Tracking 3D Human Pose with Large Root Node Uncertainty

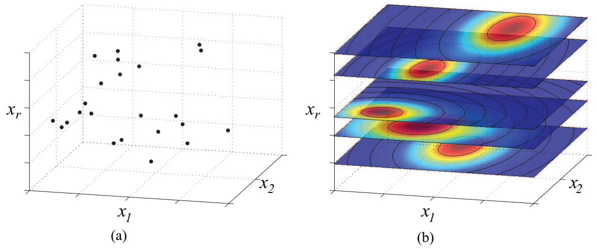
Ben Daubney and Xianghua Xie
Department of Computer Science, Swansea University
United Kingdom, SA2 8PP
{B.Daubney, X.Xie}@swansea.ac.uk

Abstract

Representing articulated objects as a graphical model has gained much popularity in recent years, often the root node of the graph describes the global position and orientation of the object. In this work a method is presented to robustly track 3D human pose by permitting greater uncertainty to be modeled over the root node than existing techniques allow. Significantly, this is achieved without increasing the uncertainty of remaining parts of the model. The benefit is that a greater volume of the posterior can be supported making the approach less vulnerable to tracking failure. Given a hypothesis of the root node state a novel method is presented to estimate the posterior over the remaining parts of the body conditioned on this value. All probability distributions are approximated using a single Gaussian allowing inference to be carried out in closed form. A set of deterministically selected sample points are used that allow the posterior to be updated for each part requiring just seven image likelihood evaluations making it extremely efficient. Multiple root node states are supported and propagated using standard sampling techniques. We believe this to be the first work devoted to efficient tracking of human pose whilst modeling large uncertainty in the root node and demonstrate the presented method to be more robust to tracking failures than existing approaches.

1. Introduction

There have been many methods proposed to estimate and track 3D human pose from a sequence of images. This is a particularly difficult task as it represents a high-dimensional problem, the consequence of which is that modeling the posterior of the pose space and searching within it is extremely challenging. Currently, the most popular solution to this problem is to find the most likely pose and propagate this into the following time instance using a temporal model [3], this then serves as a prediction or prior over the subject's configuration. The problem with this approach is that if the most likely pose found is incorrect, tracking errors will start to accumulate often leading to failure.



(a)
(b)

Figure 1. Representing the posterior using a set of samples. Whilst in standard approaches a sample typically represents the posterior at a single location (a), under the proposed method for each sample a hyperplane of the posterior is estimated conditioned on the root node state, x_r (b).

Whilst many methods have been proposed to improve the chances of finding the global maximum of the posterior [3, 6, 7, 2], a significant problem is that often the global maximum does not correspond to the correct pose for a given image. The observational likelihoods typically used, such as binary silhouettes, are weak and often ambiguous. As a result an active area of research is in learning much stronger part models [1, 8]. Whilst this will lead to improved pose estimation it is unlikely to exclusively solve this problem. Noisy observations will always be present in image data preventing the global maximum in the posterior from corresponding to the correct pose. An alternative solution to this problem is to develop methods that are capable of representing and propagating greater uncertainty in the pose space. This will ensure that tracking an incorrect mode will not result in catastrophic failure since a much wider area of the posterior will be supported. This is the focus of the paper.

There are two principal parts to this problem: The first is how to efficiently represent a greater portion of the posterior distribution. The second is how to update this at each time frame without a greater computational burden than existing methods.

Current methods to efficiently estimate pose often decompose the human body into parts and represent it as a probabilistic graph. Hidden nodes in the graph represent

This work was funded by RIVIC.

the state of each part and connections between these nodes represent prior distributions learnt over these. Pose estimation can then be performed using a Bayesian methodology, hidden nodes are treated as nuisance parameters and marginalized over. Techniques that employ this approach include Non-Parametric Belief Propagation [7, 15], Variational MAP [6] and Partitioned Sampling [2]. Often the root node of the graph describes the global position and orientation of the torso. However, we observe that in the domain of 3D pose estimation a small change in the root node (e.g. orientation) can make a large change in the posterior of the remaining parts. We believe that when there is large uncertainty in the root node marginalizing over it produces severe blurring of the posterior resulting in poor pose estimation. To prevent this rather than integrating over the root node, we support many hypothesis of its state and for each of them update the likelihood of the remaining nodes conditioned on it. For each root node hypothesis a set of Gaussian distributions are used to model the posterior distribution for all remaining parts. Effectively this allows a hyperplane of the posterior to be estimated for each root node hypothesis rather than, for example a particle filter, where a sample only measures the posterior at a single point. This is illustrated in Figure 1.

The advantage of this approach over others (e.g. [3, 2, 7, 6]) which are typically converged to a single maximum is that it allows greater uncertainty to be represented in the root node without increasing the uncertainty in the remaining nodes of the model. Whilst the aforementioned approaches could simply be iterated fewer times it has been observed that for articulated models the root node must first converge before the remaining nodes are able to do so [2], therefore allowing greater uncertainty in this using these approaches will greatly increase the uncertainty in the remaining parts of the model. Covariance Scaled Sampling [14] modeled the posterior with Gaussian distributions and searched the pose space along axis with the greatest uncertainty. However, only a few modes were supported and a Gaussian was estimated across the entire pose space meaning the observations were most likely very sparse. In contrast our approach efficiently supports of the order of a hundred different modes and a Gaussian is used to represent the posterior for each individual limb conditioned on a given root node hypothesis.

