Markerless Motion Capture of Multiple Characters Using Multiview Image Segmentation

Yebin Liu, Juergen Gall, *Member*, *IEEE*, Carsten Stoll, Qionghai Dai, *Senior Member*, *IEEE*, Hans-Peter Seidel, and Christian Theobalt

Abstract—Capturing the skeleton motion and detailed time-varying surface geometry of multiple, closely interacting peoples is a very challenging task, even in a multicamera setup, due to frequent occlusions and ambiguities in feature-to-person assignments. To address this task, we propose a framework that exploits multiview image segmentation. To this end, a probabilistic shape and appearance model is employed to segment the input images and to assign each pixel uniquely to one person. Given the articulated template models of each person and the labeled pixels, a combined optimization scheme, which splits the skeleton pose optimization problem into a local one and a lower dimensional global one, is applied one by one to each individual, followed with surface estimation to capture detailed nonrigid deformations. We show on various sequences that our approach can capture the 3D motion of humans accurately even if they move rapidly, if they wear wide apparel, and if they are engaged in challenging multiperson motions, including dancing, wrestling, and hugging.

Index Terms—Markerless motion capture, multiview video, multiple characters, image segmentation

1 Introduction

Markerless human motion capture has been studied for several decades and is still a very active field of research in computer vision [1], [2]. While a tremendous amount of progress has been made in skeleton pose estimation, for example, [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], many applications, including realistic character animation for games and movies, require capturing time-varying geometry in more detail, for example, of soft tissue or a garment. To this end, 3D surface estimation methods have been proposed, for example, [13], [14], [15], [16]. However, tracking the full geometry over time is more demanding on processing time and image quality than skeleton-based methods. Therefore, a two-pass approach has been proposed [17] that utilizes a skeleton to increase the robustness of a mesh-based method [15]. In the first pass, a skeleton is semi-automatically fit into the reconstructed visual hull for each frame. The second pass deforms a template mesh according to the estimated skeleton and refines the template to fit the silhouettes.

- *Y. Liu and Q. Dai are with the Automation Department, Tsinghua University, Shuangqing road, Beijing 100084, China. E-mail: {liuyebin, qhdai}@tsinghua.edu.cn.*
- *J. Gall is with the Perceiving Systems Department, Max Planck Institute for Intelligent Systems, Spemannstrasse 41, Tübingen 72076, Germany. E-mail: juergen.gall@tue.mpg.de.*
- *C. Stoll, H.-P. Seidel, and C. Theobalt are with the Max Planck Institute for Informatik and Saarland University, Campus E 1.4 , Saarbrücken 66123, Germany. E-mail: {stoll, hpseidel, theobalt}@mpi-inf.mpg.de.*

Manuscript received 30 July 2012; revised 19 Dec. 2012; accepted 9 Feb. 2013; published online 20 Feb. 2013.
Recommended for acceptance by C. Bregler.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2012-07-0581.
Digital Object Identifier no. 10.1109/TPAMI.2013.47.

Our work is related to the approach in [17], but instead of using a two-pass approach we estimate the skeleton pose and the mesh deformation together for a single frame. To this end, we use a body model that is a combination of a bone skeleton with joints, as well as a surface whose deformation is only loosely coupled with the skeleton motion. In this way, the skeleton provides a low-dimensional motion parameterization, which facilitates tracking of fast movements of the body, and the skeleton pose estimation benefits from the template mesh adaptation over time. Our approach exceeds the performance of related methods from the literature since both accurate skeleton and surface motion are found fully automatically. Moreover, the captured performances can be easily edited and used in animation frameworks typical for games and movies, which are almost exclusively skeleton based [18].

Another advantage of our approach is the ability to capture multiple characters simultaneously. In contrast to capturing only a single person, multiperson scenarios impose additional challenges, in particular, frequent occlusions and ambiguities in assigning commonly used features like silhouettes, color, edges, or interest points to one person. Therefore, only a very few works [19], [20], [16] have addressed this scenario and even in these works the amount of physical contact between two characters is very limited, for example, the hand shake of two people.

In this work, we go beyond the abilities of related methods because our approach captures the skeleton motion and time-varying geometry of multiple, closely interacting characters performing actions with frequent physical contact like wrestling, dancing, or hugging. To handle the high dimensionality of the pose parameters of all people and to resolve the feature-to-person assignments, we employ a probabilistic multiview image segmentation to

determine the image regions each person belongs to. To this end, we use a 3D shape prior for segmenting interacting characters that integrates the previously estimated poses and shapes. The segmentation allows us to generate separate silhouette contours and image features for each person, which drastically reduces the ambiguities. This allows us to perform pose and surface estimation efficiently and in parallel for each performer.

Preliminary versions of this paper appeared in [21] and [22]. While Gall et al. [21] introduced the approach for estimating the skeleton pose and time-varying geometry of a single character, Liu et al. [22] introduced the probabilistic multiview image segmentation framework for capturing the motion of two characters. The present paper gives a comprehensive overview of the full system and extends the previous approaches by the ability to handle more than two people. The system is also thoroughly evaluated, including a quantitative evaluation of the impact of the 3D shape prior and the color model on the segmentation accuracy, a quantitative evaluation of the impact of the quality of the template model on the pose and shape estimation, and a qualitative evaluation on 23 multiview video sequences. The sequences is comprised of 13 sequences with a single person or an animal, seven sequences with two interacting people, and three sequences with three people. The sequences were recorded with seven different camera setups and more than 20 different subjects performing a wide range of motions in a variety of clothes.

2 Related Work

Many approaches exist for human pose estimation either from images or videos [1], [2] or from depth data [23], [24]. We mention only the methods that are most related to ours. For a more detailed discussion, we refer to the books [2], [23].

Similarly to the work of Bregler and Malik [4], we represent the kinematic chain of a human skeleton by twists. In this case, the human motion can be linearized and efficiently optimized by local optimization. In the literature, several approaches for optimizing the pose parameters have been proposed. For instance, stochastic metadescent for local optimization has been used in [25]. Gavrila and Davis [3] 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 [26] 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 [27], [5], global optimization techniques like simulated annealing [28], [6] have also been proposed to overcome the limitations of local optimization. However, global optimization is still too expensive for large datasets and skeletons with many degrees of freedom.

To increase the accuracy of human body models, implicit surfaces based on metaballs [29], shape-from-silhouette model acquisition [30], or the learned SCAPE body model [31], [32] have been proposed. Most of these approaches model the human body without clothing. Balan and Black [33] use SCAPE to estimate the human body underneath clothes from a set of images. Tracking humans wearing more general apparel has been addressed in [34], 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 [13], [14] rely on the visual hull but suffer from topology changes that occur frequently in shape-from-silhouette reconstructions. Mesh-based tracking approaches, for example, [35], [15], 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 for single person tracking is most similar to the work of Vlasic et al. [17], 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 visually appealing results, a considerable amount of manual interaction is required in [17] (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. In contrast, our local-global optimization is fully automatic and also works on data of poor image quality. In [36], our approach has been further extended to estimate not only surface and pose parameters, but also the parameters of the skeleton. Another extension has been proposed in [37], where the approach is applied to depth data and camera poses are estimated in addition.

Markerless motion capture of multiple performers has only been considered in very few works. Cagniart et al. [20], [16] use a patch-based approach for surface tracking of multiple moving subjects based on the visual hull geometry. However, they do not provide skeleton motion, and the subjects are well separated and never interact closely. Guillemaut et al. [38] propose a volumetric graph-cut method for the segmentation and reconstruction of multiple players in sports scenes like football games. This approach reconstructs only a rough 3D shape of each player, which is suitable for applications like 3D television broadcast, but not for detailed performance capture. Egashira et al. [19] propose a volumetric segmentation on the visual hull of the scene to separate the persons. However, when multiple people are in physical contact, volumetric segmentation of the visual hull is not as accurate as image-based segmentation prior to 3D reconstruction.

A number of researchers have investigated methods for tracking bounding boxes of multiple humans from a single camera [39] or multiple cameras [40], [41], [42]. In the very restricted context of pedestrians and only walking motion, the skeleton motions of several people have been estimated

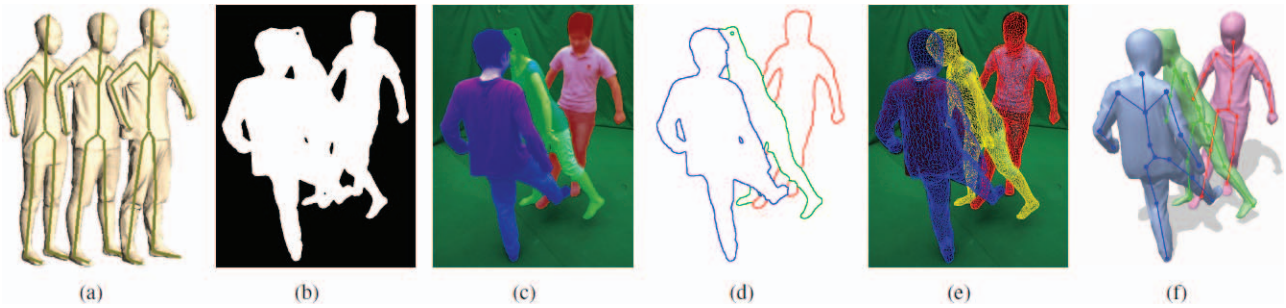


Fig. 1. Overview of our processing pipeline: (a) Articulated template models. (b) Input silhouettes. (c) Segmentation. (d) Contour labels assigned to each person. (e) Estimated surface. (f) Estimated 3D models with embedded skeletons.

in [43], [44]. Zhang and Ngan [45] present a joint object detection, segmentation, and tracking approach to segment a group of people into individual human objects and track them across the video sequence using multiview video, without the estimation of skeleton motion and surface geometry for each human.

Image segmentation techniques have been used for skeleton pose estimation of a single person in [46], [47], [48]. In [47], [48], the articulated pose of the previous frame is used as shape prior for level-set segmentation, and the pose is either estimated within an analysis-by-synthesis framework [48] or in combination with optical flow and SIFT features [47]. Graph-cut segmentation is used in [46], where a multiview foreground image segmentation is coupled with a simple stick model for pose estimation. For each time instant, the method computes the segmentation costs for all candidate poses and chooses the pose with minimal energy. However, the pose estimates may sometimes be inaccurate since the minimum cut cost does not necessary coincide with the correct pose. In the case of multiple people, this may become even more of a problem since occlusions often change the 2D topology.

