidenbladh et al. [10], Balan et al. [18] and Bandouch et al. [19].

3.2 The Observation Model

The stereo camera provides us with a set of three-dimensional points in each frame. We perform a simple but efficient segmentation of the input image by removing points that are further away than a certain background threshold. If the remaining points contain any outliers (points far away from the body), we translate these points to bring them within a given Euclidean distance of their nearest limb. This final set of points constitutes the input observation vector $\boldsymbol{y}$. An example of an observation along with a human pose estimate is shown in Fig. 2.

(a) (b)

Fig. 2. (a) A segmented stereo image of a human body. (b) A human skeleton estimate projected on the image data.

We use the observational model presented in [6]: For each time step t we generate a set of sample points $\boldsymbol{\mathcal{X}}$ of which each sample $\boldsymbol{\mathcal{X}}_i$ is to be compared with the observation in order to compute the likelihood $p(\boldsymbol{y}|\boldsymbol{\mathcal{X}}_i)$. For this, we use the nonlinear mapping h (Eq. 2) constructed as follows. Given a state vector $\boldsymbol{\mathcal{X}}_i$ and an input observation $\boldsymbol{y}$, we want to represent the state $\boldsymbol{\mathcal{X}}_i$ in observation space. The state is composed of all joint angles in a kinematic skeleton. To each bone in this skeleton we assign a cylinder with a radius corresponding to the thickness of the limb; these will serve as our skin model. We then project all points from $\boldsymbol{y}$ onto the nearest cylinder of the stick figure. As we are working with cylinders, these projections can be performed trivially. By projecting the points of $\boldsymbol{y}$ onto skeleton $\boldsymbol{\mathcal{X}}_i$ we obtain a new observation vector $\boldsymbol{\mathcal{Y}}_i$ that is comparable to $\boldsymbol{y}$ since they both have the same dimensionality and the points in the vectors correspond to each other. Thus, the likehood model can be expressed as

$$p(\boldsymbol{y}|\boldsymbol{\mathcal{X}}_i) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mathcal{Y}}_i, \lambda^2 \boldsymbol{I}) \ , \tag{4}$$

where λ^2 is a variance parameter.

3.3 Computational complexity

The computational complexity of the tracker depends on the filtering method used. For the particle filter, the computational complexity is $O(n(n_y + \log(n)))$ with n being the number of particles sampled and n_y the dimensionality of the observation space. For the UKF, the computational complexity is $O(n_x^2 n_y^2)$ due to a singular value decomposition of a $\mathbb{R}^{n_y \times n_x}$ matrix. In our experience, the performance of UKF is very competitive with that of the particle filter since $n >> n_x$.

4 Results and Evaluation

To measure the performance of the particle filter vs. the UKF we apply both filters on two image sequences of 300 frames each. Examples of the results are shown in Fig. 3. The videos are available from `http://humim.org/scia2011`. We see that the UKF provides smoother and visually more accurate results compared to the particle filter. Only when the particle filter sampling is dense (1500 particles), the quality is somewhat close visually to that of the UKF. In the first image sequence, both filters are able to track the motion reasonably. The second sequence is harder to track as body parts move close to each other and self-occlusions occur. The particle filter fails on several occasions during sequence 2. UKF proves more robust than the particle filter as it misestimates the human pose on only one occasion where an arm passes by the head closely. We believe that most of these problems are caused by our simple skin model and our observational model that for each point in the observation makes a projection onto the nearest cylindrical limb. This is very likely to cause problems when limbs are positioned close to each other.

Overall, the unimodal assumption seems to hold since the observational model is strong enough to favorize the single, correct pose by a large margin. It is possible, however, to imagine special cases in which unimodality cannot be assumed, e.g. if an entire arm is hidden behind a person's back. In this case, when updating the Kalman gain $\boldsymbol{K}$, the variance of the kinematic joints related to the arm will automatically be adjusted to reflect this uncertainty. Thus, when the variance goes up for certain joints, the tracker should try to estimate these joints differently, e.g. by relying on a predictive model.

4.1 Accuracy

To obtain a more precise basis for comparison, the tracked person is wearing physical markers that are tracked in 3D using a high precision motion capture system. These will serve as our ground truth data. In total, there are eight markers placed on the human; three markers on each arm and two on the shoulders.