Existing hierarchical methods (e.g. [11, 10]) assume that some parts can be better localized than others and attempt to exploit this structure. However, the limitation with these approaches is that if there is uncertainty in these parts the methods perform poorly. Whilst some of these approaches could be employed in our framework by simply executing them for multiple hypothesis the result would be computationally expensive. The emphasis in this work is to model a much wider volume of the posterior with little additional computation. To this end we test our method against the Sequence Importance Resampling Particle Filter (SIR-PF) and the Annealed Particle Filter (APF) using the equivalent number of image likelihood evaluations as our method. Under our scheme to update the posterior given a root node hypothesis requires the equivalent image likelihood evaluations as just 7 particles making it extremely efficient.

As the probability density function for each node is approximated by a Gaussian distribution, inference can be performed deterministically. This approach shares many similarities to the Rao-Blackwellised Particle Filter (RBPF) [4] where nodes are partitioned into two sets; root nodes and leaf nodes. Root nodes are propagated stochastically and the leaf nodes updated conditioned on these. In this work the distribution over the root node is also propagated stochastically. However, not all the other nodes in the graph are directly connected to it. The RBPF was recently used by Xu and Li to track 3D pose [16], they partitioned opposing sides of the body into root (left) and leaf (right) nodes. Given an estimate of the states of the left side of the body, motion correlation models were used to integrate over the states of the opposing side. Using our method the graph structure commonly used by existing approaches is maintained [5, 13]. The body is modeled as a tree with the root node located at the pelvis and the branches of the tree represent different limbs.

In this work we propose a method to track 3D human pose captured from multiple cameras. No high-level motion models are used to improve tracking and the method is tested against two standard approaches: Firstly, the annealed particle filter is used to show that representing only a single mode results in a technique that often falls into the wrong maxima causing tracking failure. Secondly, the SIR particle filter is employed to show that when used to model a larger uncertainty, this uncertainty is extended to all parts of the model. This is in comparison to the presented approach that is shown to be able to represent large uncertainty in the root node, making it more robust to tracking failure, without inflating the uncertainty of the remaining parts of the model. For each technique the same experimental parameters are used as are the same temporal diffusion models, though in the presented approach temporal uncertainty is propagated deterministically compared to the SIR-PF and APF where they are propagated stochastically. Experimental results are provided using the HumanEva dataset [12]. We believe this to be the first work devoted to efficient tracking of human pose whilst modeling large uncertainty in the root node and suggest it to be an important topic of research to improve tracking techniques.

2. Approach Overview & Paper Organization

To perform efficient tracking the body is decomposed into its constituent parts which allows it to be represented over a probabilistic graph. The nodes are partitioned into the root node, representing the global position and orientation of the body, and the remaining nodes representing the orientation of each part. This is defined more formally in Section 3.1. The state of each node, excluding the root node, is represented as a quaternion rotation. Learning a distribution over quaternions is difficult, however, in Section 3.2 an approximation, similar to that presented in [13], is described to achieve this. Often a distribution learnt over quaternion space must be propagated though a rotation, this will be necessary, for example, to propagate uncertainty between neighboring parts. This is performed using the Unscented Transform, which is briefly described in Section 3.3.

The posterior distribution over the root node is represented by a set of samples. For each sample, a set of Gaussians are used to represent the posterior for each part conditioned on the given root node state. The parameters of each distribution are updated in each frame using a set of deterministically selected sample points, which we describe in Section 5. Combining these with limb conditionals, that represent the prior distribution over the configuration between connected parts (Section 4), efficient probabilistic inference can be performed as described in Section 6.

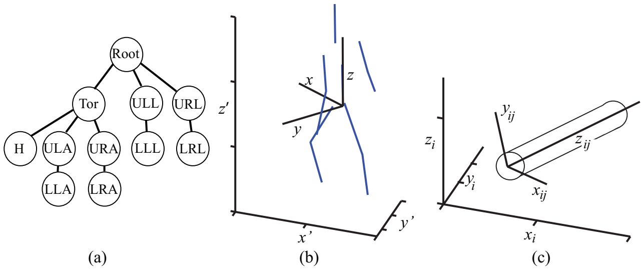
Whilst the posterior distribution over the root node is propagated through time stochastically, the distribution over all other nodes are propagated by inflating the covariances deterministically (Section 7).

Quantitative results, including a comparison between the proposed method, the APF and SIR-PF are provided in Section 8. Finally, conclusions are provided and avenues for further work are discussed in Section 9.