3 Overview

We capture one or multiple human performers using synchronized and calibrated cameras. For each input image, foreground silhouettes are extracted by background subtraction. As in [17], we aim at estimating the skeleton configuration (*pose*), consisting of the global rigid transformation of the torso and the joint angles of the skeleton, as well as nonarticulated surface deformations (*shape*) that cannot be represented by a skeleton-driven deformation. Unlike previous work, we go beyond single person tracking and capture pose and shape in the context of challenging human-human interactions with physical contact.

An outline of the processing pipeline is given in Fig. 1. Starting with the estimated poses and shapes of all people in the previous frame, the proposed algorithm estimates the poses and the shapes in the current frame based on the captured multiview images and foreground silhouettes (Fig. 1b). Since the whole space for the unknown pose and shape parameters becomes very large for multiple people, we split the tracking problem into a multiview 2D segmentation problem (Figs. 1c and 1d) and a 3D pose and shape estimation problem (Figs. 1e and 1f). The segmentation separates the people in the image domain by assigning a label to each foreground pixel. Then, based on the labeled pixels, the pose and the shape are estimated for each person independently. To facilitate understanding, we discuss the pose and shape estimation given the segmentation first in Section 4 and then introduce the multiperson motion capture approach with segmentation in Section 5.

4 Pose and Shape Estimation

The body model of each human character consists of two components, a 3D triangle mesh surface model S with 3D vertices V_i and an underlying bone skeleton as shown in Fig. 1a. The configuration of the skeleton is represented by a set of twists $\theta_j\hat{\xi}_j \in se(3)$ as in [4]. Each twist can be converted into a rigid body motion using the exponential map: $\exp(\theta_j\hat{\xi}_j) \in SE(3)$. For more details on the twist representation, we refer to [49].

Each vertex V_i is associated with a bone m with a skinning weight $\alpha_{i,m}$, where $\sum_m \alpha_{i,m} = 1$. Since each bone m is influenced by n_m out of totally N joints, the transformation of a vertex V_i with blend skinning is given by

$$T_i(\Theta)V_i = \mathrm{DLB}\big(\alpha_{i,m}; T_m(\Theta)\big)V_i, \tag{1}$$

$$T_m(\Theta) = \prod_{j=0}^{n_m} \exp\left(\theta_{\iota_m(j)}\hat{\xi}_{\iota_m(j)}\right), \tag{2}$$

where DLB computes the weighted mean of the transformations $T_m(\Theta)$ using dual quaternion skinning [50]. The mapping ι_m 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 $\Theta := (\theta_0\xi_0, \Theta_{joints}) \in \mathbb{R}^d$ that consists of the six parameters for the global twist $\theta_0\hat{\xi}_0$ and the joint angles $\Theta_{joints} := (\theta_1, \ldots, \theta_N)$. While the joints are manually placed into each mesh, the skinning weights $\alpha_{i,m}$ are automatically computed using the approach [51].

An outline of the pose and surface estimation is given in Fig. 2. Starting with the estimated mesh and skeleton from the previous frame, the skeleton pose is optimized as described in Section 4.1 such that the projection of the deformed surface fits the image data in an optimal way (Fig. 2b). Since this step only captures deformations that can be approximated by articulated surface skinning (Fig. 2c), the nonrigid surface is subsequently refined as described in Section 4.2 (Fig. 2d). The estimated refined surface and skeleton pose serve as initialization for the next frame to be

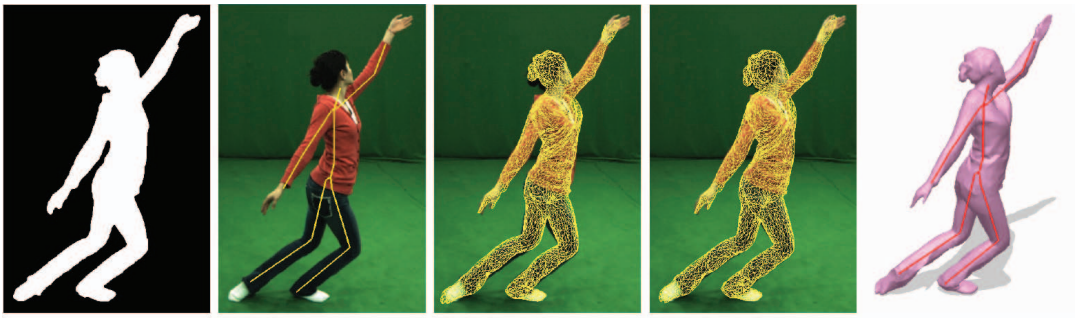


Fig. 2. Using the estimated surface of the previous frame, the pose of the skeleton (b) is optimized such that the deformed surface (c) fits the image data (a). Since skeleton-based pose estimation is not able to capture garment motion (c), the surface is refined to fit the silhouettes (d).

tracked (Fig. 2e). The approach for pose and surface estimation is summarized in Algorithm 1.

Algorithm 1: Pose and shape estimation of one person.

Data: Surface $S^t = \{V_i^t\}$ and pose Θ^t of frame t; silhouettes $\{F_c\}$ of frame $t+1$; matched SIFT features between frames t and $t+1$.
Result: Surface S^{t+1} and pose Θ^{t+1}.
begin
  Skeleton-based pose estimation (Sec. 4.1)
  Compute correspondences (V_i^t, x_i)
  Get Θ^{t+1} by solving (4)
  Compute errors $E_m(\Theta^{t+1})$ for all limbs m (5)
  if $\exists m : E_m(\Theta^{t+1}) > \bar{E}$ **then**
    Update $\Theta^{t+1} = P^{-1}(\tilde{\Theta})$ by solving (6)
  Deform surface by $V_i^{t+1,p} = T_i(\Theta^{t+1})V_i^t$ (1)
  Surface refinement (Sec. 4.2)
  Compute correspondences $(V_i^{t+1,p}, x_i)$
  Estimate surface $V_i^{t+1,r}$ by solving (10)
  Update surface by $V_i^{t+1} = \lambda V_i^{t+1,r} + (1-\lambda)V_i^{t+1,p}$

4.1 Skeleton-Based Pose Estimation

Since local pose optimization is prone to getting stuck in local minima of the energy function 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 (Section 4.1.1). In the second phase, misalignments are detected and resolved by global optimization (Section 4.1.2).

4.1.1 Local Optimization

For estimating the pose parameters Θ, 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 [52]. 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 a Plücker line $L_i = (D_i, M_i)$[1] [53], the error of a pair $(T_{m_i}(\Theta)V_i, x_i)$ is given by the norm of the perpendicular vector between the line L_i and the transformed point $T_{m_i}(\Theta)V_i$:

$$\|\Pi(T_{m_i}(\Theta)V_i) \times D_i - M_i\|_2, \tag{3}$$

where Π denotes the projection from homogeneous coordinates to nonhomogeneous coordinates. In contrast to (1), the skinning weights are not used, and m_i is the limb with the highest skinning weight, i.e., $\text{argmax}_m \alpha_{i,m}$. The resulting least-squares problem with weights w_i for the correspondences

$$\underset{\Theta}{\text{argmin}} \frac{1}{2} \sum_i w_i \|\Pi(T_{m_i}(\Theta)V_i) \times D_i - M_i\|_2^2 \tag{4}$$

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. 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 third order autoregression and β is a small constant. Since the optimization regards the limbs as rigid structures, the mesh is updated between the iterations by dual quaternion blending (1) to approximate smooth surface deformations.

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 because they are more stable under large displacements. For the first iteration, we set the weights such that $\sum_i w_i^T = \alpha \sum_i w_i^C$, with $\alpha = 2.0$. This means that the impact of the texture features is twice as high as the contour correspondences. After the first iteration, the solution will already be close to the nearest local minimum such that the texture features can be downweighted by $\alpha = 0.1$. In addition, obvious outliers are discarded by thresholding the reprojection error of the texture correspondences.

1. A Plücker line $L = (D, M)$ is determined by a unit vector D and a moment M, where $X \times D - M = 0$ for all points X on the line.

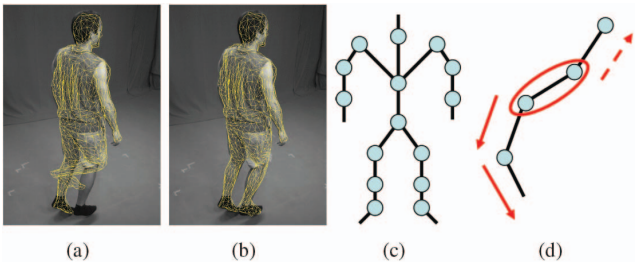


Fig. 3. Although local optimization is prone to errors, often only a single branch of the kinematic chain is affected (a). This reduces the computational burden for global optimization because it can be performed in a lower dimensional subspace to correct the estimation error (b). After detecting misaligned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

4.1.2 Particle-Based Global Optimization

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

$$E_m(\Theta)=\frac{1}{Z}\sum_{\{i;m_i=m\}}\|\Pi(T_{m_i}(\Theta)V_i)\times D_i - M_i\|_2^2, \qquad (5)$$

where only contour correspondences are used and $Z=|\{i;m_i=m\}|$. When the energy exceeds a given threshold $\bar{E}$, the affected limb is labeled as misaligned. While large values for $\bar{E}$ increase the runtime, low values increase the risk of getting stuck in a local minimum. In our experiments, we found that thresholds above 400, corresponding to a RMSE error of 20 mm, give good results. 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), as illustrated in Fig. 3. For instance, a wrong estimate of the shank might be caused by a rotation error along the axis of the thigh.

After labeling the joints of the misaligned limbs, the parameter space of the skeleton pose $\mathbb{R}^d$ is projected onto a lower dimensional search space $P(\Theta)\rightarrow\tilde{\Theta}\in\mathbb{R}^h$ with $h\leq d$ by keeping the parameters of the nonlabeled joints fixed. To find the optimal solution for $\tilde{\Theta}$, we minimize the energy

$$\underset{\tilde{\Theta}}{\operatorname{argmin}}\{E_S(P^{-1}(\tilde{\Theta}))+\gamma\, E_R(\tilde{\Theta})\}. \qquad (6)$$

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

The silhouette functional $E_S(P^{-1}(\tilde{\Theta}))$ is a modification of the Hamming distance. Using the inverse mapping $\Theta=P^{-1}(\tilde{\Theta})$ as a new pose, the surface model is deformed by (1) and projected onto the image plane for each camera view c, denoted by $\mathbf{B}_c(\Theta)$. As shown in Fig. 4b, the projection encodes the body parts of all persons.

