Global Stochastic Optimization for Robust and Accurate Human Motion Capture

Juergen Gall Thomas Brox Bodo Rosenhahn Hans-Peter Seidel

MPI–I–2007–4–008 December 2007

Authors' Addresses

Juergen Gall
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany

Thomas Brox
Department of Computer Science
University of Dresden
01162 Dresden
Germany

Bodo Rosenhahn
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany

Hans-Peter Seidel
Max-Planck-Institut für Informatik
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany

Acknowledgements

This research is partially funded by the Max-Planck Center for Visual Computing and Communication. We would like to thank Leonid Sigal and Stefano Corazza for providing the data for the `HumanEva-II` dataset.

Abstract

Tracking of human motion in video is usually tackled either by local optimization or filtering approaches. While local optimization offers accurate estimates but often looses track due to local optima, particle filtering can recover from errors at the expense of a poor accuracy due to overestimation of noise. In this paper, we propose to embed global stochastic optimization in a tracking framework. This new optimization technique exhibits both the robustness of filtering strategies and a remarkable accuracy. We apply the optimization to an energy function that relies on silhouettes and color, as well as some prior information on physical constraints. This framework provides a general solution to markerless human motion capture since neither excessive preprocessing nor strong assumptions except of a 3D model are required. The optimization provides initialization and accurate tracking even in case of low contrast and challenging illumination. Our experimental evaluation demonstrates the large improvements obtained with this technique. It comprises a quantitative error analysis comparing the approach with local optimization, particle filtering, and a heuristic based on particle filtering.

Keywords

Human Motion Capture, Stochastic Optimization, Particle Filter

Contents

1 Introduction

Techniques for markerless human motion capture with three-dimensional models that appeared in the last decade can be classified into two groups, namely filtering and optimization strategies. The filtering approaches regard the images as noisy observations of the unknown true state, that is the position, rotation and joint configuration of the human model in each frame. They assume that the dynamics of the human can be modeled by a stochastic process, usually a Markov process, and that the images are generated from the true pose by a stochastic process disturbed by noise. Depending on the underlying processes, the solutions are based on Kalman or particle filtering [14].

The optimization approaches assume the existence of a cost function based on some image features such that the true pose is a global optimum of the function. The cost function may depend on the estimates from previous frames as it occurs from Bayesian modeling where a posterior distribution for a single frame is optimized. After optimization, however, only the estimate but not the distribution is taken into account for the next frame – in contrast to filtering where the uncertainty in the estimate is propagated over time. Since standard global optimization techniques are very expensive, local optimization algorithms like gradient descent are commonly used. So far, neither filtering nor optimization performed significantly better than the other, since both strategies have advantages and disadvantages.

Filtering methods are known to be robust and can recover from errors since they can model noise and resolve ambiguities over time. Particularly, particle filters are popular due to the multimodality of the solution since they approximate a distribution instead of a single value. Furthermore, they do not require linearity of the involved model like the Kalman filter. However, the available convergence results assume that the underlying stochastic processes are known – which in practice is rarely the case. Finding the right models for human motion tracking – both for the dynamics and for the likelihood – is very difficult and so far unsolved. Instead, the weakness of the

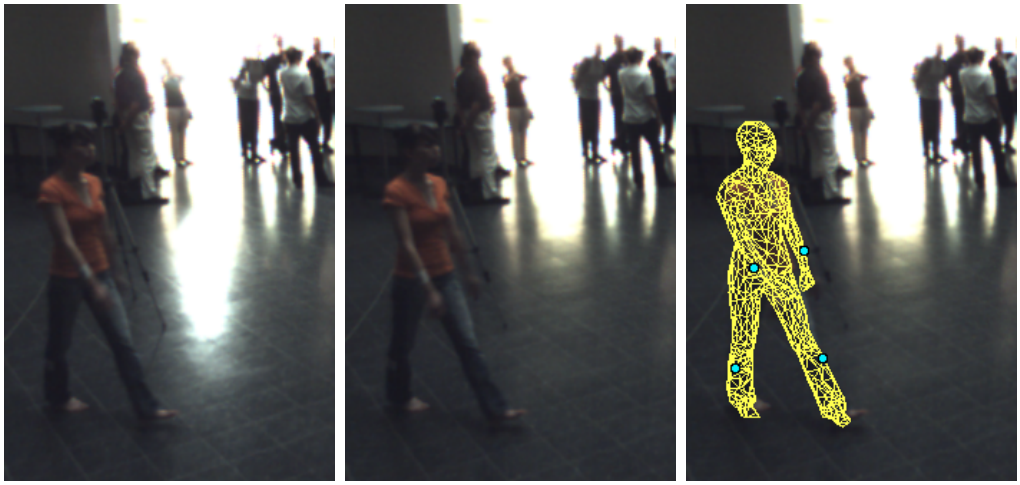


Figure 1.1: **From left to right:** *a, b)* Two successive frames of a multi-view video sequence with low contrast, rapidly changing illumination, and moving people in the background. *c)* The projected mesh shows an accurate estimate for the frame b). The cyan dots are estimates for the markers that were used for a quantitative error analysis.

models is often handled by overestimating the noise yielding a poor performance in high dimensional spaces.

Energy minimization approaches are usually more flexible with regard to the underlying model. However, local optimization suffers from local optima. This has the effect that tracking fails in case of fast motion and the method usually cannot recover from errors.

Our main contribution is to fill the gap between the filtering and local optimization approaches. To this end, we propose a tracking framework that is based on global stochastic optimization, namely interacting simulated annealing (ISA) [15]. Since ISA approximates a distribution rather than a single value, similar to a particle filter, it inherits the advantages of filtering like multimodality and robustness. However, instead of approximating the posterior distribution, the distribution of interest concentrates its mass around the global optima as illustrated in Figure 2.1. Hence, we avoid the modeling problem of filtering approaches where the types of the involved distributions affect the posterior, and thus the outcome. Whereas for global optimization, the shape of the energy function is unimportant as long as the true state is close to the global optimum, which simplifies the modeling task. Indeed, our experiments reveal a better accuracy of our framework than filtering approaches and more robustness than local optimization.

We make use of the larger flexibility in the modeling by introducing an energy function that relies on silhouettes and color, as well as some prior information on physical constraints. In contrast to other works, which regard the appearance of the human for each view as independent, we estimate a statistical appearance model of the 3D surfaces of individual body parts using histogram representations. This makes the model more robust to 3D rotations than comparable 2D models as they are used *e.g.* in [1].

