Motion Capture Using Joint Skeleton Tracking and Surface Estimation

Juergen Gall[1,2], Carsten Stoll[2], Edilson de Aguiar[2], Christian Theobalt[3], Bodo Rosenhahn[4], and Hans-Peter Seidel[2]
[1]BIWI, ETH Zurich [2]MPI Informatik [3]Stanford University [4]Leibniz-Universität Hannover
gall@vision.ee.ethz.ch {stoll,edeaguia,hpseidel}@mpi-inf.mpg.de theobalt@stanford.edu rosenhahn@tnt.uni-hannover.de

Abstract

This paper proposes a method for capturing the performance of a human or an animal from a multi-view video sequence. Given an articulated template model and silhouettes from a multi-view image sequence, our approach recovers not only the movement of the skeleton, but also the possibly non-rigid temporal deformation of the 3D surface. While large scale deformations or fast movements are captured by the skeleton pose and approximate surface skinning, true small scale deformations or non-rigid garment motion are captured by fitting the surface to the silhouette. We further propose a novel optimization scheme for skeleton-based pose estimation that exploits the skeleton's tree structure to split the optimization problem into a local one and a lower dimensional global one. We show on various sequences that our approach can capture the 3D motion of animals and humans accurately even in the case of rapid movements and wide apparel like skirts.

1. Introduction

Estimating the 3D motion of humans or animals is a fundamental problem in many applications, including realistic character animation for games and movies, or motion analysis for medical diagnostics and sport science. Ideally, one expects to both estimate an articulated rigid-body skeleton that explains the overall motion of the character, as well as the potentially non-rigid deformation of the surface, *e.g.* caused by tissue or garment. On the one end of the spectrum, many current automatic approaches track only a skeleton model which poses strong restrictions on the subject, like tight clothing. Since garment motion, for instance, is non-rigid and rarely aligned with the motion of the underlying articulated body, these algorithms often dramatically fail if the subject wears wide clothing like a dress. On the other end of the spectrum, there are methods which capture a faithfully deforming 3D surface of the subject, but do not provide an underlying skeleton.

In contrast, our approach captures both skeletal motion as well as an accurately deforming surface of an animal or human by fitting a body model to multi-view image data. Our body model is a combination of a bone skeleton with joints, as well as a surface whose deformation is only loosely coupled with the skeleton motion. We can accurately capture both detailed surface deformations and motion of wide apparel, which are essential for realistic character animations. At the same time, the skeleton provides a low-dimensional motion parametrization which facilitates tracking of fast movements. Our captured performances can be easily edited and used in animation frameworks typical for games and movies, which are almost exclusively skeleton-based. Finally, our approach exceeds the performance of related methods from the literature since both accurate skeleton and surface motion are found fully-automatically. This is achieved by the following contributions:

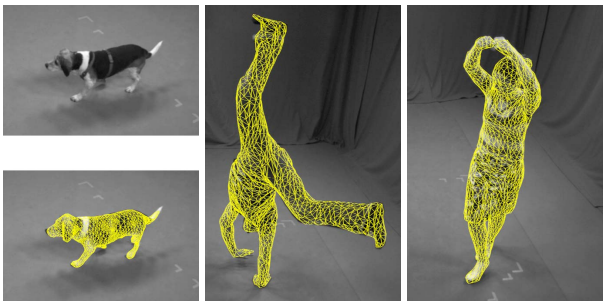


Figure 1. Our approach captures the motion of animals and humans accurately even in the case of rapid movements and wide apparel. The images show three examples of estimated surfaces that are superimposed on the images.

- Our approach recovers the movement of the skeleton and the temporal deformation of the 3D surface in an interleaved manner. To find the body pose in the current frame, we first optimize the skeletal pose and use simple approximate skinning to deform the detailed surface of the previous time step into the current time step. Once converged, the fine surface deformation at the current time step is computed without limiting the deformation to comply with the skeleton. This im-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

proves also the skeleton estimation and avoids errors caused by wide apparel since the refined surface model of the previous frame provides a good approximation of the surface at the current frame.

- Since skeleton-based pose estimation is more constrained than surface estimation, our approach is less sensitive to silhouette noise than comparable visual hull approaches and runs even on medium quality multi-view sequences like the HumanEva benchmark [23]. The reliability and accuracy of our approach is demonstrated on 12 sequences that consist of over 5000 frames with 9 different subjects (including a dog) performing a wide range of motions and wearing a variety of clothing.[1]

- Since local optimization methods get stuck in local minima and cannot recover from errors, they cannot track challenging sequences without manual interaction. In order to overcome the limitations of local optimization, we propose a novel optimization scheme for skeleton-based pose estimation. It exploits the tree structure of the skeleton to split the optimization problem into a local and a lower dimensional global optimization problem.

