HybridFusion: Real-Time Performance Capture Using a Single Depth Sensor and Sparse IMUs

Zerong Zheng[1], Tao Yu[1,2], Hao Li[3], Kaiwen Guo[4], Qionghai Dai[1], Lu Fang[5], and Yebin Liu[1(✉)]

[1] Tsinghua University, Beijing, China
liuyebin@mail.tsinghua.edu.cn
[2] Beihang University, Beijing, China
[3] University of Southern California, Los Angeles, CA, USA
[4] Google Inc., Mountain View, CA, USA
[5] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China

Abstract. We propose a light-weight yet highly robust method for real-time human performance capture based on a single depth camera and sparse inertial measurement units (IMUs). Our method combines non-rigid surface tracking and volumetric fusion to simultaneously reconstruct challenging motions, detailed geometries and the inner human body of a clothed subject. The proposed hybrid motion tracking algorithm and efficient per-frame sensor calibration technique enable non-rigid surface reconstruction for fast motions and challenging poses with severe occlusions. Significant fusion artifacts are reduced using a new confidence measurement for our adaptive TSDF-based fusion. The above contributions are mutually beneficial in our reconstruction system, which enable practical human performance capture that is real-time, robust, low-cost and easy to deploy. Experiments show that extremely challenging performances and loop closure problems can be handled successfully.

Keywords: Performance capture · Real-time · Single-view · IMU

1 Introduction

The 3D acquisition of human performances has been a challenging topic for decades due to the shape and deformation complexity of dynamic surfaces, especially for clothed subjects. To ensure high-fidelity digitalization, sophisticated multi-camera array systems [4,5,7,8,14,17,24,29,43] are preferred for professional productions. TotalCapture [13], the state-of-the-art human performance capture system, uses more than 500 cameras to minimize occlusions during

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01240-3_24) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 389–406, 2018.
https://doi.org/10.1007/978-3-030-01240-3_24

human-object interactions. Not only are these systems difficult to deploy and costly, they also come with a significant amount of synchronization, calibration, and data processing effort.

On the other end of the spectrum, the recent trend of using a single depth camera for dynamic scene reconstruction [10,12,25,31] provides a very convenient and real-time approach for performance capture combined with online non-rigid volumetric depth fusion. However, such monocular systems are limited to slow and controlled motions. While improvement has been demonstrated lately in systems like BodyFusion [44], DoubleFusion [45] and SobolevFusion [32], it is still impossible to reconstruct occluded limb motions (Fig. 1(b)) and ensure loop closure during online reconstruction. For practical deployment, such as gaming, where fast motion is expected and possibly interactions between multiple users, it is necessary to ensure continuously reliable performance capture.

Fig. 1. The state-of-the-art methods easily get failed under severe occlusions. (a, d): color references captured from Kinect (up) and a 3rd person view (down). (b, e) and (c, f): results of DoubleFusion and our method rendered in the 3rd person view. (Color figure online)

We propose HybridFusion, a real-time dynamic surface reconstruction system that achieves high-quality reconstruction of extremely challenging performances using hybrid sensors, i.e., a single depth camera and several inertial measurement units (IMUs) sparsely located on the body. Intuitively, for the cases of extremely fast or highly occluded or self-rotating limb motions, which cannot be handled by the optical sensors alone, the IMUs can provide high frame rate orientation information that help infer better human motion estimations. Moreover, they are low cost and easy to wear. For other cases, a single depth camera owns sufficient capacity to achieve robust reconstruction, so as to maintain the light-weight and convenient property of the whole system compared to multi-camera ones.

Combining IMUs with depth sensors within a non-rigid depth fusion framework is non-trivial. First, we need to minimize the effort and experience required for mounting and calibrating each IMU. We, therefore, propose a per-frame sensor calibration algorithm integrated into the tracking procedure to get accurate IMU calibration without any additional extra steps. We also extend the non-rigid tracking optimization to a hybrid tracking optimization by adding the IMU constraints. Moreover, previous tracking&fusion methods [25,45] may generate

seriously deteriorated reconstruction results for challenging motions and occlusions due to the wrongly fused geometry, which will further affect the tracking performance, and vice versa. We thus propose a simple yet effective scheme that jointly models the influence of body-camera distance, fast motions and occlusions in one metric, which guides the TSDF (Truncated Signed Distance Field) fusion to achieve robust and precise results even under challenging motions (see Fig. 1). Using such a light-weight hybrid setup, we believe HybridFusion presents the right sweet spot for practical performance capture system as it is real-time, robust and easy to deploy. Commodity users can capture high-quality body performances and 3D content for gaming, VR/AR applications at home.

Note that IMUs or even hybrid sensors have been adopted previously to improve the skeleton-based motion tracking [11,20,22,28]. Comparing with these state-of-the-art hybrid motion capture systems like [11], the superiority of HybridFusion is twofold: for one, our system can reconstruct the detailed outer surface of the subject and estimate the inner body shape simultaneously, while [11] needs a pre-defined model as input; for another, our system can track the non-rigid motion of the outer surface, while [11] outputs skeleton poses merely. By further examining the differences in the skeleton tracking solely, our system still demonstrates substantially higher accuracy. In [11] IMU readings are only used to query similar poses in a database, yet we integrate the inertial measurements into a hybrid tracking energy. The detailed model and non-rigid registration further improve the accuracy of pose estimation, since a detailed geometry model with an embedding deformation node graph better describes the motion of the user than a body model driven by a kinetic chain.

The main contributions of HybridFusion can be summarized as follows.

- **Hybrid motion tracking.** We propose a hybrid non-rigid tracking algorithm for accurate skeleton motion and non-rigid surface motion tracking in real-time. We introduce an IMU term that significantly improves the tracking performance even under severe occlusion.
- **Sensor calibration.** We introduce a per-frame sensor calibration method to optimize the relationship between each IMU and its attached body part during the capture process. Unlike other IMU-based methods [2,20,28], this method removes the requirement of explicit calibration and provides accurate calibration results along the sequence.
- **Adaptive Geometry fusion.** To address the problem that previous TSDF fusion methods are vulnerable in some challenging cases (far body-camera distance, fast motions, occlusions, etc.), we propose an adaptive TSDF fusion method that considers all the factors above in one tracking confidence measurement to get more robust and detailed TSDF fusion results.

