Gradient-Enhanced Particle Filter for Vision-Based Motion Capture

Daniel Grest and Volker Krüger

Aalborg University Copenhagen, Denmark
Computer Vision and Machine Intelligence Lab
{dag,vok}@cvmi.aau.dk

Abstract. Tracking of rigid and articulated objects is usually addressed within a particle filter framework or by correspondence based gradient descent methods. We combine both methods, such that (a) the correspondence based estimation gains the advantage of the particle filter and becomes able to follow multiple hypotheses while (b) the particle filter becomes able to propagate the particles in a better manner and thus gets by with a smaller number of particles. Results on noisy synthetic depth data show that the new method is able to track motion correctly where the correspondence based method fails. Further experiments with real-world stereo data underline the advantages of our coupled method.

1 Introduction

Motion tracking and human pose estimation are important applications in motion analysis for sports and medical purposes. Motion capture products used in the film industry or for computer games are usually marker based to achieve high quality and fast processing.

Marker-less motion capture approaches often rely on gradient based methods [13, 3, 19, 7, 10, 18]. These methods estimate the parameters of a human body model by minimizing differences between model and some kind of observations, e.g. depth data from stereo, visual hulls or silhouettes. Necessary for minimization are correspondences between model and observed data. The main problem of these correspondence based optimization methods is, that they often get stuck in wrong local minima. From this wrong estimated pose they can usually not recover.

Other approaches like particle filters [5, 11] try to approximate the probability distribution in the state space by a large number of particles (poses) and are therefore unlikely to get stuck in local minima, because they can follow and test a large number of hypotheses. However, to be sure that the "interesting" region (*typical set*) of a high-dimensional state space is properly sampled, a large amount of particles is usually necessary [15, 2].

To take the advantages of both approaches, we combine a particle filter based approach [15, 11, 6] with correspondence based gradient estimation.

The effect can be interpreted in two ways: (1) Following the local gradient within the particle filter allows to find minima (including the global minimum) with less particles and (2) enhancing the gradient descent method with multiple hypothesis helps to avoid to get stuck in local minima.

The key to our approach and the main difference to a particle filter like CONDENSATION [11] is the gradient descent for each particle and the merging of particles. Particles, which are close to each other after the gradient descent, are merged into a single particle. Then, the idea is to re-distribute (propagate) the particles in a way that takes into account the local shape of the likelihood function. That way the number of particles can be greatly reduced while maintaining the ability to follow multiple hypotheses.

We will discuss our new approach in the context of marker less motion tracking from a single stereo view with two cameras. There is a wide variety of stereo algorithms available differing in quality and processing time. Commercial stereo cameras calculate depth data on chip using a simple algorithm [20] in real-time and keep the CPU free. Other more sophisticated algorithms need up to multiple minutes per image pair [8].

We will at first discuss relevant work within the field of motion tracking. Then, introduce our body model and the motion parameterization, which are used in the correspondence based optimization. The next section briefly explains important aspects of particle filter tracking methods, which are necessary for the combination. Then, results on synthetic data and real motion sequences are given, that show the advantages of the combined motion tracking. The last section concludes the paper with a short discussion of the presented achievements.

2 Related work

Motion tracking of the human body is addressed in the literature with different methods. A recent and extensive survey of vision based human motion tracking can be found in [16]. Approaches relevant to this work arise from different directions depending on the kind of input data, e.g. depth data or images, and the number of cameras. Visual hull approaches have shown to give very accurate results [13, 3]. They build the visual hull of the person from segmented images of multiple cameras and then fit a template model to the 3D hull. Usually some kind of gradient based optimization is utilized to estimate the motion parameters of the model. Fitting the template model directly to segmented images is another possibility as done in [19], where it was shown, that the marker-less approach has an accuracy similar to marker based tracking.

Similar to the visual hull approach is [12], where multiple stereo cameras observe the motion of a person. The resulting 3D points are then used with an Extended Kalman Filter to estimate the upper body motion.

When only two or less cameras are used for motion tracking, particle filters [11] or particle filtering methods are utilized [5]. Their advantage is, that multiple tracking hypotheses can be followed, because a large number of particles approximates the posterior probability of the system's state. Therefore, the tracking is not prone to get stuck in local minima and multiple hypotheses can be followed. This ability to track multiple hypotheses is very useful, because body parts can become occluded, if only a single viewpoint is used. Then, the occluded motion has to be 'guessed' in order to track successfully, when the occluded body part becomes visible again. However, if full body motion with 30 DOF is to be estimated the number of particles can approximate the posterior distribution only within a small region in the state space. Therefore, once again it is likely to face the problem of local minima. Otherwise the number of parti-

cles has to be increased, which increases computation time. For 30 DOF the necessary amount of particles can result in a computation time of up to multiple hours per image frame of a video sequence. However a parallel processing of particles is possible. Our approach decreases the amount of necessary particles significantly and allows such processing in reasonable time.