The optimization scheme is motivated by the observation that local optimization is efficient and accurately tracks most frames of a sequence. However, it fails completely in some frames where the motion is fast or the silhouettes are noisy. The error often starts at a certain limb or branch of the skeleton and is propagated through the kinematic chain over time until the target is irrevocably lost. Our approach interferes before the error spreads. It detects misaligned limbs after local optimization and re-estimates the affected branch by global optimization. Since global optimization is only performed for few frames and for a lower dimensional search space, the approach is suitable for large data sets and high dimensional skeleton models with over 35 degrees of freedom.

2. Related Work

Marker-less human motion capture has been studied for more than 25 years and is still a very active field in computer vision [20]. From the beginning, skeleton-based pose estimation techniques have been popular where articulated bone hierarchies model a human's motion and simple shape proxies such as cylinder or superquadrics approximate the surface.

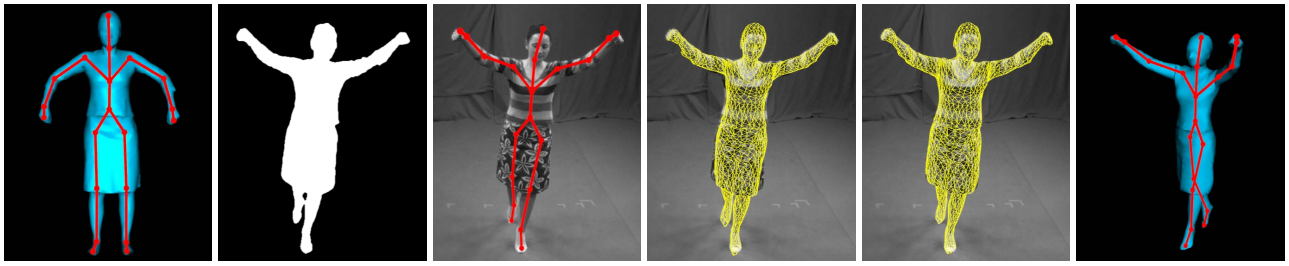
Bregler and Malik [7] represent the kinematic chain by twists and estimate the pose by local optimization. Stochastic meta descent for local optimization has been used in [18]. Gavrila and Davis [16] propose a search space decomposition where the pose of each limb is estimated in a hierarchical manner according to the kinematic chain. Starting with the torso and keeping the parameters of the other limbs fixed, the pose of each limb is locally searched in a low-dimensional space one after another. This approach, however, propagates errors through the kinematic chain such that the extremities suffer from estimation errors of preceding limbs. Drummond and Cipolla [13] iteratively propagate the distributions of the motion parameters for the limbs through the kinematic chain to obtain the maximum a posteriori pose for the entire chain subject to the articulation constraints. Besides stochastic approaches [24, 12], global optimization techniques [9, 15] have been also proposed to overcome the limitations of local optimization. However, global optimization is still too expensive for large data sets and skeletons with many degrees of freedom.

Since articulated models are not very realistic models of the human body, implicit surfaces based on metaballs [21], shape-from-silhouette model acquisition [8], or the learned SCAPE body model [1, 3] have been proposed. While these approaches model the human body without clothing, Balan and Black [2] have used SCAPE to estimate the human body underneath clothes from a set of images. Tracking humans wearing more general apparel has been addressed in [22] where a physical model of the cloth is assumed to be known.

In contrast to skeleton-based approaches, 3D surface estimation methods are able to capture time-varying geometry in detail. Many approaches like [25, 27] rely on the visual hull but suffer from topology changes that occur frequently in shape-from-silhouette reconstructions. Mesh-based tracking approaches as proposed in [11] and [10] provide frame-to-frame correspondences with a consistent topology. Fitting a mesh model to silhouettes and stereo, however, requires a large amount of correspondences to optimize the high dimensional parameter space of a 3D mesh. This, in turn, makes them more demanding on processing time and image quality than skeleton-based methods.

Our approach is most similar to the work of Vlasic et al. [28] where a two-pass approach has been proposed. In the first pass, a skeleton is geometrically fit into the visual hull for each frame. The second pass deforms a template model according to the estimated skeleton and refines the template to fit the silhouettes. Despite of visual appealing results, a considerable amount of manual interaction is required, namely up to every 20th frame, to correct the errors of the skeleton estimation. The errors are caused by fitting the skeleton to the visual hull via local optimization without taking a complete surface model or texture information into account. Moreover, their visual hull approach is sensitive to silhouette errors. In contrast, our local-global optimization makes for a fully-automatic approach that also works on data of poor image quality.