2 Related Work

The related work can be classified into two categories: IMU-based human performance capture and volumetric dynamic reconstruction. We refer readers to overview of prior works including pre-scanned template based dynamic

reconstruction [9,15,34,40,42,46], shape template based dynamic reconstruction [1,3,18,29,30] and free-form dynamic reconstruction [16,23,26,35,37] in [45].

IMU-Based Human Performance Capture. A line of research on combining vision and IMUs [11,20–22,27,28] or even using IMUs alone [41] targets at high quality human performance capture. Among all of those works, Malleson *et al.* [20] combined multi-view color inputs, sparse IMUs and SMPL model [18] in a real-time full-body skeleton motion capture system. Pons-moll *et al.* [28] used multi-view color inputs, sparse IMUs and pre-scanned user templates to perform full-body motion capture offline. The system is improved by using 6 IMUs alone [41] to reconstruct natural human skeleton motion using global optimization method, but still offline. Vlasic *et al.* [39] used the output of the inertial sensors for extended kalman filter to perform human skeleton motion capture. Tautges *et al.* [36] and Ronit *et al.* [33] both utilized sparse accelerometer data and data-driven methods to retrieve correct poses in the database. Helten *et al.* [11] used the most similar setup to our method (single-view depth information, sparse IMUs and parametric human body model). They combined generative tracker and discriminative tracker that retrieving closest poses in a dataset and perform real-time human motion tracking. However, the parametric body model cannot describe detailed surfaces of clothing.

Non-rigid Surface Integration. Starting from DynamicFusion [25], non-rigid surface integration methods get more and more popular [10,12,31] because of the single-view, real-time and template-free properties. It also inspires a branch of multi-view volumetric dynamic reconstruction methods [6,7] that achieved high quality reconstruction results. The basic idea of non-rigid surface integration is to perform non-rigid surface tracking and TSDF surface fusion iteratively, such that the surface information gets more and more complete along the scene motions when unseen surface parts get observed and tracked. To improve the reconstruction performance of DynamicFusion on human body motions, BodyFusion [44] integrated articulated human motion prior (skeleton kinematic chain structure) and constraint the non-rigid deformation and skeleton motion to be similar. DoubleFusion [45] leveraged parametric body model (SMPL [18]) in non-rigid surface integration to improve the tracking, loop closure and fusion performance, and achieved the state-of-the-art single-view human performance capture results. However, all of these methods are still incompetent to handle fast and challenging motions, especially for occluded motions.

3 Overview

Initialization. We adopt 8 IMUs that sparsely located on the upper and lower limbs of the performer as shown in Fig. 2. It is worth mentioning that unlike [20, 41] which require IMUs to be specific to model vertices, the IMUs in our system are attached to bones as we merely trust and use the orientation measurements. Such strategy greatly relaxes users' efforts to wear the sensors since they only need to ensure the IMUs are attached to the correct bones and roughly aligned

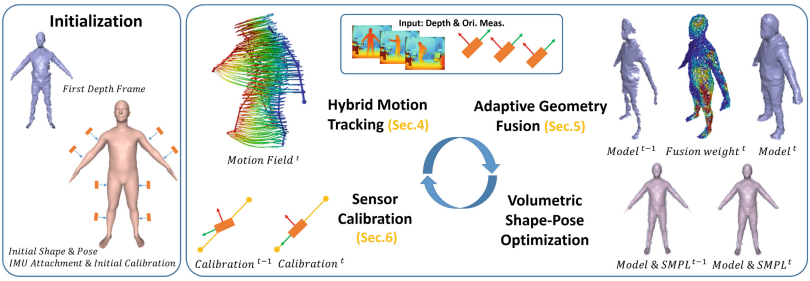

Fig. 2. Illustration of HybridFusion pipeline.

with their length directions. Here the number of IMUs is determined by the balance between performance and convenience, as further elaborated in Sect. 7.3.

The performer is required to start with a rough A-pose. After getting the first depth frame, we use it to initialize the TSDF volume by projecting the depth pixels into the volume, and then estimate the initial shape parameters β_0 and pose θ_0 using volumetric shape-pose optimization [45]. We construct a "double node graph" consisting of predefined on-body node graph and free-form sampled far-body node graph. We use θ_0 and the initial IMU readings to initialize sensor calibration. The triangle mesh is extracted from the TSDF volume with Marching Cube algorithm [19].

Main Pipeline. The lack of ground truth transformation between IMUs and their attached bones leads to unstable tracking performance in our hybrid motion tracking step. Therefore, we keep optimizing the sensor calibration frame by frame, and the calibration gets more and more accurate thanks to the increasing number of successfully tracked frames with different skeleton poses. Following [45], we also optimize the inner body shape and the canonical pose. In summary, our pipeline performs hybrid motion tracking, adaptive geometry fusion, volumetric shape-pose optimization and sensor calibration sequentially, as shown in Fig. 2. Below is a brief introduction of the main components of our pipeline.

- **Hybrid Motion Tracking.** Given the current depth map and the IMU measurements, we propose to jointly track the skeletal motion and the surface non-rigid deformation through a new hybrid motion tracking algorithm. We construct a new energy term to constrain the orientations of the skeleton bones using the orientation measurements of their corresponding IMUs.
- **Adaptive Geometry Fusion.** To improve the robustness of the fusion step, we propose an adaptive fusion method that utilizes tracking confidence to adjust the weight of TSDF fusion adaptively. The tracking confidence can be estimated according to the normal equations in the current procedure of hybrid motion tracking.
- **Volumetric Shape-Pose Optimization.** We perform volumetric shape-pose optimization after adaptive geometry fusion. Based on the updated

TSDF volume, we optimize the inner body shape and canonical pose to obtain better canonical body fitting and skeleton embedding.
– **Sensor Calibration.** Given the motion tracking results and IMU readings at current frame, we optimize the sensor calibration to acquire more accurate estimations of the transformations between IMUs and their corresponding bones, as well as more accurate transformation estimation between the inertial coordinate and the camera coordinate.

