Markerless Motion Capture of Interacting Characters Using Multi-view Image Segmentation

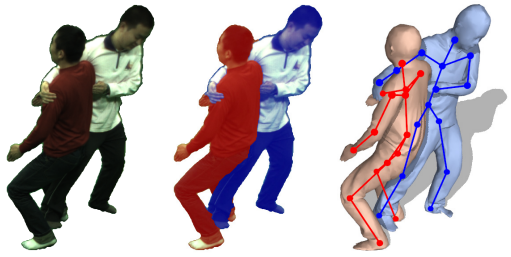
Yebin Liu[1,3] Carsten Stoll[1] Juergen Gall[2] Hans-Peter Seidel[1] Christian Theobalt[1]
[1]MPI Informatik [2]BIWI, ETH Zurich [3]Automation Department, Tsinghua University, TNList

Abstract

We present a markerless motion capture approach that reconstructs the skeletal motion and detailed time-varying surface geometry of two closely interacting people from multi-view video. Due to ambiguities in feature-to-person assignments and frequent occlusions, it is not feasible to directly apply single-person capture approaches to the multi-person case. We therefore propose a combined image segmentation and tracking approach to overcome these difficulties. A new probabilistic shape and appearance model is employed to segment the input images and to assign each pixel uniquely to one person. Thereafter, a single-person markerless motion and surface capture approach can be applied to each individual, either one-by-one or in parallel, even under strong occlusions. We demonstrate the performance of our approach on several challenging multi-person motions, including dance and martial arts, and also provide a reference dataset for multi-person motion capture with ground truth.

1. Introduction

Nowadays, motion capture is an essential acquisition technology with many applications in computer vision and computer graphics. Many human motions can only be observed in the context of human-human interactions. In some application areas like sports science, biomechanics, or character animation for games and movies, these interactions involve frequent close physical contact. Marker-based systems can in principle capture such motions of interacting subjects, but they suffer from widely known shortcomings, such as: errors due to broken marker trajectories, long setup times, and the inability to simultaneously capture dynamic shape and motion of actors in normal clothing. Further on, in multi-person sequences, frequent manual intervention is necessary. Markerless multi-view capturing algorithms overcome some of these limitations for single person motions, and succeed to reconstruct motion and time-varying geometry of people in loose apparel, like a skirt [12, 16, 25]. However, on sequences where multiple persons interact closely, markerless methods typically struggle.



(a) (b) (c)

Figure 1. Our approach captures the motion of interactive characters even in the case of close physical contact: (a) one of the 12 input images, (b) segmentation, (c) estimated skeleton and surface.

In the multi-person case, the amount of pose ambiguities increases significantly since commonly used features like silhouettes, color, edges, or interest points cannot be uniquely assigned to one person. Due to frequent occlusions, these ambiguities become even more challenging when people interact closely. It is infeasible to directly apply single-person pose optimization algorithms that rely on such features to the multi-person case, e.g., by jointly optimizing the pose parameter space of two body models. However, if each image pixel could be assigned to one of the observed persons or the background, pose estimation could be performed by a single-person tracker, even under occlusion.

In this paper, we propose a markerless motion capture method for two interacting characters that is based on robust segmentation of the input multi-view videos, see Fig. 1. After segmentation, an existing state-of-the-art single-person markerless motion capture method is adapted to track each person.

In order to resolve the pixel assignment before pose estimation, we employ multi-view image segmentation to determine the image regions each person belongs to. To this end, we introduce a novel shape prior for segmenting interacting characters that integrates the previously estimated poses and shapes. The segmentation allows us to gener-

ate 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. Our main contributions are:

- We introduce a maximum a-posteriori Markov random field (MAP-MRF) optimization framework to segment the persons in each image. We incorporate shape, pose, and local appearance information from the previously tracked frame into our energy function.
- We adapt the single-person motion capture method from [16] to reliably recover skeletal motion and detailed time-varying geometry of multiple persons, even for complex motions like dancing and martial arts. Tracking is efficient since every segmented person is captured independently.
- We provide a new multi-view human-human interaction (MHHI) dataset[1] for evaluating multi-person capture approaches. The ground truth of the dataset is obtained by an industrial marker-based capture system.

