Data-Driven Importance Distributions for Articulated Tracking

Søren Hauberg and Kim Steenstrup Pedersen

{hauberg, kimstp}@diku.dk,
The eScience Centre, Dept. of Computer Science, University of Copenhagen

Abstract. We present two data-driven importance distributions for particle filter-based articulated tracking; one based on background subtraction, another on depth information. In order to keep the algorithms efficient, we represent human poses in terms of spatial joint positions. To ensure constant bone lengths, the joint positions are confined to a non-linear representation manifold embedded in a high-dimensional Euclidean space. We define the importance distributions in the embedding space and project them onto the representation manifold. The resulting importance distributions are used in a particle filter, where they improve both accuracy and efficiency of the tracker. In fact, they triple the effective number of samples compared to the most commonly used importance distribution at little extra computational cost.

Key words: Articulated tracking · Importance Distributions · Particle Filtering · Spatial Human Motion Models

1 Motivation

Articulated tracking is the process of estimating the pose of a person in each frame in an image sequence [1]. Often this is expressed in a Bayesian framework and subsequently the poses are inferred using a particle filter [1–11]. Such filters generate a set of sample hypotheses and assign them weights according to the likelihood of the observed data given the hypothesis is correct. Usually, the hypotheses are sampled directly from the motion prior as this vastly simplifies development. However, as the motion prior is inherently independent of the observed data, samples are generated completely oblivious to the current observation. This has the practical consequence that many sampled pose hypotheses are far away from the modes of the likelihood. This means that many samples are needed for accurate results. As the likelihood has to be evaluated for each of these samples, the resulting filter becomes computationally demanding.

One solution, is to sample hypotheses from a distribution that is not "blind" to the current observation. The particle filter allows for such *importance distributions*. While the design of good importance distributions can be the deciding point of a filter, not much attention has been given to their development in articulated tracking. The root of the problem is that the pose parameters are related to the observation in a highly non-linear fashion, which makes good importance distributions hard to design. In this paper, we change the pose parametrisation and then suggest a simple approximation that allows us to design highly efficient importance distributions that account for the current observation.

1.1 Articulated Tracking using Particle Filters

Estimating the pose of a person using a single view point or a small baseline stereo camera is an inherently difficult problem due to self-occlusions and visual ambiguities. This manifests itself in that the distribution of the human pose is multi-modal with an unknown number of modes. Currently, the best method for coping with such distributions is the particle filter [12]. This relies on a prior motion model $p(\theta_t|\theta_{t-1})$ and a data likelihood model $p(\mathbf{Z}_t|\theta_t)$. Here θ_t denotes the human pose at time t and $\mathbf{Z}_t$ the observation at the same time. The particle filter approximates the posterior $p(\theta_t|\mathbf{Z}_{1:t})$ as a set of weighted samples. These samples are drawn from an *importance distribution* $q(\theta_t|\mathbf{Z}_t, \theta_{t-1})$ and the weights are computed recursively as

$$w_t^{(n)} \propto w_{t-1}^{(n)} p\left(\mathbf{Z}_t|\theta_t^{(n)}\right) r_t^{(n)} \qquad \text{s.t.} \qquad \sum_{n=1}^{N} w_t^{(n)} = 1 \ , \tag{1}$$

where the superscript (n) denotes sample index and the *correction factor* $r_t^{(n)}$ is given by

$$r_t^{(n)} = \frac{p\left(\theta_t^{(n)}|\theta_{t-1}^{(n)}\right)}{q\left(\theta_t^{(n)}|\mathbf{Z}_t, \theta_{t-1}^{(n)}\right)} \ . \tag{2}$$

In practice, it is common use the motion prior as the importance distribution, i.e. to let $q(\theta_t|\mathbf{Z}_t, \theta_{t-1}) = p(\theta_t|\theta_{t-1})$ as then $r_t^{(n)} = 1$ which simplifies development. This does, however, have the unwanted side-effect that the importance distribution is "blind" to the current observation, such that the samples can easily be placed far away from the modes of the likelihood (and hence the modes of the posterior). In practice, this increases the number of samples needed for successful tracking. As the likelihood has to be evaluated for each sample, this quickly becomes a costly affair; in general the likelihood is expensive to evaluate as it has to traverse the data.