Motion tracking from a single stereo view as in this work has been addressed before in [4], where the human is modeled with 6 cylinders and motion can be roughly tracked with 10Hz. The authors use projective methods, which are inferior to direct methods as they state themselves in [4], but are easier to implement and require less computation time (no comparison is made). In [18] a direct approach is taken, where a human model consisting of spheres (meta-balls) is fitted to stereo data silhouettes from a single view. Due to the high number of estimated parameters, the method is not real-time capable. Both methods use a correspondence based gradient descent method, which requires manual initialization and can get stuck in local minima.

In [17], 3D body tracking is done using particle propagation. The Zakai Equation is applied to model the propagation of probability density function (pdf) over time in order to reduce the number of particles.

In a previous work [10] we showed that our direct approach is able to track upper arm motion with 5Hz and can therefore compete with the projective method of [4] according to processing time. Here we present a combination with a particle filter that allows us to track very noisy arm motion and complex full body motion of the whole body from stereo data alone, even though body parts are temporarily occluded.

A nice review on Monte Carlo-based techniques can be found in [15]. Classical papers about particle filtering are Condensation[5, 11], Sequential Importance Sampling [6] and sequential Monte Carlo [14].

3 Body Model and Motion Parameterization

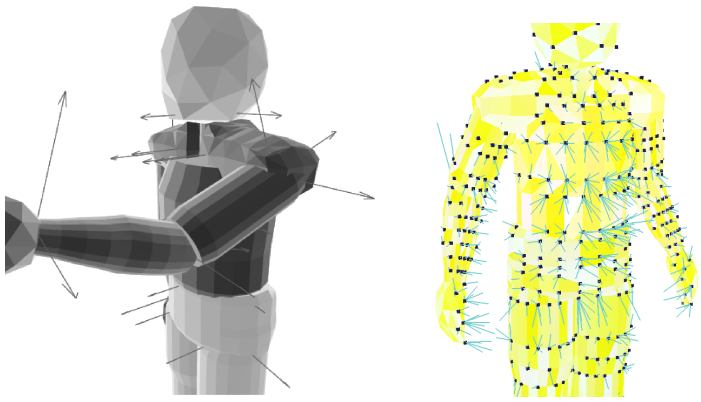

Fig. 1. Left: The body model with rotation axes shown as arrows. Right: The difference (small blue lines) between observed depth point and nearest model point (black boxes) is minimized.

The motion capabilities of the human model is based on the MPEG4 standard, with up to 180 DOF. An example model is shown in Fig. (1) left. The MPEG4 description allows to exchange body models easily and to re-animate other models with the captured motion data. The model for a specific person is obtained by silhouette fitting of a template model as described in [9].

The MPEG4 body model is a combination of kinematic chains. The motion of a point, e.g. on the hand, may therefore be expressed as a concatenation of rotations [10]. As the rotation axes are known, e.g. the flexion of the elbow, the rotation has only one degree of freedom (DOF), i.e. the angle around that axis. In addition to the joint angles, there are 6 DOF for the position and orientation of the object within the global world coordinate frame. For an articulated object with p joints we describe the transformation of the point $\boldsymbol{p}$ within the chain [10] as

$$\begin{aligned}\boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{p}) = &(\theta_x, \theta_y, \theta_z)^T + \\ &(R_x(\theta_\alpha) \circ R_y(\theta_\beta) \circ R_z(\theta_\gamma) \circ R_{\omega_1, q_1}(\theta_1) \circ \cdots \\ &\cdots \circ R_{\omega_p, q_p}(\theta_p))\boldsymbol{p}\end{aligned}$$

where $(\theta_x, \theta_y, \theta_z)^T$ is the global translation, R_x, R_y, R_z are the rotations around the global x, y, z-axes with Euler angles α, β, γ and $R_{\boldsymbol{\omega},\boldsymbol{q}}(\theta_i), i \in \{1..p\}$ denotes the rotation around the known axis with angle θ_i. The axis is described by the normal vector $\boldsymbol{\omega_i}$ and a point $\boldsymbol{q_i}$ on the axis.

Eq. 1 gives the position of a point $\boldsymbol{p}$ on a specific segment of the body (e.g. the hand) with respect to joint angles $\boldsymbol{\theta}$ and an initial body pose.

