# Recursive Bayesian Filtering for Multiple Human Pose Tracking from Multiple Cameras

Oh-Hun Kwon, Julian Tanke, and Juergen Gall

University of Bonn, Germany
ohkwon@uni-bonn.de, {tanke, gall}@iai.uni-bonn.de

**Abstract.** Markerless motion capture allows the extraction of multiple 3D human poses from natural scenes, without the need for a controlled but artificial studio environment or expensive hardware. In this work we present a novel tracking algorithm which utilizes recent advancements in 2D human pose estimation as well as 3D human motion anticipation. During the prediction step we utilize an RNN to forecast a set of plausible future poses while we utilize a 2D multiple human pose estimation model during the update step to incorporate observations. Casting the problem of estimating multiple persons from multiple cameras as a tracking problem rather than an association problem results in a linear relationship between runtime and the number of tracked persons. Furthermore, tracking enables our method to overcome temporary occlusions by relying on the prediction model. Our approach achieves state-of-the-art results on popular benchmarks for 3D human pose estimation and tracking.

## 1 Introduction

Markerless motion capture [1–9] has many applications in sports [10, 11] and surveillance [12]. Utilizing multiple calibrated cameras extends the field of view, allows to resolve ambiguities such as foreshortening and occlusions, and provides accurate 3D estimates. However, challenges still remain: large crowds and close interactions result in heavy occlusions which severely degrade the 3D tracking performance. Furthermore, most recent works [3–5, 9] cast multiple 3D human pose estimation from multiple camera views as an association problem where extracted 2D pose features have to be matched across views and across time. This way, the time complexity grows quadratic [9] or even exponential [4, 5] with the number of tracked individuals, making tracking of large numbers of persons impractical.

In this work we cast the problem of estimating multiple persons from multiple calibrated cameras as a tracking problem where each person is individually tracked using the well-known recursive Bayesian filtering method [13]. Individually tracking each person results in a linear relationship between time complexity and the number of persons in the scene. Furthermore, utilizing a tracking framework enables us to retain plausible poses even under temporary heavy occlusion. Last but not least, the Bayesian framework allows us to quantify uncertainty.

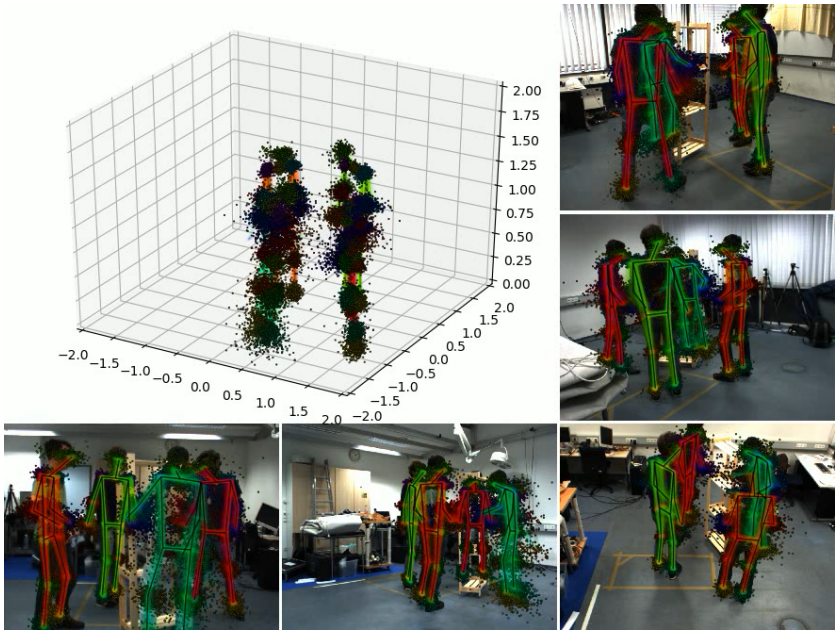


Fig. 1: Probabilistic representation for 3D pose tracking. The black points represent 3D pose predictions from the prediction step while the colored skeletons represent the pose samples after the update step. Notice that both representations model uncertainty. The final pose is the black skeleton at the center of each person.

Recursive Bayesian filtering naturally lends itself for human pose tracking from multiple cameras. It models an underlying process $z_{1:T}$, which we are interested in but which we cannot directly observe. Instead, at each time step $t$ we receive observations $o_t$ which are related to $z_t$. Bayesian filtering provides us with tools to form our best guess about $z_t$ given the observations $o_{1:t}$. For 3D human pose tracking, the unobserved hidden state $z_t$ represents the 3D pose at time $t$ while the observation $o_t$ represents the camera input at time $t$ for all cameras. Bayesian filtering utilizes a prediction step, which forecasts the current estimate in time, and an update step, which incorporates current observations into the prediction. To model uncertainty we utilize a sample-based approach for $z_t$. For the prediction step we build on recent advancements in 3D human motion anticipation [14] and utilize a sequence-to-sequence model. During the update step we process all samples in $z_t$ and make use of importance sampling, similar to the particle filter [13]. In order to reduce the number of required particles, we combine it with an optimization step to find good 3D poses. Our method achieves state-of-the-art results on common multiple person multi-camera pose estimation and tracking benchmarks.

## 2 Related Work

Methods for 2D human pose estimation can be split into top-down and bottom-up approaches. Top-down 2D human pose estimation [15, 16] first estimates 2D bounding boxes for each person and then estimates the 2D pose per detected human on a fixed resolution. Bottom-up methods [17] on the other hand estimate features that assist in assembling 2D poses. For example, part affinity fields [17] estimate vector fields that indicate the association of joints.