To use the particle filter for articulated tracking, we need a human pose representation. As is common [1], we shall use the kinematic skeleton (see fig. 1). This representation is a collection of connected rigid bones organised in a tree structure. Each bone can be rotated at the point of connection between the bone and its parent. We model the bones as having known constant length, so the angles between connected bones constitutes the only degrees of freedom in the kinematic skeleton. We collect these into one large vector θ_t representing all joint angles in the model at time t. To represent constraints on the joint angles, they are confined to a subset $\mathbf{\Theta}$ of $\mathbb{R}^N$.

From known bone lengths and a joint angle vector θ_t, the joint positions can be computed recursively using *forward kinematics* [13]. We will let $F(\theta_t)$ denote the joint positions corresponding to the joint angles θ_t. In this paper, we will make a distinction between joint *angles* and joint *positions* as this has profound impact when designing data-driven importance distributions.

1.2 Related Work

In articulated tracking, much work has gone into improving either the likelihood model or the motion prior. Likelihood models usually depend on cues such as *edges* [2–4],

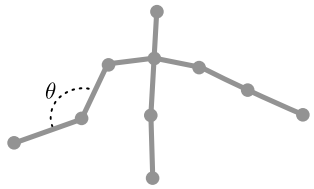


Fig. 1. An illustration of the *kinematic skeleton*. Bones are connected in a tree structure where branches have constant length. Angles between connected bones constitute the only degrees of freedom in the model.

optical flow [4, 11] or *background subtraction* [3, 5, 14–18]. Motion priors are usually crafted by learning activity specific priors, such as for *walking* [6, 7, 19, 20]. These approaches work by restricting the tracker to some subspace of the joint angle space, which makes the priors activity specific. When no knowledge of the activity is available it is common [5, 6, 18, 21] to simply let θ_t follow a normal distribution with a diagonal covariance, i.e.

$$p_{\text{gp}}(\theta_t|\theta_{t-1}) \propto \mathcal{N}(\theta_t|\theta_{t-1}, \text{diag}) \ \mathcal{U}_{\boldsymbol{\Theta}}(\theta_t) \ , \tag{3}$$

where $\mathcal{U}_{\boldsymbol{\Theta}}$ is a uniform distribution on the legal set of angles that encodes the joint constraints. Recently, Hauberg et al. [8] showed that this model causes the spatial variance of the joint positions to increase as the kinematic chains are traversed. In practice this means that with this model the spatial variance of e.g. the hands is always larger than of the shoulders. To avoid this somewhat arbitrary behaviour it was suggested to build the prior distribution directly in the spatial domain; a solution we will review in sec. 3.

In this paper we design data-driven importance distributions; a sub-field of articulated tracking where little work has been done. One notable exception is the work of Poon and Fleet [9], where a hybrid Monte Carlo filter was suggested. In this filter, the importance distribution uses the gradient of the log-likelihood, which moves the samples closer to the modes of the likelihood function (and, hence, also closer to the modes of the posterior). This approach is reported to improve the overall system performance.

In the more general filtering literature, the *optimal particle filter* [12] is known to vastly improve the performance of particle filters. This filter incorporates the observation in the importance distribution, such that samples are drawn from $p(\theta_t|\theta_{t-1}, \mathbf{Z}_t)$, where $\mathbf{Z}_t$ denotes the observation at time t. In practice, the optimal particle filter is quite difficult to implement as non-trivial integrals need to be solved in closed-form. Thus, solutions are only available for non-linear extensions to the Kalman filter [12] and for non-linear extensions of left-to-right Hidden Markov models with known expected state durations [22].

2 A Failed Experiment

Our approach is motivated by a simple experiment, which proved to be a failure. In an effort to design data-driven importance distributions, we designed a straight-forward importance distribution based on silhouette observations. We, thus, assume we have a

binary image $\mathbf{B}_t$ available, which roughly separates the human from the scene. When sampling new poses, we will ensure that joint positions are within the human segment. We model the motion prior according to eq. 3, i.e. assume that joint angles follow a normal distribution with diagonal covariance.