If the current pose is $\boldsymbol{\theta}_t$ and only relative motion is estimated the resulting Jacobian is:

$$J = \begin{bmatrix} 1 & 0 & 0 & \frac{\partial m_x}{\partial \theta_\alpha} & \frac{\partial m_x}{\partial \theta_\beta} & \frac{\partial m_x}{\partial \theta_\gamma} & \frac{\partial m_x}{\partial \theta_1} & .. & \frac{\partial m_x}{\partial \theta_p} \\ 0 & 1 & 0 & \frac{\partial m_y}{\partial \theta_\alpha} & \frac{\partial m_y}{\partial \theta_\beta} & \frac{\partial m_y}{\partial \theta_\gamma} & \frac{\partial m_y}{\partial \theta_1} & .. & \frac{\partial m_y}{\partial \theta_p} \\ 0 & 0 & 1 & \frac{\partial m_z}{\partial \theta_\alpha} & \frac{\partial m_z}{\partial \theta_\beta} & \frac{\partial m_z}{\partial \theta_\gamma} & \frac{\partial m_z}{\partial \theta_1} & .. & \frac{\partial m_z}{\partial \theta_p} \end{bmatrix} \tag{1}$$

The derivatives at zero are:

$$\left.\frac{\partial \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{p})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{0}} = \boldsymbol{\omega_j} \times (\boldsymbol{p} - \boldsymbol{q_j}) \tag{2}$$

where $j \in \{1, .., p\}$ and $\boldsymbol{q}_j$ is an arbitrary point on the rotation axis. The simplified derivative at zero is valid, if relative transforms in each iteration step of the *Nonlinear Least Squares* are calculated and if all axes and corresponding point pairs are given in world coordinates.

4 Gradient Enhanced Particle Filtering

Because our method combines the advantages of gradient based optimization methods with particle filtering, we give now a brief overview of important aspects of both. Then, we present the combined algorithm and discuss the main differences to particle filters.

4.1 Correspondence Based Pose Estimation

Correspondence based methods for pose estimation of articulated objects minimize an error function with respect to motion parameters. The human body can be modeled as an articulated object, consisting of multiple kinematic chains.

It is common to assume that the shape and size of the body model is known for the observed person, such that the minimization is only with respect to joint angles and the global transform. In that case, the kinematic chain simplifies to a chain of rotations around arbitrary axes in space Given here is a short description of the estimation algorithm, for more details see [10].

Estimating the motion of the human body from given 3D-3D correspondences $(\boldsymbol{p_i}, \tilde{\boldsymbol{p}}_{\boldsymbol{i}})$ is done here by solving a Nonlinear Least Squares Problem. The minimization yields the joint angles and the global orientation and position. For n correspondences the minimization problem is given as:

$$\min_{\boldsymbol{\theta}} \sum_{i}^{n} |\boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{p_i}) - \tilde{\boldsymbol{p}}_{\boldsymbol{i}}|^2 \quad . \tag{3}$$

To find the minimizer with the iterative *Gauss-Newton* method the Jacobian of the residual functions, Eq. (1), is necessary.

The points $\boldsymbol{p_i}$ and $\tilde{\boldsymbol{p}}_{\boldsymbol{i}}$ form a correspondence. For each observed point $\tilde{\boldsymbol{p}}_{\boldsymbol{i}}$ the closest point $\boldsymbol{p_i}$ on the model is sought. Therefore, different observed points can have the same corresponding model point. This is shown in Fig. 1 right, where each correspondence is shown as a small blue line and the model points are drawn as black boxes.

The minimization problem is solved with the dampened Gauss-Newton method[1], which is similar to the Levenberg-Marquardt[1] method. Dampening ensures that the parameter change does not increase infinitely, if the determinant of the Gram matrix $J^T J$ is close to zero, which can happen when a body part is largely occluded. The solution is found by solving iteratively:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - (J^T J + \lambda I)^{-1} J^T \boldsymbol{r}(\boldsymbol{\theta}_t) \tag{4}$$

where the Jacobian J is given in equation (1), I is the identity matrix, λ is the dampening value (set to 0.1) and $\boldsymbol{r}(\boldsymbol{\theta}_t)$ is the vector with current residuals. For each point there are three residuals, one for each component (x, y, z):

$$\boldsymbol{r_i}(\boldsymbol{\theta}_t) = \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{p_i}) - \tilde{\boldsymbol{p}}_{\boldsymbol{i}} \tag{5}$$

and $\boldsymbol{r_i} = (r_{ix}, r_{iy}, r_{iz})$.

The optimization consists of two loops: The Gauss-Newton method (GN) loops until it converges on the given set of point correspondences. Because the correspondences are not always correct, new correspondences are calculated again after convergence of the GN. Then, GN starts anew with the improved set of correspondences. This Iterative Closest Point (ICP) method [1] can be repeated until convergence. However, the gain in more then 3 ICP iterations is very small, therefore we usually apply only 2 or 3 ICP optimizations. We will use in the following the term "gradient tracking" in order to refer to this technique.

It is important to note, that the Gauss-Newton optimization is more efficient than a standard Gradient Descent and requires less control parameters. However we will refer to it as "gradient tracking", because both methods rely on the gradient.

