A Quantitative Evaluation of Video-based 3D Person Tracking

Alexandru O. Bălan Leonid Sigal Michael J. Black
Department of Computer Science, Brown University, Providence, RI 02912, USA
{alb, ls, black}@cs.brown.edu

Abstract

The Bayesian estimation of 3D human motion from video sequences is quantitatively evaluated using synchronized, multi-camera, calibrated video and 3D ground truth poses acquired with a commercial motion capture system. While many methods for human pose estimation and tracking have been proposed, to date there has been no quantitative comparison. Our goal is to evaluate how different design choices influence tracking performance. Toward that end, we independently implemented two fairly standard Bayesian person trackers using two variants of particle filtering and propose an evaluation measure appropriate for assessing the quality of probabilistic tracking methods. In the Bayesian framework we compare various image likelihood functions and prior models of human motion that have been proposed in the literature. Our results suggest that in constrained laboratory environments, current methods perform quite well. Multiple cameras and background subtraction, however, are required to achieve reliable tracking suggesting that many current methods may be inappropriate in more natural settings. We discuss the implications of the study and the directions for future research that it entails.

1. Introduction

The recovery of human pose and motion from video sequences has improved dramatically in the last five years. In particular, a variety of Bayesian methods have been developed for recovering 3D human pose [7, 13, 15, 16, 18]. Each of these methods makes different modeling choices regarding the formulation of a likelihood term and an *a priori* probability term used in the Bayesian model. These methods also vary in how they perform inference. To date, no quantitative results have been presented and each group pursuing this problem has used different image sequences. As a result, it has been impossible to compare the methods quantitatively or even to determine why one method might work better than another. To address this problem, we undertake the first quantitative evaluation of current human tracking formulations. In doing so we develop a novel set of evaluation data with ground truth human motion and a set of evaluation measures for comparing the accuracy of human motion trackers. This paper presents a quantitative analysis of current methods with the goal of teasing apart which modeling choices are important and where fundamental problems remain. Rather than developing new algorithms for 3D human motion estimation, this paper evaluates current methods and quantitatively compares different design choices. A full review of the literature is outside the scope of this paper; rather, here we focus on a representative class of Bayesian methods that use particle filtering [7, 15, 10]. For a more detailed review of the literature, see [8, 11]. The paper also presents a novel human motion dataset, baseline software, and evaluation measures that are being made available to the research community in the interests of encouraging quantitative evaluation and comparison[1].

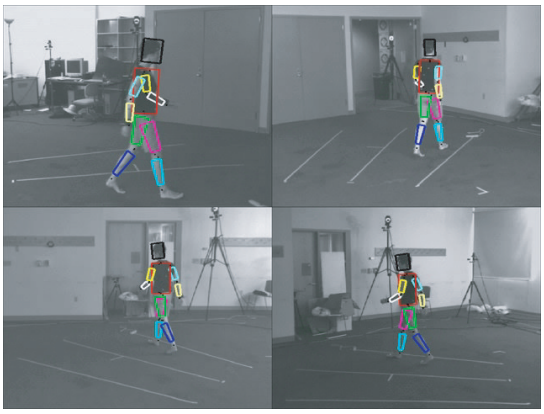


Figure 1: **Example Ground Truth Data.** The 3D body model is obtained by a commercial Vicon motion capture system. This 3D body model is shown here projected into four calibrated camera views.

Quantitative evaluation of human motion recovery requires image sequences with ground truth. We describe a novel facility that combines synchronized video capture with a commercial motion capture system. The captured data includes training data consisting of various motion capture sequences and testing data that includes multi-camera calibrated video data and the associated ground truth body poses (Figure 1).

[1] The data and software used in these experiments are available by writing to the authors.

No commonly accepted methods for quantitative comparison of video-based motion capture results have been established. We explore a number of possible measures based on the 3D Euclidean distance of various marker locations. There are a number of important issues that must be considered in choosing a measure that is "algorithm neutral".

For comparison purposes we implemented a generic particle filter framework for Bayesian tracking. We tested two variations corresponding to traditional particle filtering [10, 15] and annealed particle filtering [7]. Beyond the choice of inference method, the success of any Bayesian approach lies in the choice of likelihood and prior. Here we implemented three commonly used likelihoods that exploit background subtraction, image gradients, and Chamfer distance to edges. We also explored two priors, one that uses no temporal information [7] and one that assumes constant velocity [13].