3. Model Representation

3.1. Graphical Model

The body is represented as a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, .., v_n\}$ are the nodes of the graph and $\{v_i, v_j\} \in \mathcal{E}$ represent the edges between them. The state of each of the nodes is defined by $\mathbf{X} = \{x_1, .., x_n\}$. We partition the graph into the root node x_r and all remaining nodes $X = \{x_1, ..x_{n-1}\}$. X represents the individual parts of the body comprising of the head (H), torso (Tor), left upper arm (LUA), left lower arm (LLA), left upper leg (LUL), left lower leg (LLL) and the opposing part for each limb. The structure of the graph is shown in Figure 2 (a). The state of these parts is represented by a quaternion rotation q_i that describes the orientation of each part in the frame of reference of the body, where the base of the torso is the origin, the z-axis is the vertical and y-axis is directed across the shoulders. The root node x_r does not explicitly represent a part, its state represents the position d_r and orientation θ_r of the body in the global frame of reference, i.e. that of the motion capture suite. This allows the transformation from the body to the global frame of reference, i.e. $X' = f(X, x_r)$. This is depicted in Figure 2 (b). Further to this the state can be decomposed into a local frame of reference, where each part is represented as a rotation defined in the frame of reference of the part to which it is connected $X_{ij} = g(X, \mathcal{E})$. This is depicted in Figure 2 (c).



(a) (b) (c)

Figure 2. (a) shows the graphical structure used to represent the body. (b) shows the frame of reference of the body X in that of the motion capture suite X'. (c) shows a part represented in the frame of reference of the part to which it is conected X_{ij}.

The posterior distribution over this graph is represented by a set of samples $\mathcal{S} = \{\mathbf{X}_1, .., \mathbf{X}_m\}$, where each sample is composed of the root node x_r and the posterior distribution of each part conditioned on this $\{p(x_1|\mathcal{O}, x_r), .., p(x_{n-1}|\mathcal{O}, x_r)\}$, where $\mathcal{O} = \{O_1, .., O_{n-1}\}$ represents the set of observations for all parts. The posterior of each part is modeled using a normal distribution $p(x_i|\mathcal{O}, x_r) = \mathcal{N}(x_i; \mu_i, \Sigma_i)$. This allows each particle to be parametized by $\mathbf{X}_l = \{x_r, \mu_1, \Sigma_1, .., \mu_{n-1}, \Sigma_{n-1}\}$.

We assume that the body is constructed of rigid limbs with fixed joint positions. This is different to approaches such as [5, 13] where connections between parts are soft, so called loose limbed models. The parameters of a particle can then be used to construct a rigid body as follows: The location of the proximal and distal joints of a part are given by $l^p_j = R(\mu_i, l^p_{ij}) + l^p_i$ and $l^d_j = R(\mu_j, [0, 0, L_j]) + l^p_j$ respectively, where $R(q, x)$ rotates the vector x by the quaternion q, L_j is the length of the part and l_{ij} is the location of the proximal joint defined in its local frame of reference (the origin of X_{ij} in Figure 2 (c)). Both these parameters are constant.

3.2. Representing a Quaternion

A unit quaternion is represented by two parts, a scalar and vector part $q = q_0 + \bar{q}$, where $\bar{q} = q_x\mathbf{i} + q_y\mathbf{j} + q_z\mathbf{k}$ and $|q| = 1$. The vector part represents the direction of the axis of rotation and the scalar part the cosine of half of the rotation. By ensuring q_0 is positive a quaternion can be represented in $\mathbb{R}^3$ using only the vector components. A value in this space then represents the direction of the axis of rotation scaled by the sine of half the rotation about it. This is similar to the approximation used in [13] except here the direction was scaled by the tangent, we opted not to use this since it results in a singularity. Given a value for $\bar{q}$, the scalar component can be recovered through $q_0 = \sqrt{1 - |\bar{q}|^2}$. However, this will have a discontinuity when $|\bar{q}| = 1$, given a set of quaternions provided for training $Q = \{q_1, .., q_m\}$ a space is constructed so that they are centered about the origin of $\mathbb{R}^3$ by solving

$$\arg\max_{q_i \in Q} \frac{1}{m} \sum_{j=1}^{m} (q_i^{-1} q_j).[1, 0, 0, 0]^T. \qquad (1)$$

This is similar to the transformation used in [13] and is performed for each part so that tracking and learning probability distributions can take place in this ‘safe’ three dimensional representation of a quaternion space.

3.3. The Unscented Transform

Often it will be desirable to propagate a distribution $\mathcal{N}(x; \mu, \Sigma)$ through some non-linear function $x' = f(x)$. One method to achieve this is to use the Unscented Transform [9]. This method decomposes the Covariance of a distribution into a set of $2D$ sigma points $\bar{\Sigma} = \{\sigma_1, .., \sigma_{2D}\}$, where D is the dimension of the covariance. Each sigma point is then translated by the mean to generate a set of points that represent the mean and covariance of the original distribution. Each sigma point is calculated as

$$\begin{aligned} \sigma_d &= \mu + \sqrt{D\upsilon_d}\mathbf{e}_d, \\ \sigma_{D+d} &= \mu - \sqrt{D\upsilon_d}\mathbf{e}_d, \end{aligned} \qquad (2)$$

where υ_d and $\mathbf{e}_d$ represents the dth eigenvalue and eigenvector of the covariance matrix. Once calculated each sigma point is then propagated through the non-linear function and the new sample mean and covariance calculated from them. This method will frequently be used to propagate a probability density function through a quaternion rotation. This process will be simply defined as $\mathcal{N}(x'; \mu', \Sigma') = \mathcal{F}(q, \mathcal{N}(x; \mu, \Sigma))$ if the rotation q is applied to the distribution or vice-versa if the sigma points are applied to q.