4.2 Particle Filter

The new method borrows some ideas from particle filter methods like CONDENSATION [11]. Particle filter approaches aim at estimating the posterior probability distribution of a system state $\boldsymbol{z}_t$ at time t, the person's current pose in our case, from observations $I_1, \ldots, I_t$:

$$\begin{aligned} & p(\boldsymbol{z}_t | I_1, I_2, \ldots, I_t) \equiv p_t(\boldsymbol{z}_t) \\ & = \int_{\boldsymbol{z}_{t-1}} p(I_t|\boldsymbol{z}_t) p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) p_{t-1}(\boldsymbol{z}_{t-1}) \ . \end{aligned} \tag{6}$$

In this equation the state space is randomly sampled according to $p_{t-1}(\boldsymbol{z}_{t-1})$ and propagated according to a motion/diffusion model $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. A likelihood probability $p(I_t|\boldsymbol{z}_t)$ for each particle is calculated, which reflects how good the observations fit to the hypothesis (the position of the particle in the state space). The posterior probability is then approximated by the weight and density of particles within the state space.

The major problem with these approaches is, that the number of necessary particles usually needs to reflect the dimensionality of the state space[15, 2]. If full body motion with 30 DOF is to be estimated the particles can usually approximate the posterior distribution only within a small region in the state space.

If depth data is the only input, calculation of the likelihood requires computation of differences between observed points and model surface. One way to compute these differences is a nearest neighbor search as described in the previous section. This search is, however, expensive in terms of computation time. Therefore, methods are desirable, which reduce the number of particles and allow to distribute the reduced set of particles in the most important regions of the state space.

4.3 Combination

In order to get by with a smaller number of particles, we apply the gradient tracking (section 4.1) to each particle. Since similar particles move to the *same* optimum, they can be merged with an appropriate adaption of the particle weight. The state space posterior probability is approximated in a particle filter by the particle weights and their spatial density. A possible weight adaption is the average of merged particle weights. However, all merged particles are nearly at the same position in the state space and therefore have nearly the same weight (likelihood). As a result it is sufficient to assign them the same weight. Another possible weight adaption is the addition of weights. However, these would favor large flat valleys in the posterior probability surface, because all particles within in this valley will descent towards the same minimum. This will also lead to a clustering of particles at specific positions, which is not desired, because we want to track as many hypotheses as possible. If the merged particles are redistributed in the next time step only according to a fixed motion model and diffusion model, it is likely that they end up in the same locally convex region (valley) and again merge at the same position in state space. Each valley can be understood as one likely pose hypothesis. It is desirable to track as many hypotheses as possible with a fixed amount of particles. To increase the number of tracked hypotheses and decrease

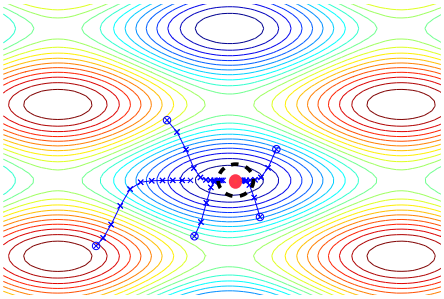

Fig. 2. Principle of the combined method. Particles whose position is within a specific area after optimization (black dashed circle) are merged into one particle (red circle).

the amount of particles per valley, the particles need to be redistributed at the next time step of Eq. (6) according tho the size and shape of their valley. This can be understood as enhancing our rather simple motion model (fixed velocity) to include the shape of the local probability surface.

In order to achieve this, we merge all particles after optimization, which are close to each other according to some distance d. In detail, if we use N particles $\boldsymbol{z}_1^{t-}, \dots \boldsymbol{z}_N^{t-}$ in our particle filter, then we have after the merging $M < N$ meaningful particles $\boldsymbol{z}_1^t, \dots \boldsymbol{z}_M^t$ left, and $N - M$ particles were merged into the remaining M particles.

Then, in order to distribute the particles in the next time step optimally, we estimate the size of the locally convex region by computing the covariance Σ_i of all those particles. Let $\boldsymbol{z}_{i_1}^{t-}, \dots \boldsymbol{z}_{i_k}^{t-}$ be the particles that merged together into the particle $\boldsymbol{z}_i^t$. The $^-$ at the top denotes the particles *before* their gradient descent and merging, t denotes the time step. The final particle $\boldsymbol{z}_i^t$ (without the $^-$) after gradient descent and merging (red circle in Fig.2) is the one with highest likelihood and is used as the mean for the covariance:

$$\Sigma_i = \frac{1}{K} \sum_{j}^{K} (\boldsymbol{z}_j^{t-} - \boldsymbol{z}_i^t)(\boldsymbol{z}_j^{t-} - \boldsymbol{z}_i^t)^T \tag{7}$$

where K is the amount of particles, which merged into $\boldsymbol{z}_i^t$. It is important to note that the covariance is calculated from the particles *before* the gradient descent and merging. Fig. 2 illustrates the merging. The blue lines show a few steps of the gradient descent for five particles. They are within a certain region (the black dashed circle) after optimization and therefore merged. The idea is to distribute particles in the next frame outside that locally convex region (valley), because otherwise they would end up again at the same position and give no additional information about the state space. Thus, at the next time step $t + 1$ N new particles are drawn from the remaining M particles of time step t according to the prior p_{t+1}. Then, each particle is propagated according to some motion model $f(\boldsymbol{z}^{t+1-})$ with added Gaussian noise. Let the particle $\boldsymbol{z}^{t+1-}$ be spawned off from the particle $\boldsymbol{z}_i^t$. The covariance of the above Gaussian is given by the covariance Σ_i of the original particle $\boldsymbol{z}_i^t$.

