A comparison of 3d model-based tracking approaches for human motion capture in uncontrolled environments.

Mohammed Shaheen[1], Juergen Gall[2], Robert Strzodka[1], Luc Van Gool[2,3], and Hans-Peter Seidel[1]
[1]MPI Informatik, Germany [2]BIWI, ETH Zurich, Switzerland [3]IBBT, ESAT-PSI, K.U.Leuven, Belgium
{mshaheen,strzodka,hpseidel}@mpi-inf.mpg.de {gall,vangool}@vision.ee.ethz.ch

Abstract

This work addresses the problem of tracking humans with skeleton-based shape models where video footage is acquired by multiple cameras. Since the shape deformations are parameterized by the skeleton, the position, orientation, and configuration of the human skeleton are estimated such that the deformed shape model is best explained by the image data. To solve this problem, several algorithms have been proposed over the last years. The approaches usually rely on filtering, local optimization, or global optimization. The global optimization algorithms can be further divided into single hypothesis (SHO) and multiple hypothesis optimization (MHO). We briefly compare the underlying mathematical models and evaluate the performance of one representative algorithm for each class. Furthermore, we compare several likelihoods and parameter settings with respect to accuracy and computation cost. A thorough evaluation is performed on two sequences with uncontrolled lighting conditions and non-static background. In addition, we demonstrate the impact of the likelihood on the HumanEva benchmark. Our results provide a guidance on algorithm design for different applications related to human motion capture.

1. Introduction

Markerless human motion capture has been studied since more than 25 years and is still a very active research area in computer vision. In this work, we concentrate on tracking algorithms for skeleton-based shape models where surface models of humans with underlying skeletons are used to find a sequence of poses that fits the image data best. For a complete overview of markerless human motion capture, we refer to the surveys [19, 20]. In contrast to the surveys, we perform a quantitative evaluation and compare the mathematical models on which filtering *(Filter)*, single hypothesis optimization *(SHO)*, and multiple hypothesis optimization *(MHO)* are based. For the experimental evaluation, one representative algorithm is selected for each class:

Figure 1. Two successive frames of a multi-view video sequence with low contrast, rapidly changing illumination, and people in the background.

- *Filter:* particle filter [21],
- *SHO:* local optimization based on twists and closest points (local) [6], fast simulated annealing (global) [10],
- *MHO:* annealed particle filter (local) [11], interacting simulated annealing (global) [13].

In addition, we evaluate several likelihoods and parameter settings with respect to accuracy and computation cost. This offers further trade-offs to tune for one or the other in order to match the needs of the application.

Most of the public available datasets as [1, 22] have been recorded in a controlled studio environment even though controlled lighting conditions or static background are not given for many applications. Hence, we perform a thorough quantitative evaluation on two sequences with low contrast, rapidly changing illumination, and people in the

background, see Figure 1. Furthermore, we provide a quantitative comparison of various likelihoods on the HumanEva benchmark [22].

The paper is structured as follows: Section 2 compares the underlying mathematical models for filtering, single hypothesis, and multiple hypothesis optimization and gives a brief overview of algorithms that belong to the three classes. The likelihoods that are used for the comparison are discussed in Section 3. The paper concludes with a brief discussion after the experimental section.

2. Models and Algorithms

In order to compare the different concepts, we use the same Bayesian formulation for optimization and filtering where $x_t \in E$ denotes the state, *i.e.* the position, orientation, and joint angles of the skeleton, and y_t denotes the observed image features at time t. We distinguish between a random variable $X : \Omega \mapsto E$ and its realization $x \in E$. For both optimization and filtering, the same likelihood function $p(Y_t = y_t | X_t = x_t)$ can be used to measure how well the shape model deformed by the skeleton pose x_t explains the image data y_t.

2.1. Filtering

Filtering models the states $X_t, \cdots, X_1$ as hidden random variables. The only observations that are available at frame t are the realizations $y_t, \cdots, y_1$ of the random variables $Y_t, \cdots, Y_1$, *i.e.* the observed image features. Under the Markov property [12], the posterior probability can be written as

$$p(X_t | Y_t = y_t, \cdots, Y_1 = y_1) = \frac{1}{Z} \underbrace{p(Y_t = y_t | X_t)}_{\text{likelihood}} \int_E \underbrace{p(X_t | X_{t-1} = x_{t-1})}_{\text{temporal transitions}} \cdot \ldots \tag{1}$$
$$\ldots \cdot p(X_{t-1} = x_{t-1} | Y_{t-1} = y_{t-1}, \cdots, Y_1 = y_1) dx_{t-1},$$

where Z is a normalization factor. Depending on the model assumptions, the posterior is given in closed-form by a Kalman filter [17] or approximated by a particle filter [16].