4. Limb Conditionals

Limb conditionals represent the edges of the graph and model the distribution $p(x_j | x_i, c_{ij})$, where c_{ij} is a connection parameter. Rather than learning a full limb conditional we follow the approximation in [13] and learn a distribution over x_{ij}, $p(x_{ij} | c_{ij})$. This is learnt over the quaternion representation described in Section 3.2, where $q_{ij} = q_i^{-1} q_j$, and the connection parameters are defined as the mean μ_{ij} and covariance Σ_{ij} of a Gaussian distribution. Given the state of x_i a prediction can be made over x_j through

$$p(x_j | x_i, c_{ij}) \approx \mathcal{F}(q_i, \mathcal{N}(x_{ij}; \mu_{ij}, \Sigma_{ij})). \qquad (3)$$

An example of the distributions predicted for the lower legs are shown in Figure 3. This shows a visualization of the distribution $p(x_j | x_i, c_{ij})$ learnt in quaternion space by projecting the sigma points into Euclidian space. As would be expected the greatest uncertainty is along the direction that the lower leg can rotate about the knee.

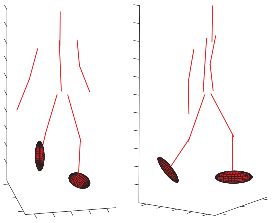


Figure 3. Visualizing the predictions of the lower legs’ state given that of the upper legs.

5. Approximating the Observational Likelihood Distribution

The observational likelihood represents the probability distribution over a part given an observation and is approximated using a Gaussian distribution $p(x_i | O_i) = \mathcal{N}(x_i; \mu_i^{obs}, \Sigma_i^{obs})$. Whilst in many signal processing tasks (e.g. tracking via radar) a measurement can be directly made and an error attached to this measurement, in the case of 3D tracking a leg position can not directly be estimated, only the image likelihood at a given location. A solution is to generate many samples of a limb’s position, weight each sample by the image likelihood and then use the weighted samples to estimate the mean and covariance of the distribution. However, this would be computationally expensive.

Instead, given a prediction of a limb’s state made from the previous frame $p(x_i^t | x_i^{t-1}, \mathcal{O}^{t-1}, x_r^{t-1}) = \mathcal{N}(x_i^t; \mu_i^t, \Sigma_i^t)$ the distribution is decomposed into a set of sigma points using (2). As well as $2D$ sigma points a copy of the mean is also maintained, so instead $2D + 1$ sigma points are selected and each scaled by $\sqrt{(D + \frac{1}{2})\upsilon_d}$. This set is defined as $\bar{\Sigma}_i^t = \{\sigma_{\{i,1\}}^t, .., \sigma_{\{i,7\}}^t\}$. Each sigma point is then projected into the image and weighted by the image likelihood at that location $w_{\{i,m\}}^t = p(O_i | \sigma_{\{i,m\}}^t)$. The weights are then normalized and the parameters $\{\mu_i^{obs}, \Sigma_i^{obs}\}$ are estimated from the weighted set of sigma points. Note that if the likelihood is uniform the distribution will remain unchanged. An example of the sample points used to represent a distribution is shown in Figure 4.

Image likelihoods are calculated using binary silhouettes. Given a binary silhouette $\mathcal{B}$ and the set of image pixels $\mathcal{P}$, pixels classified as the foreground are set to one $\mathcal{B}(\mathcal{P}_{fg}) := 1$ and those classified as the background are set to zero $\mathcal{B}(\mathcal{P}_{bg}) := 0$. Given a limb projected into the image consisting of the pixels $\mathcal{L}(x_i) \subset \mathcal{P}$, the cost is defined as $p(O_i | x_i) \propto \sum_{l \in \mathcal{L}(x_i)} \mathcal{B}(l)$. To prevent different limbs being assigned to the same mode (over counting), each con-

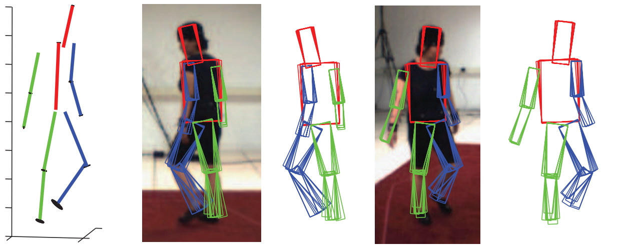


Figure 4. An example of a set of sample points used to estimate observational likelihood distributions projected into two views. They represent the distributions shown on the left.

structs a version of the binary silhouette for the opposing part $\mathcal{B}_{opp(i)}$, given by

$$\mathcal{B}_{opp(i)}(\mathcal{L}(\bar{\Sigma}_i^t) \cap \mathcal{P}_{fg}) := 0.5, \tag{4}$$

where $\bar{\Sigma}_i^t$ represents the set of sigma points. This makes it preferable for a limb to be located where the opposing limb is not predicted to be, whilst preferring this over locating a limb to a region of the image classified as the background.