A quantitative error analysis is performed to compare our approach with local optimization, particle filtering, and a state-of-the-art extension [1] of the annealed particle filter [12]. Our framework features automatic initialization and provides accurate estimates even in the case of video sequences with low contrast and challenging illumination, see Figure 1.1. Since neither excessive preprocessing nor strong assumptions are required, except a 3D model, it is a very general solution to human motion capture.

1.1 Related Work

A Kalman filter was integrated into a framework with multiple abstraction levels of the human dynamics [6]. In order to avoid the linearity assumptions of the Kalman filter, Isard and Blake [18] applied a particle filter to 2D tracking. For 3D human motion capture, particle filters were combined with Markov chains, called Hybrid Monte Carlo filtering [10], and graphical models, called nonparametric belief propagation [21, 23]. While other approaches rely on local optimization [7, 9, 17, 20], some heuristics based on particle filters were developed to combine local optimization with filtering, *e.g.* covariance scaled sampling [25], smart [5] and annealed particle filtering [12].

Global stochastic optimization as interacting simulated annealing has so far only been applied to pose estimation for still images [16], which is not suitable for tracking since it is too slow and requires accurate silhouettes. The advantages of taking the appearance into account have been shown in [1, 19] where the appearance was modeled by a mixture of Gaussians. The appearance can also be modeled in a more general manner by histograms as suggested for face tracking, see *e.g.* [3]. Furthermore, there are many approaches that learn the dynamics of special motions offline from training samples. These methods are out of scope since they apply only to a small subset of human motion patterns. Here, we only mention the Gaussian process dynamical models [26] since we also use Gaussian processes – but for online learning.

2 Pose Estimation by Global Optimization

The pose estimation for a single frame is performed by interacting simulated annealing [15] where the pose is represented by a vector containing the position, rotation, and joint angles of the 3D skeletal model. The solution is given by a distribution η_t whose mass concentrates in the region of global minima of a given energy function $V \geq 0$ as t tends to infinity, see Figure 2.1. This behavior is described by the following convergence theorem saying that for any $\epsilon > 0$

$$\lim_{t \to \infty} \eta_t \left(V \geq \sup \left\{ v \geq 0;\ V \geq v \ a.e. \right\} + \epsilon \right) = 0. \tag{2.1}$$

Although an analytical solution is not available for η_t, it can be approximated by n samples:

$$\eta_t^n := \sum_{i=1}^{n} \pi^{(i)} \delta_{x_t^{(i)}} \tag{2.2}$$

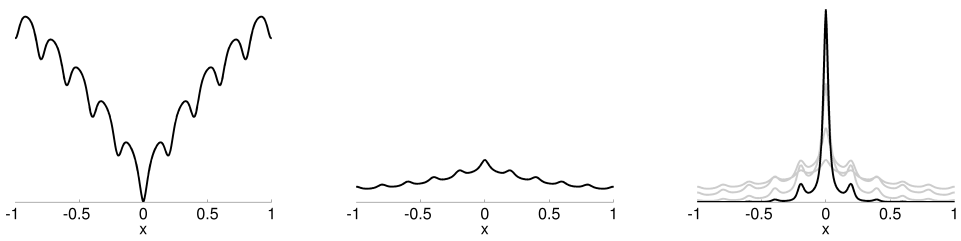


Figure 2.1: **From left to right:** *a)* Energy function V with global minimum at zero. *b)* η_1. *c)* The mass of η_t concentrates around the global minimum as t increases. For a limited number of iterations, η_t is multimodal.

Figure 2.2: The set of particles. **From left to right:** *a)* Weighting. Particles with higher weights are brighter. *b)* Selection. *c)* Mutation.

where $x_t^{(i)}$ are called particles, $\pi^{(i)}$ weights and δ denotes the Dirac measure. The optimization consists of a weighting, a selection, and a mutation operation that are iterated T times, see Figure 2.2. Finally, an estimate for the pose is obtained by the mean $\hat{x} = \int \eta_T^n(x)\, dx$.

Weighting

Assuming that a set of particles $(x_t^{(i)})_{i=1\ldots n}$ exists, each particle is weighted by the Boltzmann-Gibbs measure

$$\pi^{(i)} = \exp\left(-\beta_t\, V\left(x_t^{(i)}\right)\right), \tag{2.3}$$

where $\beta_t = (t+1)^b$ with $b = 0.7$ is an annealing scheme that increases monotonically. After normalizing the weights such that $\sum_i \pi^{(i)} = 1$, the weight indicates the probability that a particle is selected for the next step.

Selection

In a first stage, particles are accepted with probability $\pi^{(i)} / \max_k \pi^{(k)}$, *i.e.* the particle with the highest weight is always accepted. Since after this first stage only m particles are selected, additional $n - m$ particles are drawn in a second stage, replacing those from the old set. This is efficiently done by stratified resampling [13]. Due to the selection operation, similar particles

with high weights are contained several times in the new set whereas particles with low weights might disappear completely.

Mutation

In order to explore the search space, the particles are spread out according to a Gaussian K_t whose covariance matrix is proportional to the sampling covariance matrix

$$\Sigma_t \approx \frac{1}{n-1}\left(\rho\, I + \sum_{i=1}^{n}(x_t^{(i)} - \mu_t)\,(x_t^{(i)} - \mu_t)^T\right), \tag{2.4}$$

where μ_t is the average, I the identity matrix, and ρ a small positive constant that ensures that the covariance does not become singular. The computational cost is reduced by using a sparse matrix that takes only correlations of joints into account that belong to the same skeleton branch.

2.1 Energy Function

The energy of a particle x is calculated by

$$V(x) = \nu\, V_{silh}(x) + \tau\, V_{app}(x) + \upsilon\, V_{phys}(x), \tag{2.5}$$

where the parameters ν, τ, and υ control the influence of the three terms, namely silhouettes, appearance, and physical constraints that are explained in Sections 2.1.1, 2.1.2, and 2.1.3, respectively.