Our gradient enhanced particle filter can be summarized as follows: Input to our algorithm are the depth points for the current image frame and an initial pose in the beginning. The steps of the algorithm are for each frame similar to a particle filter method except for the gradient descent and the merging of particles.

1. Only at the first frame: Distribute particles according to an initial distribution in the vicinity of the given initial pose.
2. Draw new particles $\boldsymbol{z}_1^{t-}, \ldots \boldsymbol{z}_N^{t-}$ according to $p_t(\boldsymbol{z}_t)$.
3. Distribute and propagate each particle according to the covariance Σ_i of the original particle and propagate according to a diffusion/propagation model: $p(\boldsymbol{z}^{t+1-}|\boldsymbol{z}^{t-}, \Sigma) = Gauss(f(\boldsymbol{z}^{t-}), \Sigma)$. Here, the motion model consists of a deterministic motion model f plus Gaussian noise, and Σ_i defines the Gaussian covariance matrix.
4. Gradient Descent for each particle $\boldsymbol{z}_i^{t+1-}$ with 2 or 3 ICP optimizations:
   (a) Render model in current pose
   (b) Calculate visible model points
   (c) For each observed point find the closest model point, which makes a correspondence
   (d) Calculate a new pose by minimizing the differences of the correspondences
5. Assign the likelihood, calculated from the residual error.
6. Merge particles $\boldsymbol{z}_{i_1}^{t+1-}, \ldots \boldsymbol{z}_{i_k}^{t+1-}$, which are within a certain distance d to each other into one particle $\boldsymbol{z}_i^{t+1}$.
7. For each particle $\boldsymbol{z}_i^{t+1}$: calculate the covariance matrix Σ_i from all the particles $\boldsymbol{z}_{i_1}^{t+1-}, \ldots \boldsymbol{z}_{i_k}^{t+1-}$ which merged into $\boldsymbol{z}_i^{t+1}$.

The motion model $f(\boldsymbol{z}^{t-1})$ predicts the new pose of the human. In our experiments, we assume constant velocity. The distance d was chosen to be 4 degrees, such that particles are merged only if each single joint angle differs less than 4 degrees to a neighboring particle. The distance check is only applied on joint angles, not on the position of the body model in the world, because it is highly unlikely, that the same pose is estimated at different positions. This way additional control parameters are avoided.

The initial covariance for each particle is chosen to equal the distance d, such that particles are definitely distributed outside the merge area. The covariance is also reduced by 10% each frame. Without this reduction the covariance can increase indefinitely. Especially if only one particle is left in a valley, it is desirable to reduce the covariance, such that it is ensured, that nearby poses are tested.

5 Synthetic data with noise

In order to show the robustness of the new method to noisy measurements we conduct an experiment on depth data generated with openGL on a synthetic sequence. The motion involves four DOF, the elbow flexion and the shoulder flexion abduct and twisting. Three example images out of the 176 frame sequence are shown in Fig. 3. The depth data is calculated from the z-buffer values after rendering the model.

For testing, Gaussian noise with different standard deviations is added to the original depth data. Fig. 4 shows two views on the depth data and the model in starting pose.

Fig. 3. First frame (left). Frame 60 and frame 176 of the synthetic sequence.

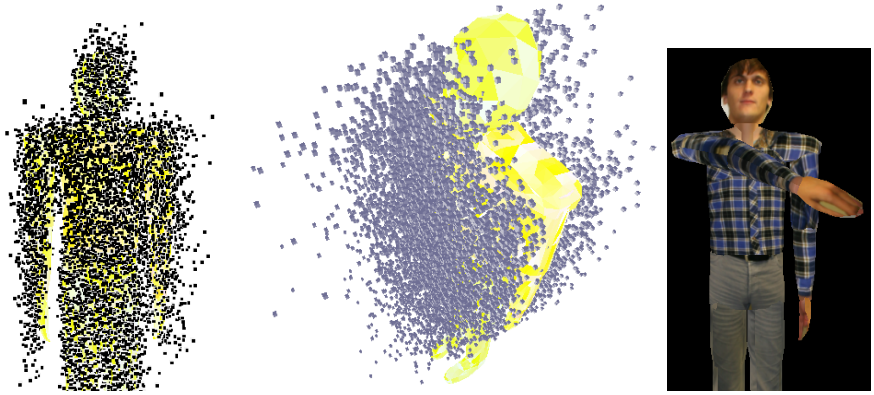

Fig. 4. Two views on the depth point cloud with Gaussian noise (deviation 15cm). Right: The standard tracking method loses track after a few frames and gets stuck in the pose shown.