4 Hybrid Motion Tracking

Since our pipeline focuses on performance capture of human, we adopt a double-layer surface representation for motion tracking, which has been proved to be efficient and robust in [45]. Similar to [9,44,45], our motion tracking is under the assumption that human motion largely follows articulated structures. Therefore, we use two kinds of motion parameterizations, skeleton motions and non-rigid node deformation. Combining IMU orientation informations, we construct a energy function for hybrid motion tracking in order to solve the two motion components in a joint optimization scheme. Given the depth map $\mathfrak{D}_t$ and inertial measurements $\mathfrak{M}_t$ of current frame t, the energy function is:

$$E_{\text{mot}} = \lambda_{\text{IMU}} E_{\text{IMU}} + \lambda_{\text{depth}} E_{\text{depth}} + \lambda_{\text{bind}} E_{\text{bind}} + \lambda_{\text{reg}} E_{\text{reg}} + \lambda_{\text{pri}} E_{\text{pri}}, \quad (1)$$

where E_{IMU}, E_{depth}, E_{bind}, E_{reg} and E_{prior} represent IMU, depth, binding, regularization and pose prior term respectively. E_{IMU} and E_{depth} are data terms that constrain the results to be consistant with IMU and depth input, E_{bind} regularizes the surface non-rigid deformation with articulated skeleton motion, E_{reg} constrains the locally as-rigid-as-possible property of the node graph and E_{prior} is used to penalize unnatural human poses. To simplify the notation, we claim that all variables in this section take their values at the current frame t, and drop their subscripts of frame index.

IMU Term. To bridge the sensors' measurements and hybrid motion tracking pipeline, we select $N = 8$ binding bones on the SMPL model (Fig. 2 Initialization) for the N inertial sensors, and these bones are denoted by $b_i^{IMU} (i = 1, \ldots, N)$. The IMU term penalizes the orientation difference between IMU readings and the estimated orientations of their attached binding bones:

$$E_{\text{IMU}} = \sum_{i \in \mathcal{S}} \left\| \mathbf{R}_{I2C} \widetilde{\mathbf{R}}_i \mathbf{R}_{S2B,i}^{-1} - \mathbf{R}\left(\mathbf{b}_i^{IMU}\right) \right\|_F^2 \quad (2)$$

where $\mathcal{S}$ is the index set of IMUs; $\widetilde{\mathbf{R}}_i$ is the orientation measurement of i-th sensor in the inertial coordinate system. $\mathbf{R}_{I2C}$ is the rotation offset between the inertial coordinate and the camera coordinate system, while $\mathbf{R}_{S2B,i}$ is the offset between the i-th IMU and its corresponding bone; more details are elaborated in Sect. 5. $\mathbf{R}(\mathbf{b}_i^{IMU})$ is the rotational part of the skeleton skinning matrix $\mathbf{G}(\mathbf{b}_i^{IMU})$, which is defined as:

$$\mathbf{G}(\mathbf{b}_i^{IMU}) = \mathbf{G}_j = \prod_{k \in \mathcal{K}_j} \exp\left(\theta_k \hat{\xi}_k\right), \quad (3)$$

where j is the index of $\mathbf{b}_i^{IMU}$ in the skeleton structure; $\mathbf{G}_j$ is the cascaded rigid transformation of jth bone; $\mathcal{K}_j$ represents parent bones indices of jth bone along the backward kinematic chain; $\exp(\theta_k \hat{\xi}_k)$ is the exponential map of the twist associated with kth bone.

Note that $\mathbf{R}_{I2C}$ and $\mathbf{R}_{B2S}$ are crucial parameters determining the effectiveness of the IMU term, and therefore they are continually optimized in our pipeline even though we can obtain sufficiently accurate estimations through initial calculation. We provide more details about calculating and optimizing $\mathbf{R}_{I2C}$ and $\mathbf{R}_{S2B}$ in Sect. 5.

The other energy terms in Eq. 1 are detailed in [44,45], as well as the efficient GPU solver for motion tracking. Please refer to these two papers for more details.

5 Sensor Calibration

On one hand, an inertial sensor gives orientation measurements in the inertial coordinate system, which is typically defined by the gravity field and geomagnetic field. On the other hand, our performance capture system runs in the camera coordinate system, which is independent of the inertial coordinate. The relationship between these two coordinates can be described as a constant mapping denoted by $\mathbf{R}_{I2C}$. Based on the mapping, we can transform all IMU outputs from inertial coordinate to the camera coordinate system, as formulated in Eq. 2. As illustrated in Fig. 3, several coordinate systems are involved in order to estimate the mapping: (1) the i-th IMU sensor coordinate system $C_{\mathbf{S}_i}$, which is aligned with the ith sensor itself, and changes when the sensor moves, (2) the inertial coordinate system $C_{\mathbf{I}}$, which remains static all the time, (3) the i-th bone coordinate system $C_{\mathbf{B}_i}$, which is aligned with the bone associated with the ith IMU sensor, and changes when the subject acts or moves, (4) the camera coordinate system $C_{\mathbf{C}}$, which also remains static. Accordingly, R_{S2B} is the transformation from $C_{\mathbf{S}}$ to $C_{\mathbf{B}}$, R_{I2C} is from $C_{\mathbf{I}}$ to $C_{\mathbf{S}}$, and their inverse transformations are denoted as R_{B2S} and R_{C2I}.

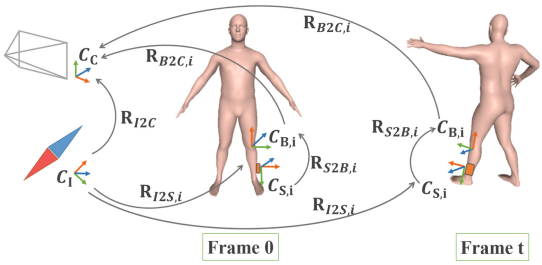

Fig. 3. Illustration of different coordinates and their relationship.