2.1.1 Silhouettes

In order to model an error function between a particle x 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 the particle as shown in Figure 2.3 a). The inconsistent areas between the silhouette and the template are then measured for each view v by

$$\begin{aligned} V_v(x) = {} & \frac{1}{2|T_v^0(x)|}\sum_{p\in T_v^0(x)}|T_v(x,p) - I_v(p)| \\ & + \frac{1}{2|I_v^0|}\sum_{p\in I_v^0}|I_v(p) - T_v(x,p)|\,, \end{aligned} \tag{2.6}$$

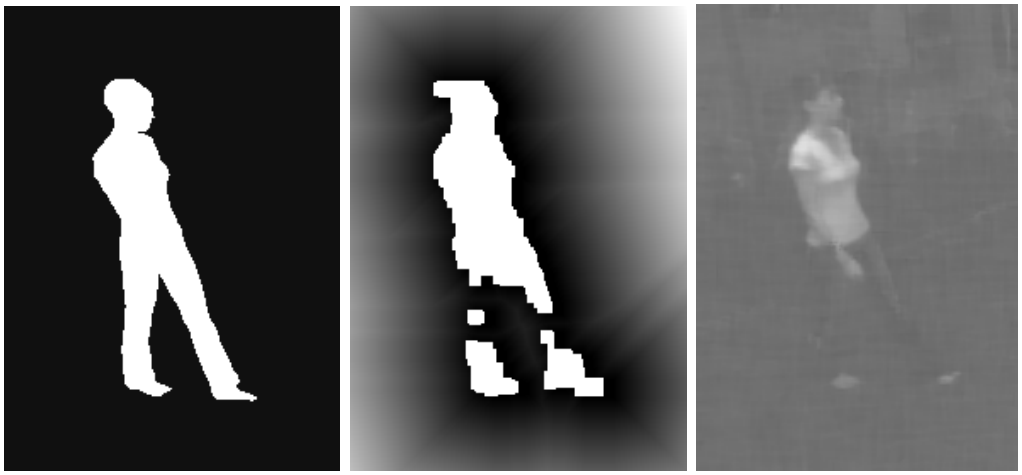


Figure 2.3: **From left to right:** *a)* Template image $T_v(x)$. *b)* Silhouette image I_v. *c)* Smoothed color channel.

where $I_v(p)$ and $T_v(x, p)$ are the pixel values for a pixel p and the sets of pixels inside the silhouettes are denoted by I_v^0 and $T_v^0(x)$. Since pixels that are far away from the silhouette should be penalized more severely, a Chamfer distance transform [4] is previously applied to I_v as shown in Figure 2.3 b). For the template, only a constant value is used due to computational efficiency. The energy term V_{silh} is finally defined as the average error of all views.

2.1.2 Appearance

Our approach for integrating color information is motivated by 2D segmentation where the separation of foreground pixels from the background relies on region statistics. Since we know the 3D model, we combine the pixel information from all views to model the statistics of different body parts rather than their separate projections to the images. For efficiency reasons, we assume the image channels u_c to be uncorrelated. Hence, the joint probability density function for a body part s can be written as

$$p_s(u) = \prod_c p_{s,c}(u_c). \tag{2.7}$$

Instead of assuming a certain family of distribution functions, we approximate the probabilities $p_{s,c}$ in a more general manner by normalized histograms $H^{(s,c)}$ where we fixed the number of bins to $K = 64$. The updating of the appearance model during tracking is explained in Section 3.

In order to measure deviations of the appearance of a particle x from the appearance model given by $H^{(s,c)}$, the particle's appearance $\tilde{H}^{(s,c)}(x)$ 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 as shown in Figure 3.1 a). Hence, a pixel p that belongs to a body part s contributes for each channel u_c to the histogram $\tilde{H}^{(s,c)}(x)$. For histogram comparison, we choose the Bhattacharya distance since it is also stable for empty bins in contrast to χ^2 or KLD [22]. The total deviation is then measured according to (2.7) by

$$V_{app}(x) = \sum_s \frac{w_s}{C} \sum_{c=1}^{C} \left(1 - \sum_{k=1}^{K} \sqrt{h_k^{(s,c)} \tilde{h}_k^{(s,c)}(x)} \right), \tag{2.8}$$

where the weights w_s reflect the size of the body parts and are determined during initialization, see Section 3.

Since the distinctiveness of the appearance model depends on the used image channels, the images are preprocessed to get a better image representation than the raw image data. We achieved good results with the CIELab color space that mimics the human perception of color differences. Since the L-channel is very sensitive to illumination changes, we used only the a- and b-channel. For small body parts like the hands where the sample sizes are rather small, image noise becomes an important issue. In order to reduce noise without smoothing over the edges that separate body parts as shown in Figure 2.3 c), we apply the edge-enhancing diffusivity function [8]

$$g(|\nabla u|^2) = \frac{1}{|\nabla u|^p + \epsilon} \tag{2.9}$$

with $\epsilon = 0.001$ and $p = 1.5$, where the smoothing is efficiently implemented by the AOS scheme [27].

2.1.3 Physical Constraints

Since human motion is subject to physical restrictions, the search can be focused on poses with higher probabilities by adding a soft constraint to the energy function. For this purpose, the probability of a skeleton deformation p_{pose} is estimated from a set of training samples y_l taken from the CMU motion database [11]. Since self-intersections between the head, the upper body, and the lower body rarely occur, the sample size L can be reduced by regarding the probabilities for the three body parts, denoted by p_{pose}^{head}, p_{pose}^{upper}, and p_{pose}^{lower}, as uncorrelated. The probability for a body part is approximated

by a Parzen-Rosenblatt estimator with a Gaussian kernel K:

$$p_{pose}(x) = \frac{1}{L\,h^d} \sum_l K\left(\frac{x - y_l}{h}\right), \tag{2.10}$$

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. Finally, we used less than 200 samples from different motions for modeling the physical constraints by

$$V_{phys}(x) = -\frac{1}{3} \ln\left(p_{pose}^{head}(x) p_{pose}^{upper}(x) p_{pose}^{lower}(x)\right). \tag{2.11}$$

3 Tracking

For tracking, the pose estimation is embedded in a framework that takes advantage of temporal coherence of sequential data. An outline of the tracking system is given in Figure 3.1 b). For the first frame, the pose is detected automatically and the appearance model is initialized as described in Section 3.1. Before estimating the pose via ISA (Section 3.3), the particles are spread in the search space, see Section 3.2. After the optimization, the appearance model is updated as discussed in Section 3.4.

3.1 Initialization

The initialization is performed automatically via stochastic optimization [16] of the energy function defined in Equation (2.5). Since the appearance of the model is unknown a priori, only the terms V_{silh} and V_{phys} for the silhouettes and physical constraints are used.