Given a sample $\mathbf{X}_l = \{x_r, \mu_1, \Sigma_1, .., \mu_{n-1}, \Sigma_{n-1}\}$, the method described in this section is used to estimate the observational likelihood for each part $\{p(x_1|O_1), .., p(x_{n-1}|O_{n-1})\}$. This requires the equivalent number of image likelihood evaluations as just 7 particles using a SIR-PF or APF.

6. Probabilistic Inference

In this section we describe how the states of the nodes are updated for each sample $\mathbf{X}_l = \{x_r, X\}$. Inference is performed by passing messages between nodes. The posterior distribution for the jth node conditioned on all observations and a given root node state is calculated as

$$p(x_j|\mathcal{O}, x_r) = p(x_j|O_j) \prod_{v_i \in \mathcal{E}(j)} p(x_j|O_i, .., O_T, x_r), \tag{5}$$

where $v_i \in \mathcal{E}(j)$ defines the set of edges connected to j and $O_i, .., O_T$ represents the set of observations for the subtree containing v_i, created by removing the edge $\{v_i, v_j\}$. Since all distributions are modeled as Gaussians the above products can be calculated in a closed form through

$$\mathcal{N}(x_j; \mu_j, \Sigma_j) = \mathcal{N}(x_j; \mu_j^{obs}, \Sigma_j^{obs}) \prod_{v_i \in \mathcal{E}(j)} \mathcal{N}(x_j; \mu_j^{\vec{ij}}, \Sigma_j^{\vec{ij}}), \tag{6}$$

where $\mathcal{N}(x_j; \mu_j^{\vec{ij}}, \Sigma_j^{\vec{ij}})$ represents the message from i to j and the product of two Gaussian distributions results in a Gaussian with parameters

$$\begin{aligned} \Sigma_k =& (\Sigma_i^{-1} + \Sigma_j^{-1})^{-1}, \\ \mu_k =& \Sigma_k(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j). \end{aligned} \tag{7}$$

Messages are calculated in two steps. Firstly, incoming messages from all nodes other than the node to which the message is being passed are combined with the local observation likelihood

$$\mathcal{N}(x_i; \mu_i^{\vec{ij}}, \Sigma_i^{\vec{ij}}) = \mathcal{N}(x_i; \mu_i^{obs}, \Sigma_i^{obs}) \prod_{v_k \in \mathcal{E}(i)/j} \mathcal{N}(x_i; \mu_i^{\vec{ki}}, \Sigma_i^{\vec{ki}}). \tag{8}$$

Secondly, this distribution is propagated through the predictive model $p(x_{ij}|c_{ij})$ so that it is defined over x_j. Given the distribution center $\mu_i^{\vec{ij}}$, a prediction can be made using (3):

$$\mathcal{N}(x_j; \mu_j^{\vec{ij}}, \Sigma_j^{\vec{ij}}) = \mathcal{F}\left(\mu_i^{\vec{ij}}, \mathcal{N}(x_{ij}; \mu_{ij}, \Sigma_{ij})\right). \tag{9}$$

Whilst this propagates the uncertainty in the predictive model, the uncertainty in the message $\Sigma_i^{\vec{ij}}$ must also be passed. This is achieved using the center of the predictive model μ_{ij}

$$\mathcal{N}(x_j; \mu_j^{err}, \Sigma_j^{err}) = \mathcal{F}\left(\mathcal{N}(x_i; \mu_i^{\vec{ij}}, \Sigma_i^{\vec{ij}}), \mu_{ij}\right). \tag{10}$$

The final message is then given by the convolution of the two of these distributions setting $\mu_j^{err} := 0$, such that the message from i to j is calculated as

$$p(x_j|O_i, .., O_T, x_r) = \mathcal{N}(x_j; \mu_j^{\vec{ij}}, \Sigma_j^{\vec{ij}} + \Sigma_j^{err}). \tag{11}$$

The posterior for each node can then be updated using (6). The root node does not send messages, however, the posterior for a given root node state is approximated as

$$p(x_r|\mathcal{O}) \approx \prod_{i=1}^{n} \sum_{j=1}^{7} p(O_i|\sigma_{\{i,j\}}). \tag{12}$$

A sample $\mathbf{X}_l$ then consists of a root node state, a set of updated Gaussian distributions using (6) and a weight equal to the posterior, $\mathbf{X}_l = \{x_r, \mu_1, \Sigma_1, .., \mu_{n-1}, \Sigma_{n-1}, w\}$. The Maximum A Posterior (MAP) pose is then selected, this is defined by the set of Gaussian centers of the sample with the highest weight.