We demonstrate the reliability of our approach with 7 different sequences consisting of over 1500 frames of multi-view video that were tracked fully automatically.

2. Related Work

Our work advances markerless motion capture approaches, that, so far, struggle to capture closely interacting actors. To enable this, we also capitalize on previous research in image segmentation.

Markerless Motion Capture Markerless human motion capture has been a very active field in computer vision for the past decades [21, 22]. Popular methods like [8, 13, 4] model a human's motion by articulated bone hierarchies. However, simple articulated models are often not able to capture the shape and motion of the human body in all detail. Statistical SCAPE body models [2, 3] represent the human body better, but time varying geometry, such as moving cloth, is not captured. Approaches like [23, 24] rely on the visual hull to reconstruct detailed geometry, but often suffer from topology changes that occur frequently in shape-from-silhouette reconstructions. Some recent approaches [10, 12] overcome these problems by using a template mesh which is tracked through the sequence. The methods of Vlasic et al. [25] and Gall et al. [16] combine the advantages of skeleton-based and mesh-based approaches. They estimate both skeleton motion and the time varying geometry.

Markerless motion-capture of multiple performers has only been considered in very few works. Cagniart et al. [10, 11] 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. In the very restricted context of pedestrians and walking motion, the skeleton motions of several persons have been estimated in [1, 17].

Segmentation for Shape Reconstruction and Tracking The joint problem of human pose estimation and multi-view segmentation has been addressed for a single person in several works [7, 9, 15]. The works [9, 15] use the previous articulated pose as shape prior for level-set segmentation and estimate the pose either within an analysis-by-synthesis framework [15] or in combination with optical flow and SIFT features [9]. Graph-cut segmentation is used in [7] where a multi-view 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 persons, this may become even more of a problem since occlusions often change the 2D topology.

There are a few recent papers which consider segmentation and tracking with more than one subject. Guillemaut et al. [18] 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. [14] propose a volumetric segmentation on the visual hull of the scene to separate the persons. However, when two persons are in physical contact, volumetric segmentation of the visual hull is not as accurate as the visual hulls of the persons segmented in the image domain.

Our approach is the first method that handles challenging human-human interactions, extracts accurate silhouettes for each person, and recovers pose and detailed time varying geometry of more than one person.

3. Overview

The performance of human interactions is captured by synchronized and calibrated cameras. Similar to [16], we acquire for each person a rigged 3d shape model comprising a bone skeleton, a triangle mesh surface model, and skinning weights for each vertex, which connect the mesh to the skeleton (Fig. 2(a)). The mesh surface of each person can be generated using a laser scanner or multi-view stereo methods. In our experiments, we use laser scans of the actors, each rigged with a skeleton with 39 degrees of freedom. Such a skeleton partitions the whole body into 15

[1]The dataset is available upon request.

(a) (b) (c) (d) (e) (f)

Figure 2. 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.

body parts where the highest skinning weight of a vertex determines its unique association to a body part. For each image, foreground silhouettes are extracted by background subtraction (Fig. 2(b)).

As in [16, 25], 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 non-articulated surface deformations (*shape*) that cannot be approximated 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. 2. Starting with the estimated poses and shapes of the two persons in the previous frame, the proposed algorithm estimates the poses and the shapes in the current frame based on the captured multi-view images and foreground silhouettes covering the two persons (the initial pose is found as in [16]) . Since the whole space for the unknown pose and shape parameters becomes very large for two persons, we split the whole tracking problem into a multi-view 2D segmentation problem (Fig. 2(c,d)) and a 3D pose and shape estimation problem (Fig. 2(e,f)). The segmentation separates the two persons in the image domain by assigning a label to each foreground pixel. It relies on a novel probabilistic shape prior derived from the previous estimated poses and shapes (Sec. 4). Then, based on the labeled pixels, the pose and the shape are estimated for each person independently (Sec. 5).

4. Multi-view Image Segmentation

The proposed multi-view segmentation of foreground pixels (Fig. 2(b)) is defined as MAP-MRF [6] based on appearance, pose, and shape information. Our energy function yields segmentations that are both efficient and robust for human motion capture under serious occlusions and ambiguous appearance.

