Graphical Models 110 (2020) 101072

Contents lists available at ScienceDirect

Graphical Models

journal homepage: www.elsevier.com/locate/gmod

Full-body motion capture for multiple closely interacting persons

Kun Li [a,*], Yali Mao [a], Yunke Liu [a], Ruizhi Shao [b], Yebin Liu [b]

[a] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[b] Department of Automation, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Multiple persons
Close interaction
Motion capture
Occlusions
Spatio-temporal constraints

ABSTRACT

Human shape and pose estimation is a popular but challenging problem, especially when asked to capture the body, hands, feet and face jointly for multiple persons with close interaction. Existing methods can only have a total motion capture of a single person or multiple persons without close interaction. In this paper, we present a fully automatic and effective method to capture full-body human performance including body poses, face poses, hand gestures, and feet orientations for closely interacting multiple persons. We predict 2D keypoints corresponding to the poses of body, face, hands and feet for each person, and associate the same person in multi-view videos by computing personalized appearance descriptors to reduce ambiguities and uncertainties. To deal with occlusions and obtain temporally coherent human shapes, we estimate shape and pose for each person with the spatio-temporal tracking and constraints. Experimental results demonstrate that our method achieves better performance than state-of-the-art methods.

1. Introduction

Human sensing [1] and modeling [2] based on images or videos are relatively essential and have extensive applications [3] such as human behavioral modeling, augmented reality, and character animation. Recent works [4–7] have shown great progress in the use of parametric model known as SMPL [8]. They employ convolutional neural network (CNN) to estimate the shape and pose parameters to achieve 3D human recovery. However, these methods are limited to estimating a single person in an image.

Reconstructing multiple persons, especially that involves close interactions with each other in natural scenes is crucial to more practical applications. But multi-person shape and pose estimation is very challenging due to serious inter-occlusions and inherent ambiguities in scenarios where multiple persons interact. Directly using the methods for single person would fail because of the incomplete information and uncertainties. Although some multi-person methods [9,10] have been proposed to deal with this problem, they can only deal with some very simple interactions. Liu et al. [11] proposed a multi-view segmentation scheme to reduce the ambiguities, which can estimate poses and surfaces of multiple persons with close interactions. But this method need manual intervention and a laser scanned template. Therefore, a fully-automatic shape and pose estimation method is required for these cases. Recently, an automatic but effective motion capture method [12] was proposed to achieve 3D shape and pose estimation for closely interacting multiple persons using multi-view images. They performed a frame-by-frame human recovery and ignored the temporal information during the shape estimation on a sequence. Consequently, there are some obvious jitters between adjacent frames. Besides, their method only focus on modeling body shape and pose without hands and face, which results in obvious mistakes on the capture of hands and face.

In this paper, we present an effective method for the full-body capture of multiple persons with close interactions, including human bodies, face poses, hand gestures and feet orientations. There are three challenges: ambiguities and uncertainties among persons, tracking difficulties due to occlusions, and motion jitters caused by frame-by-frame computation. To reduce ambiguities and uncertainties among persons, we propose to calculate discriminative person appearance descriptors for association after predicting 2D keypoints of body, face, hands and feet of each person. Then, we adopt spatio-temporal tracking to handle occlusions, and impose temporally coherent shape constraint to avoid motion jitters. Our method is simple but effective. Experimental results show that our method outperforms the state-of-the-art methods with more comprehensive capture and more accurate estimation.

Our main contributions are summarized as follows.

- We contribute a fully automatic and effective motion capture method for multiple people with close interactions which optimizes 3D human model of body, face, hands and feet simultaneously using multi-view constraints.
- We obtain multi-view association by computing discriminative person appearance descriptors, and use spatio-temporal tracking to deal with occlusions.

* Corresponding author.
E-mail addresses: lik@tju.edu.cn (K. Li), maoyali@tju.edu.cn (Y. Mao), 2019216111@tju.edu.cn (Y. Liu), jia1saurus@gmail.com (R. Shao), liuyebin@mail.tsinghua.edu.cn (Y. Liu).

https://doi.org/10.1016/j.gmod.2020.101072
Received 15 February 2020; Received in revised form 27 April 2020; Accepted 3 May 2020
Available online 22 May 2020

- We estimate the consensus shape for each person by imposing temporally coherent shape constraint, which helps to avoid motion jitters.

2. Related work

2.1. Multi-person 2D pose estimation in images

Existing methods of multi-person 2D pose estimation can be classified into two categories: bottom-up methods [13–17] which is very fast but has error associations between different persons, and top-down methods [18–20] that usually focus more on accuracy and are hence more time-consuming.