The consistency error between the segmented silhouette F_c of a person and a projection $\mathbf{B}_c(\Theta)$ of its model is measured pixelwise by

$$E_S(\Theta)=\frac{1}{|\{c\}|}\sum_c\sum_i d_{c,i}(\Theta), \qquad (7)$$

with the general bidirectional distance:

$$\begin{aligned} d_{c,i}(\Theta) = {} & I_F\big(F_{c,i},B_{c,i}(\Theta)\big)g_F\big(F_{c,i},B_{c,i}(\Theta)\big)\\ & + I_B\big(B_{c,i}(\Theta),F_{c,i}\big)g_B\big(B_{c,i}(\Theta),F_{c,i}\big). \end{aligned} \qquad (8)$$

While I is an indicator function of an error, g specifies the cost of an error. The first term measures how well the silhouette data is explained by the model. In detail, $I_F(f,b)$ is only one if f belongs to the silhouette of the person and b is not a projected body part of the person. In this case, the error is measured by $g_F(f,b)=\frac{\lambda}{Z_F}$, where $\lambda=80$ is a constant and Z_F denotes the area of the silhouette. The second term measures how well the projection is explained by the silhouette. Hence, $I_B(b,f)$ is only one if b is a body part that was visible in the previous frame and f is not part of the silhouette. In this case, $g_B(b,f)=\frac{d(f)}{Z_B}$, where $d(f)$ denotes the distance to the closest point on the silhouette F_c and Z_B the area of the visible body parts. The explicit handling of occlusions is necessary since the pose of each person is estimated individually. Furthermore, g_F uses, in contrast to g_B, a weaker, namely, constant, cost model due to efficiency.

The second term of the energy function (6) introduces a smoothness constraint by penalizing deviations from the predicted pose $\hat{\Theta}$ in the lower dimensional space:

$$E_R(\tilde{\Theta})=\|\tilde{\Theta}-P(\hat{\Theta})\|_2^2. \qquad (9)$$

Since we seek the globally optimal solution for $\tilde{\Theta}\in\mathbb{R}^h$, we use a particle-based global optimization approach [54], [6]. 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

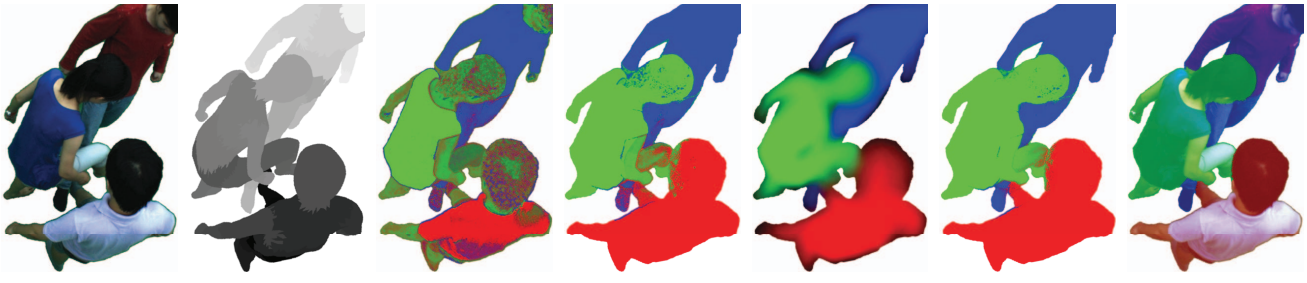


Fig. 4. Segmentation with shape and appearance information. (c)-(f) show the probability maps calculated according to different terms. (a) Input image after background subtraction. (b) Projections of the body parts B_k^j. (c) Color term using a whole body appearance model. (d) Color term using the body part appearance model. (e) Shape prior. (f) Combined shape prior and body part appearance model (17). (g) Segmentation result from (f).

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 $\tilde{\Theta}$ 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 $20 * h$ 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 percent 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.

4.2 Surface Refinement

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 Figs. 2c and 2d. 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. [15] and Vlasic et al. [17]. As in Section 4.1, we extract contour correspondences (V_i, x_i) from all views c, but we minimize the error in the image domain instead of the 3D space for better accuracy. 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. On the other hand, this gives the deformation further degrees of freedom to adapt to our constraints in the best way possible. Using a Laplacian deformation framework [55], we refine the surface S^p, obtained from skeleton-based pose estimation (Section 4.1), by solving the least-squares problem

$$\underset{S^r}{\operatorname{argmin}}\left\{\sum_{V^r \in S^r} \|LV^r - LV^p\|_2^2 + \alpha \sum_i \|P^{c_i} V_i^r - x_i\|_2^2\right\}, \quad (10)$$

where L is the cotangent Laplacian matrix [55] and V^p are the vertex positions of the previous surface S^p corresponding to V^r. While the first term preserves the differential properties of the previous mesh, the second term, weighted by α, aims at minimizing the error of the correspondences. Given the 3×4 projection matrix P^c of a camera c, split into its translation vector T^c and the remaining 3×3 transformation N^c, we can express a silhouette alignment constraint of the second term using two linear equations:

$$\begin{aligned}\left(N_1^c - x_{i,1} N_3^c\right) V_i &= -T_1^c + x_{i,1} T_3^c,\\ \left(N_2^c - x_{i,2} N_3^c\right) V_i &= -T_2^c + x_{i,2} T_3^c.\end{aligned} \quad (11)$$

Here, the subscripts l of N_l, $x_{i,l}$, and T_l correspond to the respective rows of the matrix or 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 x_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 the presence of inaccurate silhouettes. We therefore perform several iterations of the deformation using lower weights. As the silhouette points on the mesh may change after a deformation, we have to recalculate the correspondences following each deformation. In all our experiments, we performed eight 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 that is obtained by a linear vertex interpolation between the mesh from skeleton pose estimation S^p and the refined mesh S^r, i.e., $V_i = \lambda V_i^r + (1 - \lambda) V_i^p$. In general, a small value $\lambda = 0.1$ is sufficient and enforces mesh consistency.

5 Multiperson Segmentation

Before estimating the pose and shape, we label the foreground pixels according to which person they belong to (Figs. 1b and 1c). To this end, we integrate appearance, pose, and shape information into an MAP-Markov random field (MRF) [56] optimization framework to achieve segmentations that are both efficient and robust for human motion capture under serious occlusions and ambiguous appearance. The full approach for multiperson pose and surface estimation with multiperson segmentation is outlined in Algorithm 2.

Algorithm 2: Pose and shape estimation of multiple persons.

Data: Surfaces S_k^t and poses Θ_k^t of frame t for all persons k; images $\{I_c\}$ and silhouettes $\{F_c\}$ of frame $t+1$; matched SIFT features between frames t and $t+1$.
Result: Surfaces S_k^{t+1} and poses Θ_k^{t+1}.
begin
  Multi-person segmentation (Sec. 5)
  Compute 3D shape prior (22)
  foreach c **do**
    Compute $\phi(I_c|\{S_k^t\},\{\Theta_k^t\}, l_i{=}k)$ (15), (17)
    Get labels $\mathbf{L}_c$ by optimizing (12)
    Label contours (Sec. 5.3)
  Pose and shape estimation (Sec. 4)
  foreach k **do**
    Estimate S_k^{t+1} and Θ_k^{t+1} using Algorithm 1.

5.1 Multiview Image Segmentation

For determining the pixel labels in each image I, we resort to MAP inference in an MRF. Previous MRF-based image segmentation methods use standard appearance-based likelihood terms, as well as smoothness potentials. In our

work, we exploit our knowledge about the 3D shape of each performer k at the previous time instant to assign each pixel i a label $l_i = k$ by optimizing an energy of the form

$$\Psi(\mathbf{L}) = \sum_i \left(\phi(I|\{S_k\}, \{\Theta_k\}, l_i) + \sum_{j \in N_i} \gamma(I|l_i, l_j) \right). \quad (12)$$

The solution of this multilabel problem $\mathbf{L}$ is obtained by graph cuts [56].

While the unary potential $\phi(I|\{S_k\}, \{\Theta_k\}, l_i)$, which fuses appearance, pose, and shape information of all people, is specific to multiperson motion capture (Section 5.2), the pairwise potentials $\gamma(I \mid l_i, l_j) = \phi(I \mid l_i, l_j) + \psi(l_i, l_j)$, which are computed over a neighborhood N_i of 8-connected pixels, are commonly used in image segmentation.

As in [57], [46], $\phi(I \mid l_i, l_j)$ is a contrast term, which favors pixels with similar color having the same label:

$$\phi(I \mid l_i, l_j) = \begin{cases} \dfrac{\mu}{S(i,j)} \exp\left(\dfrac{-\|I_i - I_j\|^2}{2\sigma^2} \right) & \text{if } l_i \neq l_j, \\ 0 & \text{if } l_i = l_j, \end{cases} \quad (13)$$

where $\|I_i - I_j\|^2$ measures the difference in the color values of pixels i and j and $S(i,j)$ is the spatial distance between the pixels. In addition, an observation-independent smoothness prior in the form of a generalized Potts model [58] is used:

$$\psi(l_i, l_j) = \begin{cases} \kappa_{i,j} & \text{if } l_i \neq l_j, \\ 0 & \text{if } l_i = l_j. \end{cases} \quad (14)$$

5.2 Appearance, Pose, and Shape

Since the appearance of humans is often very similar, for example, skin or hair color, commonly used appearance models for image segmentation are too weak to segment several people that are very close and occlude each other. In our case, however, the poses $\Theta = \{\Theta_k\}$ and shapes $S = \{S_k\}$ of all people which have been recovered in the previous frame are strong cues that can be integrated as shape priors for segmentation. We therefore model the unary potential $\phi(I \mid S, \Theta, l_i)$ not only conditioned on the label l_i, but also on S and Θ:

$$\phi(I \mid S, \Theta, l_i) \propto -\log P(I \mid S, \Theta, l_i). \quad (15)$$

Since the appearance of the body usually comprises various colors and the color distribution of the whole body is often not discriminative enough to distinguish different persons, we use a color model for each body part. The intuition behind this is that the color distribution is usually consistent for a body part but varies strongly between different body parts, for example, while hands are typically skin colored, other parts like upper body or legs are often covered by clothes of a specific color. We therefore model the appearance locally on the surface of each person and integrate shape priors for each person into a common multiview segmentation approach. We model the person's appearance for each of the body parts B_k^j as

$$P(I \mid S, \Theta, l_i) = \sum_j P(I \mid i \in B_k^j, S, \Theta, l_i) P(i \in B_k^j \mid S, \Theta, l_i). \quad (16)$$

$P(i \in B_k^j \mid S, \Theta, l_i)$ is a shape prior modeling the probability that a pixel i belongs to body part B^j of person k. This term will be described in Section 5.2.1. Since the appearance of a pixel depends only on the body part, (16) can be simplified as

$$P(I \mid S, \Theta, l_i) = \sum_j P\big(I \mid i \in B_k^j\big) P\big(i \in B_k^j \mid S, \Theta, l_i\big). \quad (17)$$

The color term $P(I \mid i \in B_k^j) \propto P(I_i \mid H_k^j)$ measures the consistency of the color I_i of a pixel i with the color distribution H_k^j for body part B^j of person k. The color distributions H_k^j are modeled in the RGB color space using Gaussian mixture models (GMMs). Since the appearance of the person may change over time due to the change in illumination, the color distribution H_k^j is updated during tracking by estimating the GMMs from the labeled pixels of the first and the previous tracked frame.