4.1. MAP-MRF Image Segmentation

In image segmentation, I is the set of image pixels to be segmented, and N defines a neighborhood on this set (in our case 8 pixels). A configuration $\mathbf{L}$ defines a segmentation. In our case, we have a label for each person, *i.e.*, $l_i \in \{A, B\}$. The image segmentation problem can be solved by finding the least energy configuration of the MRF. Given the observed data D, a commonly used energy corresponding to configuration $\mathbf{L}$ comprises three terms:

$$\Psi(\mathbf{L}) = \sum_{i \in I} \left(\phi(D|l_i) + \sum_{j \in N_i} \left(\phi(D|l_i, l_j) + \psi(l_i, l_j) \right) \right) \quad (1)$$

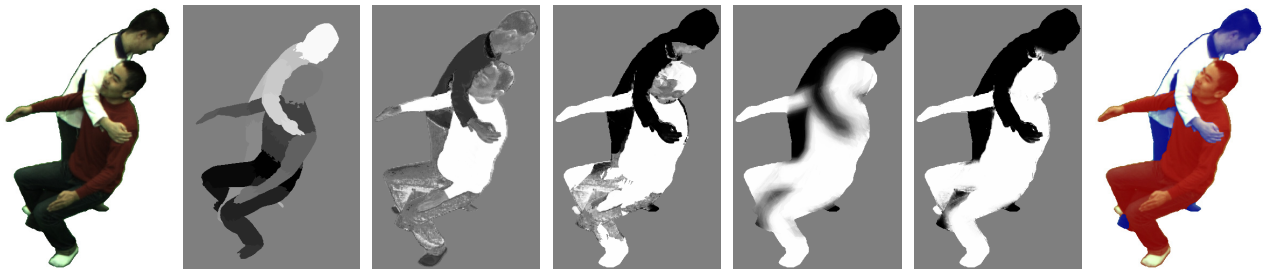
where $\phi(D|l_i)$ is a likelihood data term which imposes individual penalties for assigning a label to pixel i. While this term incorporates only appearance information for a standard segmentation problem, we propose a term that takes appearance, pose, and shape information into account. The term will be described in detail in Section 4.2. $\psi(l_i, l_j)$ is a smoothness prior taking the form of a generalized Potts model [6]. The contrast term $\phi(D|l_i, l_j)$ favors pixels with similar color having the same label. As in [5, 7], we use the contrast term

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

where $||I_i - I_j||^2$ measures the difference in the color values of pixels i and j and $S(i, j)$ the spatial distance between the pixels. Once the MAP-MRF energy function in Eq. (1) has been defined, it can be minimized via graph cuts [5].

4.2. Segmentation Using Shape and Appearance

A standard MRF for image segmentation performs poorly when segmenting images in which the appearance models of the two persons are not highly discriminative. For instance, skin and hair colors of two persons are often very similar. In our case, the poses and shapes of the two persons have been recovered in the previous frame, which are strong



(a) (b) (c) (d) (e) (f) (g)

Figure 3. Segmentation with shape and appearance information. (a) Input image after background subtraction. (b) Body parts B_k^j. (c) Probability for label A according to color term using a whole body appearance model *(white: high, black: low)*. (d) Probability using the body part appearance model. (e) Shape prior. (f) Data term Eq. (5) combining shape and color. (g) Segmentation result from (f).

cues that can be integrated as shape priors for segmentation. However, we have to pay attention to the fact that the two persons are very close and may occlude each other. Hence, we propose a functional that models the appearance locally on the surface of each person and integrates shape priors for each person.