[1] For results and data, see `www.vision.ee.ethz.ch/~gallju`.



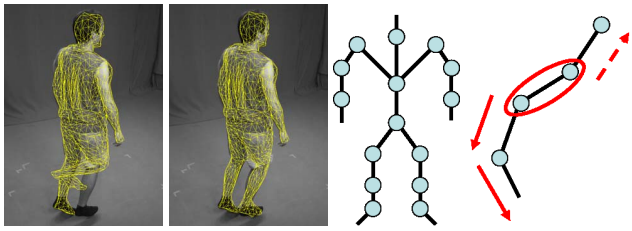
(a) Mesh model (b) Segmented images (c) Estimated skeleton (d) Deformed surface (e) Estimated surface (f) Estimated 3D Model

Figure 2. Having an articulated template model *(a)* and silhouettes *(b)* from several views, our methods tracks the skeleton and estimates the time-varying surface consistently without supervision by the user *(f)*. Using the estimated surface of the previous frame, the pose of the skeleton *(c)* is optimized such that the deformed surface *(d)* fits the image data. Since skeleton-based pose estimation is not able to capture garment motion *(d)*, the surface is refined to fit the silhouettes *(e)*.

3. Overview

The performance of an animal or human is captured by synchronized and calibrated cameras and the silhouettes are typically extracted by background subtraction or chroma-keying. Our body model comprises of two components, a 3D triangle mesh surface model $\mathcal{M}$ with 3D vertex locations V_i, and an underlying bone skeleton as shown in Figure 2 a. We assume that a 3D surface model $\mathcal{M}$ of the tracked subject in a static pose is available. It might be acquired by a static full-body laser scan or by shape-from-silhouette methods. In our experiments, we demonstrate results for both cases, but would like to note that model acquisition is outside of the scope of this paper. A kinematic skeleton is then inserted into the 3D mesh. In our case, an object-specific skeleton with usually around 36 degrees-of-freedom is generated by manually marking the joint positions. Thereafter, weights $\rho_{i,k}$ are automatically computed for each V_i which describe the association of V_i with each bone k [5]. The weights allow us to do skinning, *i.e.* a simple approximation of non-linear surface deformation based on the skeleton pose. Weighted skinning is used to interpolate the joint transformations on a per-vertex-basis. We use quaternion blend skinning [17] which produces less artifacts than linear blend skinning methods.

An outline of the processing pipeline is given in Figure 2. Starting with the estimated mesh and skeleton from the previous frame, the skeleton pose is optimized as described in Section 4 such that the projection of the deformed surface fits the image data in an globally optimal way (Figure 2 c). Since this step only captures deformations that can be approximated by articulated surface skinning (Figure 2 d), subsequently the non-rigid surface is refined as described in Section 5 (Figure 2 e). The estimated refined surface and skeleton pose serve as initialization for the next frame to be tracked.



(a) (b) (c) (d)

Figure 3. Although local optimization is prone to errors, often only a single branch of the kinematic chain is affected *(a)*. This reduces the computational burden for global optimization since it can be performed in a lower dimensional subspace to correct the estimation error *(b)*. After detecting misaligned limbs *(red circle)*, the kinematic chain is traversed *(red arrows)* to label bones and associated joints that have to be globally optimized *(c,d)*.

4. Skeleton-based Pose Estimation

Since local pose optimization is prone to errors and global pose optimization is very expensive, our method estimates poses in two phases. The first phase searches for the nearest local minimum of an energy functional that assesses the model-to-image alignment based on silhouettes and texture features. To this end, the whole articulated skeleton is optimized locally (Section 4.1). Subsequently, misaligned bones are detected by evaluating the energy E_k of each rigid body part. When the energy exceeds a given threshold, the affected limb is labeled as misaligned. In addition, the preceding limb in the kinematic chain is also labeled when the joint between the limbs has less than three degrees of freedom (*e.g.* knee or elbow). For instance, a wrong estimate of the shank might be caused by a rotation error along the axis of the thigh. Then the labeling process is continued such that all bones until the end of the branch are labeled as illustrated in Figure 3. Thereafter, the labeled bones are re-estimated by global optimization (Section 4.2).

4.1. Local Optimization

The articulated pose is represented by a set of twists $\theta_j \hat{\xi}_j$ as in [7]. A transformation of a vertex V_i which is associated with bone k_i and influenced by n_{k_i} out of totally N joints is given by

$$T_\chi V_i = \prod_{j=0}^{n_{k_i}} \exp\left(\theta_{\iota_{k_i}(j)} \hat{\xi}_{\iota_{k_i}(j)}\right) V_i, \tag{1}$$

where the mapping ι_{k_i} represents the order of the joints in the kinematic chain. Since the joint motion depends only on the joint angle θ_j, the state of a kinematic chain is defined by a parameter vector $\chi := (\theta_0 \xi_0, \Theta) \in \mathbb{R}^d$ that consists of the six parameters for the global twist $\theta_0 \hat{\xi}_0$ and the joint angles $\Theta := (\theta_1, \ldots, \theta_N)$. We remark that (1) can be extended by taking the skinning weights into account as in [4].