Bottom-up approaches first detect all joints and then assign these joints to each person. Deepcut [15] used deep features to jointly detect and label joints and associated them to individuals with correlation clustering. Deepercut [16] benefited from deeper Resnet [21] and improved the runtime of method [15]. Multiposenet [17] is a multi-task model by using the Pose Residual network to get keypoints and detection. OpenPose [13,14] used Part Affinity Fields (PAFs) to encode an unstructured paired relationships between joints and individuals, which can achieve real-time performance and provide 2D detections of human body, hands, face and feet on single image.

Top-down approaches first detect persons and then estimate pose for each person using single person pose estimation method. Papandreou et al. [19] estimated heatmap and its offset as the keypoints. CPN [20] used a framework composed of GlobalNet and RefineNet to predict easy and hard keypoints respectively. Fang et al. [18] estimated keypoints by combining different human detectors and pose estimators, which achieves good performance on human body detection.

In this paper, we aim to have a total capture of multiple persons. Therefore, OpenPose [13,14] is used as our 2D keypoint detector to jointly detect the keypoints of human body, face, hands and feet.

2.2. Pose tracking

Based on bottom-up multi-person 2D pose estimation, disordered keypoints are obtained, which can not be directly used in videos or multi-view situation. A few trackers [22,23] tried to link person detection across frames. Kim et al. [22] estimated the person appearance using Convolution Neural Network (CNN) to label the corresponding people. Tang et al. [23] presented a pair-wise feature extracted by patch matching to describe the relationships between persons. Posetrack [24] and artTrack [25] extended deepercut [16] and proposed to build a spatio-temporal graph to formulate this problem. Some work treated the tracking problem as an image-matching problem. Multi-image matching [26,27] aims to find the correspondence of a set of images. Dong et al. [26] proposed a cycle-consistent matching method based on appearance and geometry information to identify the same person in different images. Xiu et al. [28] proposed to use deepmatching as robust feature extractor and designed an efficient pose tracker based on pose flows. We extend these methods and propose spatio-temporal tracking to match the same person across views and frames by computing discriminative person appearance descriptors and pose similarities.

2.3. 3D recovery of human shape and pose

More 3D applications in real world and industrial demand to estimate human 3D pose and shape. At the beginning, the methods [29,30] only focused on the 3D pose estimation. They inferred 3D pose just from 2D features and ignored the 3D shape. Moreover, the results were usually bad when the feature detector did not work well. Simo-Serra et al. [31] proposed a Bayesian framework to jointly address 2D detection and 3D inference to get a better performance. Roberts et al. [32] simplified motion capture by editing their selected keyframes. Lifkooee et al. [33] proposed to utilize intrinsic Laplacian offsets to implement human pose transfer. At the same time, many methods based on deep learning have been proposed. DeepPose [34] was a simple but efficient cascaded pose estimation method based on Deep Neural Networks (DNN). Park et al. [35] directly estimated 3D human pose with end-to-end CNNs. They improved the efficiency of CNNs by combining 2D pose and image features, and got a more precise 3D pose.

With the use of statistic model [8,36] that consisting of shape and pose, human shape and pose can be estimated simultaneously. A few work [12,37–39] exploited different constraints to effectively regularized the fitting process for human reconstruction. Further more, many work [4–7,40] used the deep learning strategy to predict the model parameters of a single body using available 2D and 3D datasets [41–43], which simplified the process of 3D human reconstruction.

However, in more complicated multi-person cases, existing methods [9,10] can only deal with very limited situation of simple interactions without inter-occlusions. Liu et al. [11] proposed to combine instance segmentation with human reconstruction. This method achieved good performance but need manual intervention and a laser scanned template model at the beginning. Li et al. [12] proposed the first method to estimate multi-person shape and pose automatically.However, they only captured the shape and pose of human body without hands and face which are important for human interactions. In contrast, we propose a method to have a full-body motion capture including the hands and face of closely interacting persons with spatio-temporal constraints.

3. Method

Our goal is to have a full-body capture including body poses, face poses, hand gestures, and feet orientations for multiple closely interacting persons. Fig. 1 presents the details of our method. We first estimate multi-person 2D keypoints using OpenPose [14], which jointly detect the features of body, feet, hands and face. To reduce the uncertainties and ambiguities, a multi-person tracking method is designed to build the spatio-temporal correspondence. Finally, we fit a newest parametric 3D human model, SMPL-X [44], so that the projected joints are as consistent as possible with multi-view 2D detections. Besides, we also introduce a temporally coherent shape constraint to avoid motion jitter.