For skeleton-based human motion capture, Sidenbladh et al. [21] have combined a particle filter with very strong motion priors to resolve the ambiguities from monocular sequences. Motion priors have been also proposed for a Rao-Blackwellised particle filter [27]. In [24] various variants of particle filters like the unscented particle filter have been evaluated for human motion capture. In [8] hybrid Monte Carlo filtering has been applied where a Markov chain Monte Carlo technique is used within a particle filter to get better samples from the posterior. Another approach follows the idea of search space decomposition where the space is divided into independent low-dimensional subspaces [18].

2.2. Single Hypothesis Optimization

In contrast to filtering, single hypothesis optimization approaches seek for a single pose $x_t \in E$ instead of a random variable:

$$\operatorname*{argmax}_{x_t \in E} p(X_t = x_t | Y_t = y_t, \hat{x}) = \operatorname*{argmax}_{x_t \in E} \underbrace{p(Y_t = y_t | X_t = x_t)}_{\text{likelihood}} \underbrace{p(X_t = x_t | \hat{x})}_{\text{smoothness prior}}. \tag{2}$$

The optimization is usually initialized by a value $\hat{x} \in E$ that is deterministically obtained from the estimate of the previous frame x_{t-1}. The smoothness term penalizes strong deviations from the previous estimate and can be modeled as higher order smoothness term to penalize velocity or acceleration changes. When the smoothness term is uniform, the pose is completely estimated from the current image data y_t. While a good initialization is important for local optimization, global optimization like fast simulated annealing [10] forgets the initialization and converges to the global optimum. For an exhaustive list of local optimization approaches, we refer to the surveys [19, 20].

2.3. Multiple Hypothesis Optimization

Instead of working with a single hypothesis, multiple hypothesis optimization estimates a density and is initialized by a distribution $\hat{p}$. The density can be continuous or discrete, *e.g.* a sum of weighted Dirac measures, $\sum_i w_i \delta_{x_i}$, where each $x_i \in E$ can be regarded as hypothesis. The random variable X with the sought density can be modeled as:

$$X : \Omega \mapsto E \text{ such that} \tag{3}$$
$$\forall \varepsilon > 0 \quad p(X \in B_\varepsilon | Y_t = y_t, \hat{p}) = 1, \tag{4}$$

where

$$B_\varepsilon = \left\{ x_t \in E : f(x_t, y_t, \hat{p}) \geq \max_x f(x, y_t, \hat{p}) - \varepsilon \right\},$$
$$f(x_t, y_t, \hat{p}) = \underbrace{p(Y_t = y_t | X_t = x_t)}_{\text{likelihood}} \underbrace{p(X_t = x_t | \hat{p})}_{\text{smoothness prior}}. \tag{5}$$

The single hypothesis optimization can be regarded as special case where the distribution is modeled by a Dirac measure δ_x. In contrast to filtering, the random variable X relies on the last realization of Y_t and the initial distribution $\hat{p}$. However, $\hat{p}$ is not (1) since this distribution is usually unknown. While global approaches like interacting simulated annealing [13] forget anyway the initialization, local approaches assume that one of the estimated local optima is a global optimum. For instance, covariance scaled sampling [23] guides the samples to the local maxima of a distribution that is modeled by a mixture of Gaussians. Smart

particle filtering [5] follows a similar idea but uses a non-parametric kernel estimator to model the distribution. The annealed particle filter [11] is similar to interacting simulated annealing, but the distribution of the samples in the search space relies on the fluctuating survival rate of the particles.

3. Likelihoods

Aiming at high performance, we investigate low-level features for the likelihood function like silhouettes, edges, color, and an anatomic prior. They are fast to evaluate and their proper combination still achieves high accuracy. We write $V(x)$ for the negative log-likelihood $-\log p(Y_t = y_t | X_t = x)$:

$$\begin{aligned} V(x) =& \lambda_{silh}\, V_{silh}(x) + \lambda_{edge}\, V_{edge}(x) \\ &+ \lambda_{color}\, V_{color}(x) + \lambda_{anat}\, V_{anat}(x), \end{aligned} \tag{6}$$

where the λs control the influence of each term.

3.1. Silhouettes

In order to model the log-likelihood for a pose $x \in E$ and a silhouette image I_v extracted by background subtraction, a template image $T_v(x)$ is generated by projecting the surface of the human model that is translated, rotated, and deformed according to x. The inconsistent areas between the silhouette and the template are then measured for each view v by

$$\begin{aligned} V_v(x) =& \frac{1}{2Z_v^T} \sum_{p \in P_v^T} |T_v(x,p) - I_v(p)| \\ &+ \frac{1}{2Z_v^I} \sum_{p \in P_v^I} |I_v(p) - T_v(x,p)|, \end{aligned} \tag{7}$$

where $I_v(p)$ and $T_v(x,p)$ are the pixel values for a pixel p and P_v^T and P_v^I denote the sets of pixels that are used to sample from. The sums are normalized by the sample sizes Z_v^T and Z_v^I. The function $V_{silh}(x)$ is then given by the average value $V_v(x)$ of all views.