Multi-person 2D pose tracking has been an active research area and recent works achieved tremendous advancements [16, 18–21]. Early works focused on solving spatio-temporal graphs [18, 19] while more recent approaches [16, 20, 21] showed that utilizing a greedy graph matching still yields state-of-the-art results while being much faster.

Extensive progress has been made in estimating 3D poses from monocular views [22–26]. For example, the problem of inferring 3D human poses from single images is split into estimating a 2D human pose and then regressing a 3D pose [22]. However, these methods do not generalize well to unconstrained data.

Multiple 3D human pose estimation from multiple views can be cast as a matching problem where poses or joints have to be matched across views for accurate triangulation. Early works [6–8, 3] utilized a 3D pictorial structure model to extract 3D poses. However, optimizing these models is time consuming, especially when applied to multiple persons, due to the large state space. When many camera views are available, a voting mechanism [27] can be employed - assuming persons are visible in most camera views. Recently, a simple baseline [9] was proposed which independently extracts 2D poses for each view and greedily matches them using geometric cues. Furthermore, they utilize Gaussian smoothing across time to introduce temporal information. While this method is simple and fast, it suffers from an early commitment to 2D pose matches. This may lead to different predictions based on the processing order of the cameras. Dong et al. [5] solve the correspondence problem of 2D poses per camera utilizing a top-down 2D pose estimator [28] for each view. They match 2D poses across views using geometric and appearance cues solved with convex optimization with cycle-consistency constrains. While this works well when persons are easy to differentiate, e.g. when full-body poses are visible, it can result in incorrect matches in more complex scenes. Zhang et al. [4] formulate cross-view matching and temporal tracking as a 4D association graph, similar to early works in 2D pose tracking [19].

To facilitate tracking, Bayesian filtering [13] is often utilized. While linear methods such as linear quadratic estimation are well-understood they restrict the model too much for tracking complex 3D human poses. Representing the distribution of poses as a set of 3D pose samples, also known as particle filter, offers more flexibility. However, due to the high dimensionality of 3D poses additional optimization steps are typically required [29–31]. In this work we show that a simple heuristic can be utilized.

In recent years deep neural networks have been used to anticipate 3D human poses from motion capture data. Holden et al. [32] show that autoencoders

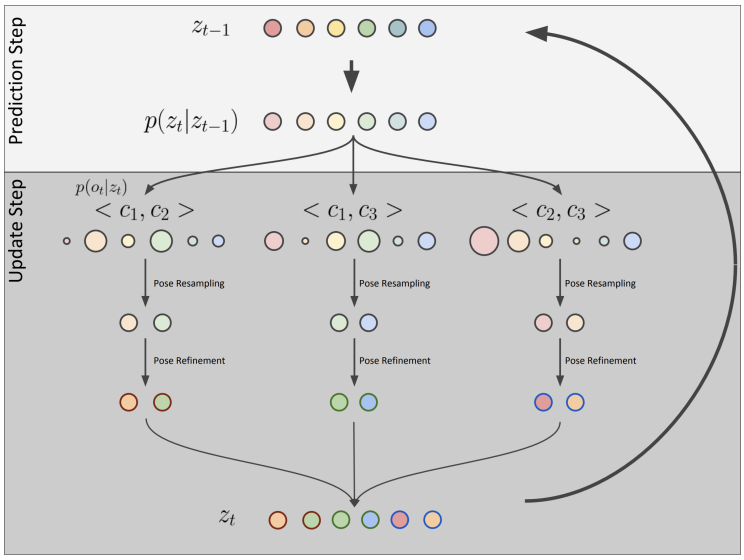


Fig. 2: Tracking procedure for tracking a single person $z$ with a set of three cameras $c_1$, $c_2$ and $c_3$. The prediction step forecast $z_{t-1}$ from time $t-1$ to $t$. In the update step each pose sample gets assigned an importance weight independently for each camera pair. The importance weights are calculated using the observations $o_t$ at time $t$. We then resample for each camera pair relative to the total number of samples and refine the poses using pose refinement. Finally, we concatenate the sub-samples for each camera pair and obtain our prediction for $z_t$.

can be utilized to learn a human motion manifold. Bütepage et al. [33] extend this idea by embedding the skeletal hierarchy structure of the data into the model. Similarly, structural RNNs [34] encode the hierarchy utilizing RNNs. The Encoder-Recurrent-Decoder (ERD) [35] auto-regressively forecasts human motion by utilizing an encoder-decoder structure for modeling human poses and an LSTM for temporal modelling while Martinez et al. [14] introduce a sequence-to-sequence architecture for human motion forecasting.

Monte-Carlo Dropout sampling [36] places a Bernoulli distribution over the weights of a neural network. This way one can generate multiple samples using different dropout masks to represent the model uncertainty.

## 3 Method

In this work we formulate the problem of estimating and tracking multiple 3D human poses from multiple calibrated cameras as a recursive Bayesian filter where the hidden states $z_t$ represent the 3D human poses and where the camera

images are the observations $o_t$ at time step $t$. More precisely, each person in the scene has a 3D pose state $z$ which is tracked independently through time, as described in Figure 2. This means that we have a Bayesian filter for each person. This has the advantage that we can easily deal with appearing and disappearing persons.