3.1. Multi-person 2D pose estimation

We use the OpenPose detector [14] in each available view, which jointly estimates 2D keypoints of body, feet, hands and face. OpenPose is a bottom-up 2D keypoint detection method using Part Affinity Fields (PAFs) to encode an unstructured paired relationships between joints and individuals, which can achieve real-time performance. For multi-person cases with close interaction, there will be some obvious mistakes on arms and legs by directly using OpenPose detector, hence we use a region-based detector [18] to rectify the errors on arms and legs.

3.2. Multi-Person pose tracking across views and frames

In Section 3.1, we got multi-person 2D keypoints, and then we need to build association across views and frames. To achieve this, we design a spatio-temporal tracking scheme. Specifically, we first adopt a pre-trained person Re-ID network [45] to obtain personalized appearance descriptors for labeling the persons in the starting frame. Then, we use temporal pose tracking for each view to get the correct orderings for multiple persons across frames. Finally, we validate the tracked results and determine the most accurate labels using spatio-temporal constraint.

Spatial tracking. To match the detected 2D poses across different views, we need an appropriate criterion to measure the likelihood that two poses belong to the same person. This problem is very similar to the work of person Re-ID (Re-identification). Therefore, we estimate the bounding box for each person, and calculate the appearance Affinity matrix to measure the similarities between bounding boxes using a publicly available Re-ID model [45]. The Re-ID network is pre-trained on massive

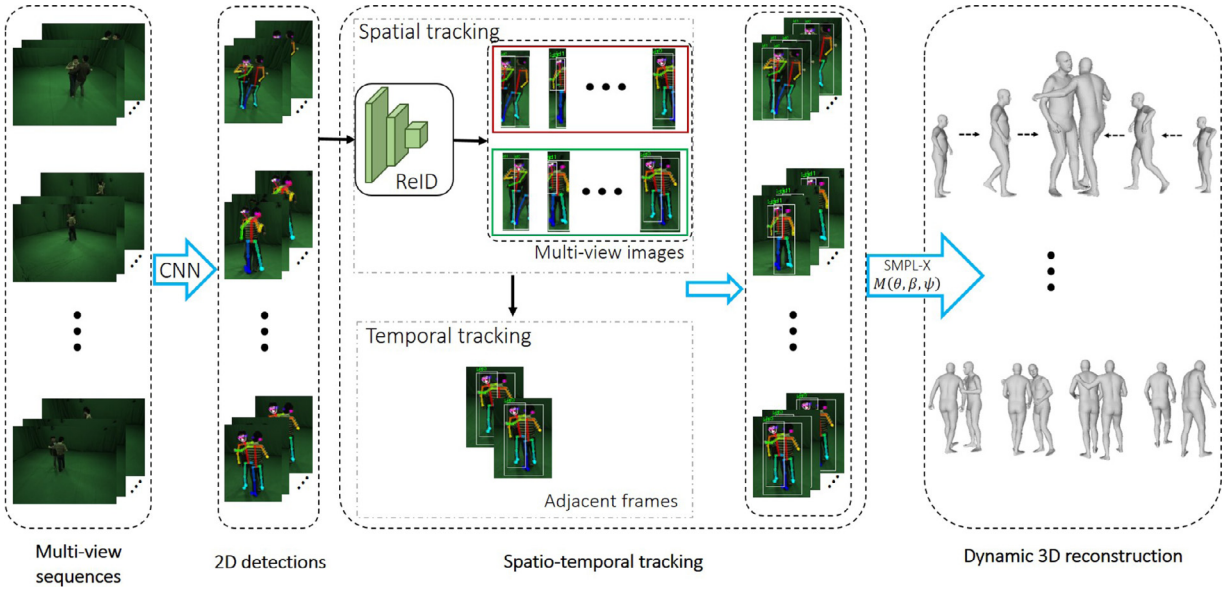

Fig. 1. The pipeline of our method. It contains of 3 components: 1) 2D pose estimation; 2) Tracking across views and frames; 3) 3D reconstruction with spatio-temporal constraints.