The simplest way to compute (7) is the Hamming distance that can be implemented by an XOR operation. In this case, I_v and $T_v(x)$ are binary images and the sampling is performed on the whole image. The normalization constants Z_v^T and Z_v^I are fixed to $\sum_{p \in P_v^{T^0}} T_v^0(p)$ where T^0 denotes the projected mesh in its default pose, *i.e.* without any deformations. We denote the simple silhouette comparison by V_{silh}.

In order to penalize pixels that are far away from the silhouette more severely, a Chamfer distance transform [4] can be applied to the images. The sampling is then performed on the sets of pixels inside of the silhouettes. The silhouette term based on the Chamfer transform is denoted by V_{silh}^{Ch}.

Since local single hypothesis optimization algorithms often rely on gradients or a Taylor approximation, the types of log-likelihoods given by Equation (7) are not optimal. An alternative are correspondence-based log-likelihoods that minimize the distance error of the correspondences between the silhouette rim $p_{v,i}$ and the vertices of the shape model $V_i(x)$ in the least squares sense [6, 15, 2]:

$$V(x) = \sum_i \| \pi_v(V_i(x)) - p_{v,i} \|_2^2, \tag{8}$$

where π_v is the projection from the world coordinate system onto the image plane for view v. The correspondences can be established by closest points search.

3.2. Edges

The log-likelihood for edges is modeled in a similar way. The Sobel operator is applied to the image I_v to extract an edge map E_v that is compared to the boundaries of the projected body parts denoted by $Edges(T_v(x))$:

$$V_{edge}(x) = \frac{1}{Y} \sum_v \sum_{p \in Edges(T_v(x))} E_v(p), \tag{9}$$

where $Y = |Edges(T_v^0)|$ is a normalization constant. The edge map can be represented by a binary image where pixels on edges are labeled with 0, and 1 otherwise. In contrast to the other features this is a one sided comparison looking for the edges of the projected template in the recorded image but not vice versa. Another variant applies a Chamfer distance transform to E_v. We denote this variant by V_{edge}^{Ch}.

3.3. Color

The color distribution of a channel c and a body part s is modeled by a normalized histogram $H^{(s,c)}$ where we fix the number of bins to $K = 64$. In order to measure deviations of a pose $x \in E$ from the color model given by the histogram $H^{(s,c)}$, the color distribution for x, denoted by $\tilde{H}^{(s,c)}(x)$, is estimated by sampling from all views. For this purpose, the triangles of the human model are used to encode the body parts of the projected surface. Hence, a pixel p that belongs to a body part s contributes for each channel c to the histogram $\tilde{H}^{(s,c)}(x)$. The total deviation is then measured by the Bhattacharya distance:

$$V_{color}(x) = \sum_s \frac{w_s}{C} \sum_{c=1}^{C} \left(1 - \sum_{k=1}^{K} \sqrt{h_k^{(s,c)} \tilde{h}_k^{(s,c)}(x)} \right), \tag{10}$$

where the weights w_s reflect the size of the body parts. To increase the distinctiveness of the color model, we use the CIELab color space that mimics the human perception of color differences. Since the L-channel is very sensitive to illumination changes, we use only the a- and b-channel. Since image noise becomes an important issue for

small body parts, we reduce noise without smoothing over the edges that separate body parts by applying the edge-enhancing diffusivity function [7]

$$g(|\nabla u|^2) = \frac{1}{|\nabla u|^p + \epsilon} \tag{11}$$

with $\epsilon = 0.001$ and $p = 1.5$, where the smoothing is efficiently implemented by the AOS scheme [25].

During tracking, the color model $H^{(s,c)}$ is adapted to the changing appearance. To this end, a normalized histogram $\hat{H}^{(s,c)}$ is generated for an estimated pose by sampling from all views. The update for bin k is then given by

$$\frac{(1-\lambda)M^{(s)}\, h_k^{(s,c)} + \lambda \hat{M}^{(s)}\, \hat{h}_k^{(s,c)}}{(1-\lambda)M^{(s)} + \lambda \hat{M}^{(s)}}, \tag{12}$$

where $M^{(s)}$ and $\hat{M}^{(s)}$ are the sample sizes for the body part s to generate H and $\hat{H}$, respectively. The parameter λ controls the speed of adaptation and the consideration of the sample sizes avoids that the statistics are distorted by a small number of samples, e.g. due to self-occlusions.

3.4. Anatomical Constraints

Anatomical constraints are modeled by the probability of a skeleton configuration p_{anat} that is estimated from 200 training samples y_l taken from the CMU motion database [9]. For efficiency reasons, we regard the probabilities for the three body parts as uncorrelated, *i.e.*

$$V_{anat}(x) = -\frac{1}{3}\log\left(p_{anat}^{head}(x) p_{anat}^{upper}(x) p_{anat}^{lower}(x)\right). \tag{13}$$

The probability for a body part is approximated by a Parzen-Rosenblatt estimator with a Gaussian kernel K:

$$p_{anat}(x) = \frac{1}{L\, h^d} \sum_l K\left(\frac{x - y_l}{h}\right), \tag{14}$$

where the d-dimensional vectors x and y_l contain only the joints for the body part. The bandwidth h is given by the maximum second nearest neighbor distance between all training samples.