The z-buffer image was sub-sampled in order to generate approximately 10000 depth points. Approximately 750 points are visible on the right arm each frame.

At a deviation rate of 15cm in depth and 1cm in the other directions the normal tracking methods loses track after a few frames and gets stuck in an arbitrary pose as shown in the right of Fig. 4.

The multi-hypotheses tracking with 20 particles is able to track the motion correctly in spite of the heavy noise. The difference to the ground truth is shown in Fig. 5. In the beginning the estimate is far off with 50 degrees however the plot shows, that the tracking recovers from the wrong local minimum. For all hypotheses the minimum difference to the ground truth is taken, because the particle with the largest likelihood does not always give the correct pose. Plotted is the absolute error in degree over the whole sequence. The elbow difference is high around frames where the arm is close to the body as at frame 60 and 105, because the correspondences established with nearest neighbor are then most ambiguous. Where the arm is far away from the body the noise

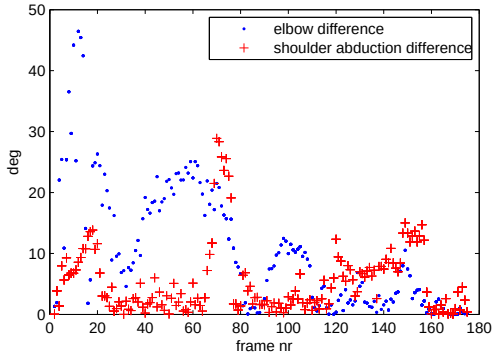

Fig. 5. Difference to ground truth with noise (deviation 15cm). The multi hypotheses tracking is able to track the whole sequence.

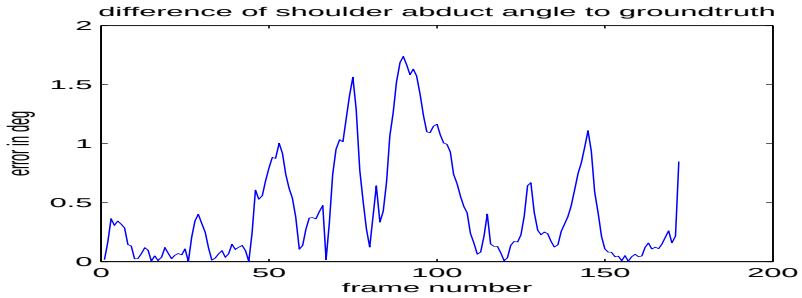

Fig. 6. Difference of the shoulder abduct angle to ground truth without noise.

on the data does not result in so many wrong correspondences. Therefore, the error is decreased. Though the error is still large for most frames, because of the heavy noise, the results show, that the new method can track the motion over the whole sequence.

Without noise the gradient tracking estimates joint angles, whose difference to the ground truth is close to zero error as shown in Fig. 6 for the abduction angle of the shoulder. The error is not zero, because the model surface is approximated by the vertices of the model's triangles. Therefore, the nearest neighbor correspondences are not perfect.

6 Real Data

In order to show the possibilities with our new method, we give further results for a video sequence, which was recorded in a motion capture lab with 8 cameras at 25fps. Two of the cameras were arranged approx. 25 cm next to each other, such that stereo depth estimation can be performed. The stereo algorithm[8] produces dense accurate

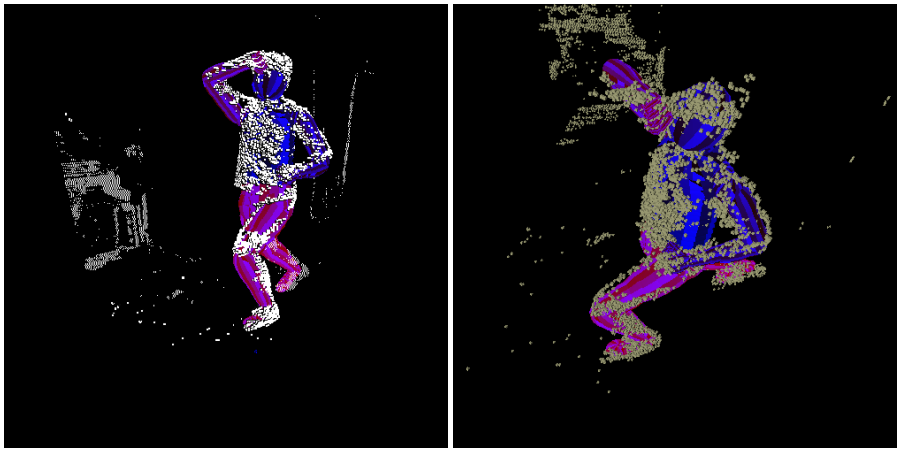

Fig. 7. Two views on the input data. Approx 10000 points are shown as white boxes.

results in non-homogenous regions within approx. one minute computation time per frame. The used effective image size is 512x512. Fig. 7 shows the depth data that is used as input. The only assumption made here is, that no scene objects are within 80cm range of the person. Also knowledge about the floor position was incorporated from camera calibration, which was conducted for the internal parameters with a small checkerboard pattern according to [21]. This calibration also yields the orientation and distance of the stereo cameras. The external parameters (orientation of the floor) were then estimated with a large checkerboard pattern lying on the floor.