We represent the shape priors conditioned on the previous estimated surfaces and poses. The skeleton pose of person k is parameterized by a vector Θ_k that defines the global position and rotation of the torso and the configuration of the joint angles. Together with the estimated surface S_k, we get the current mesh M_k by applying the corresponding transformation T_{Θ_k} to each vertex of the surface mesh S_k, shortly denoted by $M_k = T_{\Theta_k} S_k$. The estimation of the pose and the shape will be discussed in Sec. 5. In this section, we assume that Θ_k and S_k are available from the previous frame. Hence, we define the likelihood data term $\phi(D|l_i)$ not only conditioned on the label $l_i = k$ but also on the corresponding surface and pose:

$$\phi(D|l_i) \propto -\log P(D|S_k, \Theta_k, l_i = k) \quad if(l_i = k). \quad (3)$$

The color distribution is often consistent for a body part but varies strongly between different body parts, e.g., while hands are typically skin colored, other parts like upper body or legs are often covered by clothes of a specific color. Since a color distribution for the whole body is consequently not very discriminative to distinguish two persons, we model the person's appearance for each of the body parts B_k^j:

$$P(D|S_k, \Theta_k, l_i = k) = \quad (4)$$
$$\sum_j P(D|i \in B_k^j, S_k, \Theta_k, l_i = k) P(i \in B_k^j | S_k, \Theta_k, l_i = k).$$

$P(i \in B_k^j | S_k, \Theta_k, l_i)$ 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. 4.2.1. Since the appearance of a pixel depends only on the body part, Eq. (4) can be simplified as

$$P(D|S_k, \Theta_k, l_i = k) = \sum_j P(D|i \in B_k^j) P(i \in B_k^j | S_k, \Theta_k, l_i = k). \quad (5)$$

The likelihood term, namely color term $P(D|i \in B_k^j) \propto P(I_i|H_k^j)$ 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 distributions H_k^j are modeled using the images from the first time step of a sequence. Fig. 3 illustrates the respective terms used during segmentation. The advantage of per-part color modeling is shown in Fig. 3(c,d).

4.2.1 Shape Prior

The term $P(i \in B_k^j | S_k, \Theta_k, l_i = k)$ in Eq. (5) is the shape prior that provides for each pixel i not only an a-priori probability for assigning a label k to it, but it also encodes the probability to which body part B_k^j it belongs to. The simplest way to model this probability would be to diffuse the 2D image silhouette or the 2D body part map of each S_k and then combine them. Regardless of whether occlusion is taken into account, this approach struggles to model the shape prior accurately, as shown in Fig. 4(b,c).

To address this issue, we do not rely on 2D diffusion of the exact pose Θ_k and surface S_k from the previous frame, but use the posterior probability

$$P(\Theta|D, S) \propto P(D|\Theta, S) P(\Theta), \quad (6)$$

that is defined for both persons, *i.e.*, $\Theta = (\Theta_A, \Theta_B)$ and $S = (S_A, S_B)$. While the previously estimated shapes S remain unchanged, we sample new pose configurations Θ for both persons from the posterior by importance sampling [19]. The pose parameters Θ are predicted from the previously estimated poses, where the distribution $P(\Theta)$ is modeled by a Gaussian with mean corresponding to the previously estimated poses. The likelihood term, $P(D|\Theta, S)$,

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

Figure 4. 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 hand of the man) that is occluded in the former frame. In contrast, the proposed 3D shape diffusion gives a better probability in this region (green circle) (d), which leads to a better segmentation (Fig. 2(c)).

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

Figure 5. Impact of the weighting parameters λ_j (Eq. 7). (a) Shape prior without weighting. (b) Shape prior with weighting. (c) Segmentation without weighting. (d) Segmentation with weighting.

measures the consistency of the projected surfaces $\mathbf{B}_v$ with the foreground silhouettes F_v for all views v:

$$P(D|\Theta, S) \propto \exp\left(-\sum_v d(F_v, \mathbf{B}_v)\right), \tag{7}$$

$$d(F_v, \mathbf{B}_v) = \sum_i \left(g(F_{v,i}, \mathbf{B}_{v,i}) + \sum_j \lambda_j \cdot f_j(\mathbf{B}_{v,i}, F_{v,i}) \right),$$

Here, g and f_j measure the pixel-wise difference between projected surfaces and the given silhouette image. $g(x, y)$ is 1 if x is a foreground pixel and y is a background pixel. $f_j(x, y)$ is 1 if x belongs to body part j and y is a background pixel. The weighting parameters λ_j steer the impact of each body part j. In this work, we set λ_j inversely proportional to the size of each body part to equalize the impact of all body parts. Note that the likelihood takes all views and all persons into account. Fig. 4(d) shows the advantage of shape priors using the 3D shape posterior.