4. Implementation Issues

4.1. Setup

The 3d surface models are acquired by a static full-body laser scan. For subject S4 of the `HumanEva-II` dataset [22], the 3d model is already part of the dataset. The surface models are reduced to around 2000 triangles and a skeleton with 30 degrees-of-freedom is inserted into the surface mesh by manually marking the joint positions. For the deformation, we use standard linear blend skinning where the weights for linear blend skinning are extracted automatically [3]. The intial pose for each sequence is estimated with global MHO. For the comparison, the same initial pose is used for all approaches.

4.2. Temporal Transitions and Smoothness Prior

In order to obtain a reasonable model for the temporal transitions, we use a 3rd order autoregression modeled by Gaussian processes [26], *i.e.* the temporal transitions are given by a conditional Gaussian distribution $p(X_t|X_{t-1}, X_{t-2}, X_{t-3})$. We use a uniform smoothness prior for SHO and MHO, *i.e.* the optimization is driven by the likelihood. The initialization value $\hat{x}$ and the distribution $\hat{p}$ are predicted from the previous estimates and obtained by the mutation operator proposed in [14], respectively.

4.3. GPU Implementation

Since all approaches except of local SHO evaluate the log-likelihood V several thousand times per frame which takes nearly the complete computation time, we evaluate likelihoods that can be efficiently computed on a GPU. The computation of $V(x)$ includes the deformation of the human surface model with respect to x, the projection onto the image plane for each view, and the consistency measurements for the selected image features. Since building histograms for each body part as in [14] is expensive on a GPU, only the silhouette and edge term are implemented on a GPU:

Silhouettes. The template images $T_v(x)$ of each particle x are rendered on the GPU. To avoid the expensive transfer between the CPU and the GPU, the human surface model in its initial pose is stored on the GPU. For each particle x, the 3d coordinates of the mesh vertices are computed in the vertex shader. For the silhouette comparison V_{silh}, the Hamming distance is used which is less discriminative than the Chamfer transformed silhouette images. To this end, the images T_v and I_v are pixelwise compared in the fragment shader where the number of inconsistent pixel values are counted by occlusion queries.

Edges. In the binary case V_{edge}, the edges are computed as the silhouettes where the edge map E_v is pixelwise compared with the edges of the projected surface $T_v(x)$. To this end, the body parts of the projected surface $T_v(x)$ are encoded by the color of the triangles. The body part boundaries are detected by a four neighborhood color comparison. By sampling along the edges of E_v and comparing pixelwise the edge values of E_v and $T_v(x)$ in a fragment shader, the number of overlapping edge pixels are counted. The counting is performed by occlusion queries.

For the Chamfer transformed edge map V_{edge}^{Ch}, the values of E_v at the body part boundaries are accumulated in a texture target by enabling the hardware blending operation.

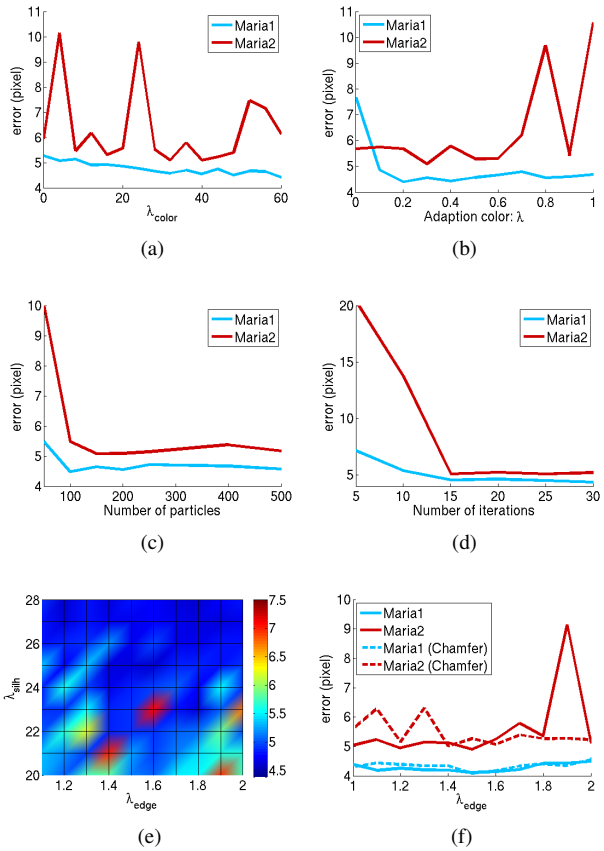


Figure 2. Tracking errors with respect to various parameters for sequences `Maria1` and `Maria2`. **From top to bottom:** *(a)* Weight for color term λ_{color}. *(b)* Adaption of color model. *(c)* Number of particles. *(d)* Number of iterations. *(e)* Weights for silhouette λ_{silh} and edge term λ_{edge}. *(f)* Binary edge map vs. Chamfer edge map.