Let $\mathcal{U}_{\mathbf{B}_t}$ denote the uniform distribution on the binary image $\mathbf{B}_t$, such that background pixels have zero probability and let $\text{proj}_{im}[F(\theta_t)]$ be the projection of joint positions $F(\theta_t)$ onto the image plane. We then define the importance distribution as

$$\tilde{q}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \propto \mathcal{N}(\theta_t|\theta_{t-1}, \text{diag})\ \mathcal{U}_{\boldsymbol{\Theta}}(\theta_t)\ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])\ . \tag{4}$$

The two first terms correspond to the motion prior and the third term ensures that sampled joint positions are within the human segment in the silhouette image. It is worth noticing that the correction factor $r_t^{(n)}$ (eq. 2) becomes constant for this importance distribution and hence can be ignored.

It is straight-forward to sample from this importance distribution using *rejection sampling* [23]: new samples can be drawn from the motion prior until one is found where all joint positions are within the human segment. This simple scheme, which is illustrated in fig. 2, should improve tracking quality. To measure this, we develop one articulated tracker where the motion prior (eq. 3) is used as importance distribution and one where eq. 4 is used. We use a likelihood model and measure of tracking error described later in the paper; for now details are not relevant. Fig. 3a and 3b shows the tracking error as well as the running time for the two systems as a function of the number of samples in the filter. As can be seen, the data-driven importance distribution *increases* the tracking error with approximately one centimetre, while roughly requiring 10 times as many computations. An utter failure!

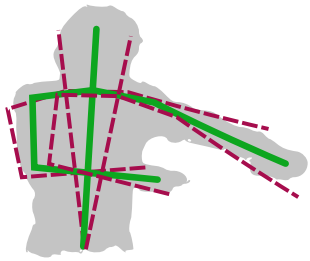


Fig. 2. An illustration of the rejection sampling scheme for simulating the importance distribution in eq. 4. The green skeleton drawn in full lines is accepted, while the two red dashed skeletons are rejected as at least one joint is outside the silhouette.

To get to the root of this failure, we need to look at the motion prior. As previously mentioned, Hauberg et al. [8] have pointed out that the spatial variance of the joint positions increases as the kinematic chains are traversed. This means that e.g. hand positions are always more variant than shoulder positions. In practice, this leads to rather large spatial variances of joint positions. This makes the term $\mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])$ dominant

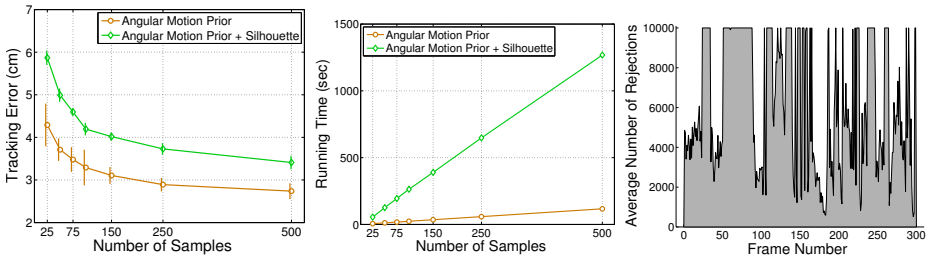


Fig. 3. Various performance measures for the tracking systems; errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error measured in centimetre. (b) The running time per frame. (c) The average number of rejections in each frame.

in eq. 4, thereby diminishing the effect of the motion prior. This explains the increased tracking error. The large running time can also be explained by the large spatial variance of the motion prior. For a sampled pose to be accepted in the rejection sampling scheme, *all* joint positions need to be inside the human silhouette. Due to the large spatial variance of the motion prior, many samples will be rejected, leading to large computational demands. To keep the running time under control, we maximally allow for 10000 rejections. Fig. 3c shows the average number of rejections in each frame in a sequence; on average 6162 rejections are required to generate a sample where all joint positions are within the human silhouette. Thus, the poor performance, both in terms accuracy and speed, of the importance distribution in eq. 4 is due to the large spatial variance of the motion prior. This indicates that we should be looking for motion priors with more well-behaved spatial variance. We will turn to the framework suggested by Hauberg et al. [8] as it was specifically designed for controlling the spatial variance of joint positions. We shall briefly review this work next.