In order to approximate $P(\Theta|D, S)$, we draw a set of samples, $\{\Theta^n\}$, from $P(\Theta)$, and weight them by

$$w_n = \frac{\exp\left(-\sum_v d(F_v, \mathbf{B}_v(\Theta^n))\right)}{\sum_n \exp\left(-\sum_v d(F_v, \mathbf{B}_v(\Theta^n))\right)}. \tag{8}$$

Hence, the probability $P(i \in B_k^j | S_k, \Theta_k, l_i = k)$ in Eq. (5) for assigning a pixel i the body part label b_k^j for person k becomes:

$$P(i \in B_k^j | S_k, \Theta_k, l_i{=}k) = \sum_n w_n \cdot \delta_{b_k^j}(\mathbf{B}_{v,i}(\Theta^n)), \tag{9}$$

$$\delta_{b_k^j}(\mathbf{B}_{v,i}(\Theta^n)) = \begin{cases} 1 & \text{if } \mathbf{B}_{v,i}(\Theta^n) = b_k^j, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where v is the corresponding view. Fig. 5 shows the advantage of weighting the body parts as in Eq. (7). Note that we are only interested in a good representation of the shape prior, and thus in the projections and not in the full pose posterior. Since several poses lead to similar projections, we achieve good results with a relatively low number of samples, despite a 78-dimensional space Θ. In our experiments, we found 300 samples enough for a reasonable approximation of the posterior Eq. (6).

4.2.2 Resolving Intersections

When the interacting persons are close to each other, the sampling from $P(\Theta)$ might generate meshes that intersect with each other in 3D. For the example shown in Fig. 1, over 80% of the samples have slight or serious intersections. Although the sampling distribution $P(\Theta)$ can be changed to generate only meshes without intersections, the additional intersection tests and constraints would make the sampling procedure expensive.

Since we are only interested in an accurate shape prior, we can apply a simple yet efficient rendering approach. Fig. 6(a) shows an example where the right hand of a person intersects the chest of the other person, removing its contribution to the data term (Fig. 6(b)). When this happens for several samples, the shape prior Eq. (9) becomes inaccurate and segmentation errors occur (Fig. 6(c,d)). However, when using front-face culling, only triangles that are not facing the camera are rendered, making the hand visible even inside of the body (Fig. 6(e)). In order to make the shape prior more robust to intersections, we generate for each sample Θ^n and view v two projections $\mathbf{B}_v$ and $\widetilde{\mathbf{B}}_v$, one with correct normals and one with inverted normals. For each pixel i, the label $\mathbf{B}_{v,i}$ is then only changed to $\widetilde{\mathbf{B}}_{v,i}$ if the labels $\mathbf{B}_{v,i}$ and $\widetilde{\mathbf{B}}_{v,i}$ correspond to two different persons. Otherwise, the label remains unchanged. As shown in Fig. 6, this procedure improves the shape prior and the corresponding segmentation.

5. Pose Tracking and Surface Estimation

After image segmentation (Fig. 2(c)), the boundary pixels of the segmented regions are associated to one of the

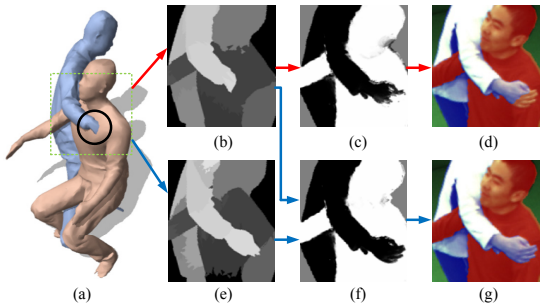


Figure 6. Resolving intersections. (a) Intersection between two persons. The hand is inside the chest. (b) Standard projection. (c) Corresponding data term and (d) Segmentation from (c). (e) Projection with front-face culling. (f) Data term combining both projections. (g) Corresponding segmentation.

persons. Contour pixels of a person that are adjacent to the background are easily assigned to the correct person. Boundary pixels in regions where two persons 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. Fig. 2(d) illustrates the contour labeling.