A Bayesian filter recursively cycles through prediction and update steps. The prediction step utilizes a prediction model $p(z_t|z_{t-1})$ which evolves the hidden state in time while the update step utilizes an observation model $p(o_t|z_t)$ which integrates measurements into the prediction. We build on recent advancements in 3D human motion anticipation [14] and model $p(z_t|z_{t-1})$ as a recurrent neural network (RNN) where uncertainty is represented by Dropout as Bayesian approximation [36]. The observation model $p(o_t|z_t)$ measures how well a 3D pose sample matches the extracted 2D joint confidence maps and part affinity fields [17, 37] for each camera view. For each tracked person, a set of 3D sample poses is used to represent the posterior $p(z_t|o_{1...t})$. A sample-based representation of the distribution [31, 30, 38] allows for a highly non-linear state space, which is required for complex human poses, while being simple to implement. In Section 3.1 we detail the prediction model while in Section 3.2 we discuss the observation model. The initialization procedure for $p(z_1|o_1)$ of each person is explained in Section 3.3. Finally, Section 3.4 explains how point samples can be obtained for each frame $t$ from the estimated posterior $p(z_t|o_1 \dots o_t)$.

### 3.1 Prediction Step

The prediction model $p(z_t|z_{t-1})$ evolves the pose state of a single tracked person in time - without taking any observation into account. The pose state $z_t$ of a person encompasses possible 3D poses which we make tractable by representing them as a fixed set of 3D pose samples. A sample is made up of 14 3D joints.

We represent $p(z_t|z_{t-1})$ as GRU [14] and we inject uncertainty by utilizing Dropout during training and inference at the final linear layer that extracts $z_t$. Dropout is crucial to generate a diverse set of forecast poses which we will discuss in our ablation studies. As $z_{t-1}$ is represented as a list of 3D pose samples, we apply the forecast for each sample independently with an independent hidden layer for the GRU for each sample. This way, $z_t$ and $z_{t-1}$ will be represented by the same number of samples while samples will be sufficiently varied due to the independent forecasting. For brevity, we define this as:

$$z_t, h_t = \text{GRU}(z_{t-1}, h_{t-1}) \tag{1}$$

where $z_t$, $h_t$, $z_{t-1}$ and $h_{t-1}$ are 3D pose samples and GRU hidden states, respectively.

The 3D poses in $z$ are in a global coordinate frame which is defined by the calibrated cameras. We transform the 3D poses into a standardized coordinate frame before forecasting. Here, the center hip joint of the poses in $z_{t-1}$ are set as the origin and the poses are rotated along the $z$ axis[1] such that the left and right

[1] assuming $z$ axis points upwards

hip joints align to the $y$ axis and such that the 3D pose faces forward along the $x$ axis. More formally, we apply the following transformation to each 3D pose

$$\hat{\mathbf{x}}_j = R^{(t-1)} \ (\mathbf{x}_j - \mathbf{x}^{(t-1)}_{\text{hiproot}}) \ \forall j \in \text{J} \tag{2}$$

where J represents all joints that make up a 3D pose and where $\mathbf{x}_j$ represents the $j$-th joint as 3D point in global coordinate space and where $\hat{\mathbf{x}}_j$ represents the same joint in normalized coordinates. The hip root joint of the pose at time $t-1$ is defined as $\mathbf{x}^{(t-1)}_{\text{hiproot}}$ and the rotation to forward-face the pose at $t-1$ is defined as

$$R^{(t-1)} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

$$\theta = \text{atan2}\big(y^{(t-1)}_{\text{righthip}} - y^{(t-1)}_{\text{lefthip}}, x^{(t-1)}_{\text{righthip}} - x^{(t-1)}_{\text{lefthip}}\big) \tag{4}$$

where $x^{(t-1)}_{\text{righthip}}$, $x^{(t-1)}_{\text{lefthip}}$, $y^{(t-1)}_{\text{righthip}}$ and $y^{(t-1)}_{\text{lefthip}}$ represent the $x$ and $y$ coordinate of the right hip and left hip, respectively. After forecasting a pose, the original position and orientation in global coordinates can be recovered by applying the transformation

$$\mathbf{x}_j = {R^{(t-1)}}^T \hat{\mathbf{x}}_j + \mathbf{x}^{(t-1)}_{\text{hiproot}} \ \forall j \in \text{J}. \tag{5}$$

The prediction model is trained with motion capture data from the Human3.6M [39] and the CMU mocap database [40] where we select 14 joints that the two datasets have in common. We utilize Adam with learning rate 0.001 and optimize over the Huber loss. The number of hidden units for the GRU is set to 2048. The dropout rate is set to 50% and a weight decay of $10^{-8}$ is added. We set the framerate to 25Hz and 30Hz, respectively, which is similar to the framerate used in the evaluation datasets.