Fig. 4 illustrates the impact of the terms used for segmentation. While the labels of the three people are illustrated by the colors red, green, and blue, the color values represent the probability of a pixel belonging to each person. Fig. 4c shows that a single color model for each person is insufficient. The skin colored regions, hair, and legs are not well associated with one person. Using the body part appearance model (Figs. 4b and 4d) improves the probability maps, but there are still some ambiguities at the legs. The shape prior (Fig. 4e) is a strong cue, although there are still many regions with low confidence, indicated by the dark colors. The probability maps and the segmentation using the full model (17) are shown in Figs. 4f and 4g. Minor ambiguities are removed by the pairwise potentials in (12).

5.2.1 Three-Dimensional Shape Prior

The shape prior $P(i \in B_k^j \mid S, \Theta, l_i)$ in (17) encodes an a priori probability for assigning a body part label B_k^j for each pixel i and therefore encodes the probability to which person k it belongs. As in previous work, this probability can be modeled by projecting each body model and diffusing the projected body parts in the 2D image domain to obtain a shape prior for all persons. This can be implemented by either projecting each person independently and combining the priors or by projecting all people together. While the first approach does not handle occlusions at all (Fig. 5b), the second approach gives zero probability to parts that were occluded but reappear in the current frame. For instance, the right arm of the woman (green) has zero probability (Fig. 5c) although the arm reappears in this frame (Fig. 5e). This shows that projecting 3D shapes to the image domain and then modeling the shape priors based on 2D distances is incorrect. We therefore model the shape prior in the 3D space and project the probabilistic 3D prior to the image domain. As shown in Figs. 5d and 5e, the 3D shape prior gives a reasonable probability map for image segmentation.

To this end, we model the shape prior using the posterior probability of the poses $P(\Theta \mid I, S)$ given the silhouette images I of all views and the estimated shapes S of the persons. To sample new pose configurations Θ in the current frame for all the persons, we use importance sampling [59]:

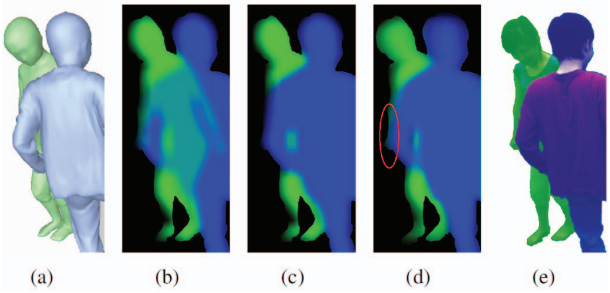


Fig. 5. Comparison of shape priors using 2D shape diffusion and 3D shape posterior. Tracked model from previous time step (a). Combining the 2D diffused shape priors for two persons yields ambiguities due to occlusions (b). When occluded pixels are removed before 2D diffusion, the obtained shape prior (c) will give zero probability to the part (right arm of the woman) that is occluded in the previous frame. In contrast, the proposed 3D shape diffusion gives a better probability in this region (red ellipse) (d), which leads to a better segmentation (e).

$$P(\Theta \mid I, S) \propto P(I \mid \Theta, S) P(\Theta), \tag{18}$$

where we take the shapes from the previous frame and rely only on linear blend skinning, as for the skeleton-based pose estimation (Section 4.1). The pose parameters Θ are predicted from a Gaussian distribution $P(\Theta)$ with mean corresponding to the previously estimated pose parameters. The likelihood term $P(I \mid \Theta, S)$ measures the importance of each sample through consistency evaluation of the projected surfaces $\mathbf{B}_c(\Theta)$ and the foreground silhouettes F_c for all views c:

$$P(I \mid \Theta, S) \propto \exp\left(- \sum_c \sum_i d'_{c,i}(\Theta) \right), \tag{19}$$

where $d'_{c,i}(\Theta)$ is the general bidirectional distance as in (8):

$$\begin{aligned} d'_{c,i}(\Theta) = {} & I'_F\big(F_{c,i}, B_{c,i}(\Theta)\big) g'_F\big(F_{c,i}, B_{c,i}(\Theta)\big) \\ & + I'_B\big(B_{c,i}(\Theta), F_{c,i}\big) g'_B\big(B_{c,i}(\Theta), F_{c,i}\big). \end{aligned} \tag{20}$$

In contrast to the pose and shape estimation that is performed for each person independently based on the labeled foreground silhouettes (Section 4), F_c contains the unlabeled foreground silhouettes of all people (Fig. 4a) and the projection $\mathbf{B}_c(\Theta)$ contains the body parts of all people (Fig. 4b). Hence, the indicator function $I'_B(b, f)$ does not need an explicit handling of occlusions and is therefore only one if b is a body part and f is not part of the silhouette. $I'_F(f, b)$ is, as previously, only one if f belongs to the silhouette of the person and b is not a projected body part of the person. The error cost functions g' are defined by $g'_F(f, b) = 1$ and $g'_B(b, f) = \frac{Z_k}{Z_b}$, where Z_b is the area of the body part b belongs to and Z_k the area of the corresponding person k. $\frac{Z_k}{Z_b}$ equalizes the impact of all body parts independent of their size to avoid having parts with small regions dominated by parts with large regions, as shown in Fig. 6.

To approximate $P(i \in B_k^j \mid S, \Theta, l_i)$, we therefore draw a set of samples $\{\Theta^n\}$ from $P(\Theta)$ and weight them by

$$w_n = \frac{\exp\left(-\sum_c \sum_i d'_{c,i}(\Theta^n)\right)}{\sum_n \exp\left(-\sum_c \sum_i d'_{c,i}(\Theta^n)\right)}. \tag{21}$$

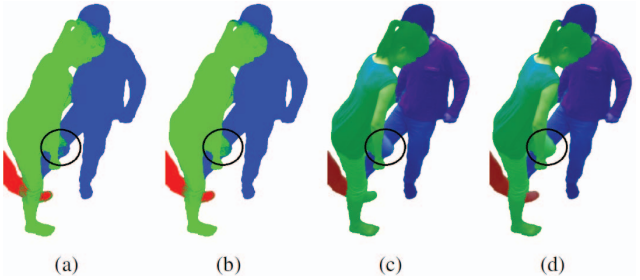


Fig. 6. Impact of the body part-dependent cost factor $\frac{Z_k}{Z_b}$ for $g'_B(b, f)$ (20). (a) Shape prior with $g'_B(b, f) = 1$. (b) Shape prior with $g'_B(b, f) = \frac{Z_k}{Z_b}$. (c) Segmentation with $g'_B(b, f) = 1$. (d) Segmentation with $g'_B(b, f) = \frac{Z_k}{Z_b}$.

Hence, the shape prior for assigning a pixel i the body part label b_k^j for person k in (17) becomes

$$P\big(i \in B_k^j \mid S, \Theta, l_i\big) = \sum_n w_n \cdot \delta_{b_k^j}(\mathbf{B}_{c,i}(\Theta^n)), \tag{22}$$

$$\delta_{b_k^j}(\mathbf{B}_{c,i}(\Theta^n)) = \begin{cases} 1 & \text{if } \mathbf{B}_{c,i}(\Theta^n) = b_k^j, \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

where c is the corresponding view. Since several poses lead to similar projections, good estimation results can be achieved with a relatively low number of samples, despite a $39 \times K$-dimensional space for Θ, with K being the number of people. In our experiments, we found 300 samples enough for a reasonable approximation of the shape prior.

5.2.2 Resolving Intersections

The evaluation of Θ^n in (21) requires the projection of the meshes. When the interacting people are close to each other, the sampling from $P(\Theta)$ might generate meshes that intersect with each other in 3D. For instance, over 80 percent of the samples have slight or serious intersections in some of the sequences shown in Fig. 11. Although we can define $P(\Theta)$ to generate only meshes without intersections, the additional intersection tests and constraints would make the sampling procedure extremely time consuming.

To obtain a reliable shape prior without intersection test, a simple yet efficient rendering approach is applied. Fig. 7a shows an example where the right hand of a person intersects the chest of the other person, removing its

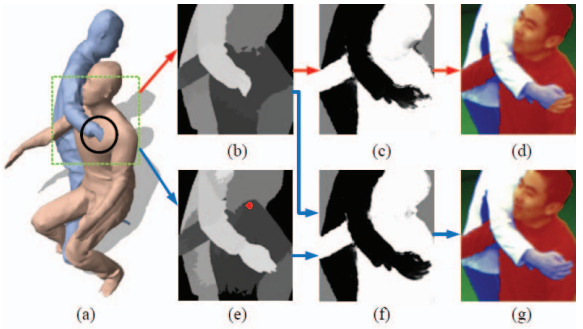


Fig. 7. Resolving intersections. (a) Intersection between two peopls. The hand is inside the chest. (b) Standard projection. (c) Corresponding data term and (d) segmentation from (c). (e) Projection with front-face culling. (f) Data term combining both projections. (g) Corresponding segmentation.

contribution to the data term (Fig. 7b). When this happens for several samples, the shape prior (22) becomes inaccurate and segmentation errors occur (Figs. 7c and 7d). However, when rendering using front-face culling, only mesh facets that are not facing the camera are rendered, making the hand visible even inside of the body (Fig. 7e). We also observe that front-face culling may produce inaccurate labeling between body parts belonging to the same person. For example, the marked red pixel on the face in Fig. 7e is inconsistently labeled as belonging to the chest. We therefore generate, for each sample Θ^n and view c, two projections $\mathbf{B}_c$ and $\widetilde{\mathbf{B}}_c$, one with normal rendering and one with front-face culling. For each pixel i, the label $\mathbf{B}_{c,i}$ is then only changed to $\widetilde{\mathbf{B}}_{c,i}$ if the labels $\mathbf{B}_{c,i}$ and $\widetilde{\mathbf{B}}_{c,i}$ correspond to two different people. Otherwise, the label remains unchanged. While this procedure does not resolve the intersection problem accurately and can create additional artifacts, it improves the shape prior and the corresponding segmentation as shown in Figs. 7f and 7g, with very low computational overhead.