3 Spatial Predictions

To design motion priors with easily controlled spatial variance, Hauberg et al. [8] first define a spatial pose representation manifold $\mathcal{M} \subset \mathbb{R}^{3L}$, where L denotes the number of joints. A point on this manifold corresponds to all spatial joint positions of a pose parametrised by a set of joint angles. More stringently, $\mathcal{M}$ can be defined as

$$\mathcal{M} = \{F(\theta) \mid \theta \in \mathbf{\Theta}\} \ , \tag{5}$$

where F denotes the forward kinematics function for the entire skeleton. As this function is injective with a full-rank Jacobian, $\mathcal{M}$ is a compact differentiable manifold embedded in $\mathbb{R}^{3L}$. Alternatively, one can think of $\mathcal{M}$ as a quadratic constraint manifold arising due to the constant distance between connected joints. It should be noted that while a point on $\mathcal{M}$ corresponds to a point in $\mathbf{\Theta}$, the metrics on the two spaces are different, giving rise to different behaviours of seemingly similar distributions.

A Gaussian-like predictive distribution on $\mathcal{M}$ can be defined simply by projecting a Gaussian distribution in $\mathbb{R}^{3L}$ onto $\mathcal{M}$, i.e.

$$p_{\text{proj}}(\theta_t|\theta_{t-1}) = \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)\right] \quad . \tag{6}$$

When using a particle filter for tracking, one only needs to be able to draw samples from the prior model. This can easily be done by sampling from the normal distribution in $\mathbb{R}^{3L}$ and projecting the result onto $\mathcal{M}$. This projection can be performed in a direct manner by seeking

$$\hat{\theta}_t = \arg\min_{\theta_t} \left\|\hat{\mathbf{x}}_t - F(\theta_t)\right\|^2 \qquad \text{s.t.} \qquad \theta_t \in \boldsymbol{\Theta} \ , \tag{7}$$

where $\hat{\mathbf{x}}_t \sim \mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)$. This is an *inverse kinematics* problem [13], where all joints are assigned a goal. Eq. 7 can efficiently be solved using gradient descent by starting the search in θ_{t-1}.

4 Data-Driven Importance Distributions

We now have the necessary ingredients for designing data-driven importance distributions. In this paper, we will be designing two such distributions: one based on silhouette data and another on depth data from a stereo camera. Both will follow the same basic strategy.

4.1 An Importance Distribution based on Silhouettes

Many articulated tracking systems base their likelihood models on simple background subtractions [3, 5, 14–18]. As such, importance distributions based on silhouette data are good candidates for improving many systems. We, thus, assume that we have a binary image $\mathbf{B}_t$ available, which roughly separates the human from the scene. When predicting new joint positions, we will ensure that they are within the human segment.

The projected prior (eq. 6) provides us with a motion model where the variance of joint positions can easily be controlled. We can then create an importance distribution similar to eq. 4,

$$q_{\text{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) \ . \tag{8}$$

While the more well-behaved spatial variance of this approach would improve upon the previous experiment, it would still leave us with a high dimensional rejection sampling problem. As this has great impact on performance, we suggest an approximation of the above importance distribution,

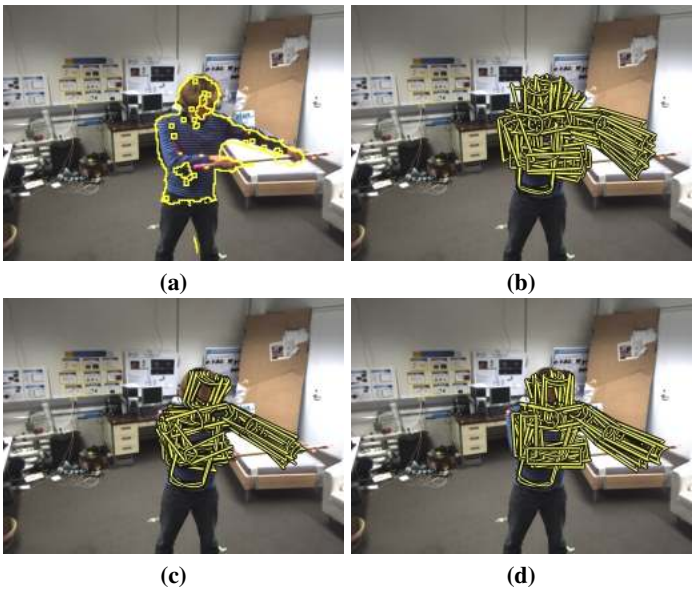
$$\begin{aligned} q_{\text{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) &\propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) && (9)\\ &= \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma)\right] \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) && (10)\\ &\approx \text{proj}_{\mathcal{M}}\left[\mathcal{N}(F(\theta_t)|F(\theta_{t-1}), \Sigma) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)])\right] \ . && (11) \end{aligned}$$