### 3.2 Update Step

To obtain the posterior $p(z_t|o_{1,\dots,t})$ for a single tracked person we need to incorporate the observations $o_t$ into the predictions $z_t$ obtained from the prediction model. For each camera we utilize Openpose [17] to extract part confidence maps and part affinity fields, similar to other multi-person multi-camera 3D pose estimation methods [9, 4]. We then calculate importance weights for each sample pose in $z_t$ and then re-sample $z_t$ based on the weights. To prevent poses that are visible in many camera views to be over-represented over poses that are visible in less cameras and to tackle false-positive detections caused by occlusion, we sample the importance weight for each camera pair independently - for all samples. The weight is calculated as follows:

$$w_{v,s} = \frac{\Phi(v,s)}{\sum_{\hat{s}}^{z_t} \Phi(v,\hat{s})} \tag{6}$$

where $v$ represents a camera pair and where $s$ represents a single 3D pose sample from $z_t$. We normalize by the scores of all samples $\hat{s}$ in $z_t$. The unnormalized

weight $\Phi(\cdot,\cdot)$ is calculated as follows:

$$\Phi(v,s) = \prod_{l \in \mathrm{L}} \sqrt{\sum_{c \in v} \phi(c,l,s)^2} + \epsilon \tag{7}$$

where L represents all limbs of a pose, as described in Openpose [17]. Each camera pair $v$ consists of two different camera views $c$. The score $\phi(\cdot,\cdot,\cdot)$ is calculated using part affinity fields $\mathrm{paf}_c$ and confidence maps $\mathrm{conf}_c$, which are obtained from Openpose [17], for a given camera $c$:

$$\phi(c,l,s) = \left( \int_{u=0}^{u=1} \max\left(0, \mathrm{paf}_c(s,l,u)\right) \, \mathrm{d}u \right) \prod_{j \in l} \mathrm{conf}_c(s,j) \tag{8}$$

where $\mathrm{paf}_c(s,l,u)$ calculates the dot product between the part affinity field for limb $l$ and the projected limb from $s$, linearly interpolated by $u$. $\mathrm{conf}_c(s,j)$ calculates the confidence score for the joint $j$ of sample $s$ for camera $c$. Finally, we resample $z_t$ for each camera pair a subset of particles to obtain the same number of initial samples as shown in Figure 2. As sampling procedure we use stochastic universal sampling.

In practice, the state space of a 3D pose is prohibitively large for a sample-based representation. However, we utilize a simple yet effective heuristic optimization called joint refinement to keep the number of samples low while obtaining accurate results. For each joint of a sample, we sample additional joint positions from a normal distribution centered at the joint. We then take the joint position with the highest confidence map score. In our ablation study we show that this significantly improves the results while it allows the numbers of samples for each person to be low.

### 3.3 Initialization Step

(a) Heatmap (b) Projected Samples (c) Output

Fig. 3: Input and output of the confidence subtraction network. The input is composed of a confidence map (a) extracted by [17] for a specific joint and projected points (b) of that joint for the tracked person. The network removes the part of the heatmap that corresponds to the tracked person.

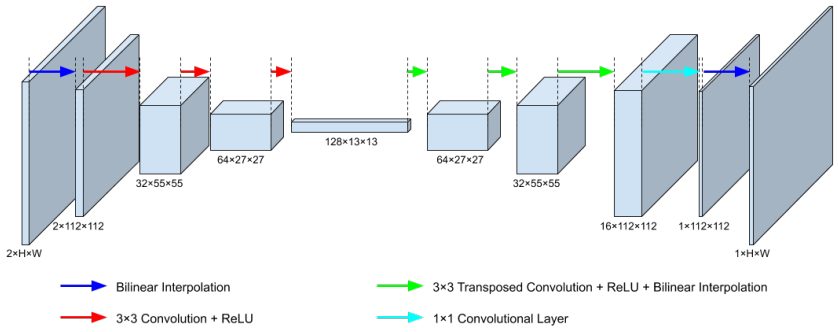


Fig. 4: Fully convolutional architecture for the confidence map subtraction network.

To facilitate multi-person 3D pose tracking, a set of currently tracked persons is kept which are all independently tracked using the prediction (Section 3.1) and update (Section 3.2) step. However, at each time step we have to check whether one or more untracked persons have entered the 3D recording volume and generate new tracks accordingly. To do so, we first remove currently tracked persons from the confidence maps for each camera using a confidence subtraction network. To remove a tracked person from a confidence map, we project the joints of all samples of that person for the given frame to that camera view (see Figure 3 (b)). We then pass the projected points as well as the confidence map to the confidence subtraction network which will return an updated confidence map without the peak of the tracked person. Figure 3 shows an example while Figure 4 details the network structure. We repeat this procedure for all tracked persons and for all camera views.

Once all tracked persons are removed from the confidence maps, we find the remaining local maxima and triangulate them pairwise if both points are close to their respective epipolar line as in [9]. To reduce the number of redundant points, we apply agglomerative hierarchical clustering with threshold $\epsilon_j$ and use the mean point of the clusters. We then build a set of 3D pose candidates by greedily matching joints based on the part affinity fields [17]. We also drop limbs that have unreasonable length. Each pose candidate is then scored using Equation (7), where $v$ contains all camera views, and the 3D pose with the highest score is selected for the new track.

As a person track is represented as a list of 3D pose samples, we utilize a stochastic generation function $z \leftarrow g(\cdot)$ which takes as input the previously selected best 3D pose candidates and generates a set of 3D pose samples $z$ that represent the distribution of the newly generated track. Once the pose samples are generated the person can be tracked using the the prediction and update steps. The new pose is removed from the confidence maps and the initialization procedure is repeated until no further person tracks are found.