5. Experiments

Uncontrolled Environment. The first two rows of Figure 5 show estimates for the sequences `Maria1` and `Maria2`. Both sequences were captured by 5 synchronized and calibrated cameras with resolution of 640x480 pixels and 50 fps. They contain a walking person in an uncontrolled environment with people in the background, low contrast, motion blur, and challenging illumination changes as shown in Figure 1. In `Maria2`, the walking person additionally swings her arms. The sequences and result videos are provided as supplemental material. The human model is a low-resolution model of a 3D scan that consists of 2000 triangles. For a quantitative error analysis, circular markers with a diameter of approximately 5 pixels were attached to the forearms and lower legs and were tracked manually.

(*pix*)	`Maria1`	`Maria2`	Time (*sec*) CPU/GPU
silh(ch) + color	4.40 ± 1.26	5.09 ± 1.43	40/-
silh	4.49 ± 1.46	5.34 ± 1.45	22.8/1.2
silh + edge	4.10 ± 1.41	4.90 ± 1.37	26.8/1.8
silh + edge(ch)	4.17 ± 1.43	5.06 ± 1.58	25.3/3.4

Table 1. Average error and standard deviation for ISA using various likelihoods *(ch: Chamfer transform)*. GPU-ISA with silhouettes and edges is as accurate as standard ISA with silhouettes and color. The achieved speed-up is in the range of 12-33.

Likelihoods. We evaluated various log-likelihoods where we fixed the parameter $\lambda_{anat} = 2.0$ (6). Figure 2 shows how sensitive global MHO is with respect to the parameters of the log-likelihood. Rows 1-2 show the results for silhouette and color, where $\lambda_{silh} = 2$, $\lambda_{color} = 40$, $\lambda = 0.3$, 200 particles, and 15 iterations are used unless otherwise stated. The diagrams show clearly that the sequence `Maria2` is more challenging for tracking due to the dynamic movement of the arms. The resulting motion blur in the images affects the appearance of the arm and explains the increase of the error for large values of λ_{color} in contrast to the `Maria1` sequence. Good values for both sequences are in a broad range from 30 to 50. The optimal value for the speed of adaption λ depends on the environment, however, Figure 2(b) shows that the error is not very sensitive to the chosen value as long as the adaption is not too fast. The optimal numbers of particles and iterations are trade-offs between accuracy and computation cost. Figures 2(c-d) show a significant decrease of the error until 100 particles and 15 iterations. More particles or iterations improve the results only marginally. The last row shows the error for parameter settings when silhouettes and edges are used. A low error is achieved for values in the top right quarter of Figures 2(e), *i.e.* $\lambda_{silh} \geq 25$ and $\lambda_{edge} \geq 1.6$. While the errors of the binary and the Chamfer edge map are similar for the `Maria` sequences, see Figure 2(f) and Table 1, the lower standard deviation in Table 2 and Figure 3 justifies the additional computational cost for the Chamfer transform.

Algorithms. For comparison of filtering, single hypothesis optimization, and multiple hypothesis optimization, we have applied the following algorithms to the sequences:

PF	standard particle filter [21]
ICP	iterative closest point approach with twists (local SHO) [6]
FSA	fast simulated annealing (global SHO) [10]
APF	annealed particle filter (local MHO) [11]
ISA	interacting simulated annealing (global MHO) [13]

The results are given in Table 3 and Figure 4. While PF, APF, FSA, and ISA use the same log-likelihood based on silhouettes and color, ICP relies on silhouettes. The anatom-

(mm)	s^{ch}+c	s+e^{ch}	s+e
Set1	34.59 ± 4.63	33.31 ± 5.17	36.60 ± 17.28
Set2	38.53 ± 6.90	42.35 ± 13.45	44.03 ± 18.66
Set3	38.07 ± 5.84	40.26 ± 10.97	40.91 ± 15.22
Time	$76sec/$	$29.4sec/$	$28sec/$
CPU/GPU	-	$4.8sec$	$2.2sec$

Table 2. Average error and standard deviation for ISA using various likelihoods *(s: silhouettes, c: color, e: edges, ^{ch}: Chamfer transform)*. While the error for the edge model is 6-7% higher, the computation time is reduced by 94-97% when a GPU implementation of the less discriminative model is used. The standard deviation of the Chamfer edge map is lower than the one for the binary edge map. The values are plotted in Figure 3.