datasets and is able to extract discriminative and descriptive appearance features that are robust to illumination and viewpoint changes. Suppose there are V cameras in the scene, and p_i detected bounding boxes in the view i. Matrix A_{ij} measures the Affinity scores of view pair (i, j). Similiar to [26], we extract the feature vectors from "pool5" layer as the descriptor for each bounding box. The sigmoid function is used to map the distances to (0, 1) after computing the Euclidean distance between the descriptors of two bounding boxes. For the first frame, we select view 0 as a reference image and then compute the similarity scores between view i and view 0. The Affinity matrix can be written as:

$$A_{0,i} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix}, \tag{1}$$

where $A_{0,i} \in \mathbb{R}^{n\times n}$, is a symmetric matrix. $n = \sum_{l\in\{0,i\}} p_l$, which is the total number of people in view 0 and view i. The element $a_{i,j} \in (0, 1)$ represents the similarity score between the ith bounding box and the jth bounding box, larger for more similar appearance. We can easily find the optimal matching by maximizing the similarity scores using the Hungarian algorithm for each pair of views. This process is shown in Fig. 2.

Temporal tracking. Temporal tracking is built by associating poses that indicate the same person across frames. We perform a frame-by-frame tracking method to combine the information of the bounding boxes and estimated 2D poses. Define P_1 and P_2 as the body poses of two adjacent frames, and denote B_1 and B_2 as the bounding boxes surrounding P_1 and P_2. The similarity metric is defined as

$$S(P_1, P_2, B_1, B_2) = \alpha P_s(P_1, P_2) + B_s(B_1, B_2), \tag{2}$$

where P_s is a function to measure the possibility of two cross-frame poses indicating the same person. We adopt the inter-frame pose distance defined in [46]:

$$P_s(P_1, P_2) = \sum_i \frac{f_2^i}{f_1^i}, \tag{3}$$

where P_1 and P_2 are the body poses of two consecutive frames, f_1^i represents the number of feature points extracted by DeepMatching [47] from the bounding box B_1^i surrounding p_1^i, and f_2^i is the number of feature points extracted from the bounding box B_2^i that match f_1^i. p_1^i and p_2^i are the ith keypoint of P_1 and P_2, respectively. The bounding boxes are 10% person bounding boxes size according to the standard PCK [41].

In Eq. (2), B_s is more like a global term compared with P_s, which includes the feature points of full body. Considering that there are some crucial feature points to identity a person that cannot be perceived by skeletons, we use the feature points extracted from bounding boxes B_1 and B_2 to describe the similarities of two poses in two adjacent frames. The similarity metric B_s is defined as

$$B_s = \frac{MI}{MU}, \tag{4}$$

$$MI = |f_1 \cap f_2|, \tag{5}$$

$$MU = |f_1 \cup f_2|, \tag{6}$$

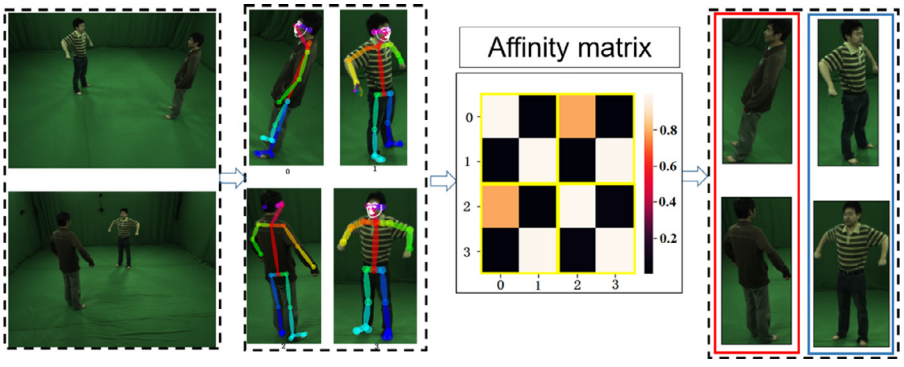

Fig. 2. Illustration of our multi-view tracking. The colors filled in Affinity matrix represents the similarity scores. The output with the same color box (red or blue) is the matched people in different views. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where *MI* are the matched feature points between B_1 and B_2, and *MU* are the total feature points extracted from B_1 and B_2.

Spatio-temporal validation. To find the most accurate labels, we then feed the previous labeled bounding boxes into the network [45] and try to maximize the similarity of the same labeled bounding boxes across all the views. For a bounding box labeled as '*l*' in view *v*, suppose the similarity scores between it and the other bounding boxes locate in the *q*th row, and the set M_p is the index set of the other views with the same label. The problem can be formulated as

$$H(B_v^l) = \max_p(mean(\sum_{k\in M_p} a_{qk})|p \in T_p), \tag{7}$$

where T_p is the total number of persons, and *p* is the *p*-th person in each view. After spatio-temporal tracking, we can get the most appropriate label of each person in each view and each frame.