5.1 Initial Sensor Calibration

We calculate an approximation of $\mathbf{R}_{I2C}$ during the initialization of our pipeline. After fitting the SMPL model to the depth image, the mapping $\mathbf{R}_{B2C,i}$: $C_{\mathbf{B}_i} \rightarrow C_{\mathbf{C}}$ is available according to $\mathbf{R}_{B2C,i} = \mathbf{R}_{t_0}\left(\mathbf{b}_i^{IMU}\right)$, where the subscript t_0 is the index of the first frame. Besides, we can also obtain the mapping from $C_{\mathbf{I}}$ to $C_{\mathbf{S}_i}$ by assigning the inverse matrix of the sensor's reading at the first frame: $\mathbf{R}_{I2S,i} = \widetilde{\mathbf{R}}_{i,t_0}^{-1}$. To transform $C_{\mathbf{I}}$ into $C_{\mathbf{C}}$ through the path $C_{\mathbf{I}} \rightarrow C_{\mathbf{S}_i} \rightarrow C_{\mathbf{B}_i} \rightarrow C_{\mathbf{C}}$, we need to know the rotation offset between the IMUs and their corresponding bone coordinate systems $\mathbf{R}_{S2B,i}$: $C_{\mathbf{S}_i} \rightarrow C_{\mathbf{B}_i}$. We assume that they are constant as the sensors are tightly attached to the limbs and we then predefine them according to the placement of the sensors. Thus, we can compute $\mathbf{R}_{I2C}$ by

$$\begin{aligned}\mathbf{R}_{I2C} &= \underset{i=1,\ldots,N}{\text{SLERP}}\left\{\left(\mathbf{R}_{I2C,i}\ ,\ w_i\right)\right\} = \underset{i=1,\ldots,N}{\text{SLERP}}\left\{\left(\mathbf{R}_{B2C,i}\mathbf{R}_{S2B,i}\mathbf{R}_{I2S,i}\ ,\ w_i\right)\right\} \\ &= \underset{i=1,\ldots,N}{\text{SLERP}}\left\{\left(\mathbf{R}_{t_0}\left(\mathbf{b}_i^{IMU}\right)\mathbf{R}_{S2B,i}\widetilde{\mathbf{R}}_{i,t_0}^{-1}\ ,\ w_i\right)\right\},\end{aligned} \tag{4}$$

where SLERP $\{\cdot\}$ is the operator of spherical linear interpolation, and w_i is the interpolation weight, which is set to $1/N$ in our experiment.

5.2 Per-Frame Calibration Optimization

Even though the influence of measurement noises tends to be diminished by averaging $\mathbf{R}_{I2C,i}$ (Sect. 5.1), the solution of the initial sensor calibration is still prone to errors due to the sparse IMU setup and the rough assignments of $\mathbf{R}_{S2B,i}$. Therefore, we propose an efficient method to continuously optimize the sensor calibration. As formulated in Sect. 4, the orientation measurements and motion estimation are related by $\mathbf{R}_{I2C}$ and $\mathbf{R}_{B2S,i}$:

$$\mathbf{R}_{I2C}\widetilde{\mathbf{R}}_i = \mathbf{R}\left(\mathbf{b}_i^{IMU}\right)\mathbf{R}_{B2S,i}^{-1}, \tag{5}$$

thus we can compute the accumulated rotations from t_0 to t as:

$$\mathbf{R}_{I2C}\widetilde{\mathbf{R}}_{i,t}\widetilde{\mathbf{R}}_{i,t_0}^{-1}\mathbf{R}_{I2C}^{-1} = \mathbf{R}_t\left(\mathbf{b}_i^{IMU}\right)\mathbf{R}_{t_0}^{-1}\left(\mathbf{b}_i^{IMU}\right). \tag{6}$$

Given the motion tracking results, we estimate the optimal rotation offset of frame t according to

$$\hat{\mathbf{R}}_{I2C} = \underset{\mathbf{R}_{I2C}}{\arg\min}\sum_{i\in\mathcal{S}}\left\|\mathbf{R}_{I2C}\widetilde{\mathbf{R}}_{i,t}\widetilde{\mathbf{R}}_{i,t_0}^{-1}\mathbf{R}_{I2C}^{-1} - \mathbf{R}_t\left(\mathbf{b}_i^{IMU}\right)\mathbf{R}_{t_0}^{-1}\left(\mathbf{b}_i^{IMU}\right)\right\|_F^2, \tag{7}$$

and then update $\mathbf{R}_{I2C}$ by blending the solution with the original value:

$$\mathbf{R}_{I2C} \leftarrow \text{SLERP}\left\{\left(\mathbf{R}_{I2C}, w\right); \left(\hat{\mathbf{R}}_{I2C}, \omega\right)\right\} \tag{8}$$

where w, ω are both interpolation weights. We set $w = 1 - \frac{1}{t}, \omega = \frac{1}{t}$ to make sure the final solution coverage to a stable global optimum. We optimize $\mathbf{R}_{S2B,i}$ in similar ways.

6 Adaptive Geometry Fusion

Similar to prior works [7,10,12,25,45], we integrate depth maps into a reference volume. To deal with the ambiguity caused by voxel collision, we follow [7,10,45] to detect collided voxels by voting the TSDF value at live frame and avoid integrating depth information into these voxels. Besides voxel collision, the surface fusion still suffers from inaccurate motion tracking, which is a factor that previous fusion methods do not consider. Inspired by previous works addressing the uncertainty of parameter estimation [38,47], we propose to fuse geometry adaptively according to the tracking confidence that measures the performance of hybrid motion tracking. Specifically, we denote x_t as the motion parameters being solved and assume it approximately follows a normal distribution:

$$p(x_t|\mathfrak{D}_t, \mathfrak{M}_t) \simeq \mathcal{N}(\mu_t, \Sigma_t), \tag{9}$$

where μ_t is the solution of motion tracking and the covariance Σ_t measures the tracking uncertainty. By assuming $p(x_t|\mathfrak{D}_t, \mathfrak{M}_t) \propto \exp(-E_{\text{mot}})$, we can approximate the covariance as

$$\Sigma_t = \sigma^2 \left(\mathbf{J}^T \mathbf{J}\right)^{-1} \tag{10}$$

where $\mathbf{J}$ is the Jacobian of E_{mot}.