In other words, we suggest imposing the data-driven restriction in the embedding space before projecting back on manifold. When the covariance Σ is block-diagonal, such that the position of different joints in embedding space are independent, this importance distribution can be written as

$$q_{\mathrm{bg}}(\theta_t|\mathbf{B}_t, \theta_{t-1}) \approx \mathrm{proj}_{\mathcal{M}} \left[\prod_{l=1}^{L} \mathcal{N}(\mu_{l,t}|\mu_{l,t-1}, \Sigma_l) \; \mathcal{U}_{\mathbf{B}_t}(\mathrm{proj}_{im}[\mu_{l,t}]) \right] \quad , \qquad (12)$$

where $\mu_{l,t}$ denotes the position of the l^{th} joint at time t and Σ_l denotes the block of Σ corresponding to the l^{th} joint. We can sample efficiently from this distribution using rejection sampling by sampling each joint position independently and ensuring that they are within the human silhouette. This is L three dimensional rejection sampling problems, which can be solved much more efficiently than one $3L$ dimensional problem. After the joint positions are sampled, they can be projected onto the representation manifold $\mathcal{M}$, such that the sampled pose respects the skeleton structure.

A few samples from this distribution can be seen in fig. 4c, where samples from the angular prior from eq. 3 is available as well for comparative purposes. As can be seen, the samples from the silhouette-driven importance distribution are much more aligned with the true pose, which is the general trend.



(a) **(b)** **(c)** **(d)**

Fig. 4. Samples from various importance distributions. Notice how the data-driven distributions generate more "focused" samples. (a) The input data with the segmentation superimposed. (b) Samples from the angular prior (eq. 3). (c) Samples from the importance distribution guided by silhouette data. (d) Samples from the importance distribution guided by depth information.

4.2 An Importance Distribution based on Depth

Several authors have also used depth information as the basis of their likelihood model. Some have used stereo [8, 10, 24] and others have used time-of-flight cameras [25]. When depth information is available it is often fairly easy to segment the data into background and foreground simply by thresholding the depth. As such, we will extend the previous model with the depth information. From depth information we can generate a set of points $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ corresponding to the surface of the observed objects. When sampling a joint position, we will simply ensure that it is not too far away from any of the points in $\mathbf{Z}$.

To formalise this idea, we first note that the observed surface corresponds to the skin of the human, whereas we are modelling the skeleton. Hence, the joint positions should not be directly *on* the surface, but a bit away depending on the joint. For instance, hand joints should be closer to the surface than a joint on the spine. To encode this knowledge, we let $\mathbf{Z}^{\oplus r_l}$ denote the set of three dimensional points where the shortest distance to any point in $\mathbf{Z}$ is less than r_l, i.e.

$$\mathbf{Z}^{\oplus r_l} = \{\mathbf{z} \mid \min_k(\|\mathbf{z} - \mathbf{z}_k\|) < r_l\} \ . \tag{13}$$

Here the r_l threshold is set to be small for hands, large for joints on the spine and so forth. When we sample individual joint positions, we ensure they are within this set, i.e.

$$\begin{aligned} q_{\text{depth}}(\theta_t|\mathbf{Z}, \theta_{t-1}) &\propto p_{\text{proj}}(\theta_t|\theta_{t-1}) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[F(\theta_t)]) \ \mathcal{U}_{\mathbf{Z}^{\oplus}}\big(F(\theta_t\big) \\ &\approx \text{proj}_{\mathcal{M}}\left[\prod_{l=1}^{L} \mathcal{N}(\mu_{l,t}|\mu_{l,t-1}, \Sigma_l) \ \mathcal{U}_{\mathbf{B}_t}(\text{proj}_{im}[\mu_{l,t}]) \ \mathcal{U}_{\mathbf{Z}^{\oplus r_l}}(\mu_{l,t})\right] \end{aligned} \tag{14}$$

where $\mathcal{U}_{\mathbf{Z}^{\oplus r_l}}$ is the uniform distribution on $\mathbf{Z}^{\oplus r_l}$. Again, we can sample from this distribution using rejection sampling. This requires us to compute the distance from the predicted position to the nearest point in depth data. We can find this very efficiently using techniques from kNN classifiers, such as k-d trees [26].