5.3 Contour Labeling

After having labeled each pixel in the input images (Fig. 1c), we assign boundary pixels of the segmented regions to the correct person (Fig. 1d). There are two types of boundary pixels to be assigned. The first type of pixels lies on the boundary between a person and the background, which can be easily assigned to the correct person. Boundary pixels in regions where two or more people overlap get the label of the person whose boundary region is closest to the camera. To this end, we evaluate the depth values of the projected models in a neighborhood of the boundary pixel and take the label with the lowest average depth.

6 Experiments

We have evaluated our approach quantitatively and qualitatively on 13 sequences with a single person or animal, seven sequences containing two people interacting with each other, and three sequences with three people. The 23 sequences consist of over 9,000 frames of multiview video. While four sequences have been newly recorded, the other sequences have been used in previous publications [13], [17], [15], [21], [1], [22]. The sequences cover a wide range of different motions, including dancing, fast fighting, and jumping. An overview of all sequences is given in Table 1. Examples of the sequences with two or three persons are shown in Fig. 11. The sequences include performances by 20 different subjects wearing casual apparel, from tight jeans and t-shirt to wide skirts. For the quantitatively evaluation, we use the HumanEva benchmark [1] and an evaluation sequence where one of the people was simultaneously tracked by a marker-based motion capture system, yielding ground-truth data. The number of cameras in each sequence varies between 4 and 12 cameras, with frame rates between 15 and 60 fps. The 3D surface models have either been acquired using a full body laser scanner or using multiview stereo reconstruction. In the experiments, we also evaluate the impact of the quality of the body model. In Section 6.1, we first evaluate the pose and shape estimation (Section 4) independently of the multiperson segmentation. The full approach for capturing multiple people interacting with each other is then evaluated in Section 6.2. More experimental results are accompanied in the submitted supplemental videos, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.47.[2]

TABLE 1
Sequences Used for Evaluation

Sequence	K[1]	F[1]	fps	C[1]	resolution
Handstand	1	401	40fps	8	1004×1004
Wheel	1	281	40fps	8	1004×1004
Dance	1	574	40fps	8	1004×1004
Skirt	1	721	40fps	8	1004×1004
Dog	1	60	40fps	8	1004×1004
New-dance[2]	1	1000	45fps	12	1296×972
Crash	2	250	45fps	11	1296×972
Couple-dance	2	300	45fps	11	1296×972
Jump	2	250	45fps	11	1296×972
Hug	2	200	45fps	11	1296×972
Hit	2	200	45fps	12	1296×972
Wrestle	2	200	45fps	12	1296×972
Fight (marker)	2	500	45fps	12	1296×972
Crossover[2]	3	200	15fps	12	1024×768
Bend[2]	3	200	15fps	12	1024×768
Hop[2]	3	200	15fps	12	1024×768
Lock [13]	1	250	25fps	8	1920×1080
Capoeira1 [15]	1	499	25fps	8	1004×1004
Capoeira2 [15]	1	269	25fps	8	1004×1004
Jazz Dance [15]	1	359	25fps	8	1004×1004
Skirt1 [15]	1	437	25fps	8	1004×1004
Skirt2 [15]	1	430	25fps	8	1004×1004
HuEvaII S4 [1]	1	1258	60fps	4	656×490

[1] *Number of people (K), frames (F), and cameras (C).* [2] *Newly captured sequences.*

6.1 Pose and Shape Estimation

We evaluated the pose and shape estimation approach (Section 4) on 13 sequences of subjects performing different motions. These sequences include low-quality sequences with few cameras, such as the HumanEvaII benchmark [1], as well as high-resolution sequences with more cameras where the subjects perform fast and challenging motions. To show that our method is not limited to capturing humans, we also tracked the motion of a small dog. Our local-global optimization approach is able to track all sequences successfully without any manual intervention. Even the challenging lock sequence [13] can be tracked fully automatically using our method, whereas the approach in [17] requires a manual pose correction for 13 out of 250 frames.

A visual comparison with a mesh-based method [15] is shown in Fig. 8. While Aguiar et al. [15] estimate the apparel but not the human pose well, in particular the orientations of the extremities like head and feet, our approach benefits from the underlying skeleton model and estimates the pose and shape accurately. To validate the benefit of coupling pose estimation and surface estimation in a direct comparison, we compared our approach with two variants. As in most previous work, the first variant performs only skeleton-based pose estimation (Section 4.1) without surface estimation. The second variant estimates only the surface (Section 4.2) and is therefore comparable to a mesh-based method, as in [15]. A visual comparison is shown in Fig. 9. As shown in Fig. 9b, the linear blend

2. Videos of the preliminary versions [21], [22] are available at http://www.youtube. com/watch?v=qCz68ukbZ7k and http://www.youtube.-com/watch?v=j4Zuj82FeLo.

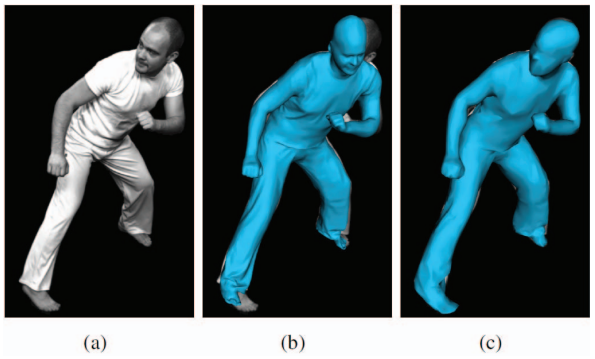


Fig. 8. Visual comparison of our approach with [15]. (a) Input image. (b) Tracked surface mesh from [15]. (c) Tracked surface mesh with lower resolution obtained by our method. Our approach estimates the human pose more accurately.

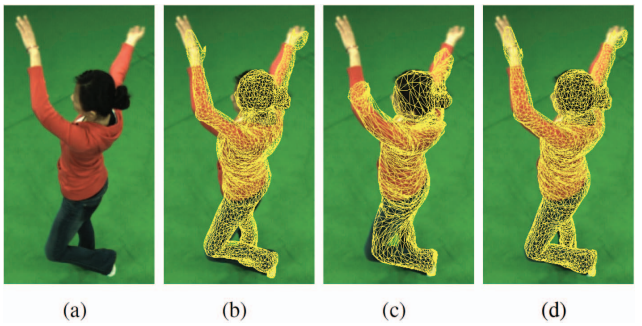


Fig. 9. Visual comparison of skeleton-based pose estimation, mesh-based surface estimation, and the proposed coupled approach. (a) Input image. (b) Estimated surface mesh with skeleton-based pose estimation. (c) Estimated surface mesh with mesh tracking. (d) Estimated surface mesh with the proposed approach.

skinning of the pose estimation does not capture the motion of the shirt and the trousers. Since the surface model does not fit perfectly, the residual error of each limb m after the local optimization, see (5), becomes larger, and the global optimization is triggered more often. In this case, the amount of global optimization increased from 8 percent using the proposed approach (Fig. 9d) to 68 percent.[3] Although the global optimization takes care that the pose does not get lost, the computation time greatly increases. The mesh-based approach without utilizing a skeleton completely fails to capture the arms, as shown in Fig. 9c. A quantitative comparison of skeleton-based pose estimation, mesh tracking, and the proposed approach is given in Section 6.2.3.

In contrast to [15] and [17], our single person motion tracking algorithm can also handle medium-resolution multiview sequences with extremely noisy silhouettes, like the HumanEvaII benchmark [1]. The dataset provides 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 sequence S4 with three subsets contains the motions walking, jogging, and balancing. The average errors for all three subsets are given in Fig. 10. 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 also compared our optimization scheme to a particle-based global optimization without local optimization. The global optimization with 15 iterations and 300 particles is not only slower, the error is also slightly higher. Although this seems to be counterintuitive, it can be explained by the different objective functions that are optimized. While local optimization minimizes the error of the contour and SIFT correspondences (4), the global optimization minimizes a very simple, pixelwise consistency measure (7). The error, however, becomes similar when both methods use only silhouettes, i.e., when SIFT features are not used by our approach, even though the objective functions are still not the same. Since our approach switches between local and global optimization and therefore between different objective functions, the standard deviation over frames is higher for our approach compared to the particle-based global optimization that estimates the pose more consistently over frames. In Section 6.2.3, we show that the difference between the two objective functions becomes smaller for high-resolution multiview sequences with less noisy silhouettes.

6.2 Multiperson Tracking

To evaluate the proposed approach for multiperson motion capture, we used 10 sequences containing two or three people closely interacting with each other. Fig. 11 shows for each sequence one frame with segmentation results and estimated skeleton poses and surface meshes. More results are shown in the supplemental video, which can be found online. Although the sequences are very challenging due to fast motions, severe occlusions, and appearance similarities, our approach provides accurate and visually appealing results. In particular, the segmentation results are very robust due to the 3D shape prior. For instance, the legs of the people in the sequences *Wrestle* and *Hug* are correctly labeled even though both people wear trousers of nearly the same color. Moreover, the feet in *Hop* and *Crossover* are assigned to the correct people despite occlusions. In some

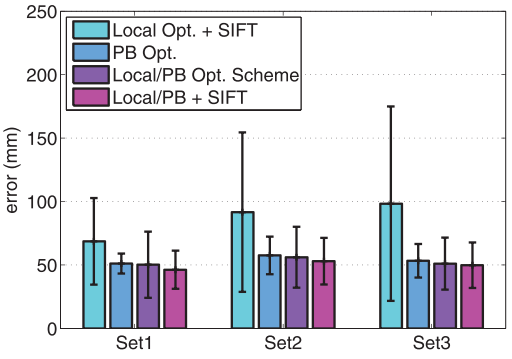


Fig. 10. Comparison of various optimization schemes. The bars show the average error and standard deviation of the joint positions of the skeleton for the S4 sequence of the HumanEva benchmark. The three sets cover the frames 2-350 (walking), 2-700 (walking + jogging), and 2-1,258 (walking + jogging + balancing). Despite of the use of SIFT features, the error of the local optimization is significantly higher in comparison to schemes that include a particle-based global optimization approach (PB).