After the pose $\hat{x}_0$ is estimated for the first frame, the histograms $H^{(s,c)}$ are generated by sampling from the images as described in Section 2.1.2. During sampling, the range of each feature channel is also determined and divided into uniform bins. Furthermore, the weights w_s in Equation (2.8) are given by the sample size for each body part s after normalizing such that $\sum_s w_s = 1$.

3.2 Mutation

After estimating the pose $\hat{x}_t$, the particles $x_t^{(i)}$ congregate around the global optima for frame t. Since this set is not well distributed for estimating the pose in the next frame, a mutation step spreads the particles in the search space. For this purpose, a 3rd order autoregression is used to predict the pose from the previous estimates, *i.e.* $x_{t+1}^{pred} = f(\hat{x}_{t:3})$ where we denote the last three estimates by $\hat{x}_{t:3} = (\hat{x}_t, \hat{x}_{t-1}, \hat{x}_{t-2})$. The function f can be learned

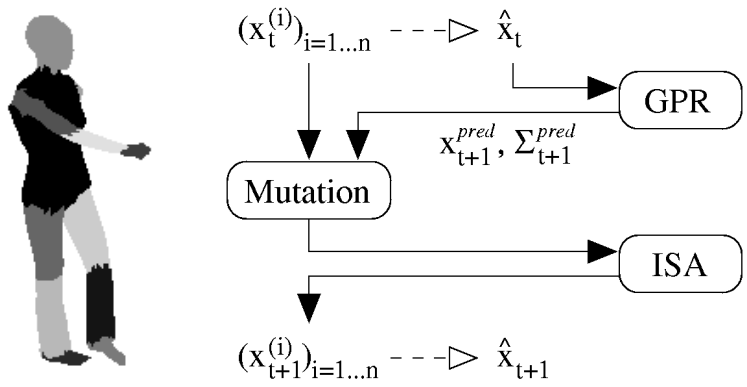


Figure 3.1: **From left to right:** *a)* Human model with $2K$ triangles. The triangles encode the body parts. *b)* Outline of the tracking system. While the particle set $(x_t^{(i)})_i$ represents the distribution of the solution, the mean $\hat{x}_t$ provides a single estimate for the pose. The pose for the next frame x_{t+1}^{pred} is predicted by Gaussian process regression (GPR), and an additional mutation operator spreads the particles in the search space. The pose is then estimated by stochastic optimization (ISA). The system is closed in the sense that any uncertainty that arises from the prediction and estimation is preserved in terms of Σ_{t+1}^{pred} and $(x_{t+1}^{(i)})_i$.

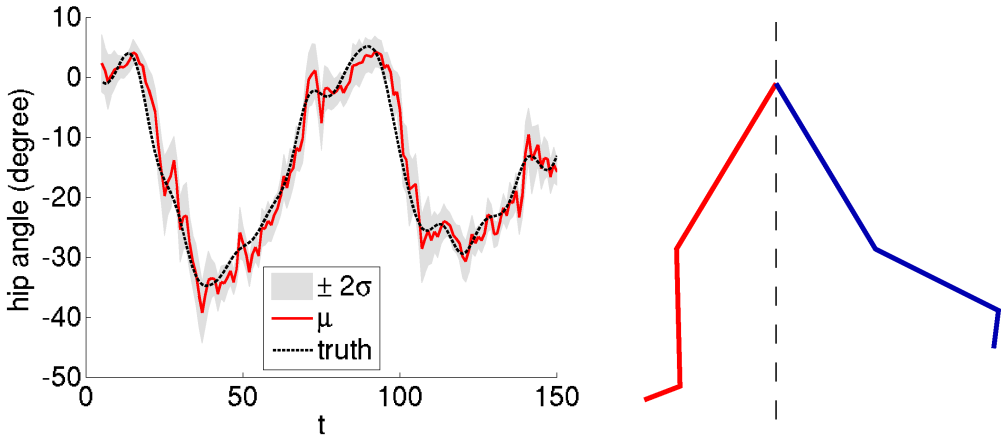


Figure 3.2: **From left to right:** *a)* Prediction of a joint angle by GPR. Predicted Gaussian distribution with μ and σ. *b)* Mutation operator. The left branch *(red)* is reconstructed from the right branch *(blue)* by mirroring the first joint.

online from the history of estimates given by the equations $\hat{x}_{t-r+1} = f(\hat{x}_{t-r:3})$ for $r = 1 \dots R$. The regression is implemented by Gaussian processes [28] that fit very well in our framework since the prediction is given by a Gaussian distribution with mean x_{t+1}^{pred} and covariance matrix Σ_{t+1}^{pred}, see Figure 3.2 a). To simplify matters, we briefly summarize only the one-dimensional prediction by Gaussian processes where the set of training data is given by $\hat{x}_R = (\hat{x}_{t-1:3}, \dots, \hat{x}_{t-R:3})^T$ and $f(\hat{x}_R) = (f(\hat{x}_{t-1:3}), \dots, f(\hat{x}_{t-R:3}))^T$.

The predictive distribution for the last three estimates $\hat{x}_{t:3}$ is obtained by the conditional Gaussian distribution $p(\hat{x}_{t+1}|\hat{x}_{t:3}, \hat{x}_R, f(\hat{x}_R))$ with mean and variance

$$x_{t+1}^{pred} = k(\hat{x}_{t:3}, \hat{x}_R)^T \mathbf{K}^{-1} f(\hat{x}_R), \tag{3.1}$$

$$(\sigma_{t+1}^{pred})^2 = k(\hat{x}_{t:3}, \hat{x}_{t:3}) - k(\hat{x}_{t:3}, \hat{x}_R)^T \mathbf{K}^{-1} k(\hat{x}_{t:3}, \hat{x}_R). \tag{3.2}$$

The covariance matrix for the training data $\mathbf{K}$ is modeled by the general covariance function

$$k(\hat{x}_{r:3}, \hat{x}_{s:3}) = a_0 \exp\left(-\frac{1}{2}\sum_{j=0}^{2} a_{j+1} \left(\hat{x}_{r-j} - \hat{x}_{s-j}\right)^2\right) + \sum_{j=0}^{2} a_{j+4}\, \hat{x}_{r-j}\hat{x}_{s-j} + \sigma_{noise}^2 \delta_{rs}, \tag{3.3}$$

where the hyperparameters a_j and σ_{noise}^2 are learned offline by minimizing the log likelihood as proposed in [28]. Due to computational efficiency, all parameters of the search space are assumed to be independent which yields a one-dimensional prediction for each degree of freedom.