For estimating the parameters χ, a sufficient set of point correspondences between the 3D model V_i and the current frame x_i is needed. For the local optimization, we rely on silhouette contours and texture. Contour correspondences are established between the projected surface and the image silhouette by searching for closest points between the respective contours. Texture correspondences between two frames are obtained by matching SIFT features [19]. In both cases, the 2D correspondences are associated with a projected model vertex V_i yielding the 3D-2D correspondences (V_i, x_i). In the contour case, x_i is the point on the image contour closest to the projected vertex location v_i in the current frame. In the texture case, x_i is the 2D location in the current frame that is associated with the same SIFT feature as the projected vertex V_i in the previous frame. Since each 2D point x_i defines a projection ray that can be represented as Plücker line $L_i = (n_i, m_i)$ [26], the error of a pair $(T_\chi V_i, x_i)$ is given by the norm of the perpendicular vector between the line L_i and the transformed point $T_\chi V_i$:

$$\left\| \Pi\left(T_\chi V_i\right) \times n_i - m_i \right\|_2, \tag{2}$$

where Π denotes the projection from homogeneous coordinates to non-homogeneous coordinates. Using Equations (1) and (2), one obtains the weighted least squares problem

$$\underset{\chi}{\operatorname{argmin}} \frac{1}{2} \sum_i w_i \left\| \Pi\left(T_\chi V_i\right) \times n_i - m_i \right\|_2^2 \tag{3}$$

that can be solved iteratively and linearized by using the Taylor approximation $\exp(\theta\hat{\xi}) \approx I + \theta\hat{\xi}$, where I denotes the identity matrix. In order to stabilize the optimization, the linear system is regularized by $\beta\theta_j = \beta\hat{\theta}_j$ where $\hat{\theta}_j$ is the predicted angle from a linear 3rd order autoregression and β is a small constant. The pose $\hat{\chi}$ represented by all $\hat{\theta}_j$ can be regarded as a conservative prediction for the current frame. Since the optimization regards the limbs as rigid structures, the mesh is updated between the iterations by quaternion blending [17] to approximate smooth surface deformation.

While contour correspondences are all weighted equally with $w_i^C = 1$, the texture correspondences have higher weights w_i^T during the first iteration since they can handle large displacements. For the first iteration, we set the weights such that $\sum_i w_i^T = \alpha \sum_i w_i^C$ with $\alpha = 2.0$, *i.e.* the impact of the texture features is twice as high as the contour correspondences. After the first iteration, the solution already converges to the nearest local minimum such that the texture features can be down-weighted by $\alpha = 0.1$. In addition, obvious outliers are discarded by thresholding the re-projection error of the texture correspondences.

After the local optimization has converged to a solution χ, the error for each limb is evaluated individually. Since each correspondence is associated with one limb k, the limb-specific energy is obtained by

$$E_k(\chi) = \frac{1}{K} \sum_{\{i;k_i=k\}} \left\| \Pi\left(T_\chi V_i\right) \times n_i - m_i \right\|_2^2, \tag{4}$$

where only contour correspondences are used and $K = |\{i; k_i = k\}|$. If at least one limb exceeds the predefined upper bound of the energy function, the second phase of the optimization, global optimization, is initiated.

4.2. Global Optimization

After labeling the joints of the misaligned limbs as illustrated in Figure 3, the parameter space of the skeleton pose $\mathbb{R}^d$ is projected onto a lower dimensional search space $P(\chi) \rightarrow \tilde{\chi} \in \mathbb{R}^m$ with $m \leq d$ by keeping the parameters of the non-labeled joints fixed. In order to find the optimal solution for $\tilde{\chi}$, we minimize the energy

$$\underset{\tilde{\chi}}{\operatorname{argmin}} \left\{ E_S(P^{-1}(\tilde{\chi})) + \gamma\, E_R(\tilde{\chi}) \right\}. \tag{5}$$

While the first term measures the silhouette consistency between the projected surface and the image, the second term penalizes strong deviations from the predicted pose and serves as a weak smoothness prior weighted by $\gamma = 0.01$.