Once all joint positions have been sampled, they are collectively projected onto the manifold $\mathcal{M}$ of possible poses. A few samples from this distribution is shown in fig. 4d. As can be seen, the results are visually comparable to the model based on background subtraction; we shall later, unsurprisingly, see that for out-of-plane motions the depth model does outperform the one based on background subtraction.

5 A Simple Likelihood Model

In order to complete the tracking system, we need a system for computing the likelihood of the observed data. To keep the paper focused on prediction, we use a simple vision system [8] based on a consumer stereo camera[1]. This camera provides a dense set of three dimensional points $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ in each frame. The objective of the vision system then becomes to measure how well a pose hypothesis matches the points. We

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

assume that points are independent and that the distance between a point and the skin of the human follows a zero-mean Gaussian distribution, i.e.

$$p(\mathbf{Z}|\theta_t) \propto \prod_{k=1}^{K} \exp\left(-\frac{\min\left[D^2(\theta_t, \mathbf{z}_k),\ \tau\right]}{2\sigma^2}\right) \ , \tag{15}$$

where $D^2(\theta_t, \mathbf{z}_k)$ denotes the squared distance between the point $\mathbf{z}_k$ and the skin of the pose θ_t and τ is a constant threshold. The minimum operation is there to make the system robust with respect to outliers.

We also need to define the skin of a pose, such that we can compute distances between this and a data point. Here, the skin of a bone is defined as a capsule with main axis corresponding to the bone itself. Since we only have a single view point, we discard the half of the capsule that is not visible. The skin of the entire pose is then defined as the union of these half-capsules. The distance between a point and this skin can then be computed as the smallest distance from the point to any of the half-capsules.

6 Experimental Results

We now have two efficient data-driven importance distributions and a likelihood model. This gives us two systems for articulated tracking that we now validate by comparison with one using the standard activity independent prior that assumes normally distributed joint angles (eq. 3) as importance distribution. We use this motion prior as reference as it is the most commonly used model. As ground truth we will be using data acquired with an optical marker-based motion capture system.

We first illustrate the different priors on a sequence where a person is standing in place while waving a stick. This motion utilises the shoulders a lot; something that often causes problems for articulated trackers. As the person is standing in place, we only track the upper body motions.

In fig. 5 we show attained results for the different importance distributions; a film with the results are available as part of the supplementary material. Visually, we see that the data-driven distributions improve the attained results substantially. Next, we set out to measure this gain.

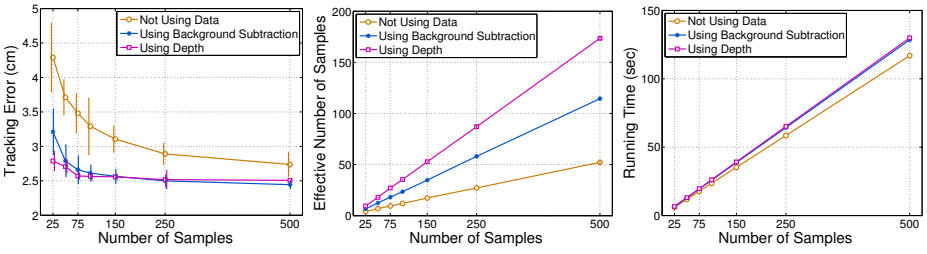
To evaluate the quality of the attained results we position motion capture markers on the arms of the test subject. We then measure the average distance between the motion capture markers and the capsule skin of the attained results. This measure is then averaged across frames, such that the error measure becomes

$$\mathcal{E} = \frac{1}{TM} \sum_{t=1}^{T} \sum_{m=1}^{M} D(\hat{\theta}_t, \mathbf{v}_m) \ , \tag{16}$$

where $D(\hat{\theta}_t, \mathbf{v}_m)$ is the orthogonal Euclidean distance between the m^{th} motion capture marker and the skin at time t. The error measure is shown in fig. 6a using between 25 and 500 particles. As can be seen, both data-driven importance distributions perform substantially better than the model not utilising the data. For a small number of samples,

(a) **(b)** **(c)**

Fig. 5. Results from trackers using 150 particles with the different importance distributions. The general trend is that the data-driven distributions improve the results. (a) The angular prior from eq. 3. (b) The importance distribution guided by background subtraction. (c) The importance distribution guided by depth information.