Since the dynamics are learned online, the prediction adapts to the current motion but it also might be corrupted by tracking errors in the past. Hence, we shift only 40% of the particles according to x_{t+1}^{pred}, another 40% is kept as it is and 20% are mutated. The mutation is motivated by evolutionary algorithms where a larger variety among a population helps to recover from errors. We propose a human specific mutation operator that is useful when only one of two legs or arms is well estimated due to occlusions. In order to reconstruct its counterpart, we imitate the behavior of humans to use their arms or legs to balance. For this purpose, the first joint of the kinematic branch is mirrored while the other joint angles remain unchanged as illustrated in Figure 3.2 b). Even though the mutated particles will be mostly rejected after the first iterations of the optimization, they support the tracker in recovering from errors. Finally, all particles are propagated by a zero-mean Gaussian distribution with covariance matrix proportional to Σ_{t+1}^{pred}.

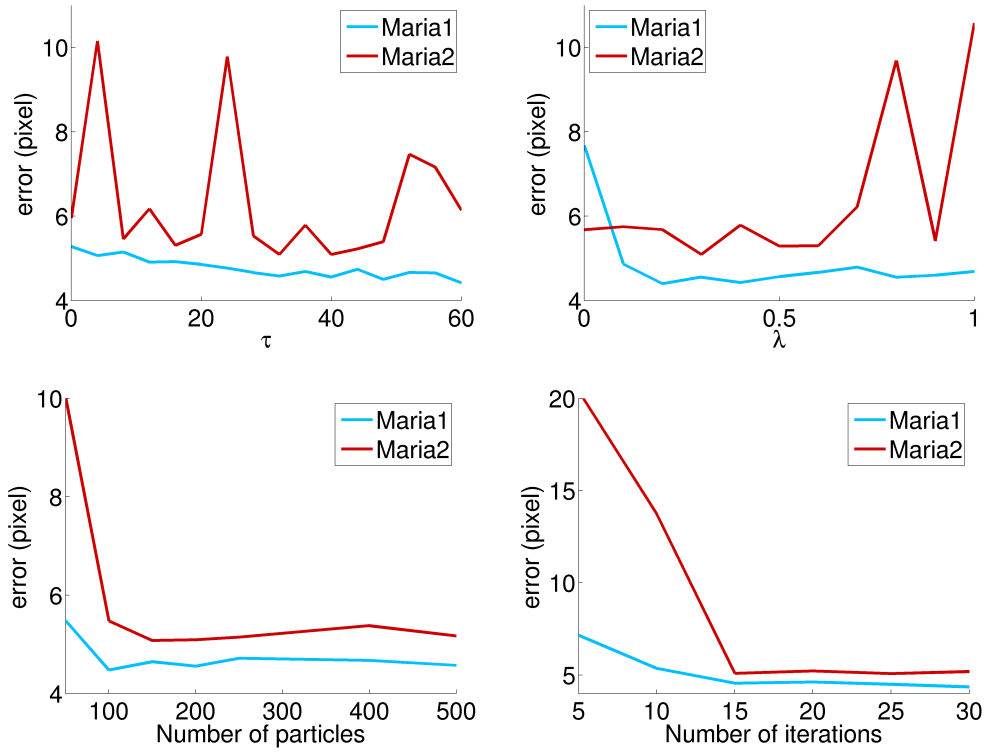


Figure 3.3: Tracking errors with respect to various parameters for sequences `Maria1` and `Maria2`. **From top left to bottom right:** *a)* Weight for appearance term V_{app}. *b)* Speed of adaption. *c)* Number of particles. *d)* Number of iterations.

3.3 Estimation

Having a well distributed set of particles, the pose is estimated by interacting simulated annealing as described in Section 2. Since many applications expect a single estimate for each frame, the weighted mean $\hat{x}_{t+1}$ is returned as estimate.

3.4 Update

After estimation, the covariance matrices for the regression are updated in Equations (3.1) and (3.2) by adding $\hat{x}_{t+1}$ to the history of estimates. Furthermore, the histograms $H^{(s,c)}$ are adapted to the changing appearance. First, a normalized histogram $\hat{H}^{(s,c)}$ is generated for $\hat{x}_{t+1}$ by sampling from all views.

The update for bin k is then given by

$$\frac{(1-\lambda)M^{(s)}\, h_k^{(s,c)} + \lambda\hat{M}^{(s)}\, \hat{h}_k^{(s,c)}}{(1-\lambda)M^{(s)} + \lambda\hat{M}^{(s)}}, \tag{3.4}$$

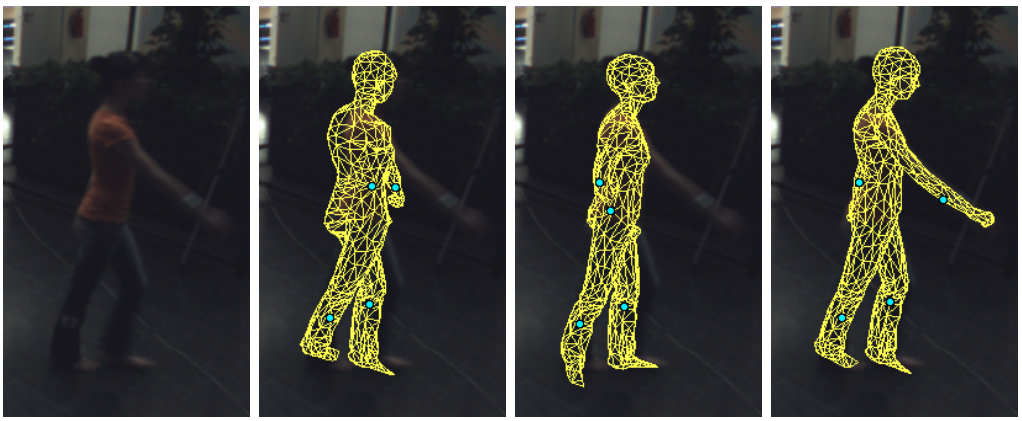
where $M^{(s)}$ and $\hat{M}^{(s)}$ are the sample sizes for the body part s to generate H and $\hat{H}$, respectively. The parameter λ controls the speed of adaptation and the consideration of the sample sizes avoids that the statistics are distorted by a small number of samples, *e.g.* due to self-occlusions.