The silhouette functional $E_S(P^{-1}(\tilde{\chi}))$ is a modification of the Hamming distance. Using the inverse mapping $\chi = P^{-1}(\tilde{\chi})$ as new pose, the surface model is deformed by quaternion blend skinning and projected onto the image plane for each camera view c. The consistency error for a single view is then obtained by the pixel-wise differences between the projected surface $S_c^p(\chi)$ in model pose χ and the binary silhouette image S_c:

$$\begin{aligned} E_S^c(\chi) = &\frac{1}{area(S_c^p)} \sum_p \left| S_c^p(\chi)(p) - S_c(p) \right| \\ &+ \frac{1}{area(S_c)} \sum_q \left| S_c(q) - S_c^p(\chi)(q) \right|, \end{aligned} \tag{6}$$

where the sums with respect to p and q are only computed over the silhouette areas of $S_c^p(\chi)$ and S_c, respectively. In order to penalize pixel mismatches that are far away from the silhouette, a Chamfer distance transform is previously applied to the silhouette image. The silhouette term E_S is finally the average of E_S^c over all views.

The second term of the energy function (5) introduces a smoothness constraint by penalizing deviations from the predicted pose $\hat{\chi}$ in the lower dimensional space:

$$E_R(\tilde{\chi}) = \|\tilde{\chi} - P(\hat{\chi})\|_2^2. \tag{7}$$

Since we seek for the globally optimal solution for $\tilde{\chi} \in \mathbb{R}^m$, we use a particle-based global optimization approach [14, 15]. The method is appropriate to our optimization scheme since the computational effort can be adapted to the dimensions of the search space and the optimization can be initiated with several hypotheses. It uses a finite set of particles to approximate a distribution whose mass concentrates around the global minimum of an energy function as the number of iterations increases. In our setting, each particle represents a single vector $\tilde{\chi}$ in the search space that can be mapped to a skeleton pose by the inverse projection P^{-1}. The computational effort depends on two parameters, namely the number of iterations and the number of particles. While the latter needs to be scaled with the search space, the number of iterations can be fixed. In our experiments, we have used 15 iterations and $20 * m$ particles with a maximum of 300 particles. These limits are necessary to have an upper bound for the computation time per frame. Furthermore, the optimization is performed on the whole search space when more than 50% of the joints are affected. It usually happens when the torso rotation is not well estimated by the local optimization which is however rarely the case.

The initial set of particles is constructed from two hypotheses, the pose after the local optimization and the predicted pose. To this end, we uniformly interpolate between the two poses and diffuse the particles by a Gaussian kernel.

5. Surface Estimation

Since quaternion blend skinning is based on the overly simplistic assumption that the surface deformation is explained only in terms of an underlying skeleton, the positions of all vertices need to be refined to fit the image data better as illustrated in Figures 2 d and e. To this end, we abandon the coupling of vertices to underlying bones and refine the surface by an algorithm that is related to the techniques used by de Aguiar et al. [10] and Vlasic et al. [28]. We also use a Laplacian deformation framework (see [6] for a comprehensive overview) to move the silhouette rim vertices of our mesh (vertices that should project onto the silhouette contour in one camera) towards the corresponding silhouette contours of our images. In contrast to previous work we do not formulate deformation constraints in 3D, *i.e.* we do not require contour vertices on the model $\mathcal{M}$ to move towards specific 3D points found via reprojection. Instead, we constrain the projection of the vertices to lie on 2D positions on the image silhouette boundary. This makes the linear system to be solved for the refined surface more complex, as we have to solve for all three dimensions concurrently rather than sequentially, as is possible in the previous works. But on the other hand this gives the deformation further degrees of freedom to adapt to our constraints in the best way possible. We reconstructed the refined surface by solving the least-squares system

$$\underset{v}{\operatorname{argmin}} \left\{ \|LV - \delta\|_2^2 + \alpha \|C_{sil} V - q_{sil}\|_2^2 \right\}. \tag{8}$$

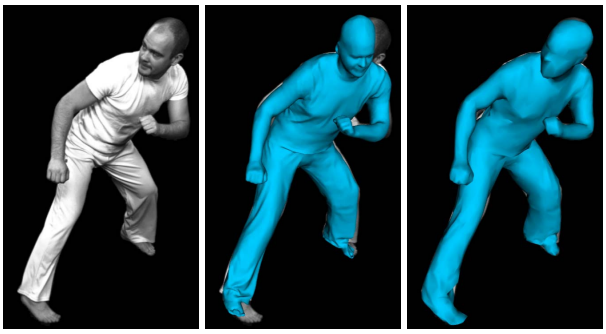
Here, L is the cotangent Laplacian matrix and δ are the differential coordinates of our current mesh with vertices V [6]. The second term in our energy function defines the silhouette constraints and their weighting factor α. Matrix C_{sil} and vector q_{sil} are assembled from individual constraints that take the following form: Given the 3×4 projection matrix M^ℓ of a camera ℓ, split into its translation vector T^ℓ and the remaining 3×3 transformation N^ℓ, the target screen space coordinates $v_i = (v_{i,u}, v_{i,v})$ and the 3D position V_i of a vertex on the 3D silhouette rim of $\mathcal{M}$, we can express a silhouette alignment constraint using two linear equations:

$$\begin{aligned} (N_1^\ell - v_{i,u} N_3^\ell) V_i &= -T_1^\ell + v_{i,u} T_3^\ell \\ (N_2^\ell - v_{i,v} N_3^\ell) V_i &= -T_2^\ell + v_{i,v} T_3^\ell \end{aligned} \tag{9}$$

Here the subscripts of N_i and T_i correspond to the respective rows of the matrix/entry of the vector. These equations force the vertex to lie somewhere on the ray going through the camera's center of projection and the pixel position v_i. Since the error of this constraint is depth-dependent and thus not linear in the image plane, we weight each constraint such that the error is 1 for a single pixel difference at the original vertex position.

Enforcing too high weights for our constraints may lead to an overadaptation in presence of inaccurate silhouettes. We therefore perform several iterations of the deformation, using lower weights. As the silhouette rim points may change after a deformation, we have to recalculate them following each deformation. In all our experiments we performed 8 iterations and used weights of $\alpha = 0.5$.

The estimation for the next frame is then initiated with the estimated skeleton and an adapted surface model which is obtained by a linear vertex interpolation between the mesh from skeleton-pose estimation $V_i^{t,p}$ and the refined mesh $V_i^{t,r}$, *i.e.* $V_i^{t+1} = \lambda V_i^{t,r} + (1 - \lambda) V_i^{t,p}$. In general, a small value $\lambda = 0.1$ is sufficient and enforces mesh consistency



(a) (b) (c)

Figure 4. Visual comparison of our approach with [10]. (a) Input image. (b) Tracked surface mesh from [10]. (c) Tracked surface mesh with lower resolution obtained by our method. While [10] handles loose clothes better, our approach estimates the human pose more reliably.

We finally remark that the surface estimation uses 2D constraints while the skeleton-based pose estimation (Section 4) uses 3D constraints. In both cases 3D constraints can be computed faster, but 2D constraints are more accurate. We therefore resort to 3D constraints during skeleton-based pose estimation which only produces an approximate pose and surface estimate, but use 2D constraints during refinement where accuracy matters.

6. Experiments

For a quantitative and qualitative evaluation of our approach, we have recorded new sequences and used public available datasets for a comparison to the related methods [10] and [28]. Altogether, we demonstrate the reliability and accuracy of our method on 12 sequences with 9 different subjects. An overview of the sequences is given in Table 1. The number of available camera views ranges from 4 to 8 cameras and the 3D surface models have been acquired by a static full body laser scan or by a shape-from-silhouette method, or by the SCAPE model. While our newly recorded sequences have been captured with 40Hz at $1004x1004$ pixel resolution, the other sequences are recorded with the settings: 25Hz and $1920x1080$ pixel resolution [25], 25Hz and $1004x1004$ pixel resolution [10], or 60Hz and $656x490$ pixel resolution [23]. Despite of the different recording settings, the sequences cover various challenging movements from rapid capoeira moves over dancing sequences to a handstand where visual hull approaches are usually prone to topology changes. Furthermore, we have addressed scale issues by capturing the motion of a small dog and wide apparel by three skirt sequences where skeleton-based approaches usually fail. The last column in Table 1 gives the achieved dimensionality reduction of the search space for global optimization and indicates the reduced computation

Sequence	Frames	Views	Model	%DoF
Handstand	401	8	Scan	3.3
Wheel	281	8	Scan	0.2
Dance	574	8	Scan	4.0
Skirt	721	8	Scan	0.2
Dog	60	8	Scan	98.3
Lock [25]	250	8	S-f-S	33.9
Capoeira1 [10]	499	8	Scan	3.4
Capoeira2 [10]	269	8	Scan	11.8
Jazz Dance [10]	359	8	Scan	43.8
Skirt1 [10]	437	8	Scan	7.2
Skirt2 [10]	430	8	Scan	6.5
HuEvaII S4 [23]	1258	4	SCAPE	79.3

Table 1. Sequences used for evaluation. The first 5 sequences are newly recorded. The other sequences are public available datasets. The sequences cover a wide range of motion, apparel, subjects, and recording settings. The last column gives the average dimensionality of the search space for the global optimization in percentage of the full search space.