3.3. Shape and pose estimation for multiple persons

Given the tracked 2D poses of body, hands, feet, and face, we estimate the 3D human shape and pose for each time instant by combing multi-view clues. We use SMPL-X [44], a newly proposed unified deformation model, as our underlying shape representation, and optimize the parameters of the model with multi-view 2D keypoints. Considering that independent optimization for each frame will lead to motion jitters, we introduce a temporal constraint for a consistent shape for each person.

Unified model. SMPL-X is an expressive statistical human model by integrating the SMPL model [8] with FLAME face model [48] and MANO hand model [49]. It has $N = 10475$ vertices and $K = 54$ joints, which is about twice that of SMPL. SMPL-X is defined by a function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta|\times|\beta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$, where $\beta \in \mathbb{R}^{|\beta|}$ contains the shape parameters of the body, face and hands, θ contains the pose parameters of jaw joint, finger joints and the body joints, and $\psi \in \mathbb{R}^{|\psi|}$ has the facial expression parameters. With this unified model, we can simply obtain the shapes and poses of multi-persons by optimizing the parameters (θ, β, ψ) for each person. Please refer to [44] for more details about the model.

Optimization. Although many previous work [4,5,7] have achieved great success in human shape and pose estimation, they are limited to dealing with a single person in an image. Hence, we cannot directly use these methods for our problem which includes serious occlusions and ambiguities caused by close interactions. Instead, we estimate the shape and pose for each person at each time instant using spatio-temporal constraints. Given the detected 2D keypoints $\{J_{est}^1, J_{est}^2, \cdots, J_{est}^{|V|}\}$ for different views *V*, we formulate the problem of fitting SMPL-X to multi-view keypoints as an optimization problem. We define the energy function as

$$E_M(\beta, \theta, \psi) = \sum_{v=1}^{|V|} E_J(\beta, \theta; K_v, J_{est}^v) + E_p(\beta, \theta, \psi) + \lambda_c E_c + \mu E_t, \tag{8}$$

where K_v contains the camera parameters of view *v*. E_J is a joint fitting data term, E_p is a prior term, E_c is a collision term that penalizes self-collisions and penetrations of several body parts, and E_t is a temporal term.

For the data term, it is defined by a re-projection loss which is used to minimize the weighted distance between the 2D projection of 3D joint $R_\theta(J(\beta))_i$ and the detected 2D keypoint $J_{est,i}^v$ for each joint *i*. It is formulated as

$$E_J(\beta, \theta; K_v, J_{est}^v) = \textstyle\sum_{\text{joint } i} \gamma_i \omega_i \rho(\Pi_{K_v}(R_\theta(J(\beta))_i) - J_{est,i}^v), \tag{9}$$

where $J(\cdot)$ is a function that transforms rest vertices into rest joints, $R_\theta(\cdot)$ is a global rotation function that converts the rest joints to the posed 3D joints according to the pose θ, Π_{K_v} denotes a projection function of the *v*-th view, ω_i is the detected confidence score of the *i*-th joint, and γ_i are per-joint weights for annealed optimization (described later). Considering the noise in detections, we use a robust Geman-McClure error function [46] defined as

$$\rho_\sigma(e) = \frac{e^2}{\sigma^2 + e^2}, \tag{10}$$

where *e* is the residual error, and σ is a robustness constant set to be 100.

The prior term E_p that jointly penalizes the body, face and hands is learned from massive training data. It is formulated as

$$E_p(\beta, \theta, \psi) = \lambda_{\theta_b} E_{\theta_b}(\theta_b) + \lambda_{\theta_f} E_{\theta_f}(\theta_f) + \lambda_{m_h} E_{m_h}(m_h) + \lambda_\beta E_\beta(\beta) + \lambda_\varepsilon E_\varepsilon(\psi), \tag{11}$$

where θ_b, θ_f, and m_h are the pose vectors of body, face, and two hands, respectively. The body pose $\theta_b = dec(Z)$ is a function of latent vectors $Z \in \mathbb{R}^{32}$ that follow the normal distribution. $\theta = (\theta_b, \theta_f, m_h)$ contains the pose parameters of the whole body. The terms E_{θ_f}, E_{m_h}, E_β and E_ε are simple L_2 priors for face pose, hand pose, body shape and facial expression, respectively. A VAE-based body pose prior E_{θ_b} is also used for a reasonable axis-angle representation for human body pose. Different from that of [37] which directly optimizes θ_b, the VAE-based body pose prior imposes the quadratic penalty on the latent vector *Z* to regularize a normal distribution of the latent space [44]. λ_* represents the weights for each term that helps the optimization in the use of annealing scheme.