7. Temporal Diffusion

Given a sample $\mathbf{X}_l^t = \{x_r^t, \mu_1^t, \Sigma_1^t, .., \mu_{n-1}^t, \Sigma_{n-1}^t, w\}$, in this section we describe how the posterior over a node $p(x_j^t|\mathcal{O}^t, x_r^t) = \mathcal{N}(x_j; \mu_j^t, \Sigma_j^t)$ is used to estimate a prior in the following frame using

$$p(x_j^{t+1}|\mathcal{O}^t, x_r^t) \approx p(x_j^{t+1}|x_j^t) p(x_j^t|\mathcal{O}^t, x_r^t). \tag{13}$$

To achieve this $p(x_j^{t+1}|x_j^t)$ is approximated in a similar way as the limb conditionals in Section 4. This is represented by a zero mean diffusion model such that

$$p(x_j^{t+1}|x_j^t) \approx \mathcal{F}\left(\mu_i^t, \mathcal{N}(\dot{x}_{ij}, 0, \dot{\Sigma}_{ij})\right), \tag{14}$$

where the model is learnt over $\dot{q}_{ij} = \left(q_{ij}^t\right)^{-1} q_{ij}^{t+1}$. Given that $p(x_j^{t+1}|x_j^t) = \mathcal{N}(x_j^t, \mu_j^{diff}, \Sigma_j^{diff})$ calculated through (14), the prior is given by $p(x_j^{t+1}|\mathcal{O}^t, x_r^t) = \mathcal{N}(x_j, \mu_j^t, \Sigma_j^t + \Sigma_j^{diff})$, where the original covariance has been inflated by Σ_j^{diff}. In Figure 5 this method is shown applied to three consecutive frames where the covariance growth across the frames is cumulative.

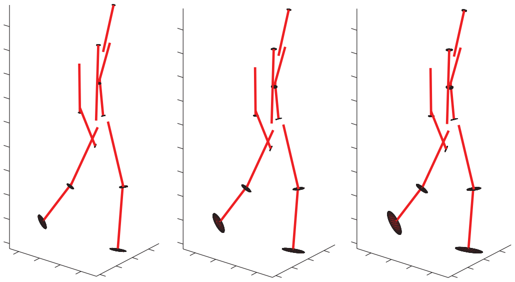


Figure 5. Example showing temporal diffusion applied to the covariances of the model for a given root node location.

The posterior over the root node is represented by the set of root node states and posterior weights taken from each sample, $p(x_r^t|\mathcal{O}^t) \approx \{x_{r,l}^t, w_l\}_{l=1}^m$. This distribution is also propagated using a zero mean diffusion model, though this is performed stochastically.

At each frame resampling is performed so areas of the posterior with a high likelihood are tracked over those with a low likelihood. Methods from annealing are used to adjust the posterior such that $w_l' = (w_l)^\beta$. A value of β can be selected such that the particle survival rate α can be estimated over the entire set of particles as described in [3]. Given a set of particles tracked over t frames the survival rate will decrease according to α^t. To allow the same survival rate to be maintained over a fixed time interval, α is set according to $\alpha = \exp \frac{\ln \alpha_c}{N_t}$, where α_c is the desired cumulative survival rate per second and N_t is the frame rate. This is used so that the uncertainty over the root node can be consistent regardless of the frame rate. A larger value of α_c will allow the distribution of the particles to represent a larger area of the posterior than a smaller value.

8. Experiments and Results

The presented method was tested using the HumanEva-I dataset which contains a scene captured from multiple views synchronized with motion capture data [12]. The 'Train' partition consisting of only motion capture data was used for training across all subjects performing walking and jogging actions and the first 300 frames of the 'Validation' partition was used for testing. Three views were used corresponding to the color cameras and foreground/background segmentation was performed using the Matlab code provided with the data set using default settings.

The presented approach was tested against two existing methods, the APF and the SIR-PF. The APF allows the presented method to be tested against an approach that converges to a single mode. Whilst the SIR-PF can be used to examine how existing approaches behave when permitted to represent a larger area of the posterior. The APF used 5 layers of 160 particles and the SIR-PF used a single layer of 800 particles. The presented method used 114 particles since calculating the posterior for each requires the equivalent image likelihood evaluations as 7 SIR-PF/APF particles.

For the APF α was set to 0.5 for each layer of annealing, therefore the survival rate per frame over all 5 layers was 0.03. This ensured the APF converged to a single mode. For the SIR-PF and the proposed method α_c was set to 0.01, at a frame rate of 60Hz this produces a survival rate per frame of 0.93. This allowed the SIR-PF and the proposed method to represent a much larger area of the posterior than the APF.

Limb limits were learnt from the training data and used to discard unlikely poses for all methods. For the APF pose was estimated using the expectation value of the samples and for the SIR-PF and the proposed method the MAP estimate was used.

It was noted that often the errors were dominated by left/right leg ambiguities, to overcome this during resampling an extra copy of a particle was occasionally maintained with the legs swapped. This was performed stochastically according to $p(swap) \propto \theta_{legs}$, where θ_{legs} is the angle between opposing upper legs. The total number of samples was still constant. Whilst this approach is rather adhoc it did alleviate the problem to some degree, most notably for the APF, though a far superior solution would be to employ a dynamic motion model. This was applied to all methods.

In Figure 6 the set of particles can be seen used to represent the posterior for the proposed method and the SIR-PF. As can be seen if the particle filter is used to represent a large uncertainty, this uncertainty is present in all parts of the model. This is in contrast to the proposed method where the posterior for each part is updated conditioned on the root node value of the particle. A large uncertainty in the root node is represented without increasing the uncertainty in the remaining parts.