The following Methods section describes the ground truth data collection, the Bayesian formulation of the tracker, and the evaluation measures. Tracking results are then presented that explore the space of design choices and provide insight into tracking failures. A discussion follows in which we summarize the state of the field and suggest the areas that need the most attention. In particular we suggest that in controlled laboratory environments with three or more calibrated cameras, good lighting, stationary backgrounds, a single subject, and only self-occlusion, current methods work remarkably well. While constrained, this is exactly the kind of environment in which current commercial, marker-based, systems operate. We found that both standard and annealed particle filters worked well in practice. While the annealed filter was more accurate it has a significant failure mode when dealing with ambiguous data. The experimental results suggest that much of the success of current methods is due to good background subtraction information. This seriously limits the applicability of these methods outside the controlled setting.

2. Methods

2.1 Ground Truth Data

To evaluate video-based human tracking we built an experimental environment for capturing 3D human motion and synchronized video imagery simultaneously. Ground truth motion is captured by a commercial Vicon system (Vicon Motion Systems Ltd, Lake Forest, CA) that uses reflective markers and six 1M-pixel cameras to recover the three-dimensional pose and motion of human subjects. Video data is captured simultaneously from four Pulnix TM6710 cameras (JAI Pulnix, Sunnyvale, CA). These are grayscale progressive scan cameras with a resolution of 644×488 pixels and a frame rate of 120Hz (though to achieve better image quality we captured video at 60Hz). Video streams are captured and stored to disk in real-time using a custom PC-based system built by Spica Technologies (Maui, HI). The Vicon system is calibrated using Vicon's proprietary software while the video cameras are calibrated using the Camera Calibration Toolbox for Matlab [2]. Offline, the coordinate frames of the two systems are aligned and temporal synchronization is achieved by tracking visible markers in both systems. Figure 1 shows the 3D body model captured by the Vicon system projected into four calibrated camera views. As is common in the literature, the body is modeled as a 3D kinematic tree of truncated cones with 31 parameters comprising the position and orientation of the torso and the relative joint angles between limbs.

To simultaneously capture video and pose our subjects wore "street clothes" on which we attached standard retro-reflective markers. These markers occupy an insignificant portion of the visible image (see Figure 4 **(a)**) and, consequently, their presence is unlikely to impact video-based tracking performance.

Using standard marker-based motion capture protocols, subjects were measured and a 3D body model was fit to these measurements. Subjects performed a variety of common motions in a $3 \times 3 \times 2\,m^3$ viewing volume. The database currently contains 3 subjects and approximately 3 minutes of calibrated video. Additional motion capture data (without video) is available for modeling human motions. To allow training and evaluation the database is divided into separate training and testing sets.

2.2 Bayesian Filtering Formulation

We pose the tracking problem as one of estimating the *posterior* probability distribution $\mathcal{P}_t^+ \equiv p(\mathbf{x}_t|\mathbf{y}_{1:t})$ for the state $\mathbf{x}_t$ of the human body at time t given a sequence of image observations $\mathbf{y}_{1:t} \equiv (\mathbf{y}_1, \ldots, \mathbf{y}_t)$. Assuming a first-order Markov process $(p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}))$ with a sensor Markov assumption $(p(\mathbf{y}_t|\mathbf{x}_{1:t}, \mathbf{y}_{1:t-1}) = p(\mathbf{y}_t|\mathbf{x}_t))$, a recursive formula for the posterior can be derived:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}. \quad (1)$$

The integral in Eq. 1 computes the *prediction* $\mathcal{P}_t^-$ using the previous posterior and the *temporal dynamics* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. The prediction is weighted by the *likelihood*, $p(\mathbf{y}_t|\mathbf{x}_t)$, of the new image observation conditioned on the pose estimate.

Non-parametric approximate methods represent distributions by a set of N random samples or particles with associated normalized weights $\{\mathbf{x}_t^{(i)}, \pi_t^{(i)}\}_{i=1}^N$, which are propagated over time using temporal dynamics and assigned new weights according to the likelihood function. This is the

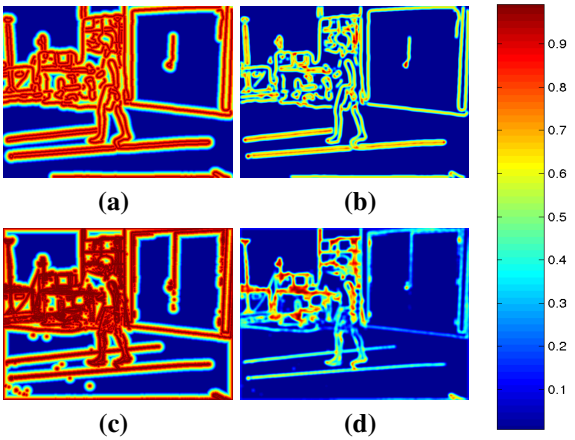


Figure 2: **Example Edge Maps.** **(a)** Canny edges using Chamfer distance (CC); **(b)** Smoothed Canny edges (CS); **(c)** Gradient edges using Chamfer distance (GC); **(d)** Smoothed Gradient edges (GS). They have been remapped between 0 and 1 to represent edge probability.