(a) **(b)** **(c)**

Fig. 6. Various performance measures for the tracking systems using different importance distributions on the first sequence. Errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error $\mathcal{E}$. (b) The effective number of samples N_{eff}. (c) The running time per frame.

the model based on depth outperforms the one based on background subtraction, but for 150 particles and more, the two models perform similarly.

In the particle filtering literature the quality of the Monte Carlo approximation is sometimes measured by computing the *effective number of samples* [12]. This measure can be approximated by

$$N_{eff} = \left(\sum_{n=1}^{N} w_t^{(n)} \right)^{-1} , \tag{17}$$

where $w_t^{(n)}$ denotes the weight of the n^{th} sample in the particle filter. Most often this measure is used to determine when resampling should be performed; here we will use it to compare the different importance distributions. We compute the effective number of samples in each frame and compute the temporal average. This provides us with a measure of how many of the samples are actually contributing to the filter. In fig. 6b we show this for the different importance distributions as a function of the number of particles. As can be seen, the data-driven importance distributions gives rise to more effective samples than the one not using the data. The importance distribution based on background subtraction gives between 1.6 and 2.2 times more effective samples than the model not using data, while the model using depth gives between 2.3 and 3.3 times more effective samples.

We have seen that the data-driven importance distributions improve the tracking substantially as they increase the effective number of samples. This benefit, however, comes at the cost of an increased running time. An obvious question is then whether this extra cost outweigh the gains. To answer this, we plot the running times per frame for the tracker using the different distributions in fig. 6c. As can be seen, the two data-driven models require the same amount of computational resources; both requiring approximately 10% more resources than the importance distribution not using the data. In other words, we can triple the effective number of samples at 10% extra cost.

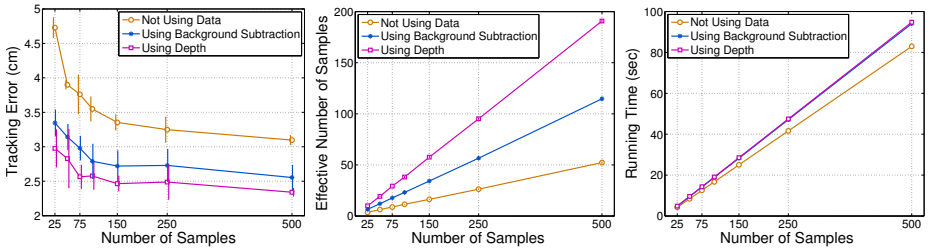
We repeat the above experiments for a different sequence, where a person is moving his arms in a quite arbitrary fashion; a type of motion that is hard to predict and as such also hard to track. Example results are shown in fig. 7, with a film again being available as part of the supplementary material. Once more, we see that the data-driven importance distributions improve results. The tracking error is shown in fig. 8a; we see that the importance distribution based on depth consistently outperforms the one based on background subtraction, which, in turn, outperforms the one not using the data. The effective number of samples is shown in fig. 8b. The importance distribution based on background subtraction gives between 1.8 and 2.2 times more effective samples than the model not using data, while the model using depth gives between 2.8 and 3.6 times more effective samples. Again a substantial improvement at little extra cost.

7 Conclusion

We have suggested two efficient importance distributions for use in articulated tracking systems based on particle filters. They gain their efficiency by an approximation that allows us to sample joint positions independently. A valid pose is then constructed by a projection onto the manifold $\mathcal{M}$ of possible joint positions. While this projection might

(a) **(b)** **(c)**

Fig. 7. Results from trackers using 150 particles with the different importance distributions. The general trend is that the data-driven distributions improve the results. (a) The angular prior from eq. 3. (b) The importance distribution guided by silhouette data. (c) The importance distribution guided by depth information.