4 Experiments

The first two rows of Figure 4.2 show estimates for the sequences `Maria1` and `Maria2`. Both sequences were captured by 5 synchronized and calibrated cameras with resolution of 640×480 pixels and 50 fps. They contain a walking person in a natural environment with people in the background, low contrast, motion blur, and challenging illumination changes as shown in Figure 1.1. 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 2K triangles. For a quantitative error analysis, circular markers with a diameter of approx. 5 pixels were attached to the forearms and lower legs and were tracked manually.

In our experiments, we fixed the parameters $\nu = 2.0$ and $\upsilon = 2.0$ in Equation (2.5) and we evaluated how sensitive our approach is with respect to the appearance parameters τ and λ as plotted in Figure 3.3. Unless otherwise stated, we used $\tau = 40$, $\lambda = 0.3$, 200 particles, and 15 iterations. 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, as shown in Figure 4.1, affects the appearance of the arm and explains the increase of the error for large values of τ 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 3.3 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 3.3 c) and d) show a significant decrease of the error until 100 particles and 15 iterations yielding a computation time of 4 seconds per image. Larger number of particles and iterations improve the results only marginally.

For comparison with filtering and local optimization, we applied a standard particle filter and an iterative closest point approach [2] to the sequences.



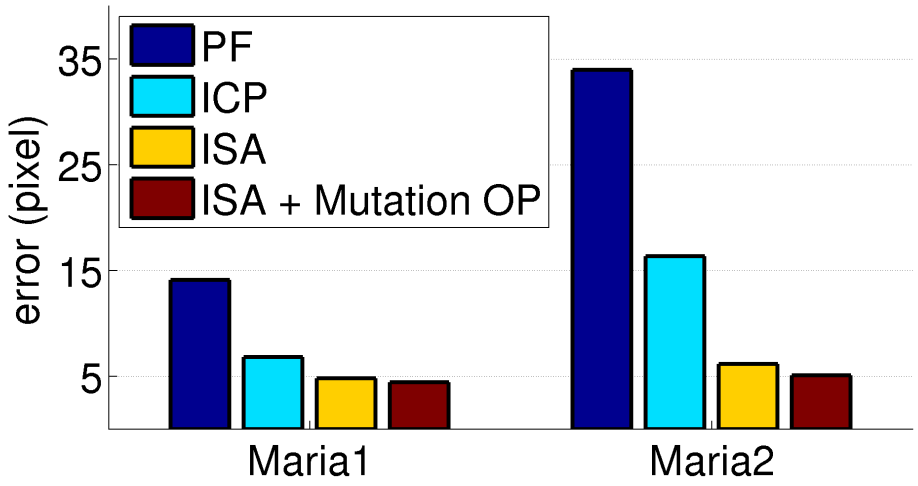


Figure 4.1: A quantitative comparison with a particle filter (PF) and local optimization (ICP). **Top:** Frame 115. Estimates for frame 115 of `Maria2` by PF, ICP, and ISA *(from left to right)*. The barely visible right arm is only correctly estimated by our approach. **Bottom:** While the estimates of the particle filter are imprecise and ICP gets stuck in local optima, our approach using ISA provides accurate estimates for both sequences. For sequence `Maria2`, the error increases only slightly whereas the error for PF and ICP increases by a factor of two. The comparison also reveals the positive effect of the mutation operator depicted in Figure 3.2 b).

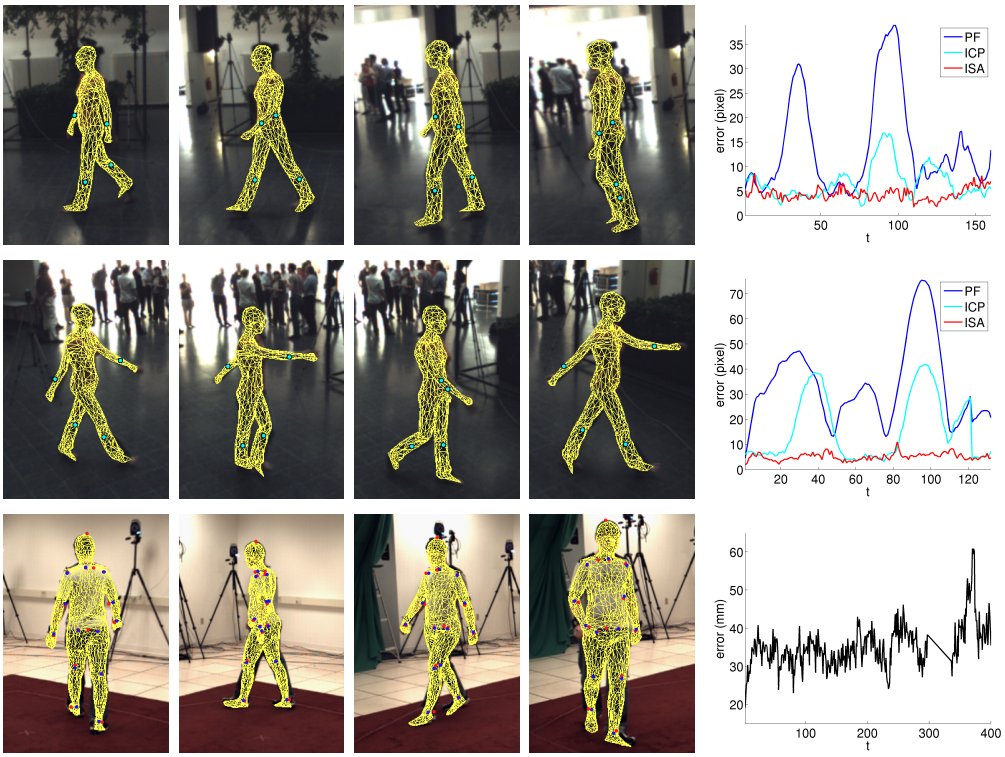


Figure 4.2: Error analysis. **Row 1:** Estimates for frames 68, 91, 114, and 137 of `Maria1`. **Row 2:** Estimates for frames 37, 56, 75, and 94 of `Maria2`. **Row 3:** Estimates for frames 80, 160, 240, and 320 for subject $S4$ of `HumanEva-II`. The frames $298 - 335$ are neglected for the error analysis since the ground truth is corrupted for these frames.

	ISA	APF [1]
error (mm)	35.02 ± 5.73	> 60

Figure 4.3: The comparison with an annealed particle filter that uses similar cues reveals an error reduction by more than 40%.