basis of the Sequential Importance Resampling (SIR) algorithm, or Condensation [1, 10]. A variation to SIR is given by the Annealed Particle Filter (APF) introduced for human tracking by Deutscher *et al.* [7]. This approach searches for peaks in the posterior distribution using simulated annealing. One characteristic of this process however is that it tends to concentrate all of the particles into one mode of the distribution rather than representing the full posterior, thus deviating from the formal Bayesian framework. In this paper we consider both SIR and APF.

2.2.1 Likelihoods

For each particle a likelihood measure needs to be computed that will estimate how well the projection of a given body pose fits the observed image. The most common approaches use edges and silhouettes.

Edge-based Likelihood Functions. We explore two ways of detecting edges in images. The first uses image gradients that have been thresholded to obtain binary maps [7] while the second uses *Canny edges* [4]. An edge distance map is then constructed for each image to determine the proximity of a pixel to an edge. Again we consider two alternatives. The first involves convolving the binary edge map with a Gaussian kernel [7], while the second computes a robust Chamfer distance from each pixel to the closest edge [9]. The four types of edge maps are shown in Figure 2. The log-likelihood is then estimated by projecting into the image sparse points along the edges of all cylinders of the model and computing the mean square error (MSE) of the edge map responses: $\log p(\mathbf{y}_t|\mathbf{x}_t) \propto \frac{1}{|\{\xi\}|}\sum_{\xi}\left(1-M(\xi)\right)^2$, where $\{\xi\}$ is the set of projected points and M is the distance map. The reader is referred to [7] for a more detailed discussion.

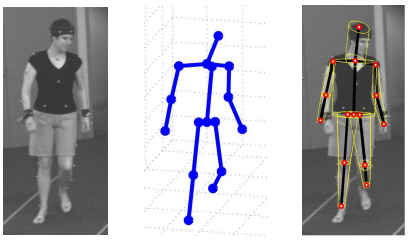
Figure 3: **Example Silhouette Maps.** In realistic scenes, silhouettes are rarely perfect.

Silhouette-based Likelihood Function. Silhouette maps have been generated by learning a Gaussian mixture model for each pixel over 1000 background images and comparing the background pixel probability to that of an uniform foreground model (Figure 3). The log-likelihood of a pose is then estimated by taking a number of visible points on each limb and projecting them into the image. The MSE between the predicted and observed silhouette values for these points is computed [7].

2.2.2 First- and Second-Order Stochastic Dynamics

Predictions from the posterior are made using temporal models. The simplest model applicable to generic motions assumes no dynamics (first-order): $\mathbf{x}_t^- = \mathbf{x}_{t-1}$ [7], while a more specific model for smooth motions assumes constant angular velocity. Although it violates the first-order Markov assumption we take the fairly standard approach and implement the second-order model by augmenting the state at time t with the previous state; that is, $\mathbf{x}_t^* \equiv [\mathbf{x}_t, \mathbf{x}_{t-1}]^T$ [13]. The prediction is obtained by $\mathbf{x}_t^- = 2\mathbf{x}_{t-1} - \mathbf{x}_{t-2}$. In both cases the predictions are diffused using normally distributed random noise to account for uncertainties. The noise is drawn from a Gaussian with diagonal covariance where the standard deviation of each body angle equal to the maximum absolute inter-frame angular difference [7].

In comparison with the work of Deutscher *et al.* [7], our training set appears more restrictive and focuses primarily on walking motions. This likely results in smaller standard deviations for most joints and this, in turn, restricts the practical size of the state spacing making particle filtering methods more effective. In addition, we implemented a hard prior that eliminates any particle corresponding to implausible body poses to reduce the search space. In particular, we check for angles exceeding anatomical joint limits and for inter-penetrating limbs [18]; these changes to [7] significantly improve tracking performance for the standard particle filter.



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

Figure 4: **Markers. (a)** VICON markers; **(b)** Body model inferred; **(c)** 15 joint locations used for computing $3D$ error.

2.3 Error Measures

2.3.1 3D Error for a Single Particle

The appropriate error measure to evaluate the goodness of a pose estimate may vary with the application domain and the tracking algorithm. Here we concentrate on a 3D error measure and consider 15 virtual markers $m \in \mathcal{M}$ placed on the body: one for pelvis, neck and head, and two for shoulders, elbows, wrists, hips, knees and ankles (Figure 4(c)).