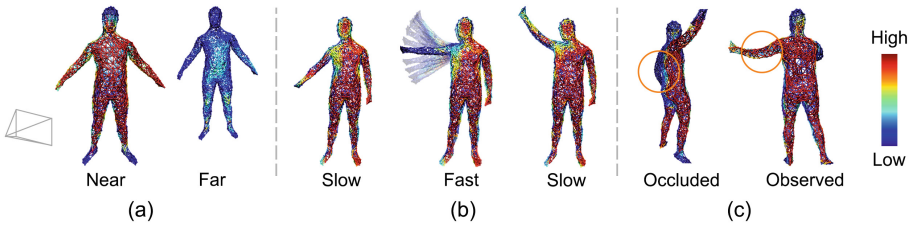

Fig. 4. Visualization of the estimated per-node tracking confidence in 3 scenarios: large body-camera distance (a), fast motions (b) and occlusions (c).

We regard the diagonal of Σ_t^{-1} as the confidence vector of the solution μ_t, which contains the confidence of both skeleton tracking and non-rigid tracking parameters calculated by our hybrid motion tracking algorithm. Since the TSDF fusion step only needs node graph to perform non-rigid deformation [25], we merge the two types of motion tracking confidence together to get a more accurate estimation of hybrid tracking confidence for each node. Therefore, the tracking confidence $C_{track}(\mathbf{x}_k)$ corresponding to a node $\mathbf{x}_k$ can be computed as

$$C_{track}(\mathbf{x}_k) = (1-\lambda) \min\left(\frac{\text{diag}(\bar{\Sigma}_t^{-1})_{\mathbf{x}_k}}{\eta_{\mathbf{x}_k}}, 1\right) + \lambda \sum_{j \in \mathcal{B}} w_{j,x_k} \min\left(\frac{\text{diag}(\bar{\Sigma}_t^{-1})_{\mathbf{b}_j}}{\eta_{\mathbf{b}_j}}, 1\right) \tag{11}$$

where $\mathcal{B}$ is the index set of bones; $\text{diag}(\bar{\Sigma}_t^{-1})_{\mathbf{x}_k}$ and $\text{diag}(\bar{\Sigma}_t^{-1})_{\mathbf{b}_j}$ are the averaged covariance values of all ICP iterations corresponding to the kth node and jth bone respectively. w_{j,x_k} is the skinning weight associated with $\mathbf{x}_k$.

To better illustrate the tracking confidence, we classify the performance capture scenarios that will adversely impact the tracking performance into 3 categories (far body-camera distance, fast motions and occlusions) and visualized the estimated tracking confidence of each node in these scenarios in Fig. 4. Since the quality of depth input is inversely proportional to body-camera distance and the low quality depth will significantly deteriorate the tracking and fusion performance, the tracking confidence of all nodes declines when the body is far from the camera (Fig. 4(a)); Moreover, the nodes under fast motions also have low tracking confidence (Fig. 4(b)), as the tracking performance for fast motions is usually worse than slow motions due to the blurred depth input and lack of correspondences; Last, for single-view capture system, occlusions will lead to lack of observations and worse tracking performance of corresponding body parts. Thus, the tracking confidence of occluded nodes decreases as in Fig. 4(c).

After calculating the tracking confidence, we perform adaptive geometry fusion as follows. For a voxel v, $\mathbf{D}(v)$ denotes the TSDF value of the voxel, $\mathbf{W}(v)$ denotes its accumulated fusion weight, $\mathbf{d}(v)$ is the projective signed distance function (PSDF) value, and $\omega(v)$ is the fusion weight of v at current frame:

$$\omega'(v) = \sum_{\mathbf{x}_k \in \mathcal{N}(v)} C_{track}\left(\mathbf{x}_k\right), \quad \omega(v) = \begin{cases} 0 & \omega'(v) < \tau, \\ \omega'(v) & \text{otherwise.} \end{cases} \tag{12}$$

Finally, the voxel is updated by

$$\mathbf{D}(v) \leftarrow \frac{\mathbf{D}(v)\mathbf{W}(v) + \mathbf{d}(v)\omega(v)}{\mathbf{W}(v) + \omega(v)}, \quad \mathbf{W}(v) \leftarrow \mathbf{W}(v) + \omega(v) \tag{13}$$

where $\mathcal{N}(v)$ is the collection of the KNN deformation nodes of voxel v, and τ is a threshold controlling the minimum integration weight.

7 Experiments

We evaluate the performance of our proposed method in this section. In Sect. 7.1 we present details on the setup of our system and report the main parameters of our pipeline. Then we compare our system with the state-of-the-art method both qualitatively and quantitatively in Sect. 7.2. We also provide evaluations of our main contributions in Sect. 7.3.

Figure 5 demonstrates the reconstructed dynamic geometries and the inner body shapes on several motion sequences, including sports, dancing and so on. From the results we can see that our system is able to reconstruct various kinds of challenging motions and inner body shapes using a single-view setup.

Fig. 5. Example results reconstructed by our system. In each grid, the left image is the color reference; the middle one is the fused surface geometry; and the right one is the inner body shape estimated by our system. (Color figure online)

7.1 System Setup

For the hard-ware setup, we use Kinect One and Noitom Legacy suite as the depth sensor and inertial sensors respectively. Our system runs in real-time (33 ms per frame) on a NVIDIA TITAN X GPU and an Intel Core i7-6700K CPU. The majority of the running time is spent on the joint motion tracking (23 ms) and the adaptive geometric fusion (6 ms). The sensor calibration optimization takes 1 ms while the shape-pose optimization takes 3 ms.