In Table 1 the average errors using the train partition are shown for each subject walking. As can be seen over all subjects the proposed method outperforms both the APF and the SIR-PF. If the root node can be tracked with high accuracy it would be expected that the APF would outperform the proposed method, since the APF uses 800 particles compared to the presented approach that uses the equivalent of 7 to update the posterior for a single root node hypothesis. This can be seen in Table 1 where for S2 and S3 the APF outperforms the presented method. However, the principal claim in this work is that by representing a larger area

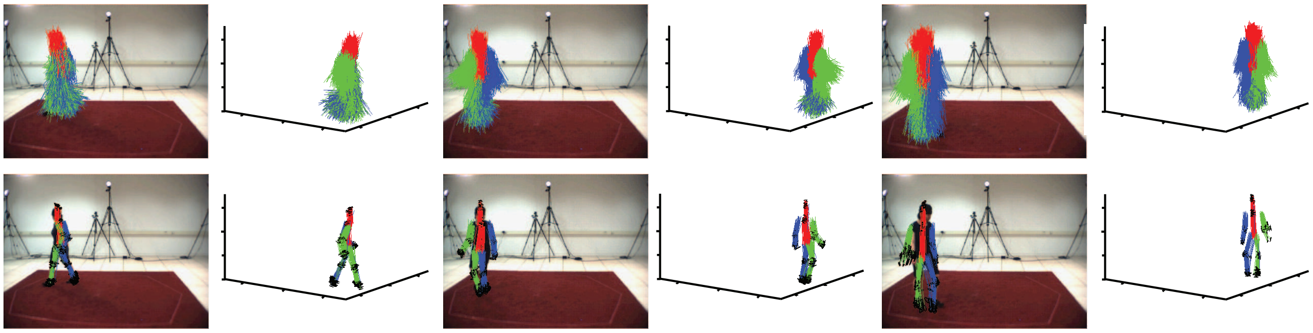


Figure 6. Example frames showing the distribution of the samples using the SIR-PF (top) and the proposed method (bottom) with $\alpha_c = 0.01$. The covariances for each sample have also been plotted for the proposed method.

of the posterior our method is less prone to tracking failure than the APF. This is evidently clear from S1 where the root node uncertainty is larger and the proposed method significantly outperformed the others.

Table 1. Pose estimation errors measured in mm using 3 cameras.

Method	S1	S2	S3	Average
APF	194.2	75.0	87.7	118.9 ± 65.5
SIR-PF	105.1	93.0	109.2	102.5 ± 8.4
Proposed	87.3	95.2	98.5	93.7 ± 5.8

This can further be seen in Figure 8 where an example of the tracking error in each frame is shown for the proposed method and the APF. During the first 60 frames when the root node is accurately tracked the error is lower for the APF, however, beyond this the APF fails whilst the proposed method is able to continue tracking the subject. To further illustrate this behavior we compare the discussed methods using just 2 camera views. Fewer camera views will result in more ambiguous observations and in these circumstances it will be beneficial to be able to represent greater uncertainty until these ambiguities can be resolved. The results are presented in Table 2. As expected the APF is more prone to tracking failure and our method outperforms both the SIR-PF and the APF. We further experimented using three cameras but at different frame rates. The error for each averaged across all subjects are presented in Figure 9. At lower frame rates, when there is greater movement by the subject across consecutive frames, the APF becomes more prone to falling into the wrong maxima and the presented method continues to outperform both techniques across all frame rates, highlighting its superiority.

Table 2. Pose estimation errors measured in mm using 2 cameras.

Method	S1	S2	S3	Average
APF	200.7	120.0	117.9	146.2 ± 47.2
SIR-PF	105.1	105.2	120.7	110.4 ± 8.9
Proposed	89.3	108.7	113.5	103.8 ± 12.8

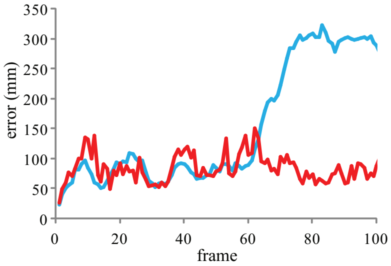


Figure 8. Tracking error in each frame for the APF (blue) and the proposed method (red).

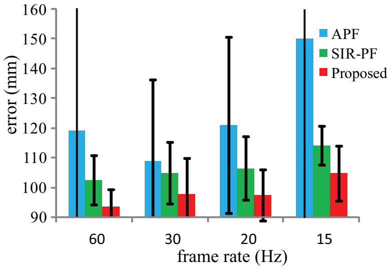


Figure 9. Tracking errors for walking using each method applied to a different frame rate.

Example frames showing the MAP estimate pose using the presented method are shown in Figure 7 where the estimated pose closely resembles that of the subject in each frame. We choose not to apply temporal smoothing hence the MAP estimate has a slightly jittery appearance across consecutive frames. Whilst the quantitative errors between the proposed method and the SIR-PF are relatively close, qualitatively the tracking is significantly poorer for the SIR-PF. In Figure 10 example frames are shown comparing the MAP solution using the SIR-PF compared to the proposed method. As can be seen the poses estimated by the SIR-PF are notably worse than those estimated by the proposed method.