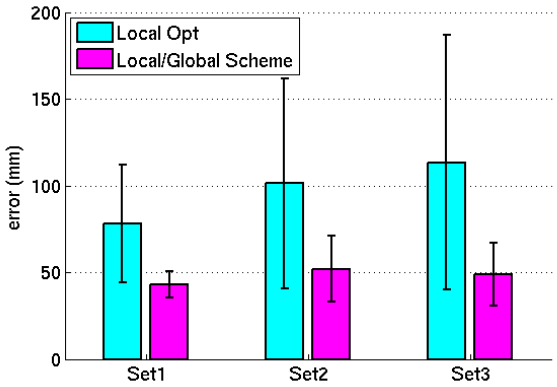


Figure 5. Comparison of our optimization scheme with local optimization. The bars show the average error and standard deviation of the joint positions of the skeleton for the S4 sequence of the HumanEva benchmark. The three sets cover the frames $2-350$ (walking), $2-700$ (walking+jogging), and $2-1258$ (walking+jogging+balancing). While our approach recovers accurately the joint positions over the whole sequence, the error for local optimization is significantly larger.

time. On the newly recorded sequences, the surface estimation requires 1.7 seconds per frame (spf), the local optimization 3 spf, and the global optimization 14 seconds for each DoF (maximal 214 spf). For the skirt sequence, the average computation time for all steps is 9 spf whereas global optimization without local optimization takes 216 spf using 15 iterations and 300 particles.

The examples in Figure 6 show that our approach accurately estimates both skeleton and surface deformation.

Even the challenging lock sequence [25] can be tracked fully automatically whereas the approach [28] requires a manual pose correction for 13 out of 250 frames. A visual comparison with a mesh-based method [10] is shown in Figure 4. Since this method does not rely on a skeleton, it is free of skinning artifacts and estimates apparel surfaces more accurately. The prior skeleton model in our approach, on the other hand, makes pose recovery of the extremities more accurate.

In contrast to [10] and [28], our algorithm can also handle medium-resolution multi-view sequences with extremely noisy silhouettes like the HumanEvaII benchmark [23]. The dataset provides a ground truth for 3D joint positions of the skeleton that has been obtained by a marker-based motion capture system that was synchronized with the cameras. The sequences S4 with three subsets contains the motions walking, jogging, and balancing. The average errors for all three subsets are given in Figure 5. The plot shows that our method provides accurate estimates for the skeleton pose, but it also demonstrates the significant improvement of our optimization scheme compared to local optimization. We finally remark that the jazz dance sequence contains some inaccurate estimates for the feet. The errors do not result from the optimization itself, but a silhouette problem in the data. Therefore the functional being optimized, which is dominated by this error-corrupted term, may lead to problems. This could be improved by using additional cues like edges.

7. Conclusion

We have presented an approach that recovers skeleton pose and surface motion fully-automatically from a multi-view video sequence. To this end, the skeleton motion and the temporal surface deformation are captured in an interleaved manner that improves both accurate skeleton and detailed surface estimation. In addition, we have introduced a novel optimization scheme for skeleton-based pose estimation that makes automatic processing of large data sets feasible. It reduces the computational burden for global optimization in high dimensional spaces by splitting the skeleton-specific optimization problem into a local optimization problem and a lower dimensional global optimization problem. The reliability of our approach has been evaluated on a large variety of sequences including the HumanEva benchmark. The proposed method exceeds the performance of related methods since it allows both accurate skeleton estimation for subjects wearing wide apparel and surface estimation without topology changes for fast movements and noisy silhouettes. This simplifies the acquisition of marker-less motion capture data for applications like character animation and motion analysis.

The work was partially funded by the Cluster of Excellence on Multimodal Computing and Interaction.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005.

[2] A. Balan and M. Black. The naked truth: Estimating body shape under clothing. In *European Conf. on Computer Vision*, pages 15–29, 2008.

[3] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *IEEE Conf. Comp. Vision and Patt. Recog.*, 2007.

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

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

[6] M. Botsch and O. Sorkine. On linear variational surface deformation methods. *IEEE Trans. on Visualization and Computer Graphics*, 14(1):213–230, 2008.

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

[8] G. 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.

[9] 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. *Annals of Biomedical Engineering*, 34(6):1019–1029, 2006.

[10] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3), 2008.

[11] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. *IEEE Conf. Comp. Vision and Patt. Recog.*, 2007.

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

[13] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *Int. Conf. on Computer Vision*, pages 315–320, 2001.

[14] 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.

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

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

[17] L. Kavan, S. Collins, J. Žára, and C. O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Trans. Graph.*, 27(4), 2008.