The weights of energy terms serve to balance the impact of different tracking cues, where the weight of IMU term is set to 5.0, while the other energy weights are identical to [16]. More specifically, the strategy of assigning λ_{IMU} is to ensure that (1) the IMU term can produce rough pose estimations, when there is a lack of correspondences (fast motion and/or occlusion), and (2) the IMU term does not affect the tracking adversely, when enough correspondences are available. Note that $\lambda_{depth} = 1.0$ and $\lambda_{bind} = 1.0$ initially, and the binding term will be gradually relaxed so as to capture the detailed non-rigd motion of the surface. The weights of the regularization term and prior term are fixed to 5.0 and 0.01 respectively, avoiding undesirable results.

7.2 Comparison

We compare against the state-of-the-art method, DoubleFusion [45] on 4 sequences, as shown in Fig. 6. The tracking performance of our system clearly

outperforms DoubleFusion especially under severe occlusions. To make quantitative comparison, we capture several sequences using the Vicon and our system simultaneously. Both systems are synchronized by flashing the infrared LED. We calibrate these two systems spatially by manually selecting the corresponding point pairs and calculate their transformation. After that, we transform the marker positions from the Vicon coordinate into the camera coordinate at the first frame, followed by tracking their motions using the motion field and comparing the per-frame positions with the Vicon-detected ground-truth. We do the same tests on DoubleFusion. Figure 7 presents the curves of per-frame maximum error of DoubleFusion and our method on one sequence. We also list the average errors over the entire sequence in Table 1. From the numerical results we can see that our system achieve the higher tracking accuracy than DoubleFusion.

Table 1. Average numerical errors on the entire sequence.

Method	DoubleFusion	HybridFusion
Avg. of Max. Err. (m)	0.0854	0.0655

We also compare our skeleton tracking performance against the state-of-the-art hybrid tracker, [11], using its published dataset. As depicted in Table 2, our system maintains more accurate and stable performance for skeleton tracking, inducing much smaller tracking errors than [11].

Table 2. Average joint tracking error and standard deviation in millimeters (compared with [11]).

Sequence	D1	D2	D3	D4	D5	D6
Helten *et al.* [11]	35.7(24.9)	47.4(31.4)	44.4(33.8)	34.7(25.4)	59.1(45.3)	56.2(41.6)
Ours	20.9(15.2)	27.6(19.6)	27.0(17.6)	15.5(15.6)	43.5(33.6)	40.9(27.5)

7.3 Evaluation

Sensor Calibration. In Fig. 8, we evaluate the proposed per-frame sensor calibration on a simple sequence. Figure 8(c) is the surface reconstruction results only using initial calibration results as described in Sect. 5.1, without the per-frame calibration optimization step (Sect. 5.2). We can see that the joint motion tracking performance suffers from the inaccuracy of the initial calibration results. Moreover, the erroneous motion tracking performance will lead to erroneous surface fusion results (ghost hands and legs). With the per-frame calibration optimization algorithm, our system can generate accurate motion tracking and surface fusion results as shown in Fig. 8(d).

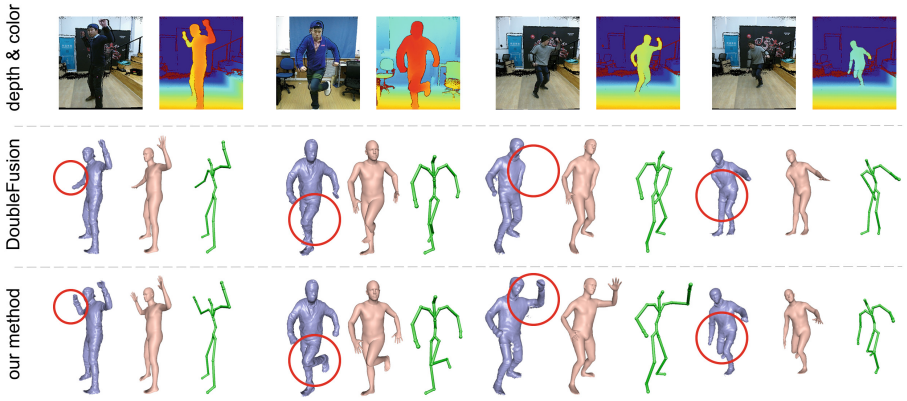

Fig. 6. Qualitative comparison against DoubleFusion. 1st row: Color and depth image as reference. 2nd and 3rd rows: The results reconstructed by DoubleFusion and our system respectively. (Color figure online)

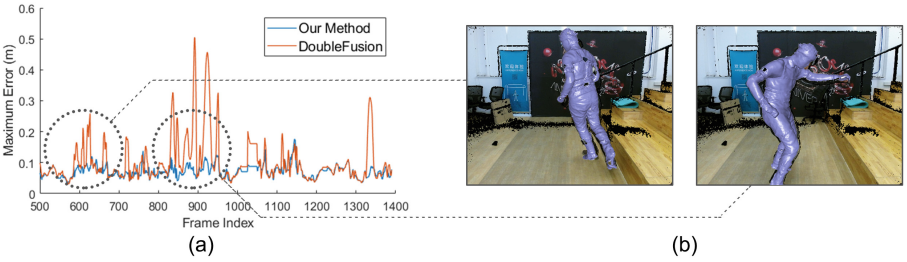

Fig. 7. Quantitative comparison on tracking accuracy against DoubleFusion. (a): The curves of maximum position error. (b): The results of our system on two time instances.

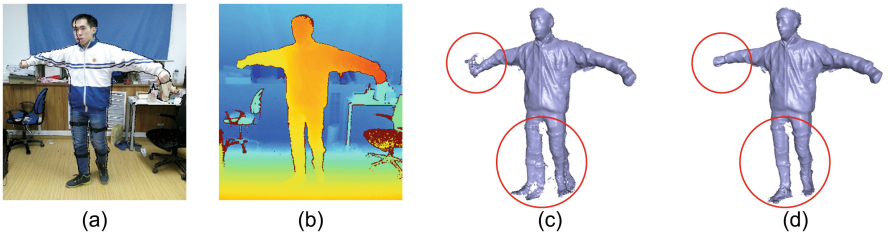

Fig. 8. Evaluation of per-frame sensor calibration optimization. (a), (b): Color and depth images as reference. (c): The reconstruction results without calibration optimization. (d): The reconstruction results with calibration optimization. (Color figure online)

Adaptive Geometry Fusion. We also evaluate the effectiveness of the adaptive geometric fusion method. We captured several sequences in three challenging scenarios for detailed surface fusion, which include far body-camera distance, body-part occlusion and fast motion. We then compare our adaptive geometry fusion method against previous fusion method used in [10,25,44,45]. In Fig. 9, the results of the previous fusion method are presented on the left side of each sub-figure, while the reconstruction results with adaptive fusion are shown on the right. As shown in Fig. 4, the fusion weights in our system can be automatically adjusted (set to a very small value or skip the fusion step) in all the situations, resulting in more plausible and detailed surface fusion results.