The same energy model was used. For the particle filter, we employed the weighting function (2.3) with $\beta_t = 1$. This is similar to the assumption that the likelihood is proportional to a product of normal densities. The particles are predicted as described in Section 3.2 without using the mutation operator since it is not supported by a filtering framework, *i.e.* 50% of the particles are shifted according to the predicted mean and the remaining 50% are directly selected. The number of particles was set to 3000, which yields the same computational effort as our approach with 200 particles and 15 iterations. The results are plotted in Figure 4.1. The global stochastic optimization approach clearly outperforms both the local optimization and the filtering. While the huge error of the particle filter indicates the weakness of the likelihood and dynamics, ICP gets stuck in local optima. It is remarkable that the error for PF and ICP increases by a factor of two for `Maria2` whereas our approach performs well for both sequences, namely $4.40mm \pm 1.26$ (`Maria1`, $\lambda = 0.2$) and $5.09mm \pm 1.43$ (`Maria2`, $\lambda = 0.3$). The error for each frame is given in Figure 4.2.

We also applied our approach to the dataset `HumanEva-II` [24] to measure the absolute 3D tracking error. The available model is not perfect since it does not contain the clothing of the subject $S4$. Since the lighting conditions are controlled, we set $\lambda = 0$. Nevertheless, we achieve accurate estimates with 250 particles as shown in row 3 of Figure 4.2. Since the set-up and movement of the sequence, namely walking in a circle, is similar to the one used in [1], we compare the results in Table 4.3. Furthermore, our implementation with 19 seconds per image is faster than the 90 seconds reported in [1].

5 Conclusion

We have shown that global stochastic optimization is a promising alternative to existing filtering and local optimization approaches for markerless human motion capture. A quantitative comparison with local optimization and particle filtering revealed that our tracking framework gives much better results even for challenging scenes where the silhouette information is unreliable. Local optimization may perform better than global optimization for sequences where local optima are not essential – but this is rarely the case in natural environments. Since the framework is easy to implement and requires neither excessive preprocessing nor strong assumptions, it is a very general solution to human motion tracking that can be specialized further. It might also be applied to other problems where filtering or local optimization perform poorly. For future work, we intend to reduce the computation time further by exploiting the parallel structure of ISA and graphics hardware.

Bibliography

[1] A. Balan and M. Black. An adaptive appearance model approach for model-based articulated object tracking. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 758–765, 2006.

[2] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. on Patt. Analysis and Machine Intell.*, 14(2):239–256, 1992.

[3] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 232–237, 1998.

[4] G. Borgefors. Distance transformations in digital images. *Comput. Vision Graph. Image Process.*, 34(3), 1986.

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

[6] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, 1997.

[7] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 8–15, 1998.

[8] T. Brox, M. Rousson, R. Deriche, and J. Weickert. Unsupervised segmentation incorporating colour, texture, and motion. In *Comp. Analysis of Images and Patterns*, volume 2756 of *LNCS*, pages 353–360. Springer, 2003.

[9] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part ii: Applications to human modeling and markerless motion tracking. *Int. J. of Computer Vision*, 63(3):225–245, 2005.

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

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

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

[13] R. Douc, O. Cappe, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Int. Symp. on Image and Signal Processing and Analysis*, pages 64–69, 2005.

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

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

[16] J. Gall, B. Rosenhahn, and H.-P. Seidel. Clustered stochastic optimization for object recognition and pose estimation. In *Patt. Recog.*, volume 4713 of *LNCS*, pages 32–41. Springer, 2007.

[17] D. Gavrila and L. Davis. 3-d model-based tracking of humans in action: a multi-view approach. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 73–80, 1996.

[18] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *European Conf. on Comp. Vision*, volume 1, pages 343–356, 1996.

[19] A. Jepson, D. Fleet, and T. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on Patt. Analysis and Machine Intell.*, 25(10):1296–1311, 2003.

[20] I. Kakadiaris and D. Metaxas. Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 81–87, 1996.

[21] M. Lee and R. Nevatia. Human pose tracking using multi-level structured models. In *European Conf. on Comp. Vision*, pages 368–381, 2006.

[22] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Int. Conf. on Computer Vision*, pages 1165–1172, 1999.

[23] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 421–428, 2004.

[24] 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.

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

[26] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. *IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 238–245, 2006.

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

[28] C. Williams and C. Rasmussen. Gaussian processes for regression. In *Advances in Neural Inf. Proc. Systems*, 1996.

Below you find a list of the most recent technical reports of the Max-Planck-Institut für Informatik. They are available via WWW using the URL `http://www.mpi-inf.mpg.de`. If you have any questions concerning WWW access, please contact `reports@mpi-inf.mpg.de`. Paper copies (which are not necessarily free of charge) can be ordered either by regular mail or by e-mail at the address below.

Max-Planck-Institut für Informatik
Library
attn. Anja Becker
Stuhlsatzenhausweg 85
66123 Saarbrücken
GERMANY
e-mail: `library@mpi-inf.mpg.de`

.