For the whole human body, it is inevitable to have physically impossible self-collisions and penetrations during the optimization process. Therefore, we use Bounding Volume Hierarchies (BVH) [50] as the collision detection. Based on [51,52], to find a collision between two triangles f_t and f_s, a point-to-point distance is used in the collision term:

$$E_c(\theta) = \sum_{(f_s(\theta), f_s(\theta))\in C} \left\{ \sum_{v_s\in f_s} \| - \Psi_{f_t}(v_s) n_s\|^2 + \sum_{v_t\in f_t} \| - \Psi_{f_s}(v_t) n_t\|^2 \right\}, \tag{12}$$

where *C* is the set of colliding triangles, *v* represents vertex, and Ψ is the local 3D distance field defined by triangles *C* and their normals *n*.

Simply fitting a SMPL-X model to multi-view 2D detections cannot produce temporally consistent shapes for the same person in different frames, which will result in motion jitters. Hence, we add a temporal constraint on the shape parameter β, which is defined as

$$E_t = \|\beta_t - \beta_{t-1}\|. \tag{13}$$

Consistent shape for the same person brings temporally stable bone lengths which is beneficial to pose estimation.

Implementation details. Our algorithm is implemented in Pytorch and we solve our optimization problem using Limited-memory BFGS (L-BFGS) [53] optimizer. The learning rate of our algorithm is set to be 1.0. Similar to [37], we optimize Eq. (8) in a multi-stage manner to avoid local minima by starting with high regularization for body and then gradually increasing the influence of hands and face in three steps. The weights γ_i in Eq. (9) corresponding to body, hands and face in three steps are set to be (1.0, 0.0, 0.0), (1.0, 0.1, 0.0) and (1.0, 2.0, 2.0), respectively. The weights λ_* except λ_c in Eq. (11) gradually decrease for a better fitting. However, λ_c increases to impose higher regularization on the collisions that will deteriorate with more influence of hands and face. As for the weight μ, larger μ brings more temporally consistent shape. Therefore, we set it to be 50 in our experiment.

4. Experiments

In this section, we first evaluate the proposed method with ablation study in Section 4.2 on a public multi-view human-human interaction (MHHI) dataset [54] (Section 4.1), and then compare our method with the state-of-the-art methods quantitatively and qualitatively in Section 4.3. Finally, we give the detailed running times of our method in Section 4.4.