3. The sequence contains 1,000 frames and tracking the sequence with one person would require to estimate $39 \times 1{,}000$ parameters in total. Due to the local-global optimization scheme, only a percentage of these parameters need to be estimated by global optimization, whereas the other parameters are estimated by local optimization.

Fig. 11. Markerless motion capture results. First row: Input images after background subtraction; second row: their corresponding segmentation results; third row: estimated surfaces and skeletons. From left to right: *Hop*, *Crossover*, *Bend*, *Wrestle*, *Hug*, *Hit*, *Jump*, *Crash*, *Couple dance*, and *Fight*.

cases, however, the segmentation is not perfect in one of the multiview frames due to very fast motions or color similarities that cannot be resolved by the shape prior, for example, foot in *Hit* and *Jump*. However, this happens only at very few frames, and the motion capture method is robust enough to deal with small inaccuracies in segmentation. Our method can also successfully capture pose and deforming surface geometry of people in loose apparel, for example, in *Couple dance*.

6.2.1 Impact of Segmentation

To show the importance of the segmentation (Section 5) for tracking multiple people, we compared our approach with a variant where the poses and shapes are estimated (Section 4) based on unsegmented foreground silhouettes. A visual comparison is shown in Fig. 12. Without segmentation, the data-to-model associations become ambiguous, yielding estimation errors (Fig. 12b). In particular, interactions with close physical contact and severe occlusions are problematic. Since the errors originate from problems in the underlying energy function of the pose estimation, even global optimization strategies cannot resolve them. Furthermore, relying only on global optimization would be very expensive due to the very high dimensional search space for multiple people. In contrast, the proposed approach correctly and efficiently determines shape and pose of both people, as local optimization succeeds in finding the correct poses for most frames (Fig. 12c).

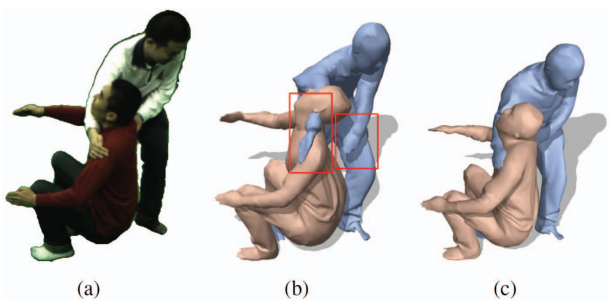


Fig. 12. (a) Input image after background subtraction. (b) Motion capture without segmentation. (c) Motion capture with segmentation. Without segmentation, features are assigned to the wrong model, which leads to significant errors.

6.2.2 Accuracy of Segmentation

For a quantitative evaluation of the segmentation, we manually labeled every 10th frame of all cameras for the sequences *Wrestle* and *Crossover*. Although these manual segmentations are not 100 percent accurate, they serve as ground-truth data for evaluation. The tracking accuracy of our approach depends on the quality of the segmentation, while in turn the segmentation depends on the tracking accuracy due to the shape prior. A high segmentation accuracy therefore also indicates accurate tracking results. Our method achieves pixel labeling accuracies of 98.4 and 98.9 percent on the sequences *Wrestle* and *Crossover*. These high values indicate that our approach is very successful in correctly segmenting the persons in the videos and thus also in tracking their motion.

We also evaluated the impact of the number of samples n used for approximating the 3D shape prior (22). We tracked both sequences with varying n and calculated the labeling accuracies, as can be seen in Fig. 13. The segmentation accuracy increases with the number of samples, but above around 200 samples the benefit of additional samples becomes negligible. Therefore, we conservatively set $n = 300$ for all our tracking experiments.

While we show qualitatively the impact of the appearance model and the shape prior in Fig. 4, we also quantitatively evaluated the impact of the terms on the segmentation accuracy. The results in Fig. 14 show that the appearance itself is too weak to obtain accurate segmentations. While the shape prior on its own generates reasonably accurate segmentations, only the proposed model, which combines appearance and shape, achieves very accurate results on both sequences.

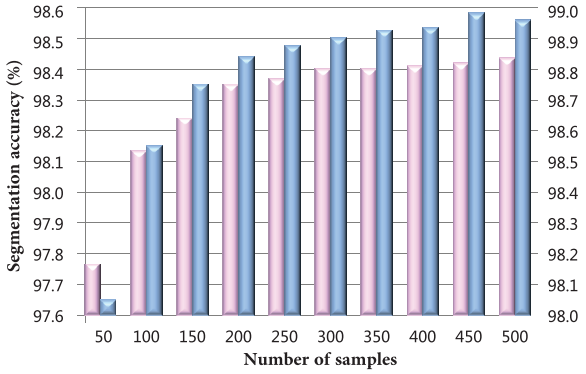


Fig. 13. Segmentation accuracy varies with the number of samples n. The red bars and blue bars show the accuracies for the sequence *Wrestle* and *Crossover*, respectively.

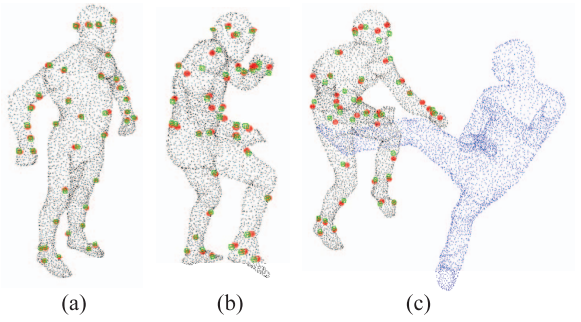


Fig. 15. Illustration of tracking accuracy. (a) Associating the markers with the vertices in the first reconstructed frame. (b) and (c) Position comparison of marker points (green points) and the corresponding 3D vertices (red points) on two of the temporal frames, with surface point cloud overlay.

6.2.3 Accuracy of Shape and Pose Estimation

For a quantitative evaluation of the shape and pose estimation, 38 markers were attached to one of the people whose motion was captured with a commercial PhaseSpace marker-based motion capture system, as shown in Fig. 15. The marker-based system was synchronized with the multiview video setup. As in all other sequences, the proposed markerless motion tracking and segmentation method is applied to the raw video data without exploiting any special knowledge about the markers in the scene. The untextured black motion-capture suit and the fast and complex motion make it challenging to track this sequence. As an error measure, we take the average distance between the markers and their corresponding vertices across all 500 frames of the evaluation sequence. This measure is more precise than the skeleton-based evaluation used for the HumanEva benchmark [1] because it captures all errors on the surface, including twists.

The average error with standard deviation is given in Fig. 16. In contrast to Fig. 10, the particle-based global pose estimation yields a slightly lower error than our more efficient optimization scheme that combines local optimization with global optimization. The plot also quantitatively evaluates the benefit of coupling pose and surface estimation as is quantitatively shown in Fig. 9. While mesh-based surface estimation without skeleton pose estimation similar to [15] performs poorly, the surface adaptation used in our approach improves the skeleton-based pose estimation. We also compare our approach to the recent work [60].

6.3 Impact of Template Models

We also thoroughly evaluated the impact of the template models on the tracking performance. In our evaluation sequence, the 3D mesh templates of the people were obtained using a laser scanner, which provides accurate and detailed geometry for the mesh templates (Figs. 17a and 17e). However, the mesh templates can also be obtained using other 3D reconstruction techniques such as multiview stereo or statistical human body models, for example, the SCAPE model [31] or [61]. These methods capture less accurate geometry and may contain reconstruction errors. To investigate the impact of the model accuracy on tracking, we generated smoothed versions (Figs. 17b, 17c, and 17f) and fitted a statistical human body model (Fig. 17d) to the scanned template mesh.

The average tracking error and standard deviation for four different combinations of the smoothed template models are given in Table 2. Geometric details are important for body parts that are approximately axially symmetric, like the head and the arm. Therefore, the error

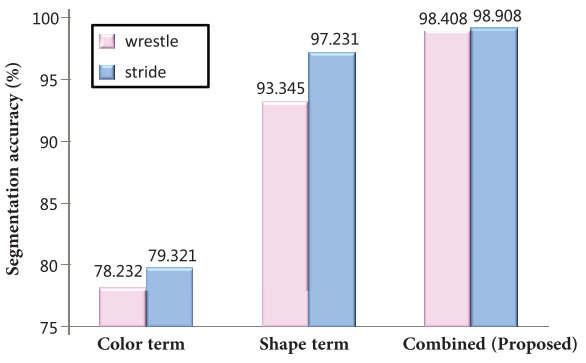


Fig. 14. Segmentation accuracy using only an appearance model, only the shape prior, or both. The red bars and blue bars show the accuracies for the sequences *Wrestle* and *Crossover*, respectively. The shape prior has been approximated with 300 samples.

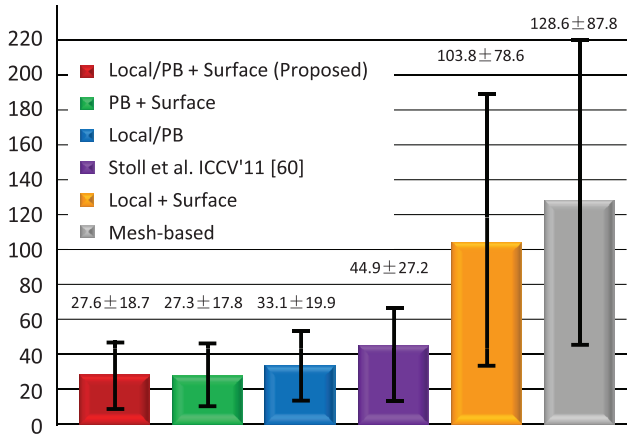


Fig. 16. Mean and standard deviation of the tracking error for the sequence shown in Fig. 15. Our proposed optimization achieves nearly the same performance as the particle-based global optimization at much lower computational cost and clearly outperforms local optimization. Without surface estimation, the average error is 5.5 mm higher for our approach, whereas using only surface estimation performs poorly. Our approach is also more accurate than the method in [60].