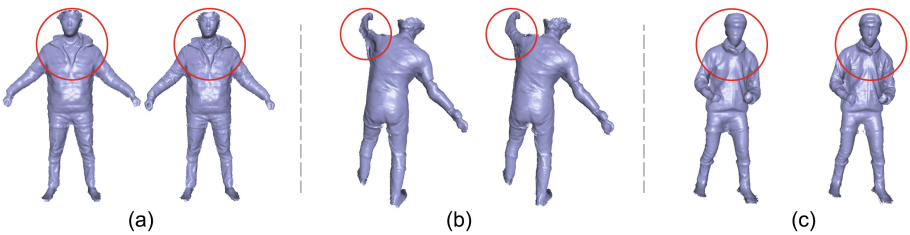

Fig. 9. Evaluation of adaptive fusion under far body-camera distance (a), occlusions (b) and fast motions (c). In each sub-figure, the left mesh is fused by previous fusion method and the right one is fused using our adaptive fusion method.

Challenging Loop Closure. In order to evaluate the performance of our system on challenging loop closure, we capture several challenging turning around motions. The results are shown in Fig. 10. As we can see, DoubleFusion fails to track the motion of the performer's arms and legs when they are occluded by the body and finally generates unsatisfactory loop closure results. In contrast, our system is able to track those motions under severe occlusions, generating complete and plausible models with such challenging turning around motions.

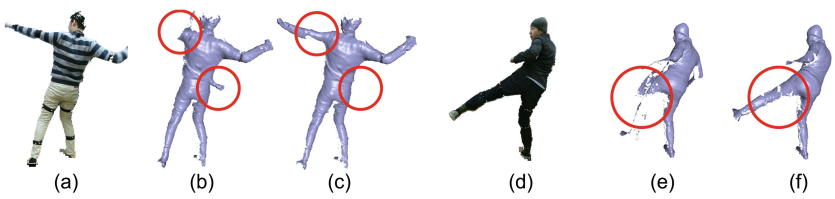

Fig. 10. Evaluation of the performance of our system on loop closure. We show the results in different frames. (a, d): Color reference. (b, e): The results reconstructed by DoubleFusion. (c, f): The results generated by our system. (Color figure online)

The Number of IMUs. To better evaluate our contributions, we also make experiments on the number of IMUs used in hybrid motion tracking. In Fig. 11,

the performer wears the full set of Noitom Legacy suite containing 17 IMUs attached on different body parts and performs several challenging motion such as leapfrogging, punching and so on. Regarding the tracking results with 17 IMUs as the ground-truth, we can get an estimation of tracking errors using different sensor setups. In Fig. 11, we present the average position error of joints using different numbers of IMUs. This experiment proves that using 8 IMUs (less than a half of the full set) with a single depth camera can achieve accurate tracking while preserving the convenience for usage.

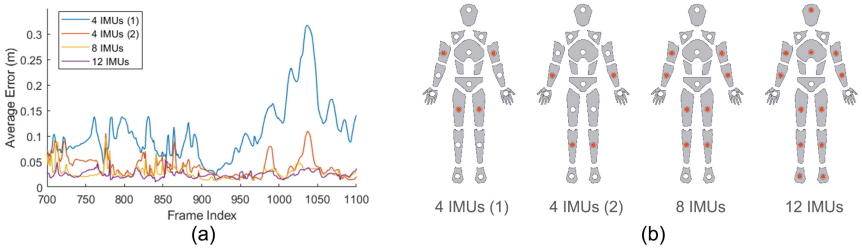

Fig. 11. Evaluation of the number of IMUs. (a): The curves of average position error of joints under different configurations. (b): Illustration of the 4 IMU configurations.

8 Discussion

Conclusion. In this paper, we have presented a practical and highly robust real-time human performance capture system that can simultaneously reconstruct challenging motions, detailed surface geometries and plausible inner body shapes using a single depth camera and sparse IMUs. We believe the practicability of our system enables light-weight, robust and real-time human performance capture, which makes it possible for users to capture high-quality 4D performances even at home. The real-time reconstructed results can be used in both AR/VR, gaming and virtual try-on applications.

Limitations. Our system cannot reconstruct very accurate surface mesh when people wearing very wide cloth because the cloth deformations are too complex for our sparse node-graph deformation model. Also, human-object interactions are very challenging, using divide-and-conquer scheme may provide plausible results. Although the IMUs we used are relatively small and easy to wear, it may still limit body motions. However, as the IMUs are getting more and more small and accurate, we believe the system setup can be even easier in the future.