The initial pose of the person is provided manually. Estimated are 24 DOF, these are in detail 3 at each shoulder, one elbow angle, 3 at each hip, one angle for each knee, one angle at the ankle and 6 parameters for the global orientation and position.

The estimation time on a 2Ghz intel Core2 Duo (1 CPU) was about 2 seconds per particle and approximately 10000 data points. For 1000 data points and 14 DOF the computation time is about 200ms per particle.

Fig. 8 shows a few frames from the resulting estimation. The multi-hypotheses tracking with 100 particles is able to track the whole sequence of 180 frames (first 3 rows in the Fig.), even though one arm and one leg are almost completely occluded temporarily. Approximately 10000 data points from the complete set of 40000 reliable data points are used per frame, resulting in a Jacobian of size 30000×24.

The gradient tracking method (row 4) is able to track correctly up to 130 frames, but loses track when the person turns and the body parts become occluded (second image last row). The gradient tracking method is lost in that case and is unable to recover.

7 Conclusions

We presented a new method for motion tracking, that combines gradient based optimization from correspondences and motion tracking with particle filters. The combined approach allows to track arm motion in spite of heavy noise, where a normal gradient

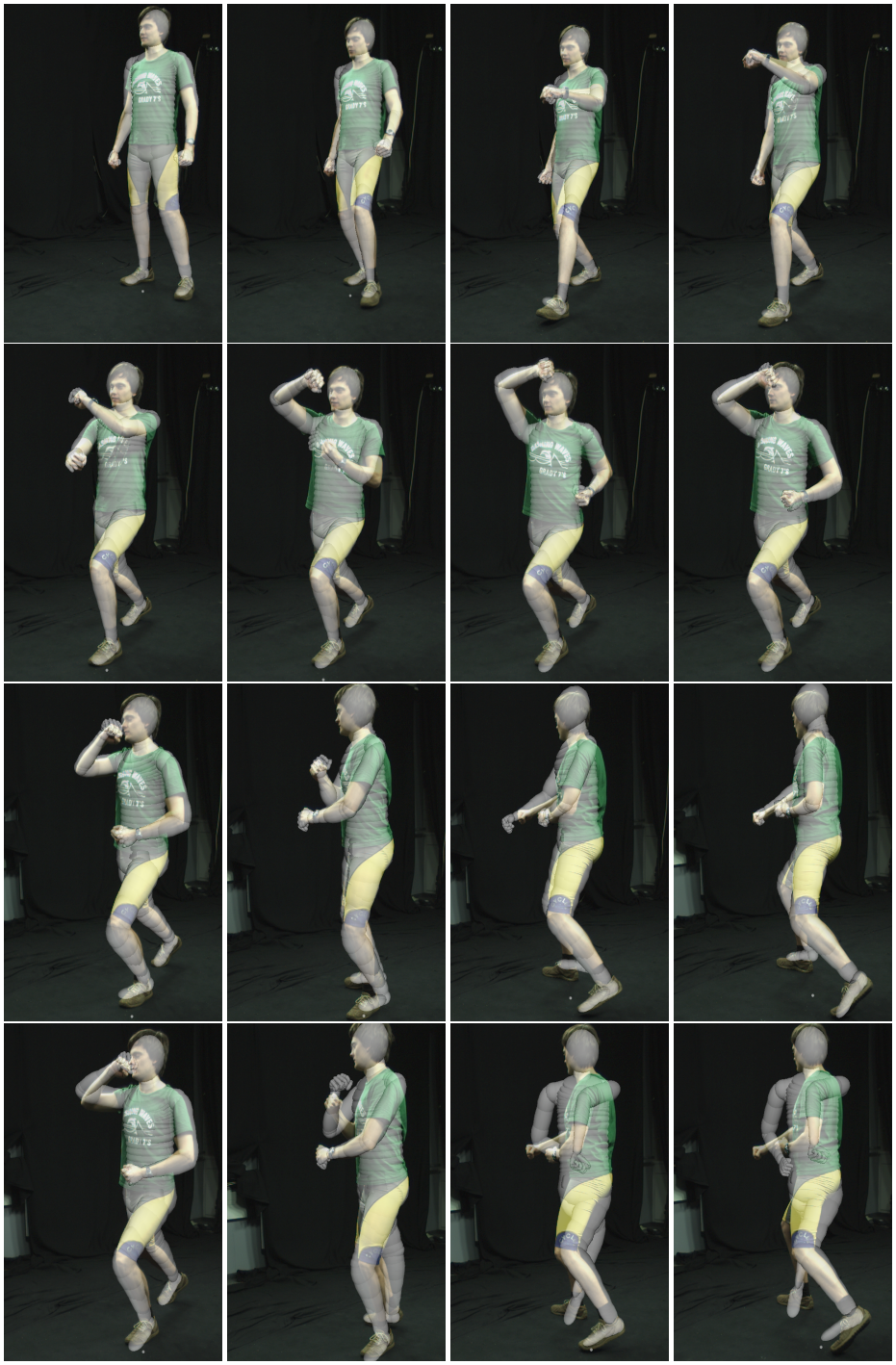

Fig. 8. Estimation results of the new combined method with 24 DOF (first 3 rows). The figure shows the original images together with the projected model in the estimated pose (overlayed in white). The last row shows the estimation results for the ”gradient tracking” for the same images as in the third row. The ”gradient tracking” loses track, because the right arm gets largely occluded, and is unable to recover.

descent method fails. Further results on stereo video sequences showed that motion with 24 DOF can be tracked from a single viewpoint. At frames where the normal tracking gets stuck in local minima and thus loses track, the new method recovers and estimates correct poses.

The main contribution of the new method is the enhanced motion model, which includes the shape of the locally convex regions of the probability surface by estimation of the covariance matrix from merged particles. In that way the number of particles is used more efficiently and allows to track a higher number of hypotheses.

We will exploit in the future further aspects of the new method: (1) the likelihood probability of the particle filter allows to easily include image information into the tracking, as for example motion areas from temporal image differences, (2) increase speed by parallel processing of particles and (3) include motion recognition probabilities into the motion model.