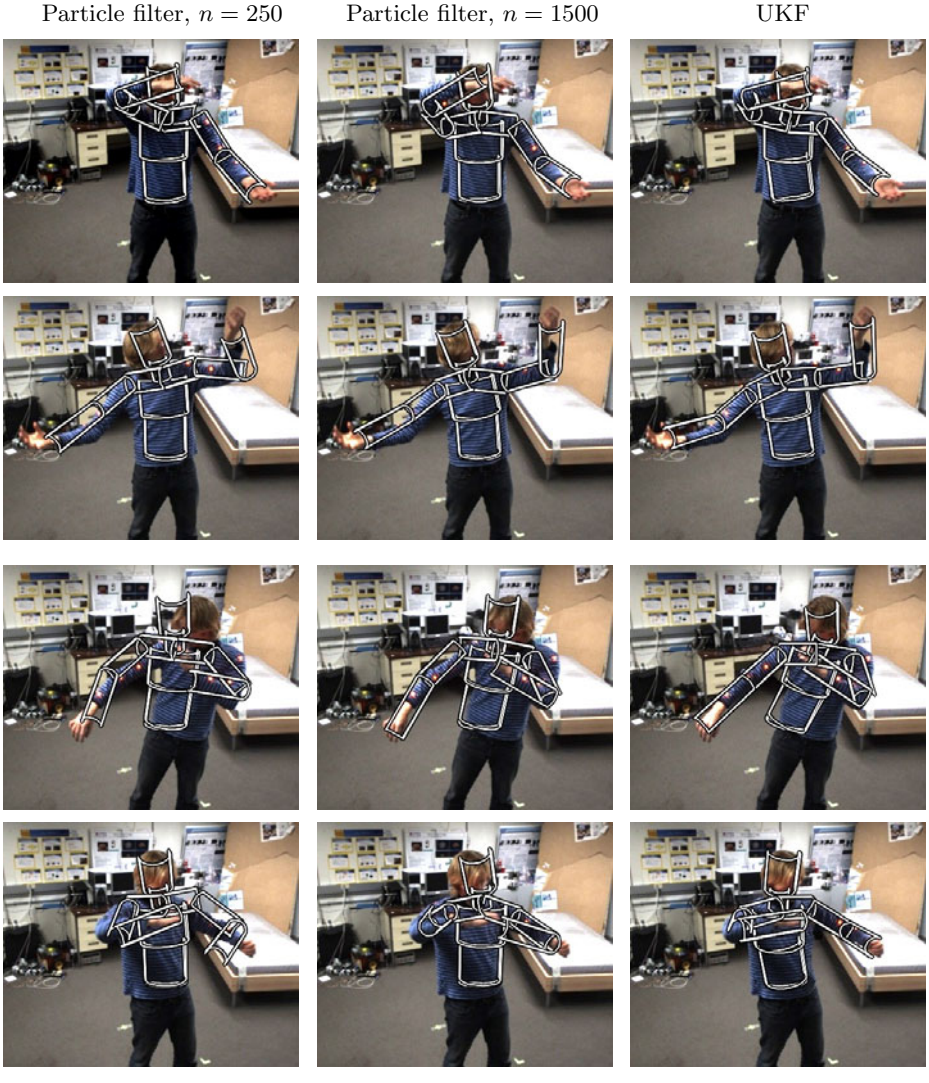
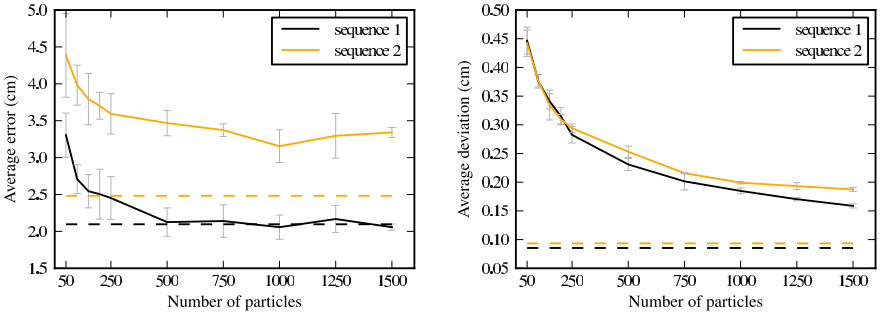


Fig. 3. The human skeleton estimated by the different filters is superimposed over selected frames from two videoes. The images in the upper two rows comes from video 1 while the two bottom rows are taken from video 2. The full videoes are available at `http://humim.org/scia2011`. Both particle filters have visible difficulties tracking the motion in the sequence as they seem less prone against self-occlusions and closely positioned body parts (which happens more often in sequence 2 than in sequence 1).



(a) (b)

Fig. 4. (a) Average error of the tracking filters. The particle filter is represented by the solid lines and the UKF by the dashed lines. The vertical error bars represent two times the standard deviation caused by the Monte Carlo sampling. The deviation is measured over several trials of the particle filter. (b) Smoothness of the filters measured by the average deviation of absolute joint positions between time steps. Low values indicate smooth trajectories. The solid lines represent the particle filter and the dashed lines represent the UKF.

To quantify the tracking quality we measure how well the markers fit with the estimated poses. For each marker, we make a projection onto the nearest limb; just as we did in the observational model. The Euclidean distance between the projection $\boldsymbol{o}$ and the marker point $\boldsymbol{m}$ can then be used as error measure. To determine the error from all eight markers of a state $\boldsymbol{x}$ over all time steps T, we calculate the average error:

$$\mathcal{E}(\boldsymbol{x}_{1:T}) = \frac{1}{8T} \sum_{t=1}^{T} \sum_{j=1}^{8} ||\boldsymbol{m}_{t,j} - \boldsymbol{o}_{t,j}|| \quad . \tag{5}$$

The resulting average error for the different filters are shown in Fig. 4a. It is clear that the UKF performs just as good or better than particle filters with a dense sampling. Furthermore, it is noteworthy that the monte carlo sampling of the particle filter results in significant deviations in accuracy when repeating the tracking. In this regard, the deterministic algorithm of the UKF offers more reliable results.