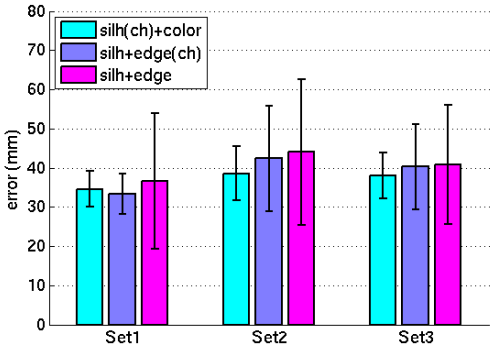


Figure 3. `HumanEva-II` ($S4$). Average error and standard deviation. The three sets contain the frames (2-350), (2-700), and (2-1258) of $S4$. The values and timings are given in Table 2.

ical constraints are used for all approaches. The number of particles and iterations was set to 3000 for PF and FSA, respectively, which yields the same computational effort as APF and ISA with 200 particles and 15 iterations, *i.e.* 40 seconds per frame (Table 1). While the log-likelihood, the number of particles, and the number of iterations were fixed for comparison, the other involved parameters of the algorithms were chosen as the best one obtained by dense grid search. For the comparison, we show only the best results for each algorithm.

The annealing approaches clearly outperform the local optimization and the filtering. While the huge error of the PF indicates the weakness of the likelihood and dynamics, ICP gets stuck in local optima. FSA provides similar results as ISA for `Maria1`, however, the error significantly increases for `Maria2` whereas ISA performs well for both sequences, see Table 3. Since single hypothesis optimization like FSA cannot handle ambiguities, it lacks the robustness of ISA. The error for each frame is given in row 3 of Figure 5.

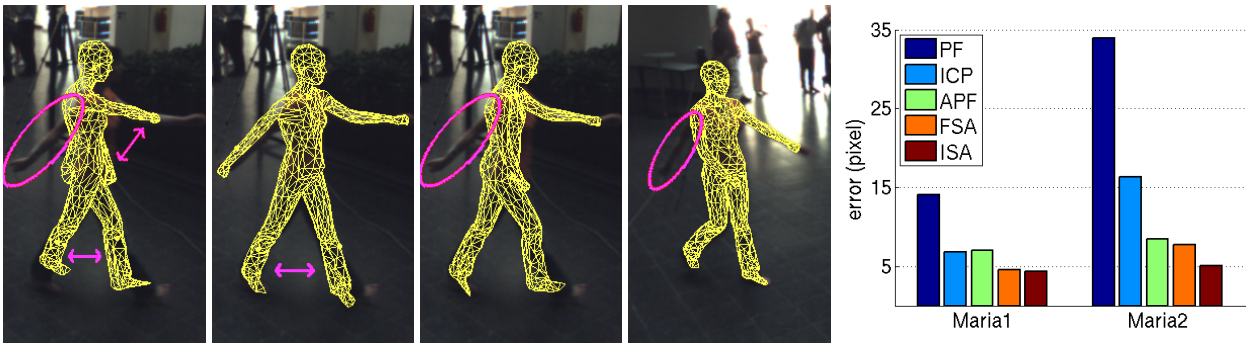
HumanEva. We also compared the likelihoods on the `HumanEva-II` [22] benchmark to measure the absolute 3D tracking error. Since the lighting conditions are controlled, we set $\lambda = 0$ for the color model of ISA. The average errors are given in Table 2 and Figure 3. The error per frame is shown in Figure 5(m). In a controlled environment, the color model performs better than the edge model that is faster to compute on a GPU. However, the less discriminative edge model is preferable where computation time matters. The better performance of the Chamfer edge map can be explained by the larger object size in pixels compared to `Maria`. In this case, inaccuracies of the shape model have a stronger impact on the overlap between the binary edge map and the projection even for the true pose.

6. Conclusion

A quantitative comparison has shown that multiple hypothesis optimization performs better than single hypothesis optimization and global approaches are better than filtering or local approaches for markerless human motion capture with 3d models. Particularly in an uncontrolled environment, local optima are unavoidable which favors global optimization methods. The filtering method failed to estimate accurate poses for the arms and legs. Since it estimates the integral over time (1), it needs a good approximation of the distributions for each frame. However, when the models for the temporal transitions and the likelihood are weak or the number of particles is low, this cannot be satisfied anymore. Even though multiple hypothesis optimization methods might also suffer from a low number of particles, they have shown to work well already with 100 particles in a 30-dimensional search space. This comes from two facts. First, the estimated distributions concentrate on a small region of the search space (4) such that good estimates are already obtained with a few particles. Second, the optimization is performed only on the current frame which means that the impact of previous estimates is reduced, particularly for global MHO. However, the approximated distribution from the previous frame helps to recover from ambiguities which favors MHO over SHO for human motion capture.

The comparison of the likelihoods revealed that in an uncontrolled environment the low-level features silhouettes and edges are sufficient enough to obtain accurate human poses at low computational cost since these features can be efficiently evaluated on a GPU. In a controlled environment like the `HumanEva` benchmark, more expensive features like color histograms per body part are more accurate, but the less discriminative edge features provide a good trade-off between accuracy and computation time for applications.