For each particle $\mathbf{x}_t^{(i)}$, the full pose error $\delta(\mathbf{x}_t^{(i)}, \tau_t)$ is computed as the average distance in millimeters of all virtual markers $m \in \mathcal{M}$ with respect to the true pose τ_t

$$\delta(\mathbf{x}_t^{(i)}, \tau_t) = \frac{\sum_{m \in \mathcal{M}} ||m(\mathbf{x}_t^{(i)}) - m(\tau_t)||}{|\mathcal{M}|} \qquad (2)$$

where $m(y)$ returns the 3D location of marker m for the body model y. We also compute the individual error for the pelvis and head. Individual errors for the lower and upper arms and legs are averaged over the left and right sides of the body. Please note that we do not consider an error measure directly based on joint angles deviations from the true pose. Computing average errors in angle representations is complicated by the presence of multiple solutions that give rise to the same pose.

2.4. Posterior Distribution Error

When evaluating Bayesian tracking methods there are a number of issues to consider. The goal is to be able to compare algorithms even though they may have different representations of the posterior distribution. A Kalman filter method [1] maintains a uni-modal distribution parameterized by its mean and covariance. Traditional particle filtering methods such as SIR use point masses sparsely distributed in a high dimensional space. Annealed particle filtering tends to concentrate its particles into one very narrow and peaked mode.

Not having access to the true posterior, we can only hope to see a high probability for the true pose and low probabilities everywhere else. However, one of the noted advantages of particle filtering is the fact that it can represent inherent ambiguities by maintaining multiple modes. Such an algorithm should not be penalized for having non-zero probability in regions that are far from the true pose as long as the true pose is "well represented" by the posterior.

We will now discuss a number of error measurement choices for the posterior error with respect to the true pose τ_t. We use $\Delta_\gamma(\mathcal{P}_t^+, \tau_t)$ to denote the error of the posterior $\mathcal{P}_t^+$, where γ represents different choices.

The most obvious choice is to sample from the posterior and compute the average error Δ_W over the sampled poses. For particle methods, when the number of samples is sufficiently large, this error converges to a weighted average over the particles errors (Eq. 3) where the error of each particle is weighted by its normalized likelihood $\pi_t^{(i)}$:

$$\Delta_W(\mathcal{P}_t^+, \tau_t) = \sum_i \pi_t^{(i)} \cdot \delta(\mathbf{x}_t^{(i)}, \tau_t). \qquad (3)$$

Algorithms that maintain a wider posterior will score worse under this measure even with the mode in the right place. This measure will favor for instance APF over SIR, since APF has narrow peaks. It also causes the error to have a positive lower bound that varies with the width of the posterior distribution. Hence this measure depends on input parameters and makes it hard to compare algorithms.

One may also compute the error of the expected pose

$$\Delta_E(\mathcal{P}_t^+, \tau_t) = \delta\left(\sum_i \pi_t^{(i)} \cdot \mathbf{x}_t^{(i)}, \tau_t\right). \qquad (4)$$

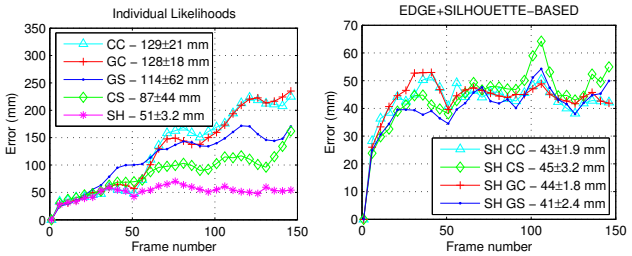
When the posterior distribution is multi-modal, the expected (mean) particle may fall in between modes and therefore provide a poor approximation of the posterior error by itself.

Alternatively, we can estimate the error of the most likely pose in the posterior distribution. For methods using particles, the MAP estimate is approximated by the particle with the largest weight

$$\Delta_{MAP}(\mathcal{P}_t^+, \tau_t) = \delta(\mathbf{x}_t^{(j)}, \tau_t), \quad \pi_t^{(j)} = \max_i(\pi_t^{(i)}). \qquad (5)$$

This measure can be unfair for particle filtering methods since the posterior probability is a function of both the likelihood weights and the density of the particles. The MAP estimate may fail to evaluate the error in regions that have a high posterior probability due to high particle density but small individual weights.