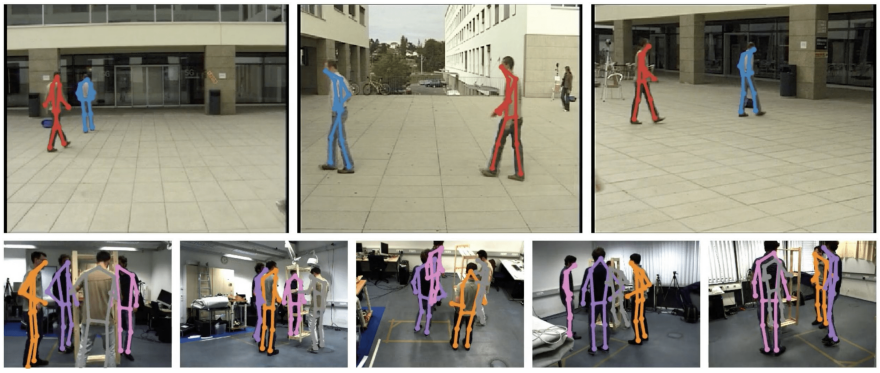

Fig. 5: Qualitative results from the Campus [6, 41] (top row) and Shelf [6] (bottom row) dataset.

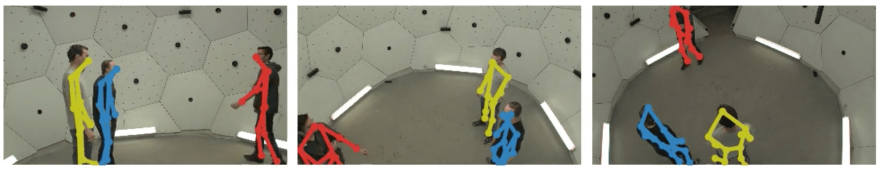

Fig. 6: Qualitative results showing the first three hd-cameras of the CMU Panoptic studio [42].

We model $g$ as a three-layer feed-forward Bayesian neural network [36] which takes as input a pose vector and outputs a pose vector. As a person might only be partially visible $g$ also fills in missing joints. This is facilitated by adding a binary vector to the input pose vector which indicates if a joint is missing. As dropout is utilized during inference, $g$ generates a diverse set of 3D pose samples. The network is trained with motion capture data from Human3.6M [39] and from the CMU mocap database [40], similar to Section 3.1. During training, random joints are removed from the pose to encourage the model to fill in missing joints. The model has three layers with 2048 hidden units each and is optimized over the Huber loss using SGD with a learning rate of 0.001, weight decay of $10^{-6}$ and dropout of 75%.

### 3.4 Inference

Using multiple samples to represent a 3D pose allows for robust tracking. However, when extracting 3D poses a final single pose is required. To obtain a final pose from $z_t$ for a tracked person at frame $t$, we calculate the weighted average for all samples using Equation (6) where $v$ contains all cameras.

| | Campus | | | | | Shelf | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A1(48) | A2(188) | A3(136) | aAvg | gAvg | A1(279) | A2(37) | A3(161) | aAvg | gAvg |
| Belagiannis et al. [6] | 82.01 | 72.43 | 73.72 | 76.05 | 74.14 | 66.05 | 64.97 | 83.16 | 71.39 | 71.75 |
| +Belagiannis et al. [7] | 83.00 | 73.00 | 78.00 | 78.00 | 76.12 | 75.00 | 67.00 | 86.00 | 76.00 | 78.09 |
| Belagiannis et al. [43] | 93.45 | 75.65 | 84.37 | 84.49 | 81.14 | 75.26 | 69.68 | 87.59 | 77.51 | 79.00 |
| Ershadi-Nasab et al. [3] | 94.18 | 92.89 | 84.62 | 90.56 | 90.03 | 93.29 | 75.85 | 94.83 | 87.99 | 92.46 |
| Dong et al. [5] | 97.40 | 90.10 | 89.40 | 92.30 | 90.79 | 97.20 | 79.50 | 96.50 | 91.07 | 95.59 |
| *Dong et al. [5] | 97.60 | 93.30 | **98.00** | **96.30** | **95.57** | 98.80 | 94.10 | 97.80 | 96.90 | 98.10 |
| +Tanke et al. [9] | **98.00** | 91.00 | **98.00** | 95.67 | 94.46 | 99.21 | 93.51 | 97.14 | 96.62 | 98.07 |
| +Zhang et al. [4] | - | - | - | - | - | 99.00 | **96.20** | 97.60 | 97.60 | 98.31 |
| +Ours | 97.35 ± 0.40 | **93.44** ± 0.04 | 97.43 ± 0.18 | 96.07 ± 0.13 | 95.40 ± 0.07 | **99.49** ± 0.06 | 95.81 ± 0.37 | **97.83** ± 0.00 | **97.71** ± 0.13 | **98.64** ± 0.05 |

Table 1: Quantitative comparison with state-of-the-art methods using percentage of correctly estimated parts (PCP) on the Campus and Shelf datasets. *A1* to *A3* represent the three actors while the number in parentheses represents the number of ground-truth frames. We report both actor-wise ($aAvg$) as well as global average ($gAvg$) PCP. Models utilizing temporal information are marked with $+$ while appearance information is marked with $*$. As our method is probabilistic, we report results as mean ± standard deviation, which is calculated over 10 runs using different random seeds.