The work was partially funded by the Cluster of Excellence on Multimodal Computing and Interaction and the Max Planck Institute for Visual Computing.



(a) PF (b) ICP (c) APF (d) FSA (e) Mean error

Figure 4. A quantitative comparison of filtering, single and multiple hypothesis optimization: particle filter (PF), local optimization (ICP), annealed particle filter (APF), fast simulated annealing (FSA), interacting simulated annealing (ISA). Estimates for frame 94 or 18 of `Maria2` are shown for PF, ICP, APF, and FSA (*a-d*). The estimates for ISA are given in row 2 of Figure 5. While the global MHO (ISA) tracks the sequence without significant errors, the other approaches fail to estimate the barely visible right arm or swap the legs. *(e)* The global optimization approaches FSA and ISA perform best. For sequence `Maria2`, the errors obtained by ISA approaches increase only slightly and are significant lower than the one for APF and FSA, cf. Table 3.

error (*pix*)	PF (s^{ch}+c)	ICP (s)	APF (s^{ch}+c)	FSA (s^{ch}+c)	ISA (s^{ch}+c)	ISA (s+e)
`Maria1` $\lambda = 0.2$	14.09 ± 9.95	6.81 ± 3.45	6.96 ± 2.74	4.56 ± 1.28	4.40 ± 1.26	4.10 ± 1.41
`Maria2` $\lambda = 0.3$	33.96 ± 16.55	16.33 ± 12.88	8.40 ± 4.98	7.71 ± 4.69	5.09 ± 1.43	4.90 ± 1.37

Table 3. Average error and standard deviation. For `Maria2`, the error is reduced by 42% and 36% compared to APF and FSA, respectively. A comparison of the likelihoods is given in Table 1 *(s: silhouettes, c: color, e: edges, ^{ch}: Chamfer transform)*.

References

[1] A. Balan, L. Sigal, and M. Black. A quantitative evaluation of video-based 3d person tracking. In *IEEE Workshop on VS-PETS*, pages 349–356, 2005.

[2] L. Ballan and G. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *Int. Symp. on 3DPVT*, 2008.

[3] I. Baran and J. Popović. Automatic rigging and animation of 3d characters. *ACM Trans. on Graphics*, 26(3):72, 2007.

[4] G. Borgefors. Distance transformations in digital images. *Comp. Vision, Graphics, and Image Processing*, 34(3), 1986.

[5] M. Bray, E. Koller-Meier, and L. V. Gool. Smart particle filtering for high-dimensional tracking. *Comp. Vision and Image Understanding*, 106(1):116–129, 2007.

[6] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *Int. J. of Comp Vision*, 56(3):179–194, 2004.

[7] T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In *Comp. Analysis of Images and Patterns*, pages 353–360, 2003.

[8] K. Choo and D. Fleet. People tracking using hybrid monte carlo filtering. In *Int. Conf. on Comp. Vision*, pages 321–328, 2001.

[9] CMU. Graphics lab motion capture database, 2009. http://mocap.cs.cmu.edu.

[10] S. Corazza, L. Mündermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. *A. of Biomedical Engineering*, 34(6):1019–1029, 2006.

[11] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *Int. J. of Comp. Vision*, 61(2):185–205, 2005.

[12] A. Doucet, N. D. Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[13] J. Gall, J. Potthoff, C. Schnörr, B. Rosenhahn, and H.-P. Seidel. Interacting and annealing particle filters: Mathematics and a recipe for applications. *J. of Math. Imaging and Vision*, 28(1):1–18, 2007.

[14] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture – a multi-layer framework. *Int. J. of Comp. Vision*, 2008.