Once the extracted contours of the persons are labeled, mesh-to-image correspondences can be extracted for each person k. Now skeleton poses Θ_k and surface deformations S_k can be estimated by local optimization as in [16] (Fig. 2(e,f)). Since the contours are labeled, correspondences are only established between points that are associated to the same person. We also use texture correspondences between temporal successive frames obtained by matching SIFT features [20]. Having the labels for the silhouettes, the matching becomes more reliable since only features with the same label are matched.

As shown in [16], the local optimization gets sometimes stuck in a local minimum and global optimization is then performed to correct the estimation error. As the energy term that has been proposed in [16] for global optimization does not handle occlusions, we use a modified measure for the consistency error between the projected surface $\mathbf{B}_v(\Theta_k)$ in model pose Θ_K and the segmented silhouette F_v:

$$E^v(\Theta_k) = \frac{1}{area(F_v^k)} \sum_i g(F_{v,i}^k, \mathbf{B}_{v,i}^k) \qquad (11)$$
$$+ \frac{1}{area(J_v^k)} \sum_i \sum_{j \in J_v^k} f_j(\mathbf{B}_{v,i}^k, F_{v,i}^k),$$

where J_v^k is the set of visible body parts for camera v in the previously estimated frame. After global optimization, the skeleton is skinned and then surface geometry is optimized [16] based on the contour information of individual persons.

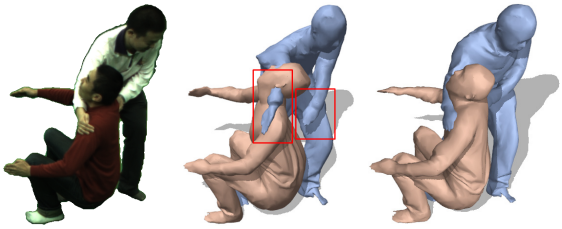


Figure 7. Input image after background subtraction (a), motion tracking without segmentation (b) and with segmentation (c). Without segmentation, features are assigned to the wrong model which leads to significant errors.

6. Results

We recorded 7 test sequences consisting of over 1500 frames. The data was recorded with 12 cameras at a resolution of 1296×972 pixels and at a framerate of 44fps. The initial segmentations were generated by background subtraction. The sequences consist of a wide range of different motions, including dancing, fighting, and jumping, see Figs. 1, 9, and accompanying video. The motions were performed by 5 different persons wearing casual clothing. We also recorded an evaluation sequence where one of the performers was simultaneously tracked by a marker-based motion capture system, yielding ground-truth data for a quantitative evaluation.

Impact of Feature-to-Person Assignment Tracking both persons with the method of [16] without the proposed feature-to-person assignment is prone to errors, see Fig. 7(b). In particular interactions with close physical contact and severe occlusions are problematic due to ambiguous data-to-model associations. Since the errors originate from the underlying energy function for pose estimation, even global optimization strategies cannot resolve the problems caused by wrong associations. Furthermore, relying only on global optimization would be very expensive. In contrast, our segmentation-based approach enables the tracker to correctly and efficiently determine shape and pose, as local optimization succeeds to find the correct poses for most frames, Fig. 7(c). Relying only on color or shape prior for segmentation is not sufficient, and may also lead to tracking errors (see supplementary video).

Segmentation and Tracking Our approach enables us to fully-automatically reconstruct skeletal pose and shape of two people, even if they are as closely interacting as in a martial arts fight, during a leap-frog jump, or while dancing, see Fig. 9 and accompanying video. Despite of severe occlusions, our method successfully captures pose and deforming surface geometry of people in loose apparel, see Fig. 9(e). In some cases, segmentation may lead to small

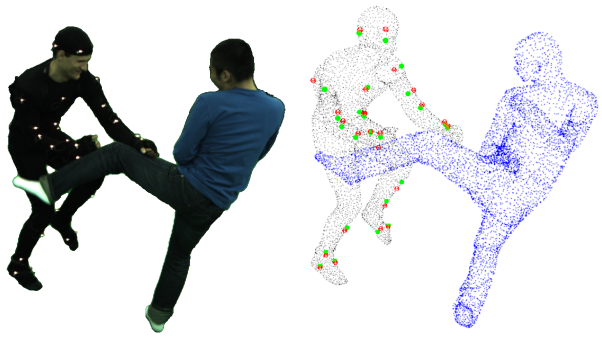
errors in one of the multi-view frames due to very fast motions (Fig. 9(a)) or color similarity (Fig. 9(e)) that cannot be resolved by the shape prior. However, this happens only at very few frames and the motion capture method is robust enough to deal with small inaccuracies in segmentation. The whole system for motion capture takes 3 to 6 minutes (higher motion speed triggers global optimization and costs more time) for a frame that consists of 12 images on a standard PC using unoptimized code.