Alternatively, we propose a measure Δ_R (Eq. 6) that is appropriate for comparing the results of algorithms that use Monte Carlo sampling in high dimensions. We use an *optimistic* error that is the minimum error over any pose that is present in the sampled posterior. More specifically, we take the minimum error over all particles regardless of their



(a) **(b)**

Figure 5: **Different Likelihood Functions.** SH $\rightarrow$ silhouette maps; $\mathcal{XY}$ $\rightarrow$ edge maps, where $\mathcal{X} \in \{$**C**anny,**G**radient$\}$ and $\mathcal{Y} \in \{$**C**hamfer,**S**moothed$\}$; **(a)** Likelihood using only one edge or silhouette map; **(b)** Likelihood based on the silhouette map plus one edge map.

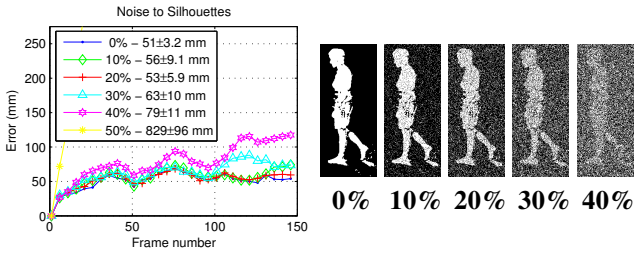


Figure 6: **Noisy Silhouettes.** B^* (with no edge likelihood term) performance degrades as silhouette noise increases. Noise level is given as a percentage of corrupted pixels (50% indicates random binary noise).

weight. We also include the error of the expected particle which may not be in the particle set

$$\Delta_R(\mathcal{P}_t^+, \tau_t) = \min \left\{ \begin{array}{l} \Delta_E(\mathcal{P}_t^+, \tau_t), \\ \min_i \left(\delta(\mathbf{x}_t^{(i)}, \tau_t) \right). \end{array} \right. \quad (6)$$

This error measurement allows fair comparison between different algorithms by providing a lower bound on the error. This measure is only informative in high dimensional problems such as body tracking where the samples in the posterior cover a small part of the search domain. In this case, it is unlikely for the true pose to be sampled by chance.

For methods that have a parametric representation (e.g. the Kalman filter) or that use a large number of particles relative to the size of the state space, we can estimate the optimistic error $\widehat{\Delta}_R$ by considering only a small set of $n \ll N$ samples drawn according to the posterior distribution (e.g. Monte Carlo sampling). Whenever possible, sampling should be done without replacement for a better error estimate

$$\widehat{\Delta}_R(\mathcal{P}_t^+, \tau_t) = \min \left\{ \begin{array}{l} \Delta_E(\mathcal{P}_t^+, \tau_t), \\ \min_{x_t^{(j)} \sim \mathcal{P}_t^+, 1 \leq j \leq n} \left(\delta(\mathbf{x}_t^{(j)}, \tau_t) \right). \end{array} \right. \quad (7)$$

3. Experiments

We have run our experiments on a portion of a circular walking sequence that contains a full 180° turn. Fixing the number of likelihood evaluations at 1000 per frame, the best results were obtained using annealed filtering with 200 particles, 5 layers of annealing, a likelihood based on silhouettes and smoothed gradients for edges, no temporal dynamics, and discarding particles corresponding to infeasible body poses (penetrating limbs or joint angles outside of the allowed range). We refer to this combination of tracking parameters as the *base configuration* B^*. The computation time is about 45 seconds per frame on a standard PC with software written in Matlab. The tracker has been initialized with the ground truth. We have evaluated the results using the measure from Eq. 7, sampling 10 different particles from the posterior. We performed 10 trials of each experiment. We plot the mean error at each frame and compute an average error over all 150 frames. We used a t-test to formally compare tracking parameter choices; hence, we report the mean and standard deviation of the average error of each method[2]. Tracking results are shown in Figure 11.

Comparing Likelihood Functions. First we ran B^* using either an edge- or a silhouette-based likelihood. Figure 5(a) shows that likelihoods using silhouettes are more powerful than edges, confirmed by the t-test. Moreover, smoothed Canny edges were statistically superior to any edge map using Chamfer distance, while any other pair was statistically insignificant. Combining silhouette maps with edge maps [7] improves tracking, but the choice of any particular edge map is not statistically significant.

We added binary noise to the silhouette maps to evaluate tolerance to poor background subtraction (Figure 6). Here we evaluate B^* without the edge term to isolate the effect of noise on the silhouettes. We observed a statistically significant decrease in performance starting at 30% noise level, however tracking results are remarkably stable even at 40%.

Comparing Temporal Dynamics and Priors. Reducing the domain of the allowable poses by imposing joint angle limits and disallowing limb inter-penetration benefits tracking significantly (Figure 7(a)).