9. Conclusions

In this paper a method has been presented to represent and track a much larger region of the posterior distribution than existing methods are able to by increasing uncertainty in the root node. This has been achieved by stochastically tracking the root node and estimating the posterior over the remaining parts of the model conditioned on each root node hypothesis. A method was presented to do this requiring the equivalent number of image likelihood evaluations as

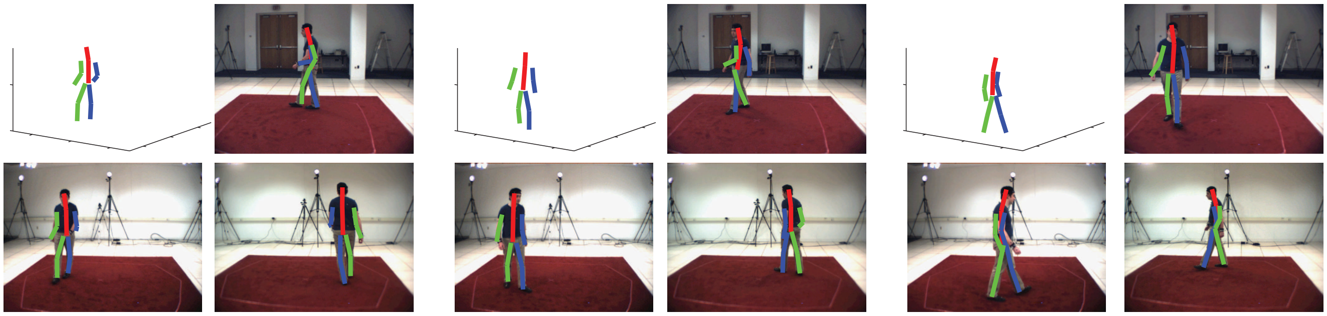


Figure 7. Example frames showing the MAP 3D pose using the proposed method projected into each camera view.

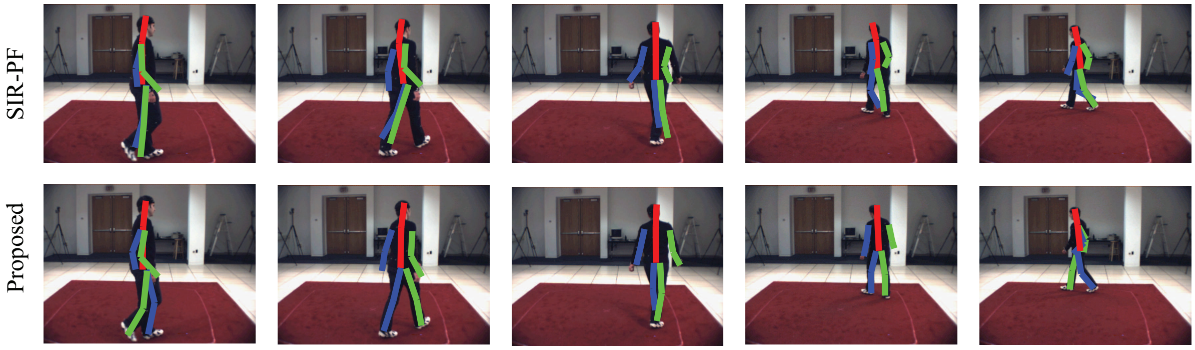


Figure 10. Comparison of pose estimation between the SIR-PF (top row) and proposed method (bottom row).

just 7 particles if using a SIR-PF or APF. It was shown that when existing methods are used to represent greater uncertainty, this uncertainty is increased across all parts of the body unlike the proposed method. Furthermore, compared to the APF that represents just a single mode, the presented approach was shown to be less prone to tracking failure. This was confirmed by quantitative results using the HumanEva data set and demonstrates that for 3D human tracking, greater robustness is achieved by supporting a much larger uncertainty in the root node. In future work we will investigate incorporating more complex dynamic models into our framework.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 1

[2] J. Deutscher, A. Davidson, and I. Reid. Automatic partitioning of high dimensional search space associated with articulated body motion capture. In *CVPR*, pages 669–676, 2001. 1, 2

[3] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, pages 185–205, 2005. 1, 2, 6

[4] A. Doucet, N. d. Freitas, K. P. Murphy, and S. J. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 176–183, 2000. 2

[5] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, pages 55–79, 2005. 2, 3

[6] G. Hua and Y. Wu. Variational maximum a posteriori by annealed mean field analysis. *PAMI*, 27(11):1747–1761, 2005. 1, 2

[7] M. Isard. Pampas: Real-valued graphical models for computer vision. In *CVPR*, pages 613–620, 2003. 1, 2

[8] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1

[9] S. Julier, J. Uhlmann, and H. Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference*, volume 3, pages 1628–1632, 1995. 4

[10] M. W. Lee and R. Nevatia. Human pose tracking using multi-level structured models. In *ECCV*, pages 368–381, 2006. 2

[11] R. Navaratnam, A. Thayananthan, P. Torr, and R. Cipolla. Hierarchical part-based human body pose estimation. In *BMVC*, 2005. 2

[12] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2009. 2, 6

[13] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, 2004. 2, 3, 4

[14] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *CVPR*, 2001. 2

[15] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *CVPR*, 2003. 2

[16] X. Xu and B. Li. Learning motion correlation for tracking articulated human body witha rao-blackwellised particle filter. In *ICCV*, 2007. 2