Quantitative Evaluation In the evaluation sequence, 38 markers were attached to one of the participating subjects whose motions are captured with a commercial PhaseSpace^{TM} marker-based motion capture system. The marker-based system runs synchronously with the multi-view video setup that records the same scene. 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 markers in the scene. The untextured black motion-capture suit and the fast and complex motion make it challenging to track this sequence. After tracking and surface reconstruction, 38 vertices on the mesh surface are associated with the 38 tracked markers (pairing is done in the first reconstructed frame). Fig. 8 shows one of the captured frames with tracked markers and their corresponding mesh vertices overlaid. Already by visual inspection one can clearly see that our reconstructed mesh-vertices are almost identical to the reference markers. The average distance between the markers and their corresponding vertices across all 500 frames of the evaluation sequence is $29.61mm$ with a standard deviation of $25.50mm$. This distance also includes errors introduced by the marker-based system itself, as we are using raw marker positions that have not been post-processed.

Limitations Currently, our approach is designed for two-person tracking, but an extension to the multi-person case is feasible and will be investigated in the future. The segmentation approach can also be modified to handle general scene backgrounds. In certain situations, segmentation errors arise that may lead to pose inaccuracies. For instance, our segmentation and tracking method may fail when the hands from two persons touch as neither appearance nor shape information are sufficient to uniquely identify the performer, see Fig. 9(e). This issue may be resolved at the cost of computation time by explicitly modeling 3D-mesh intersections. Runtime performance can also be improved by using lower resolution meshes in shape prior calculation.

7. Conclusion

In this paper, we proposed a segmentation-based markerless motion capture method that enables us to track skeleton motion and detailed surface geometry of interacting persons. The segmentation approach is based on a new probabilistic shape and appearance model that enables reliable image segmentation of the two persons even under challenging occlusions. This robust multi-view segmentation enables us to reliably and accurately capture shape and pose of each actor. To our knowledge, this is the first method to fully-automatically track people in close interaction.



(a) (b)

Figure 8. Illustration of tracking accuracy. (a) Input image after background subtraction. (b) Position comparison of marker points (green points) and the corresponding 3D vertices (red points), with surface point cloud overlay.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

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

[3] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 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] Y. Boykov and M. Jolly. Iterative graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, pages 105–112, 2001.

[6] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *CVPR*, pages 648–655, 1998.

[7] M. Bray, P. Kohli, and P. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *ECCV*, pages 642–655, 2006.

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

This work has been developed within the Max-Planck-Center for Visual Computing and Communication collaboration, and has been co-financed by the Intel Visual Computing Institute.