Contrary to our expectations, our implementation using constant angular velocity dynamics performs worse than zero angular velocity (fig 7(b-c)). Moreover, it does not track more than twenty frames without constraining the domain of the allowable poses. We observe that it is more

[2]We used a 2-tail non-directional t-test with a pooled estimate of the standard deviation, having 18df and a 95% significance level ($\alpha = 0.05$).

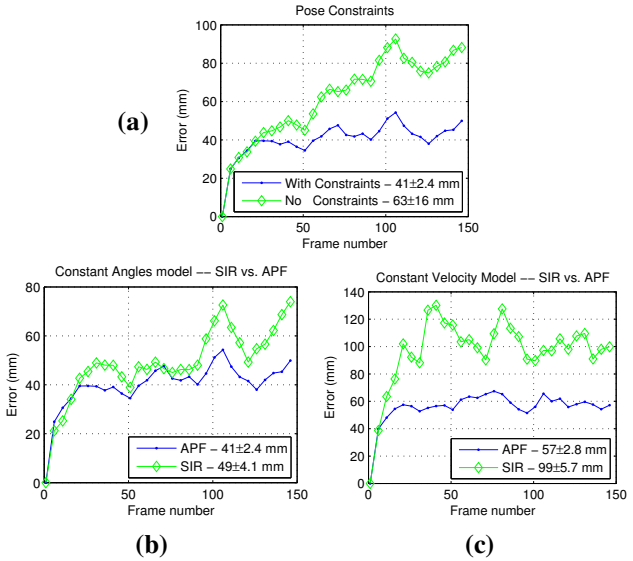


Figure 7: **Priors and Temporal Dynamics.** **(a)** B^* with and without imposing constraints on the joint angles; **(b)** 1^{st}- and **(c)** 2^{nd}-order dynamics; APF corresponds to B^*; SIR runs with 1000 particles; same tracking parameters as B^*.

prone to accumulation of error and conclude that better prior models are needed to overcome this problem.

Comparing Regular and Annealed Particle Filtering. Since the bulk of the computation involves evaluating the likelihood, we keep the number of likelihood evaluations the same when comparing APF and SIR (Figure 7(b-c)). Hence, the number of particles used for SIR (*i.e.* 1000) is the product of the number of layers and the number of particles per layer in the annealed method. Based on a 1-tailed t-test, APF performs significantly better than SIR at 95% confidence level, regardless of the prior model used. Effectively however, the difference in performance is not great.

Varying the Number of Particles. We observe in Figure 8(b) a linear trend on the log scale of the number of particles with respect to performance error of the APF. This suggests that the number of particles needed to linearly improve the accuracy of particle filters grows exponentially.

Varying the Number of Camera Views. Figure 9 shows that at least three cameras are necessary for tracking, and a forth one does not significantly improve tracking accuracy. It is worth noting however that in our setup, cameras 1 and 4 are facing each other and therefore the silhouettes are expected to be almost the same albeit reflected (see Figure 1). Tracking fails on average after 65 frames using two cameras, while monocular tracking fails after 40.

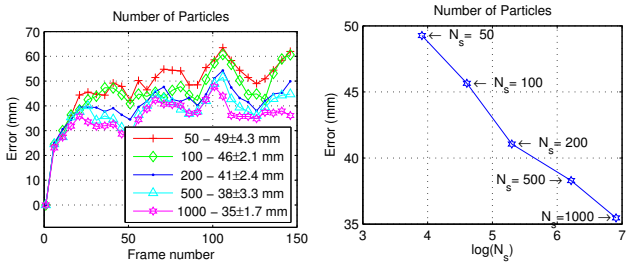


Figure 8: **Different Number of Particles.** **(a)** Performance results using B^* with different number of particles N_s; **(b)** Results plotted on a log scale of the number of particles exhibit a linear relationship.

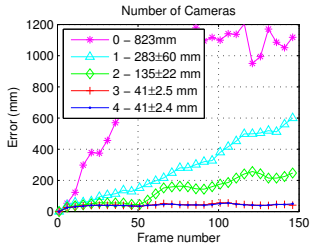


Figure 9: **Different Number of Cameras.** Zero cameras means a uniform likelihood function was used. We are not able to reliably track with B^* more than 65 frames using less than three camera views.

Error for Limbs. Body extremities such as lower arms and lower legs are the hardest body parts to track while the pelvis is tracked fairly consistently (Figure 10(a)).