(a) **(b)** **(c)**

Fig. 8. Various performance measures for the tracking systems using different importance distributions on the second sequence. Errorbars denote one standard deviation of the attained results over several trials. (a) The tracking error $\mathcal{E}$. (b) The effective number of samples N_{eff}. (c) The running time per frame.

seem complicated it merely correspond to a least-squares fit of a kinematic skeleton to the sampled joint positions. As such, the suggested importance distributions are quite simple, which consequently means that the algorithms are efficient and that they actually work. In fact, our importance distributions triple the effective number of samples in the particle filter, at little extra computational cost. The simplicity of the suggested distributions also makes them quite general and easy to implement. Hence, they can be used to improve many existing tracking systems with little effort.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding **108** (2007) 4–18
2. Sminchisescu, C., Triggs, B.: Kinematic Jump Processes for Monocular 3D Human Tracking. In: In IEEE International Conference on Computer Vision and Pattern Recognition. (2003) 69–76
3. Duetscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: CVPR, Published by the IEEE Computer Society (2000) 2126
4. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. The International Journal of Robotics Research **22** (2003) 371
5. Kjellström, H., Kragić, D., Black, M.J.: Tracking people interacting with objects. In: IEEE CVPR. (2010)
6. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: ECCV. Volume II of LNCS 1843., Springer (2000) 702–718
7. Urtasun, R., Fleet, D.J., Fua, P.: 3D People Tracking with Gaussian Process Dynamical Models. In: IEEE CVPR. (2006) 238–245
8. Hauberg, S., Sommer, S., Pedersen, K.S.: Gaussian-like spatial priors for articulated tracking. In Daniilidis, K., Maragos, P., , Paragios, N., eds.: ECCV. Volume 6311 of LNCS., Springer (2010) 425–437
9. Poon, E., Fleet, D.J.: Hybrid monte carlo filtering: Edge-based people tracking. IEEE Workshop on Motion and Video Computing **0** (2002) 151
10. Hauberg, S., Pedersen, K.S.: Stick it! articulated tracking using spatial rigid object priors. In: ACCV 2010, Springer-Verlag (2010)
11. Brubaker, M.A., Fleet, D.J., Hertzmann, A.: Physics-based person tracking using the anthropomorphic walker. International Journal of Computer Vision **87** (2010) 140–155
12. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. Statistics and computing **10** (2000) 197–208
13. Erleben, K., Sporring, J., Henriksen, K., Dohlmann, H.: Physics Based Animation. Charles River Media (2005)
14. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: IEEE CVPR. (2007) 1–8
15. Vondrak, M., Sigal, L., Jenkins, O.C.: Physical simulation for probabilistic motion tracking. In: CVPR, IEEE (2008)
16. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. International Journal of Computer Vision **87** (2010) 75–92
17. Bandouch, J., Beetz, M.: Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models. In: Computer Vision Workshops (ICCV Workshops). (2009)

18. Balan, A.O., Sigal, L., Black, M.J.: A quantitative evaluation of video-based 3d person tracking. Visual Surveillance and Performance Evaluation of Tracking and Surveillance **0** (2005) 349–356
19. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian Process Dynamical Models for Human Motion. IEEE PAMI **30** (2008) 283–298
20. 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 (2008) 1705–1712
21. Bandouch, J., Engstler, F., Beetz, M.: Accurate human motion capture using an ergonomics-based anthropometric human model. In: AMDO '08: Proceedings of the 5th international conference on Articulated Motion and Deformable Objects, Springer-Verlag (2008) 248–258
22. Hauberg, S., Sloth, J.: An efficient algorithm for modelling duration in hidden markov models, with a dramatic application. J. Math. Imaging Vis. **31** (2008) 165–170
23. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
24. Ziegler, J., Nickel, K., Stiefelhagen, R.: Tracking of the articulated upper body on multi-view stereo image sequences. In: IEEE CVPR. (2006) 774–781
25. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: IEEE CVPR. (2010) 755–762
26. Arya, S., Mount, D.M.: Approximate nearest neighbor queries in fixed dimensions. In: Proc. 4th ACM-SIAM Sympos. Discrete Algorithms. (1993) 271280