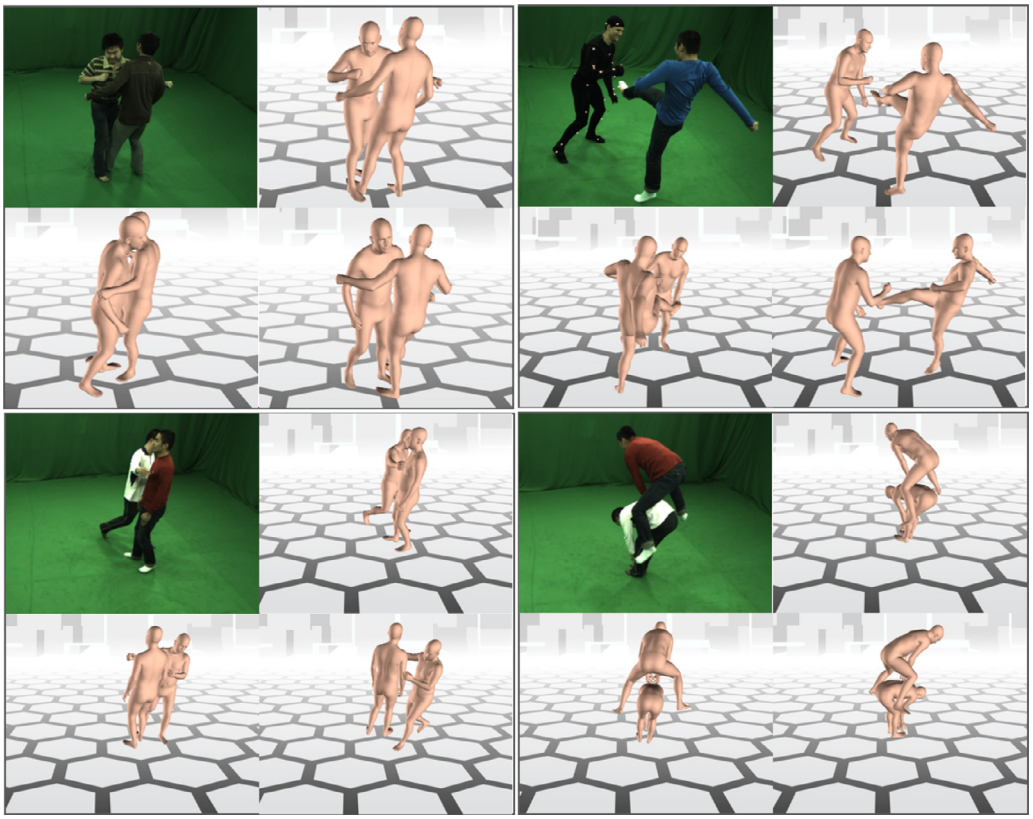

Fig. 3. 3D reconstruction of human body, face, feet, and hands for multiple people with close interactions. The presented results projected to 3 different views are examples of 4 sequences respectively.

4.1. Evaluation dataset

MHHI is a dataset that includes 7 sequences with different motions. It was recorded by 12 synchronized and calibrated cameras with the image resolution of 1296 × 967. Each motion sequence contains more than 200 frames. There are four challenging available public sequences (*Crash, Jump, Wrestle* and *Fight*), categorized into marker-based capture data and markerless capture data. The *Fight* sequence is a marker-based motion capture sequence that can be used for quantitative evaluation. In this sequence, 38 markers are attached to a person whose motion is captured by a commercial marker-based motion capture system PhaseSpace^{TM} as the ground truth. We quantitatively evaluate our method on this sequence which contains complex and extreme poses as well as fast motion. Some examples of our reconstruction results on the four available sequences are shown in Fig. 3.

4.2. Ablation study

We first explore how the results are affected by different components. Table 1 gives the comparison results of without and with tracking on the *Fight* dataset. As shown in this table, tracking plays an important role in multi-person shape and pose estimation. In the case of no tracking, the detected 2D keypoints have no correspondences. Therefore, the optimization across views and frames will be invalid, which brings a worse result. If only using the temporal tracking, the results will get a little improved because of considering temporal information. Larger improvement is achieved by using our spatio-temporal tracking. This benefits from our discriminative person appearance descriptor and the elegant design of spatio-temporal tracking. We also compare the results using original OpenPose detection and the updated 2D detection. The updated operation does not contribute as much as tracking, but it makes the estimation more stable with smaller standard deviation. Fig. 4 shows the visual results of without and with tracking. We can see obvious improvement by using our spatio-temporal tracking.

Table 1
Quantitative evaluation of without tracking (N. track), with temporal tracking (T. track), and with spatio-temporal tracking (S.T. track). O: using original OpenPose detection; O+: using updated OpenPose detection.

	N. track	T. track	S.T. track	
			O	O+
Mean(mm)	765.91	241.56	31.79	**30.35**
Std(mm)	420.00	187.34	10.27	**7.89**

We further compare the face reconstruction results of using SMPL-X model [44] and a representative face model FLAME [48] on the MHHI dataset in Fig. 5. FLAME is an existing lightweight method that can reconstruct an expressive face and the whole head.We fit FLAME to the same 2D keypoints as ours for a fair comparison by using their official code.[1] Because multi-person close interaction scenario contains severe occlusions and the image resolution of human faces is not high, directly using the existing 3D face reconstruction method cannot generate satisfied results. The FLAME method shows expressive results for some visible and frontal views, but fails to generate correct results for occlusion cases. On the contrast, our method obtains better results for both persons in various views.

4.3. Qualitative and quantitative evaluation

To our best knowledge, very little work can achieve 3D shape and pose estimation for multiple closely interacting persons. Method [12] is

[1] https://github.com/TimoBolkart/TF_FLAME.