## 4 Experiments

### 4.1 Quantitative Comparison

We provide a quantitative comparison to recent state-of-the-art methods using the Campus [6, 41] as well as the Shelf [6] dataset. Qualitative results on this datasets can be seen in Figure 5. As metric we use percentage of correct parts (PCP) in 3D [10] and we adopt the head position alignment utilized in [5] as well as the temporal Gaussian smoothing described in [9]. Furthermore, we report PCP averaged over the actors (aAvg) and PCP averaged over the actors weighted by the number of visible frames (gAvg), which was first discussed in [43]. Weighting by the number of visible frames (gAvg) provides a more accurate measure as it does not overemphasize actors which appear only in very few frames. Table 1 presents our results. For the the Shelf dataset we achieve state-of-the-art results while we achieve highly competitive results on the Campus dataset. We argue that the top-down pose estimation model and the appearance model of [5] are beneficial when the full bodies are visible and the scenes are relatively uncluttered, as it is the case with the Campus dataset (Figure 5 top row). However, in more complex scenes where bodies are only partially visible and with large background clutter and occlusions, such as Shelf, the appearance model does not help as much. Here, temporal information is crucial to recover from occlusions.

| Method | MOTA | Precision | Recall | MOTA | Precision | Recall |
|---|---|---|---|---|---|---|
| | Average | | | Nose | | |
| Tanke et al. [9] | 0.82 | 91.0 | 91.1 | 0.84 | 91.7 | 91.8 |
| Ours | **0.87** | **93.3** | **94.1** | **0.94** | **96.6** | **97.5** |
| | Left Wrist | | | Right Wrist | | |
| Tanke et al. [9] | 0.82 | **91.2** | 91.3 | 0.86 | **93.0** | 93.1 |
| Ours | **0.83** | 91.1 | **91.9** | 0.86 | 92.6 | **93.4** |
| | Left Foot | | | Right Foot | | |
| Tanke et al. [9] | 0.81 | 90.5 | 90.6 | 0.77 | 88.6 | 88.7 |
| Ours | **0.90** | **94.6** | **95.5** | **0.84** | **91.5** | **92.3** |

Table 2: Tracking scores MOTA [44], precision and recall for sequence 160422_ultimatum1 of the CMU Panoptic Studio [42].

### 4.2 Tracking

For evaluating the tracking performance of our method, we utilize the MOTA [44] score as well as precision and recall. We cannot evaluate tracking on the Shelf or Campus dataset as some of the ground-truth annotations are missing, which results in a large number of false positives. Instead we evaluate on the CMU Panoptic studio [42], which utilizes the same human pose keypoints [45] as our method and which provides unique identifiers for each person in the scene. We use the sequence 160422_ultimatum1 from frames 300 to 1300 as in [9] since it contains different interacting persons that enter and leave the scene. A sample scene can be seen in Figure 6. To ensure occlusions, we utilize only the first three hd-cameras and we consider a track as correct if its prediction is within 10cm of the ground-truth. For measuring the tracking accuracy, we utilize the nose, left/right wrist and left/right foot. Our results are presented in Table 2. We observe that our model significantly outperforms [9] for feet and nose since these keypoints are for some frames only visible in one camera as shown in Figure 6. Our method can recover these cases.

### 4.3 Ablation

Our ablation results are presented in Table 3. Removing tracking and only using the pose initialization algorithm described in Section 3.3 at each frame results in very strong results for the Shelf dataset while the performance drops significantly for the Campus dataset. The reason for this is that the pose initialization works better when multiple views are present (5 for Shelf, 3 for Campus) while tracking helps when a person is temporally visible in only one or two views. Removing pose resampling during the update step and instead using a fixed set of samples for each camera pair results in a significant performance drop. One of the biggest factors for the strong performance of our method is the joint refinement as the sample-based representation of 3D poses does not permit enough samples to

| | Campus | | | | | Shelf | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A1(48) | A2(188) | A3(136) | aAvg | pAvg | A1(279) | A2(37) | A3(161) | aAvg | pAvg |
| only Pose Initialization | 91.85 ± 1.33 | 92.94 ± 0.22 | 69.96 ± 0.95 | 84.92 ± 0.68 | 84.40 ± 0.46 | 99.51 ± 0.14 | 94.03 ± 0.83 | 97.69 ± 0.05 | 97.07 ± 0.31 | 98.47 ± 0.13 |
| w/o Pose Resampling | 87.29 ± 7.86 | 90.57 ± 0.33 | 88.27 ± 5.43 | 88.71 ± 3.74 | 89.31 ± 2.62 | 97.47 ± 1.10 | 88.95 ± 1.40 | 97.83 ± 0.09 | 94.75 ± 0.70 | 96.93 ± 0.70 |
| w/o Joint Refinement | 47.33 ± 24.53 | 74.78 ± 14.91 | 57.52 ± 5.16 | 59.88 ± 11.89 | 64.93 ± 9.81 | 90.96 ± 1.95 | 73.11 ± 5.39 | 91.89 ± 1.92 | 85.32 ± 2.24 | 89.89 ± 1.27 |
| w/o Pose Prediction | 70.56 ± 12.91 | 82.93 ± 6.40 | 73.19 ± 5.22 | 75.56 ± 4.23 | 77.77 ± 4.01 | 99.16 ± 0.04 | 65.22 ± 11.09 | 97.73 ± 0.03 | 87.37 ± 3.70 | 96.04 ± 0.86 |
| Pose Prediction : $\mathcal{N}(0, 0.01^2)$ | 75.83 ± 27.52 | 80.84 ± 6.47 | 70.49 ± 10.29 | 75.72 ± 11.99 | 76.41 ± 8.38 | 99.12 ± 0.04 | 69.24 ± 12.09 | 97.71 ± 0.04 | 88.69 ± 4.03 | 96.33 ± 0.94 |
| Pose Prediction : w/o dropout | 90.42 ± 0.73 | 92.12 ± 0.12 | 97.28 ± 0.15 | 93.27 ± 0.27 | 93.79 ± 0.15 | 99.29 ± 0.04 | 94.78 ± 1.09 | 97.76 ± 0.00 | 97.28 ± 0.37 | 98.43 ± 0.09 |
| Joint Refinement : Gradient Ascent | 96.15 ± 0.10 | 92.34 ± 0.11 | 97.13 ± 0.22 | 95.21 ± 0.07 | 94.58 ± 0.12 | 99.40 ± 0.05 | 93.92 ± 0.68 | 97.80 ± 0.03 | 97.04 ± 0.23 | 98.43 ± 0.07 |
| Proposed | 97.35 ± 0.40 | 93.44 ± 0.04 | 97.43 ± 0.18 | 96.07 ± 0.13 | 95.40 ± 0.07 | 99.49 ± 0.06 | 95.81 ± 0.37 | 97.83 ± 0.00 | 97.71 ± 0.13 | 98.64 ± 0.05 |