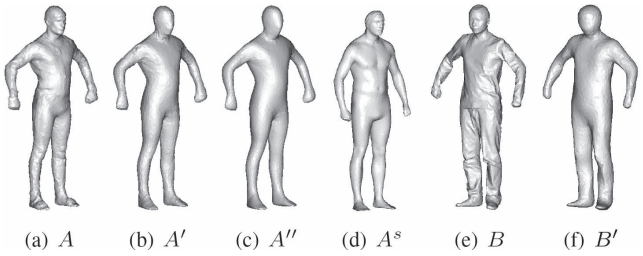


Fig. 17. Tracking with different 3D template models. (a) Scan of subject A. (b) Smoothed mesh A'. (c) Further smoothed mesh A''. (d) Fitted SCAPE model A^s. (e) Scan of subject B. (f) Smoothed mesh B'. The tracking results for the smoothed meshes are shown in Table 2. The tracking results for the SCAPE model with different parameters are given in Fig. 18.

TABLE 2
Comparison of Tracking Performances
Using Differently Smoothed Template Models

Models	$A+B$	$A'+B$	$A'+B'$	$A''+B'$
Error(mm)	27.6±18.7	30.2±18.9	31.0±19.0	31.2±19.0

The corresponding models are shown in Fig. 17.

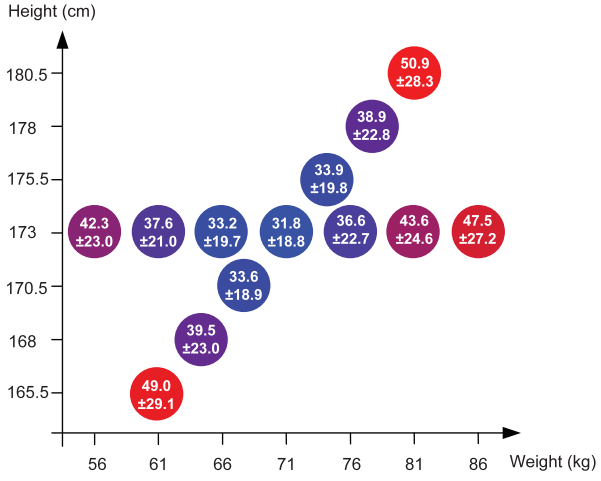


Fig. 18. Comparison of tracking performances using different 3D template models. Each circle is a sample of the shape parameter space of the fitted SCAPE model A^s shown in Fig. 17d. The mean and standard deviation of the tracking errors are shown in the center of each sample.

slightly increases by smoothing the template meshes. While the error is only measured for subject A, it is interesting to note that the quality of the model B has also some impact on the tracking accuracy of subject A. Based on the fitted statistical body model A^s (Fig. 17d), we also modified the weight and height parameters of the model. The tracking errors are given in Fig. 18. Although statistical body models do not capture apparel, the error of A^s (31.8 mm) is only slightly higher than using a smoothed model. The change on the weight and height of the model, however, substantially degrades the tracking performance. This can be explained by the mismatch of the skeleton of the model with the skeleton of the subject. While smoothing or fitting a body model to the scan mainly affects the surface mesh, changing height and weight also affects the skeleton. Although our method still produces accurate tracking results in comparison to related work, it shows that the body model can have a big impact on the tracking accuracy.

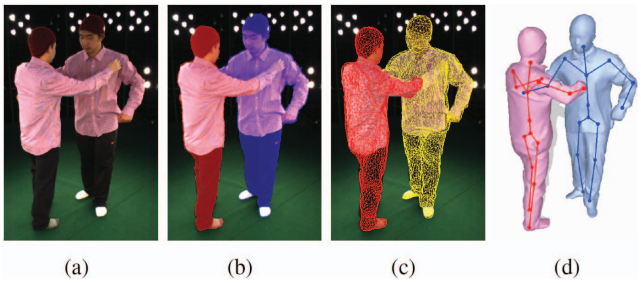


Fig. 19. Two people wearing the same clothes. (a) One of the input images. (b) The right arm of one person is wrongly segmented. (c) Estimated surface meshes. (d) Reconstructed models with skeletons.

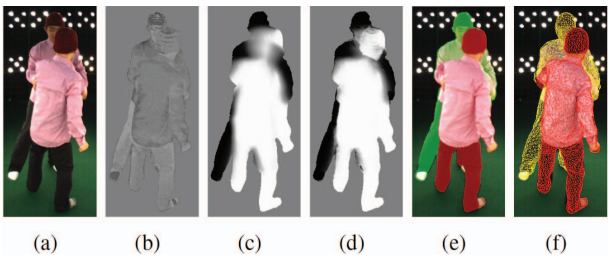


Fig. 20. (a) Another example of two persons wearing the same clothes. (b) Color term. (c) Shape term. (d) Combined shape prior and body part appearance model. (e) The right arm of the occluded person is wrongly segmented. (f) Estimated surface meshes.

6.4 Computation Time

The local optimization for pose estimation takes about 3 seconds per frame for a single person in an 8-camera setup. The global optimization takes about 12 seconds for each dimension that is optimized and up to a maximum of about 168 seconds per frame. The surface estimation takes 2 seconds. Depending on the difficulty of the performed motion, the runtime per frame for a single person without segmentation thus varies from about 6 seconds to 2.5 minutes.

When tracking multiple subjects, the runtime is mainly limited by the calculation of the shape prior, which takes about 1.5 minutes per person per frame. However, the computation time of the global optimization and the shape prior could be drastically reduced to a few seconds by using a GPU [62]. The image segmentation takes 10 seconds for a frame composed of 12 images. The whole system for capturing the motion of two persons and 12 cameras takes 3 to 6.5 minutes per frame on a standard PC and 4.5 to 8 minutes for three persons and 12 cameras.

6.5 Limitations

Currently, our approach assumes that the clothes of the captured actors are at least slightly different, which is usually the case in everyday life; see the first row of Fig. 11. When the clothes of the people are exactly the same, as in Figs. 19 and 20, the color term Fig. 20b is not able to discriminate between the people and the shape prior fails to resolve the ambiguities when the persons are in contact. As result, some body parts are wrongly labeled, as shown in Figs. 19b and 20e, and the pose and shape estimation are erroneous. Our segmentation and tracking method may also fail when the hands of two people touch, as neither appearance nor shape information are sufficient to uniquely identify the person for each pixel. For instance, the hands of the people in the sequence *Couple dance* are not correctly

tracked (Fig. 11). This issue may be resolved at the cost of computation time by explicitly modeling body parts and intersections. The detail of geometry that can be captured is also limited by the image resolution and the used image features, namely, silhouettes and SIFT features. We also assume that foreground silhouettes are available or can be easily extracted. An extension of the segmentation to general scene backgrounds, however, is feasible. Finally, runtime performance can be improved by using a GPU or lower resolution meshes for the shape prior.

7 Conclusion

We have proposed an approach that advances the state of the art in markerless human motion capturing because it is the first approach that captures skeleton motion and detailed surface geometry of two or more closely interacting persons. To keep the complexity of the problem tractable, we have divided the task into several subproblems that are solved for each frame one by one, but that depend on each other over the entire sequence. For motion capture, we first estimate the articulated motion by a skeleton-based approach using a combination of local and global optimization. The residual nonarticulated motion is then estimated by a mesh-based approach. To capture the motion of multiple people, we first solve the feature-to-model assignment problem by segmentation and then estimate the pose and shape of each person independently. We have further shown that the proposed 3D shape prior is a better model for segmentation than commonly used 2D shape priors.

Acknowledgments

This work was supported by the National Basic Research Project (No. 2010CB731800) and the Project of NSFC (No. 61073072, 60932007, U0935001, and 61021063), and the Intel Visual Computing Institute in Saarbrücken, Germany. Juergen Gall was supported by the DFG Emmy Noether program (GA 1927/1-1).

References

[1] L. Sigal, A. Balan, and M. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *Int'l J. Computer Vision,* vol. 87, pp. 4-27, 2010.

[2] *Visual Analysis of Humans—Looking at People,* T.B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, eds. Springer, 2011.

[3] D. Gavrila and L. Davis, "3-D Model-Based Tracking of Humans in Action: A Multi-View Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 73-80, 1996.

[4] C. Bregler, J. Malik, and K. Pullen, "Twist Based Acquisition and Tracking of Animal and Human Kinematics," *Int'l J. Computer Vision,* vol. 56, no. 3, pp. 179-194, 2004.

[5] J. Deutscher and I. Reid, "Articulated Body Motion Capture by Stochastic Search," *Int'l J. Computer Vision,* vol. 61, no. 2, pp. 185-205, 2005.

[6] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, "Optimization and Filtering for Human Motion Capture—A Multi-Layer Framework," *Int'l J. Computer Vision,* vol. 87, no. 1, pp. 75-92, 2010.

[7] L. Ballan and G. Cortelazzo, "Marker-Less Motion Capture of Skinned Models in a Four Camera Set-Up Using Optical Flow and Silhouettes," *Proc. Int'l Conf. 3D Data Processing, Visualization and Transmission,* 2008.

[8] R.P. Horaud, M. Niskanen, G. Dewaele, and E. Boyer, "Human Motion Tracking by Registering an Articulated Surface to 3-D Points and Normals," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 158-164, Jan. 2009.

[9] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T.P. Andriacchi, "Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation," *Int'l J. Computer Vision,* vol. 87, nos. 1/2, pp. 156-169, 2010.

[10] R. Li, T.-P. Tian, S. Sclaroff, and M.-H. Yang, "3D Human Motion Tracking with a Coordinated Mixture of Factor Analyzers," *Int'l J. Computer Vision,* vol. 87, nos. 1/2, pp. 170-190, 2010.

[11] L. Bo and C. Sminchisescu, "Twin Gaussian Processes for Structured Prediction," *Int'l J. Computer Vision,* vol. 87, nos. 1/2, pp. 28-52, 2010.

[12] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient Regression of General-Activity Human Poses from Depth Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 415-422, 2011.

[13] J. Starck and A. Hilton, "Model-Based Multiple View Reconstruction of People," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 915-922, 2003.