Acknowledgement. This work is supported by the National Key Foundation for Exploring Scientific Instrument of China No. 2013YQ140517; the National NSF of China grant No. 61522111, No. 61531014, No. 61233005, No. 61722209 and No. 61331015. Hao Li was supported by the ONR YIP grant N00014-17-S-FO14, the CONIX Research Center, an SRC program sponsored by DARPA, the U.S. ARL under contract number W911NF-14-D-0005, Adobe, and Sony.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM Trans. Graph. **24**(3), 408–416 (2005)
2. Baak, A., Helten, T., Müller, M., Pons-Moll, G., Rosenhahn, B., Seidel, H.-P.: Analyzing and evaluating markerless motion tracking using inertial sensors. In: Kutulakos, K.N. (ed.) ECCV 2010. LNCS, vol. 6553, pp. 139–152. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35749-7_11
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
4. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless garment capture. In: ACM TOG. vol. 27, p. 99. ACM (2008)
5. Brox, T., Rosenhahn, B., Gall, J., Cremers, D.: Combined region and motion-based 3D tracking of rigid and articulated objects. IEEE TPAMI **32**(3), 402–415 (2010)
6. Dou, M., et al.: Motion2fusion: real-time volumetric performance capture. ACM Trans. Graph. **36**(6), 246:1–246:16 (2017)
7. Dou, M., et al.: Fusion4D: real-time performance capture of challenging scenes. ACM TOG **35**(4), 114 (2016)
8. Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR, pp. 1746–1753. IEEE (2009)
9. Guo, K., Xu, F., Wang, Y., Liu, Y., Dai, Q.: Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In: ICCV, pp. 3083–3091 (2015)
10. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, Albedo and motion reconstruction using a single RGBD camera. ACM Trans. Graph. (TOG) **36**(3) (2017)
11. Helten, T., Muller, M., Seidel, H.P., Theobalt, C.: Real-time body tracking with one depth camera and inertial sensors. In: The IEEE International Conference on Computer Vision (ICCV), December 2013
12. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: real-time volumetric non-rigid reconstruction. In: ECCV (2016)
13. Joo, H., Simon, T., Sheikh, Y.: Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: CVPR. IEEE (2018)
14. Leroy, V., Franco, J.S., Boyer, E.: Multi-view dynamic shape refinement using local temporal integration. In: ICCV. IEEE (2017)
15. Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction. In: ACM TOG. vol. 28, p. 175. ACM (2009)
16. Liao, M., Zhang, Q., Wang, H., Yang, R., Gong, M.: Modeling deformable objects from a single depth camera. In: ICCV (2009)
17. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2720–2735 (2013)
18. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)

19. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1987, pp. 163–169. ACM, New York, NY, USA (1987)
20. Malleson, C., Volino, M., Gilbert, A., Trumble, M., Collomosse, J., Hilton, A.: Real-time full-body motion capture from video and IMUs. In: 2017 Fifth International Conference on 3D Vision (3DV) (2017)
21. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: European Conference on Computer Vision, September 2018
22. von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and IMUs. Trans. Pattern Anal. Mach. Intell. PAMI **38**(8) (2016)
23. Mitra, N.J., Flöry, S., Ovsjanikov, M., Gelfand, N., Guibas, L.J., Pottmann, H.: Dynamic geometry registration. In: SGP, pp. 173–182 (2007)
24. Mustafa, A., Kim, H., Guillemaut, J., Hilton, A.: General dynamic scene reconstruction from multiple view video. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015, pp. 900–908 (2015)
25. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: reconstruction and tracking of non-rigid scenes in real-time. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
26. Pekelny, Y., Gotsman, C.: Articulated object reconstruction and markerless motion capture from depth video. In: CGF. vol. 27, pp. 399–408. Wiley Online Library (2008)
27. Pons-Moll, G., et al.: Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In: IEEE International Conference on Computer Vision (ICCV), pp. 1243–1250, November 2011
28. Pons-Moll, G., Baak, A., Helten, T., Müller, M., Seidel, H.P., Rosenhahn, B.: Multisensor-fusion for 3D full-body human motion capture. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2010
29. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: seamless 4D clothing capture and retargeting. ACM Trans. Graph. (Proc. SIGGRAPH) **36**(4), 73:1–73:15 (2017). Two first authors contributed equally
30. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: a model of dynamic human shape in motion. ACM Trans. Graph. (Proc. SIGGRAPH) **34**(4), 120:1–120:14 (2015)
31. Slavcheva, M., Baust, M., Cremers, D., Ilic, S.: KillingFusion: Non-rigid 3D reconstruction without correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
32. Slavcheva, M., Baust, M., Ilic, S.: SobolevFusion: 3D reconstruction of scenes undergoing free non-rigid motion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
33. Slyper, R., Hodgins, J.: Action capture with accelerometers. In: Gross, M., James, D. (eds.) Eurographics/SIGGRAPH Symposium on Computer Animation. The Eurographics Association (2008)
34. Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. In: SIGGRAPH, SIGGRAPH 2007. ACM, New York (2007)
35. Süßmuth, J., Winter, M., Greiner, G.: Reconstructing animated meshes from time-varying point clouds. In: CGF. vol. 27, pp. 1469–1476. Blackwell Publishing Ltd. (2008)

36. Tautges, J., et al.: Motion reconstruction using sparse accelerometer data. ACM Trans. Graph. **30**(3), 18:1–18:12 (2011)
37. Tevs, A., et al.: Animation cartography-intrinsic reconstruction of shape and motion. ACM TOG **31**(2), 12 (2012)
38. Tkach, A., Tagliasacchi, A., Remelli, E., Pauly, M., Fitzgibbon, A.: Online generative model personalization for hand tracking. ACM Trans. Graph. **36**(6), 243:1–243:11 (2017)
39. Vlasic, D., et al.: Practical motion capture in everyday surroundings. In: Proceedings of SIGGRAPH 2007. ACM (2007)
40. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM TOG. vol. 27, p. 97. ACM (2008)
41. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: automatic 3D human pose estimation from sparse IMUs. Computer Graphics Forum, Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics), pp. 349–360 (2017)
42. Xu, W., et al.: MonoPerfCap: human performance capture from monocular video. ACM TOG **37**, 27 (2017)
43. Ye, G., Liu, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Performance capture of interacting characters with handheld kinects. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 828–841. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_59
44. Yu, T., et al.: BodyFusion: real-time capture of human motion and surface geometry using a single depth camera. In: The IEEE International Conference on Computer Vision (ICCV). ACM, October 2017
45. Yu, T., et al.: DoubleFusion: real-time capture of human performance with inner body shape from a depth sensor. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
46. Zollhöfer, M., et al.: Real-time non-rigid reconstruction using an RGB-D camera. ACM TOG **33**(4), 156 (2014)
47. Zou, D., Tan, P.: CoSLAM: collaborative visual slam in dynamic environments. IEEE Trans. Pattern Anal. Mach. Intell. **35**(2), 354–366 (2013)