Table 3: Ablation study using percentage of correctly estimated parts (PCP) on the Campus and Shelf datasets. *A1* to *A3* represent the three actors while the number in parentheses represents the number of ground-truth frames. We report both actor-wise ($aAvg$) as well as global average ($gAvg$) PCP.

accurately represent such high dimensional data. Removing the pose prediction model and just utilizing a zero velocity model also results in a significant performance loss. Replacing the zero-velocity model with a normal distribution for pose prediction does not significantly improve the results. Replacing the heuristic joint refinement algorithm described in Section 3.2 with a gradient ascent based algorithm results in a slight performance drop. We argue that the local optimization gets stuck in local optima while the heuristic can jump over them and find even better pose configurations.

### 4.4 Parameters

The effects of the hyperparameters are shown in Figure 7. The Dropout rates of both the prediction model and the initialization model $g$ are determined to obtain a reasonable approximation of uncertainty. If it is too small, the uncertainty is underestimated. For large values, the generated samples are too diverse, making the approach inefficient. The number of pose samples is important to ensure a sufficient representation of the pose distribution. However, a too high number of pose samples impedes sometimes the discovery of newly appearing persons and thus degrades the overall quality of the results. The distance threshold $\epsilon_j$ of the hierarchical clustering for merging joints influences the quality of the triangulated 3D joint positions to initialize poses. While a high value $\epsilon_j$ merges 3D joints of different persons, more redundant 3D joints would remain with a lower threshold. Using many samples for joint refinement encourages that each joint is located in regions with high part confidences. When the number is too large, it reduces the variety of the samples which weakens the tracking quality.

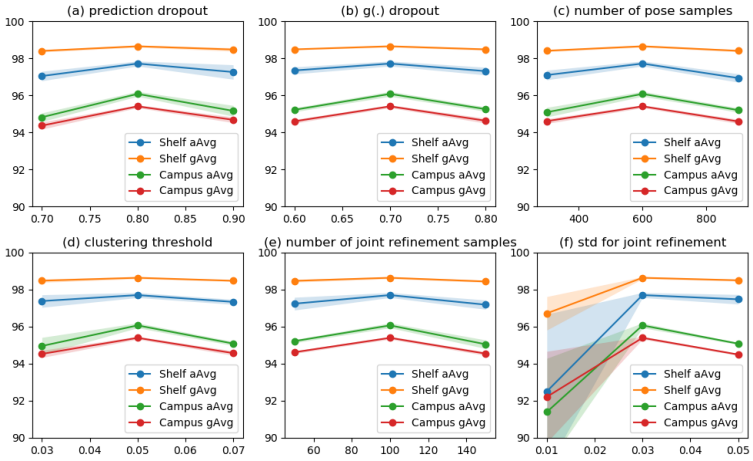


Fig. 7: Evaluation of hyperparameters. PCP is evaluated while varying the hyperparameters. With each setting, the experiments are performed 10 times. The solid line indicates the mean value of the PCP and the colored area is the coverage determined by the standard deviation. (a) The dropout rate of the prediction model. (b) The dropout rate of the model $g(\cdot)$. (c) The number of pose samples. (d) Distance parameter to merge joints using the hierarchical clustering. (e) The number of samples for joint refinement. (f) Standard deviation for joint refinement.

Similarly, a high standard deviation for the joint refinement allows to search a large 3D space for each joint. If it is too large, the joints might move to the wrong position.

### 4.5 Runtime Analysis

In Table 4 we compare the runtime of the approach [9] with our approach on the same machine, using an Intel Core i7-7700 3.60GHz and a Nvidia GeForce 1080ti. We evaluate the runtime on the Shelf dataset, which uses five cameras and which has 2, 3 or 4 persons in the scene. Both [9] and our method are implemented in Python, utilizing the output of the official OpenPose [17] implementation which processes an image in 35ms. While our approach needs more time than [9] for two actors, the runtime scales better as the number of actors increases. While the runtime of [9] increases quadratically as the number of actors increases, our approach requires 26ms for each additional actor, i.e. the runtime increases linearly as the number of actors increases.

| | 2 Actors | 3 Actors | 4 Actors |
|---|---|---|---|
| Tanke et al. [9] | 0.023s | 0.045s | 0.104s |
| Ours | 0.062s | 0.088s | 0.114s |

Table 4: Time analysis for the Shelf dataset with respect to the number of actors. The time for the prediction and update steps of our method are measured with 300 sampled 3D poses per person and 50 sampled points for joint refinement.