MPI-I-2007-RG1-002	T. Hillenbrand, C. Weidenbach	Superposition for Finite Domains
MPI-I-2007-5-003	F.M. Suchanek, G. Kasneci, G. Weikum	Yago : A Large Ontology from Wikipedia and WordNet
MPI-I-2007-5-002	K. Berberich, S. Bedathur, T. Neumann, G. Weikum	A Time Machine for Text Search
MPI-I-2007-5-001	G. Kasneci, F.M. Suchanek, G. Ifrim, M. Ramanath, G. Weikum	NAGA: Searching and Ranking Knowledge
MPI-I-2007-4-007	R. Herzog, V. Havran, K. Myszkowski, H. Seidel	Global Illumination using Photon Ray Splatting
MPI-I-2007-4-006	C. Dyken, G. Ziegler, C. Theobalt, H. Seidel	GPU Marching Cubes on Shader Model 3.0 and 4.0
MPI-I-2007-4-005	T. Schultz, J. Weickert, H. Seidel	A Higher-Order Structure Tensor
MPI-I-2007-4-004	C. Stoll	A Volumetric Approach to Interactive Shape Editing
MPI-I-2007-4-003	R. Bargmann, V. Blanz, H. Seidel	A Nonlinear Viseme Model for Triphone-Based Speech Synthesis
MPI-I-2007-4-002	T. Langer, H. Seidel	Construction of Smooth Maps with Mean Value Coordinates
MPI-I-2007-4-001	J. Gall, B. Rosenhahn, H. Seidel	Clustered Stochastic Optimization for Object Recognition and Pose Estimation
MPI-I-2007-2-001	A. Podelski, S. Wagner	A Method and a Tool for Automatic Veriication of Region Stability for Hybrid Systems
MPI-I-2007-1-002	E. Althaus, S. Canzar	A Lagrangian relaxation approach for the multiple sequence alignment problem
MPI-I-2007-1-001	E. Berberich, L. Kettner	Linear-Time Reordering in a Sweep-line Algorithm for Algebraic Curves Intersecting in a Common Point
MPI-I-2006-5-006	G. Kasnec, F.M. Suchanek, G. Weikum	Yago - A Core of Semantic Knowledge
MPI-I-2006-5-005	R. Angelova, S. Siersdorfer	A Neighborhood-Based Approach for Clustering of Linked Document Collections
MPI-I-2006-5-004	F. Suchanek, G. Ifrim, G. Weikum	Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents
MPI-I-2006-5-003	V. Scholz, M. Magnor	Garment Texture Editing in Monocular Video Sequences based on Color-Coded Printing Patterns
MPI-I-2006-5-002	H. Bast, D. Majumdar, R. Schenkel, M. Theobald, G. Weikum	IO-Top-k: Index-access Optimized Top-k Query Processing
MPI-I-2006-5-001	M. Bender, S. Michel, G. Weikum, P. Triantafilou	Overlap-Aware Global df Estimation in Distributed Information Retrieval Systems
MPI-I-2006-4-010	A. Belyaev, T. Langer, H. Seidel	Mean Value Coordinates for Arbitrary Spherical Polygons and Polyhedra in $\mathbb{R}^3$
MPI-I-2006-4-009	J. Gall, J. Potthoff, B. Rosenhahn, C. Schnoerr, H. Seidel	Interacting and Annealing Particle Filters: Mathematics and a Recipe for Applications
MPI-I-2006-4-008	I. Albrecht, M. Kipp, M. Neff, H. Seidel	Gesture Modeling and Animation by Imitation

MPI-I-2006-4-007	O. Schall, A. Belyaev, H. Seidel	Feature-preserving Non-local Denoising of Static and Time-varying Range Data
MPI-I-2006-4-006	C. Theobalt, N. Ahmed, H. Lensch, M. Magnor, H. Seidel	Enhanced Dynamic Reflectometry for Relightable Free-Viewpoint Video
MPI-I-2006-4-005	A. Belyaev, H. Seidel, S. Yoshizawa	Skeleton-driven Laplacian Mesh Deformations
MPI-I-2006-4-004	V. Havran, R. Herzog, H. Seidel	On Fast Construction of Spatial Hierarchies for Ray Tracing
MPI-I-2006-4-003	E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, H. Seidel	A Framework for Natural Animation of Digitized Models
MPI-I-2006-4-002	G. Ziegler, A. Tevs, C. Theobalt, H. Seidel	GPU Point List Generation through Histogram Pyramids
MPI-I-2006-4-001	A. Efremov, R. Mantiuk, K. Myszkowski, H. Seidel	Design and Evaluation of Backward Compatible High Dynamic Range Video Compression
MPI-I-2006-2-001	T. Wies, V. Kuncak, K. Zee, A. Podelski, M. Rinard	On Verifying Complex Properties using Symbolic Shape Analysis
MPI-I-2006-1-007	H. Bast, I. Weber, C.W. Mortensen	Output-Sensitive Autocompletion Search
MPI-I-2006-1-006	M. Kerber	Division-Free Computation of Subresultants Using Bezout Matrices
MPI-I-2006-1-005	A. Eigenwillig, L. Kettner, N. Wolpert	Snap Rounding of Bézier Curves
MPI-I-2006-1-004	S. Funke, S. Laue, R. Naujoks, L. Zvi	Power Assignment Problems in Wireless Communication
MPI-I-2005-5-002	S. Siersdorfer, G. Weikum	Automated Retraining Methods for Document Classification and their Parameter Tuning
MPI-I-2005-4-006	C. Fuchs, M. Goesele, T. Chen, H. Seidel	An Emperical Model for Heterogeneous Translucent Objects
MPI-I-2005-4-005	G. Krawczyk, M. Goesele, H. Seidel	Photometric Calibration of High Dynamic Range Cameras
MPI-I-2005-4-004	C. Theobalt, N. Ahmed, E. De Aguiar, G. Ziegler, H. Lensch, M.A. Magnor, H. Seidel	Joint Motion and Reflectance Capture for Creating Relightable 3D Videos
MPI-I-2005-4-003	T. Langer, A.G. Belyaev, H. Seidel	Analysis and Design of Discrete Normals and Curvatures
MPI-I-2005-4-002	O. Schall, A. Belyaev, H. Seidel	Sparse Meshing of Uncertain and Noisy Surface Scattered Data
MPI-I-2005-4-001	M. Fuchs, V. Blanz, H. Lensch, H. Seidel	Reflectance from Images: A Model-Based Approach for Human Faces
MPI-I-2005-2-004	Y. Kazakov	A Framework of Refutational Theorem Proving for Saturation-Based Decision Procedures
MPI-I-2005-2-003	H.d. Nivelle	Using Resolution as a Decision Procedure
MPI-I-2005-2-002	P. Maier, W. Charatonik, L. Georgieva	Bounded Model Checking of Pointer Programs
MPI-I-2005-2-001	J. Hoffmann, C. Gomes, B. Selman	Bottleneck Behavior in CNF Formulas
MPI-I-2005-1-008	C. Gotsman, K. Kaligosi, K. Mehlhorn, D. Michail, E. Pyrga	Cycle Bases of Graphs and Sampled Manifolds
MPI-I-2005-1-007	I. Katriel, M. Kutz	A Faster Algorithm for Computing a Longest Common Increasing Subsequence
MPI-I-2005-1-003	S. Baswana, K. Telikepalli	Improved Algorithms for All-Pairs Approximate Shortest Paths in Weighted Graphs
MPI-I-2005-1-002	I. Katriel, M. Kutz, M. Skutella	Reachability Substitutes for Planar Digraphs
MPI-I-2005-1-001	D. Michail	Rank-Maximal through Maximum Weight Matchings
MPI-I-2004-NWG3-001	M. Magnor	Axisymmetric Reconstruction and 3D Visualization of Bipolar Planetary Nebulae
MPI-I-2004-NWG1-001	B. Blanchet	Automatic Proof of Strong Secrecy for Security Protocols
MPI-I-2004-5-001	S. Siersdorfer, S. Sizov, G. Weikum	Goal-oriented Methods and Meta Methods for Document Classification and their Parameter Tuning