4 Discussion and Future Work

While we treat the Vicon data as the "ground truth" it is worth noting that the "true" human motion is somewhat elusive. Our model of the body is only an approximation to the the human body (though it is fairly typical of the state of the art). Additionally, the synchronization between the motion capture and the video is estimated from data and likely has estimation errors that are difficult to quantify. Finally, while the marker locations are estimated to millimeter accuracy they may move relative to the rigid structures (bones) of the body and hence, even the highest quality motion capture data can only provide an approximation to the true limb locations.

Here we made the common assumption that the body can be represented by rigid limbs (e.g. truncated cones) connected by revolute joints. We make no attempt to fit the limbs shape to the image measurements. Note that Smichisescu and Triggs [18] use superquadrics which may fit the observations better. More generally, one may want to fit a

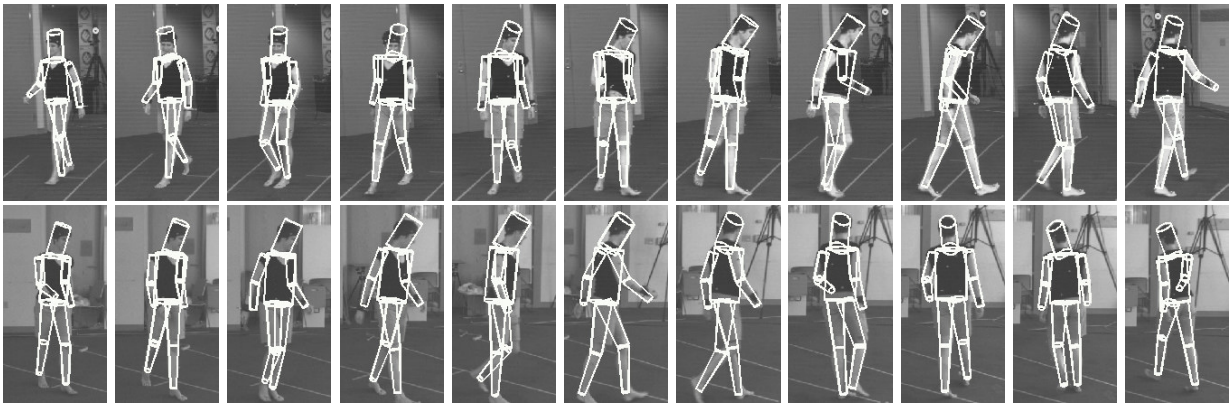


Figure 11: **Tracking Results using B^*.** Every 15^{th} frame from cameras 2 (top) and 3 (bottom).

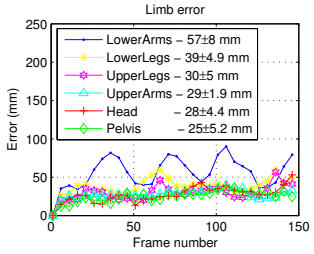


Figure 10: **Individual Limb Error.** Tracking using B^*.

deformable limb model to the data [12]. Evaluating whether this improves tracking would be interesting future work.

Our evaluation measures are appropriate for the 3D motion capture task but may be less appropriate for methods that infer 3D pose from monocular images. In such situations, ambiguities in depth may make the 3D solution very bad (under our measures) while the joint angles or the image projection of the 3D points are actually quite good. Consequently it is worth developing additional measures for these monocular cases.

One of the main conclusions of this study is that current methods rely heavily on background subtraction. Even when systems combine a variety of cues with background subtraction, these other cues may be serving a secondary role. We found that edge information helps in the precise localization of limbs (i.e. they improve accuracy) but alone the tested edge likelihoods are not sufficient for reliable tracking. We also note that the exact form of the edge term did not significantly change the results.

One of the key problems facing human tracking then is the formulation of reliable image likelihood models that apply across a wide range of imaging conditions and do not rely on the knowledge of static backgrounds. Other likelihoods have been proposed for human tracking and should be evaluated in our framework. For example, more principled edge likelihoods have been formulated using phase information [13] and the learned statistics of filter responses [14]. Non-edge-based methods include optical flow [3, 15, 19, 20], flow occlusion/disocclusion boundaries [18], image templates [5], and principal component-based models of appearance [17]. The effect of the prior should also be explored further, particularly in the case of monocular tracking where strong priors [16] are likely to be important in improving tracking results.

Further analysis should consider the importance of the optimization method beyond the two explored here. For example, experiments with hybrid Monte Carlo sampling [6], partitioned sampling [7], or covariance-scaled sampling [18] should be pursued.

Note that most models have parameters that must be tuned; for example, limb lengths, likelihoods, priors, and details of the optimization methods such as the number of particles. The existence of ground truth allows these parameters to be set in a principled way by optimizing inference over the training data.