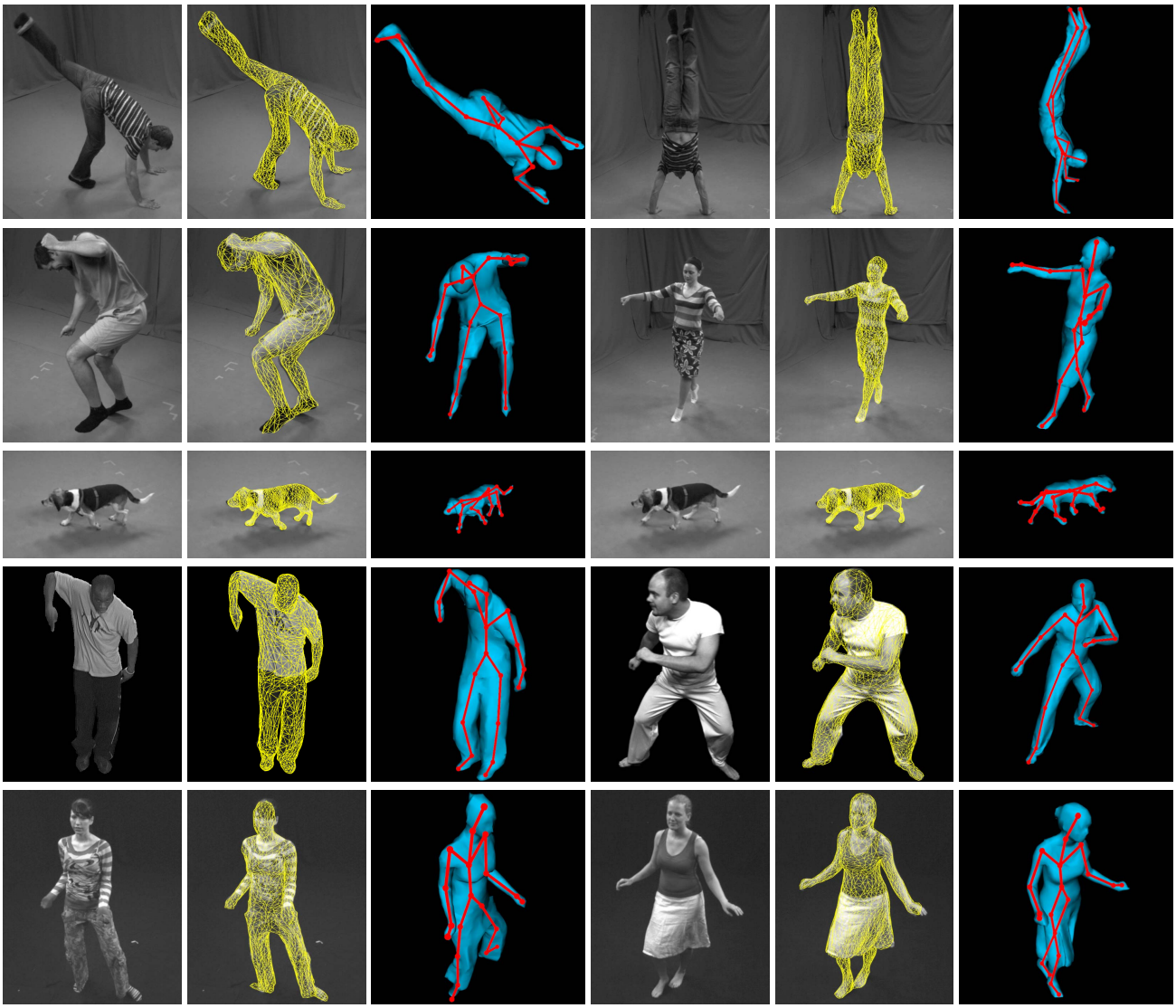


Figure 6. Input image, adapted mesh overlay, and 3D model with estimated skeleton from a different viewpoint respectively.

[18] R. Kehl, M. Bray, and L. van Gool. Full body tracking from multiple views using stochastic sampling. In *IEEE Conf. Comp. Vision and Patt. Recog.*, pages 129–136, 2005.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004.

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

[21] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1182–1187, 2003.

[22] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.-P. Seidel. A system for articulated tracking incorporating a clothing model. *Mach. Vision Appl.*, 18(1):25–40, 2007.

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

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

[25] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *Int. Conf. on Computer Vision*, 2003.

[26] J. Stolfi. *Oriented Projective Geometry: A Framework for Geometric Computation*. Academic Press, Boston, 1991.

[27] K. Varanasi, A. Zaharescu, E. Boyer, and R. Horaud. Temporal surface tracking using mesh evolution. In *European Conf. on Computer Vision*, pages 30–43, 2008.

[28] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3):1–9, 2008.