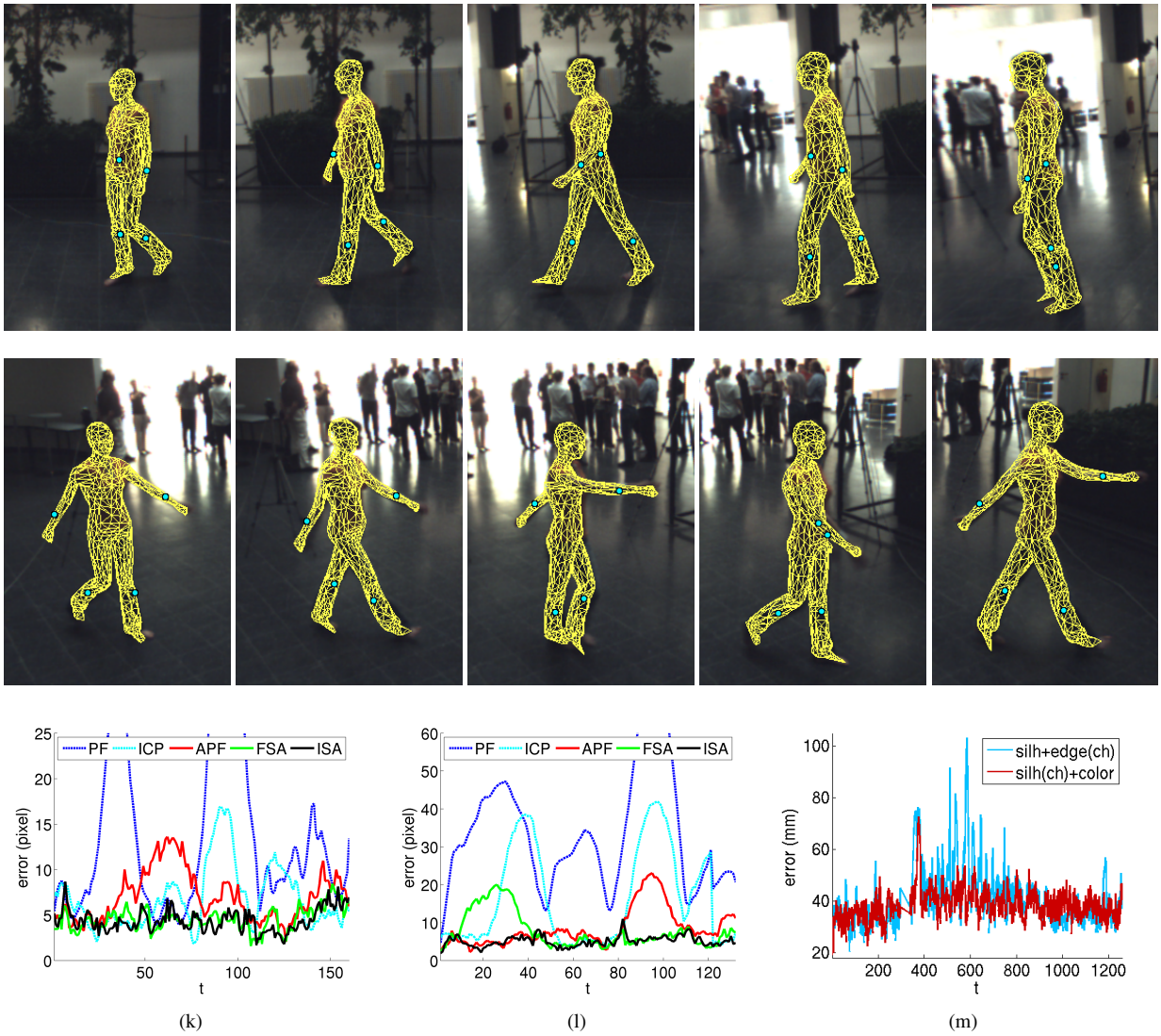
[15] J. Gall, B. Rosenhahn, and H.-P. Seidel. Drift-free tracking of rigid and articulated objects. In *IEEE Conf. on Comp. Vision and Pattern Recognition*, 2008.

[16] N. Gordon, D. Salmond, and A. Smith. Novel approach to non-linear/non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

[17] R. Kalman. A new approach to linear filtering and prediction problems. *Trans. of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[18] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conf. on Comp. Vision*, pages 3–19, 2000.

[19] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Comp. Vision and Image Understanding*, 104(2):90–126, 2006.



(k) (l) (m)

Figure 5. Error analysis for ISA. **Row 1:** Estimates for frames 45, 68, 91, 114, and 137 of `Maria1`. **Row 2:** Estimates for frames 18, 37, 56, 75, and 94 of `Maria2`. **Row 3:** *(k)* `Maria1`. The global optimization approaches FSA and ISA perform best. *(l)* `Maria2`. Only ISA is able to track the sequence without significant errors. *(m)* 3D tracking error for subject $S4$ of `HumanEva-II`. ISA with silhouettes and Chamfer edge map (s+e^{ch}) performs similar to ISA with Chamfer silhouettes and color, but running (frames $350-700$) causes some error peaks for the less discriminative silhouette and edge model. This is reflected by the higher standard deviation in Figure 3.

[20] R. Poppe. Vision-based human motion analysis: An overview. *Comp. Vision and Image Understanding*, 108(1-2):4–18, 2007.

[21] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conf. on Comp. Vision*, pages 702–718, 2000.

[22] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006.

[23] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *Int. J. of Robotics Research*, 22(6):371–391, 2003.

[24] P. Wang and J. Rehg. A modular approach to the analysis and evaluation of particle filters for figure tracking. In *IEEE Conf. on Comp. Vision and Pattern Recognition*, pages 790–797, 2006.

[25] J. Weickert, B. T. H. Romeny, and M. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. on Image Processing*, 7:398–410, 1998.

[26] C. Williams and C. Rasmussen. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*, 1996.

[27] X. Xu and B. Li. Learning motion correlation for tracking articulated human body with a rao-blackwellised particle filter. In *Int. Conf. on Comp. Vision*, 2007.