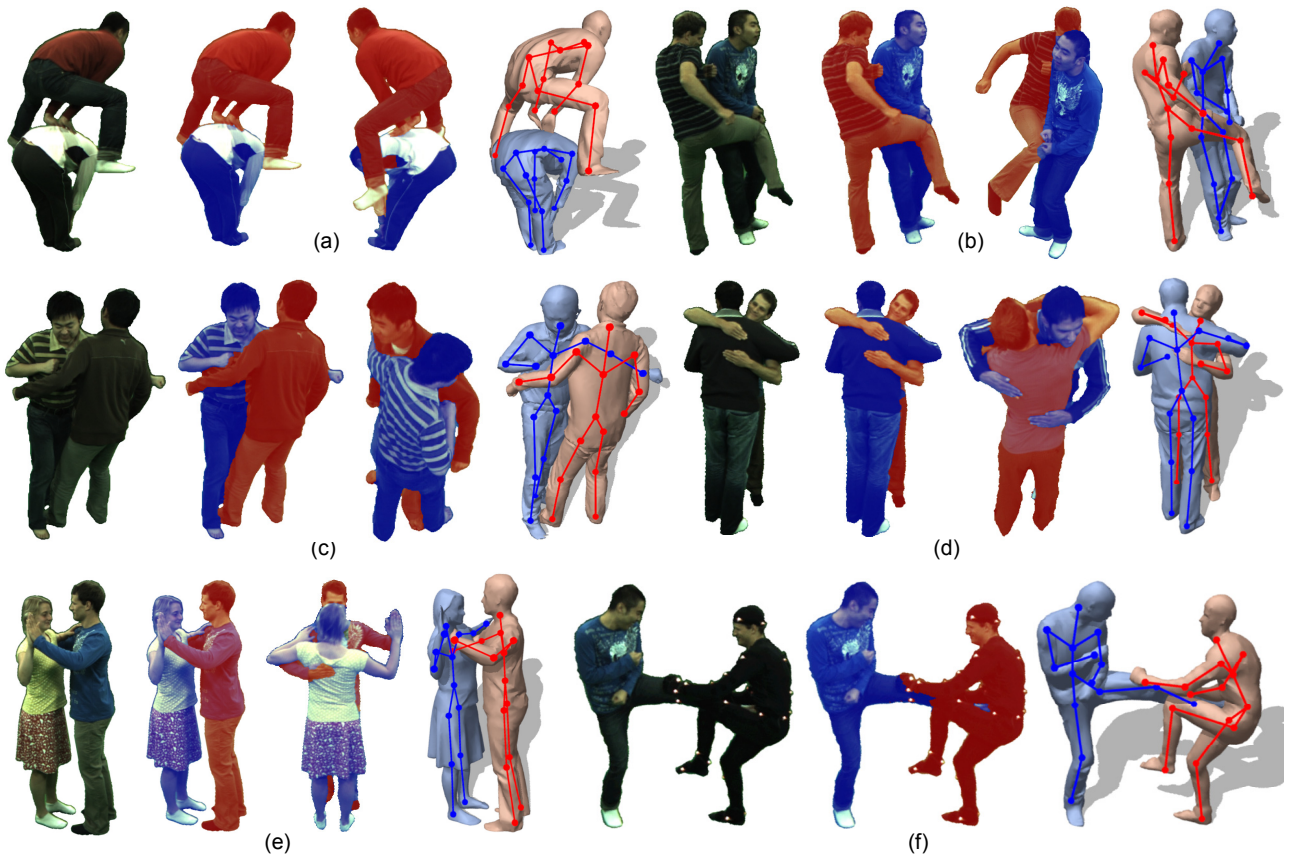


Figure 9. Input images after background subtraction, their corresponding segmentation results, and surface reconstruction and tracked skeleton for several sequences.

[9] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region- and motion-based 3d tracking of rigid and articulated objects. *TPAMI*, 32(3):402–415, 2010.

[10] C. Cagniart, E. Boyer, and S. Ilic. Free-from mesh tracking: a patch-based approach. In *CVPR*, 2010.

[11] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, 2010.

[12] 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, 2008.

[13] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[14] H. Egashira, A. Shimada, D. Arita, and R. Taniguchi. Vision-based motion capture of interacting multiple people. In *ICIAP*, pages 451–460, 2009.

[15] J. Gall, B. Rosenhahn, and H.-P. Seidel. Drift-free tracking of rigid and articulated objects. In *CVPR*, 2008.

[16] J. Gall, C. Stoll, E. Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753, 2009.

[17] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. van Gool. Articulated multibody tracking under egomotion. In *ECCV*, 2008.

[18] J.-Y. Guillemaut, J. Kilner, and A. Hilton. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complext dynamic scenes. In *ICCV*, pages 809–816, 2009.

[19] J. Hammersley and D. Handscomb. *Monte Carlo methods*. Methuen, London, 1964.

[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

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

[22] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87:4–27, 2010.

[23] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *ICCV*, pages 915– 922, 2003.

[24] K. Varanasi, A. Zaharescu, E. Boyer, and R. Horaud. Temporal surface tracking using mesh evolution. In *ECCV*, pages 30–43, 2008.

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