While the annealed particle filter consistently outperformed the standard filter the practical difference was slight. This may be due to the constrained training data which limits the range of human motions. It has been observed that particle filtering can perform well, even with high-dimensional body models, when strong priors are enforced to focus the particles to valid regions of the space [15]. Our results are contrary to those reported by Deutscher *et al.* [7] where the particle filter performed much worse than the annealed filter. Again, we posit that this is due to their use of a much broader prior model. In summary: good image data means that one can use a weak prior and a simple algorithm (SIR). When the data becomes less rich or the prior less constraining the algorithm must work harder (APF).

One problem with the annealed approach emerged in the experiments. When silhouette data was ambiguous or noisy the annealed particle filter sometimes got "stuck" in the wrong interpretation. The annealing forces the method

to represent one of the modes in the data. This is exactly the wrong behavior when the interpretation of the data is ambiguous and this is one of the reasons standard particle filters have become popular. Thus in choosing a particular method one must consider the quality of the input data. In a controlled motion capture laboratory, the annealed particle filter works well but in a less constrained environment a more general filter that can model ambiguity may be needed (along with a better prior as suggested above).

While this study suggests current methods are quite reliable in controlled settings we are unable to predict when marker-less motion capture will be a commercially viable alternative to marker-based systems. If greater accuracy is required, one would want to use cameras of higher image quality and higher accuracy however we have not studied the effect of image resolution on accuracy and can make no predictions regarding what resolution might be needed in practice. The limiting issue would appear to be the computational challenge of processing full video streams (rather than simple marker positions). Current commercial systems run in real-time while any vision-based method based on particle filtering is currently far from real-time (the unoptimized Matlab method here takes 45 seconds/frame).

5. Conclusions

We have presented the first quantitative evaluation of Bayesian methods for the 3D tracking of humans in video. In particular we explored the effect of various likelihood terms (using background subtraction and edge measures) and prior models. Additionally we compared standard particle filtering with annealed particle filtering. We found that in the case of three or more cameras that both optimization methods worked well and that the key component for successful tracking was good background subtraction.

One of the contributions of this work is the creation of novel ground truth sequences and an evaluation measure for quantitatively comparing methods. The results suggest directions for future work on human tracking. In particular, to move beyond the controlled laboratory environment, better likelihood models are needed that do not rely heavily on known, static backgrounds. Furthermore, coping with monocular sequences will likely require better priors. Here we considered methods that assume manual initialization yet fully automatic systems are necessary to recover from tracking failures. Despite the limitations of current methods, the accuracy and reliability is such that an engineering effort to build a highly accurate markerless motion capture system seems within reach.

Acknowledgments. This work was supported by a gift from Intel Corporation. We thank Michelle Lee, Johnathan Bankard and David Erickson for their assistance in the motion capture lab.

References

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Sig. Proc.*, 50(2):174–188, 2002.

[2] J.-Y. Bouguet. Camera calibration toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/.

[3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, pp. 8–15, 1998.

[4] J. Canny. A computational approach to edge detection. *PAMI*, 8(6):679–698, 1986.

[5] T-J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. *CVPR*, v. 1, pp. 239–245, 1999.

[6] K. Choo and D. Fleet. People tracking using hybrid monte carlo filtering. *ICCV*, v. 2, pp. 321–328, 2001.

[7] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2004.

[8] D. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[9] D. Gavrila and L. Davis. 3-D model-based tracking of humans in action: A multi-view approach. *CVPR*:73–80, 1996.

[10] M. Isard and A. Blake. Condensation – Conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.

[11] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, 18:231–268, 2001.

[12] R. Plänkers and P. Fua. Articulated soft objects for video-based body modeling. *ICCV*, v. 1, pp. 394–401, 2001.

[13] E. Poon and D. Fleet. Hybrid Monte Carlo filtering: Edge-based people tracking. *IEEE Workshop on Motion and Video Computing*, pp. 151–158, 2002.

[14] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. *IJCV*, 54(1–3):183–209, 2003.

[15] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, v. 2, pp. 702–718, 2000.

[16] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *ECCV*, pp. 784–800, 2002.

[17] H. Sidenbladh, F. De la Torre, and M. Black. A framework for modeling the appearance of 3D articulated figures. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 368–375, 2000.

[18] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6), 371–391, 2003.

[19] S. Wachter and H. Nagel. Tracking of persons in monocular image sequences. *CVIU*, 74(3):174–192, 1999.

[20] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki. Incremental tracking of human actions from multiple views. *CVPR*, pp. 2–7, 1998.