4.2 Motion Smoothness

When looking at the image sequences, it becomes clear that the UKF tracking produces smoother and more realistic motions whereas the skeleton generated by the particle filter tends to shake between time steps. To quantify this smoothness, we introduce the following measure. For each time step t we take the absolute position $\boldsymbol{a}_{t,j}$ of each joint j in the human skeleton and measure the movement

from the previous time step. The smoothness measure is then calculated as the average deviation of all joints J over T time steps.

$$\mathcal{S}(\boldsymbol{x}_{1:T}) = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} ||\boldsymbol{a}_{t,j} - \boldsymbol{a}_{t-1,j}|| \tag{6}$$

The results of our filters are shown in Figure 4b. UKF is clearly superior with a low deviation between time steps. One could image that another flavor of the particle filter such as the annealed particle filter will give more smooth results. However, filters that rely on Monte Carlo methods will always exhibit some jittering. This reveals another advantage of using the deterministic UKF.

5 Conclusion

In this paper we have shown that not only is the unscented Kalman filter applicable for articulated tracking, it also provides superior results in terms of accuracy and smoothness compared to the particle filter using substantially fewer likelihood evaluations. For this to be possible it is, however, essential that the likelihood model is mostly unimodal. For general monocular situations this is not the case, but it seems to be a reasonable assumption when working with stereo data. This observation makes practical articulated tracking systems much more plausible.

In this paper we used a simple likelihood model based on a simple skin model. For particle filters, this simplicity is essential as we need to be able to evaluate the likelihood fast due to the vast number of particles required. When using UKF, more involved likelihood models are possible as we only need to evaluate it a few times due to the low number of sigma points. Thus, in the future, we will consider more realistic skin models in the observational model. Other future work includes an automatic initialization of the tracker as well as an extension of the implementation to work with full body models as this will extend the use of the tracking system. Finally, we need a more elaborate evaluation of the tracker on more sequences of varied complexity.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding 108, 4–18 (2007)
2. Hauberg, S., Lapuyade, J., Engell-Nørregård, M., Erleben, K., Steenstrup Pedersen, K.: Three dimensional monocular human motion analysis in end-effector space. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 235–248. Springer, Heidelberg (2009)
3. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 126–133. IEEE Comput. Soc, Los Alamitos (2000)

4. Bandouch, J., Beetz, M.: Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In: IEEE International Workshop on Human-Computer Interaction, vol. 2 (2009)
5. Kjellström, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: CVPR 2010: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010)
6. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 425–437. Springer, Heidelberg (2010)
7. Julier, S., Uhlmann, J.: A new extension of the Kalman filter to nonlinear systems. In: Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, vol. 3, p. 26 (1997)
8. Wan, E., Van Der Merwe, R.: The unscented Kalman filter for nonlinear estimation. In: Proceedings of Symposium, pp. 153–158 (2000)
9. Lu, Z., Carreira-Perpinan, M., Sminchisescu, C.: People Tracking with the Laplacian Eigenmaps Latent Variable Model. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA (2008) 1705–1712
10. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3D human figures using 2D image motion. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 702–718. Springer, Heidelberg (2000)
11. Elgammal, A.M., Lee, C.S.: Tracking People on a Torus. IEEE Transaction on Pattern Analysis and Machine Intelligence 31, 520–538 (2009)
12. Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for people tracking from small training sets. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 403–410 (2005)
13. Loy, G., Eriksson, M., Sullivan, J., Carlsson, S.: Monocular 3D reconstruction of human motion in long action sequences. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 442–455. Springer, Heidelberg (2004)
14. Sminchisescu, C., Triggs, B.: Kinematic Jump Processes for Monocular 3D Human Tracking. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 69–76 (2003)
15. Ziegler, J., Nickel, K., Stiefelhagen, R.: Tracking of the articulated upper body on multi-view stereo image sequences. In: CVPR 2006: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 774–781. IEEE Computer Society, Washington, DC, USA (2006)
16. Stenger, B., Mendonca, P.R.S., Cipolla, R.: Model-based hand tracking using an unscented kalman filter. In: Proc. British Machine Vision Conference, vol. I, pp. 63–72 (2001)
17. Cappé, O., Godsill, S., Moulines, E.: An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE 95, 899–924 (2007)
18. Balan, A.O., Sigal, L., Black, M.J.: A Quantitative Evaluation of Video-based 3D Person Tracking. In: Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 349-356. IEEE Computer Society, Los Alamitos (2005)
19. Bandouch, J., Engstler, F., Beetz, M.: Accurate human motion capture using an ergonomics-based anthropometric human model. In: Proc. of the 5th Int. Conf. on Articulated Motion and Deformable Objects, pp. 248–258. Springer, Heidelberg (2008)