[14] K. Varanasi, A. Zaharescu, E. Boyer, and R. Horaud, "Temporal Surface Tracking Using Mesh Evolution," *Proc. European Conf. Computer Vision,* pp. 30-43, 2008.

[15] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance Capture from Sparse Multi-View Video," *ACM Trans. Graphics,* vol. 27, article 98, 2008.

[16] C. Cagniart, E. Boyer, and S. Ilic, "Probabilistic Deformable Surface Tracking from Multiple Videos," *Proc. European Conf. Computer Vision,* 2010.

[17] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated Mesh Animation from Multi-View Silhouettes," *ACM Trans. Graphics,* vol. 27, no. 3, pp. 1-9, 2008.

[18] C. Stoll, J. Gall, E. de Aguiar, S. Thrun, and C. Theobalt, "Video-Based Reconstruction of Animatable Human Characters," *ACM Trans. Graphics,* vol. 29, no. 6, article 139, 2010.

[19] H. Egashira, A. Shimada, D. Arita, and R. Taniguchi, "Vision-Based Motion Capture of Interacting Multiple People," *Proc. Int'l Conf. Image Analysis and Processing,* pp. 451-460, 2009.

[20] C. Cagniart, E. Boyer, and S. Ilic, "Free-Form Mesh Tracking: A Patch-Based Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.

[21] J. Gall, C. Stoll, E. Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion Capture Using Joint Skeleton Tracking and Surface Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1746-1753, 2009.

[22] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless Motion Capture of Interacting Characters Using Multi-View Image Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1249-1256, 2011.

[23] A. Fossati, J. Gall, H. Grabner, X. Ren, and K. Konolige, *Consumer Depth Cameras for Computer Vision—Research Topics and Applications.* Springer, 2012.

[24] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient Human Pose Estimation from Single Depth Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* doi: 10.1109/TPAMI.2012.241, 2012.

[25] R. Kehl, M. Bray, and L. van Gool, "Full Body Tracking from Multiple Views Using Stochastic Sampling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 129-136, 2005.

[26] T. Drummond and R. Cipolla, "Real-Time Tracking of Highly Articulated Structures in the Presence of Noisy Measurements," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 315-320, 2001.

[27] C. Sminchisescu and B. Triggs, "Estimating Articulated Human Motion with Covariance Scaled Sampling," *Int'l J. Robotics Research,* vol. 22, no. 6, pp. 371-391, 2003.

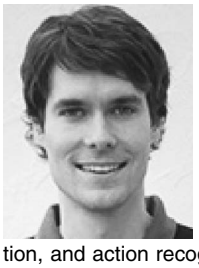
[28] S. Corazza, L. Mündermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. Andriacchi, "A Markerless Motion Capture System to Study Musculoskeletal Biomechanics: Visual Hull and Simulated Annealing Approach," *Annals Biomedical Eng.,* vol. 34, no. 6, pp. 1019-1029, 2006.

[29] R. Plankers and P. Fua, "Articulated Soft Objects for Multiview Shape and Motion Capture," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 9, pp. 1182-1187, Sept. 2003.

[30] G. Cheung, S. Baker, and T. Kanade, "Shape-from-Silhouette across Time Part II: Applications to Human Modeling and Markerless Motion Tracking," *Int'l J. Computer Vision,* vol. 63, no. 3, pp. 225-245, 2005.

[31] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape Completion and Animation of People," *ACM Trans. Graphics,* vol. 24, no. 3, pp. 408-416, 2005.
[32] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker, "Detailed Human Shape and Pose from Images," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[33] A. Balan and M. Black, "The Naked Truth: Estimating Body Shape under Clothing," *Proc. European Conf. Computer Vision,* pp. 15-29, 2008.
[34] B. Rosenhahn, U. Kersting, K. Powell, R. Klette, G. Klette, and H.-P. Seidel, "A System for Articulated Tracking Incorporating a Clothing Model," *Machine Vision Applications,* vol. 18, no. 1, pp. 25-40, 2007.
[35] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel, "Marker-Less Deformable Mesh Tracking for Human Shape and Motion Capture," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2007.
[36] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof, "Simultaneous Shape and Pose Adaption of Articulated Models Using Linear Optimization," *Proc. European Conf. Computer Vision,* pp. 724-737, 2012.
[37] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt, "Performance Capture of Interacting Characters with Handheld Kinects," *Proc. European Conf. Computer Vision,* pp. 828-841, 2012.
[38] J.-Y. Guillemaut, J. Kilner, and A. Hilton, "Robust Graph-Cut Scene Segmentation and Reconstruction for Free-Viewpoint Video of Complex Dynamic Scenes," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 809-816, 2009.
[39] A.M. Elgammal and L.S. Davis, "Probabilistic Framework for Segmenting People under Occlusion," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 145-152, 2001.
[40] S.M. Khan and M. Shah, "Tracking Multiple Occluding People by Localizing on Multiple Scene Planes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 3, pp. 505-519, Mar. 2009.
[41] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera People Tracking with a Probabilistic Occupancy Map," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 2, pp. 267-282, Feb. 2008.
[42] K. Kim and L.S. Davis, "Multi-Camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering," *Proc. European Conf. Computer Vision,* pp. 98-109, 2006.
[43] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D Pose Estimation and Tracking by Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.
[44] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. van Gool, "Articulated Multibody Tracking under Egomotion," *Proc. European Conf. Computer Vision,* 2008.
[45] Q. Zhang and K.N. Ngan, "Segmentation and Tracking Multiple Objects under Occlusion from Multiview Video," *IEEE Trans. Image Processing,* vol. 20, no. 11, pp. 3308-3313, Nov. 2011.
[46] P. Kohli, J. Rihan, M. Bray, and P. Torr, "Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts," *Int'l J. Computer Vision,* vol. 79, pp. 285-298, 2008.
[47] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers, "Combined Region- and Motion-Based 3D Tracking of Rigid and Articulated Objects," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 3, pp. 402-415, Mar. 2010.
[48] J. Gall, B. Rosenhahn, and H.-P. Seidel, "Drift-Free Tracking of Rigid and Articulated Objects," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[49] R.M. Murray, S.S. Sastry, and L. Zexiang, *A Mathematical Introduction to Robotic Manipulation.* CRC Press, Inc., 1994.
[50] L. Kavan, S. Collins, J. Žára, and C. O'Sullivan, "Skinning with Dual Quaternions," *Proc. Symp. Interactive 3D Graphics and Games,* pp. 39-46, 2007.
[51] I. Baran and J. Popović, "Automatic Rigging and Animation of 3D Characters," *ACM Trans. Graphics,* vol. 26, no. 3, article 72, 2007.
[52] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.
[53] J. Stolfi, *Oriented Projective Geometry: A Framework for Geometric Computation.* Academic Press, 1991.
[54] J. Gall, J. Potthoff, C. Schnoerr, B. Rosenhahn, and H.-P. Seidel, "Interacting and Annealing Particle Filters: Mathematics and a Recipe for Applications," *J. Math. Imaging and Vision,* vol. 28, no. 1, pp. 1-18, 2007.
[55] M. Botsch and O. Sorkine, "On Linear Variational Surface Deformation Methods," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 1, pp. 213-230, Jan./Feb. 2008.
[56] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 1068-1080, June 2008.
[57] Y. Boykov and M. Jolly, "Iterative Graph Cuts for Optimal Boundary and Region Segmentation of Objects in n-d Images," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 105-112, 2001.
[58] Y. Boykov, O. Veksler, and R. Zabih, "Markov Random Fields with Efficient Approximations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 648-655, 1998.
[59] J. Hammersley and D. Handscomb, *Monte Carlo Methods.* Methuen, 1964.
[60] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, "Fast Articulated Motion Tracking Using a Sums of Gaussians Body Model," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 951-958, 2011.
[61] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A Statistical Model of Human Pose and Body Shape," *Computer Graphics Forum,* vol. 2, no. 28, 2009.
[62] M. Shaheen, J. Gall, R. Strzodka, L. Van Gool, and H.-P. Seidel, "A Comparison of 3D Model-Based Tracking Approaches for Human Motion Capture in Uncontrolled Environments," *Proc. Workshop Applications of Computer Vision,* 2009.

Yebin Liu received the BE degree from the Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He was a research fellow in the Computer Graphics Group of the Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor at Tsinghua University. His research areas include computer vision and computer graphics.



Juergen Gall received the PhD degree in computer science from Saarland University and the Max Planck Institut für Informatik in 2009. From 2009 until 2012, he was a postdoctoral researcher in the Computer Vision Laboratory, ETH Zurich. Since 2012, he has been a senior research scientist at the Max Planck Institute for Intelligent Systems in Tübingen. His research interests include interacting particle systems, markerless human motion capture, object detection, and action recognition. He is a member of the IEEE.

Carsten Stoll received the PhD degree in computer science from Saarland University and the Max Planck Institut für Informatik in 2009. Since 2010, he has headed the research group "Optical Performance Capture" at the Max Planck Center for Visual Computing and Communication in Saarbrücken. His research interests include human performance capture and geometric modeling.

Qionghai Dai received the BS degree in mathematics from Shanxi Normal University, China, in 1987, and the ME and PhD degrees in computer science and automation from Northeastern University, China, in 1994 and 1996, respectively. Since 1997, he has been with the faculty of Tsinghua University, Beijing, China, where he is currently a professor and the director of the Broadband Networks and Digital Media Laboratory. His research areas include video communication, computer vision, and graphics. He is a senior member of the IEEE.

Hans-Peter Seidel is the scientific director and chair of the computer graphics group at the Max Planck Institut für Informatik and a professor of computer science at Saarland University. He has received grants from a wide range of organizations, including the German National Science Foundation (DFG), the German Federal Government (BMBF), the European Community (EU), NATO, and the German-Israel Foundation. In 2003, he received the Leibniz Preis, the most prestigious German research award, from the German Research Foundation (DFG).

Christian Theobalt is an associate professor of computer science at the Max Planck Institut für Informatik and Saarland University, Saarbrücken, Germany. Most of his research deals with algorithmic problems that lie on the boundary between the fields of computer vision and computer graphics, such as dynamic 3D scene reconstruction and markerless motion capture. For his work, he has received several awards, including the Otto Hahn Medal of the Max-Planck Society in 2007, the EUROGRAPHICS Young Researcher Award in 2009, and the German Pattern Recognition Award in 2012.