## 5 Conclusion

In this paper we have presented a novel tracking algorithm based on the well-known recursive Bayesian filtering framework and on recent advancements in 2D human pose estimation and 3D human motion anticipation. Our approach tracks multiple persons, initializes newly appearing persons, and recovers occluded joints. Our approach achieves state-of-the-art results for 3D human pose estimation as well as for 3D human pose tracking. In the future our approach could be extended using an appearance model similar to [5]. Furthermore, we could include a smoothing step which would improve 3D pose predictions backwards through time, utilizing the model uncertainty.

## Acknowledgment

The work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 – 390732324, GA 1927/8-1, and the ERC Starting Grant ARCA (677650).

## References

1. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: Conference on Computer Vision and Pattern Recognition. (2011)
2. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. Transactions on Pattern Analysis and Machine Intelligence (2013)
3. Ershadi-Nasab, S., Noury, E., Kasaei, S., Sanaei, E.: Multiple human 3d pose estimation from multiview images. Multimedia Tools and Applications (2018)
4. Zhang, Y., An, L., Yu, T., Li, X., Li, K., Liu, Y.: 4d association graph for realtime multi-person motion capture using multiple video cameras. In: Conference on Computer Vision and Pattern Recognition. (2020)
5. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3d pose estimation from multiple views. In: Conference on Computer Vision and Pattern Recognition. (2019)
6. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Conference on Computer Vision and Pattern Recognition. (2014)

7. Belagiannis, V., Wang, X., Schiele, B., Fua, P., Ilic, S., Navab, N.: Multiple human pose estimation with temporally consistent 3d pictorial structures. In: European Conference on Computer Vision. (2014)
8. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. Transactions on Pattern Analysis and Machine Intelligence (2016)
9. Tanke, J., Gall, J.: Iterative greedy matching for 3d human pose tracking from multiple views. In: German Conference on Pattern Recognition. (2019)
10. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: Conference on Computer Vision and Pattern Recognition. (2013)
11. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: British Machine Vision Conference. (2013)
12. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: International Conference on Computer Vision. (2015)
13. Särkkä, S.: Bayesian filtering and smoothing. Cambridge University Press (2013)
14. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Conference on Computer Vision and Pattern Recognition. (2017)
15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. (2016)
16. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European conference on Computer Vision. (2018)
17. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Conference on Computer Vision and Pattern Recognition. (2017)
18. Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B.: Arttrack: Articulated multi-person tracking in the wild. In: Conference on Computer Vision and Pattern Recognition. (2017)
19. Iqbal, U., Milan, A., Gall, J.: Posetrack: Joint multi-person pose estimation and tracking. In: Conference on Computer Vision and Pattern Recognition. (2017)
20. Doering, A., Rafi, U., Leibe, B., Gall, J.: Multiple human pose estimation with temporally consistent 3d pictorial structures. In: European Conference on Computer Vision. (2020)
21. Doering, A., Iqbal, U., Gall, J.: Joint flow: Temporal flow fields for multi person tracking. British Machine Vision Conference (2018)
22. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: International Conference on Computer Vision. (2017)
23. Iqbal, U., Doering, A., Yasin, H., Krüger, B., Weber, A., Gall, J.: A dual-source approach for 3d human pose estimation from single images. Computer Vision and Image Understanding (2018)
24. Iqbal, U., Molchanov, P., Breuel Juergen Gall, T., Kautz, J.: Hand pose estimation via latent 2.5 d heatmap regression. In: European Conference on Computer Vision. (2018)
25. Kostrikov, I., Gall, J.: Depth sweep regression forests for estimating 3d human pose from images. In: British Machine Vision Conference. (2014)
26. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: International Conference on 3D Vision. (2018)

27. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., et al.: Panoptic studio: A massively multiview system for social interaction capture. Transactions on Pattern Analysis and Machine Intelligence (2017)
28. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Conference on Computer Vision and Pattern Recognition. (2018)
29. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. International Journal of Computer Vision (2005)
30. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. International Journal of Computer Vision (2010)
31. Yao, A., Gall, J., Gool, L.V., Urtasun, R.: Learning probabilistic non-linear latent variable models for tracking complex activities. In: Advances in Neural Information Processing Systems. (2011)
32. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. Transactions on Graphics (2016)
33. Bütepage, J., Black, M.J., Kragic, D., Kjellstrom, H.: Deep representation learning for human motion prediction and classification. In: Conference on Computer Vision and Pattern Recognition. (2017)
34. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Conference on Computer Vision and Pattern Recognition. (2016)
35. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: International Conference on Computer Vision. (2015)
36. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. (2016)
37. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. Transactions on Pattern Analysis and Machine Intelligence (2019)
38. Muñoz-Salinas, R., Medina-Carnicer, R., Madrid-Cuevas, F.J., Carmona-Poyato, A.: Particle filtering with multiple and heterogeneous cameras. Pattern Recognition (2010)
39. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. Transactions on Pattern Analysis and Machine Intelligence (2014)
40. : CMU Mocap Database. http://mocap.cs.cmu.edu/ (0)
41. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. Pattern Analysis and Machine Intelligence (2007)
42. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: International Conference on Computer Vision. (2015)
43. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. Transactions on Pattern Analysis and Machine Intelligence (2015)
44. Bernardin, K., Elbs, A., Stiefelhagen, R.: Multiple object tracking performance metrics and evaluation in a smart room environment. In: International Workshop on Visual Surveillance. (2006)
45. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. (2014)