References

1. Edwin K.P. Chong and Stanislaw H. Zak. *An Introduction to Optimization, Second Edition*. Wiley, 2001.
2. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.
3. E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Reconstructing Human Shape and Motion from Multi-View Video. In *2nd European Conference on Visual Media Production (CVMP)*, London, UK, 2005.
4. D. Demirdjian, T. Ko, and T. Darrell. Constraining Human Body Tracking. In *Proceedings of ICCV*, Nice, France, October 2003.
5. J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *CVPR*, volume 2, 2000.
6. A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. 10:197–209, 2000.
7. Lars Mündermann et al. Validation of a markerless motion capture system for the calculation of lower extremity kinematics. In *Proc. American Society of Biomechanics*, Cleveland, USA, 2005.
8. L. Falkenhagen. Hierarchical block-based disparity estimation considering neighbourhood constraints. In *International workshop on SNHC and 3D Imaging*, 1997.
9. D. Grest, D. Herzog, and R. Koch. Human Model Fitting from Monocular Posture Images. In *VMV*, Nov. 2005.
10. D. Grest, J. Woetzel, and R. Koch. Nonlinear Body Pose Estimation from Depth Images. In *Proc. of DAGM*, Vienna, Sept. 2005.
11. M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Compputer Vision*, 29:5–28, 1998.
12. R. Stiefelhagen J. Ziegler, K. Nickel. Tracking of the Articulated Upper Body on Multi-View Stereo Image Sequences. In *CVPR*, volume 1, pages 774–781, New York, June 2006. IEEE Computer Society.
13. Roland Kehl, Matthieu Bray, and Luc J. Van Gool. Full Body Tracking from Multiple Views Using Stochastic Sampling. In *Proc. CVPR*, pages 129–136, 2005.
14. J.S. Liu and R. Chen. Sequential monte carlo for dynamic systems. 93:1031–1041, 1998.
15. D. MacKay. *Learning in Graphical Models, M.Jordan (ed.)*, chapter Introduction to Monte Carlo Methods, pages 175–204. MIT Press, 1999.

16. T. Moeslund, A. Hilton, and V. Krueger. A Survey of Advances in Vision-based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, 104(2-3):90–127, 2006.
17. H. Moon, R. Chellappa, and A. Rosenfeld. 3d object tracking using shape-encoded particle propagation. 2001.
18. R. Plänkers and P. Fua. Model-Based Silhouette Extraction for Accurate People Tracking. In *Proc. of ECCV*, pages 325–339. Springer-Verlag, 2002.
19. B. Rosenhahn, U. Kersting, D. Smith, J. Gurney, T. Brox, and R. Klette. A System for Marker-Less Human Motion Estimation . In W. Kropatsch, editor, *DAGM*, Wien, Austria, Sept. 2005.
20. Videre. Videre Design: Stereo On A Chip. www.videredesign.com, 2007.
21. Z. Zhang. Flexible Camera Calibration By Viewing a Plane From Unknown Orientations. In *ICCV*, pages 666–673, Corfu, Greece, 1999.