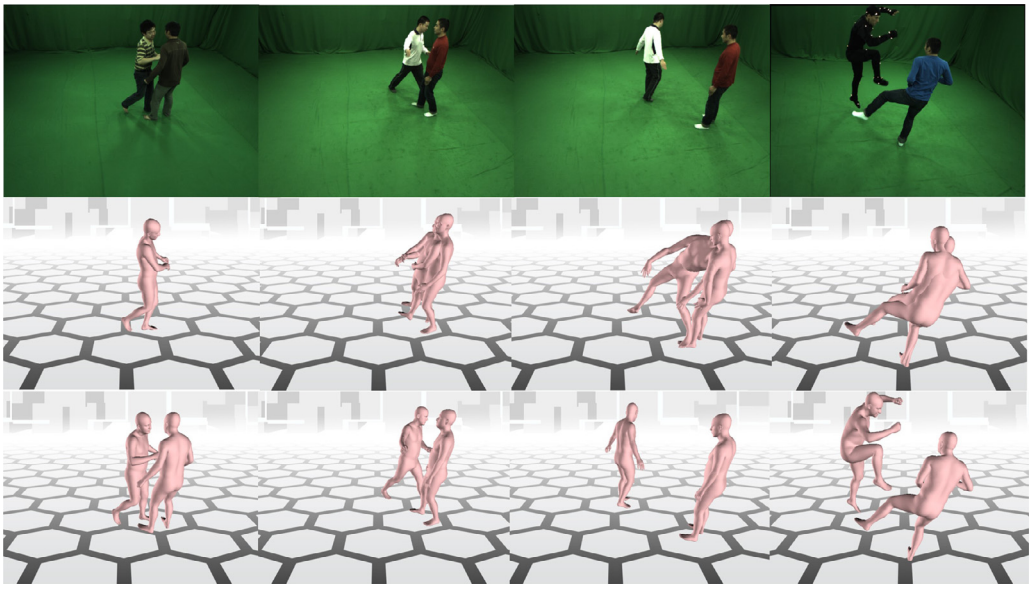

Fig. 4. 3D reconstruction results on four sequences (first row) without tracking (second row) and with spatio-temporal tracking (third row).

Fig. 5. Comparison of face reconstruction of FLAME [48] (middle) against our method (bottom).

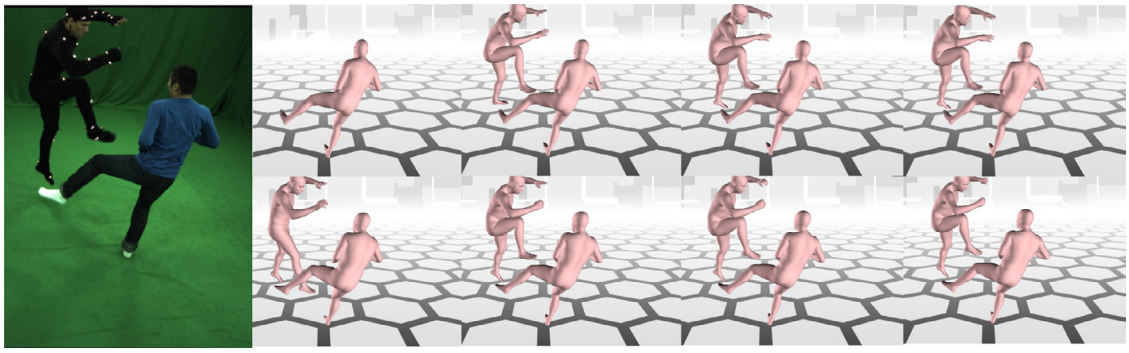

Fig. 6. The reconstruction results with method [12] (top row) and with our method (bottom row) using 2, 4, 8, 12 views from left to right.

Fig. 7. 3D reconstruction results for four sequences (top) by method [12] (middle), and our method (bottom).

a very recent method for reconstructing multiple people involving close interactions. This method uses multi-view cues to compensate incompletions and ambiguities due to occlusions, which has achieved the state-of-the-art results. In order to give a fair comparison, we evaluate our method on the *Fight* dataset similar to [12]. Quantitative evaluation results with different views are given in Table 2. The selection of cameras is based on their indices in the dataset. We calculate the mean distance with standard deviation between 38 tracked markers and its paired 3D vertices that are matched in the first reconstructed frame across all 500 frames. As shown in Table 2, our method outperforms the state-of-the-

Table 2
Comparison on different number of views. All errors are in millimeters.

Number of views	[12]		Ours	
	Mean	Std	Mean	Std
2 views	242.27	985.75	63.93	57.35
4 views	58.42	177.56	37.88	17.67
8 views	48.57	10.06	32.73	11.44
12 views	43.30	9.45	**30.35**	**7.89**

Fig. 8. Total capture results for multiple persons by using method [39] (middle row) and our method (bottom row).

art method [12] on both mean error and standard deviation metrics, which demonstrates that our method achieves more accurate shape and pose estimation. It is worth noting that our method performs much better than method [12] especially with sparse views. This verifies the effectiveness of our approach on better dealing with ambiguities, occlusions and motion jitters. Some visual results are shown in Fig. 6. It can be seen that our method can basically reconstruct the two persons in the case of two views while method [12] only reconstructs one person.

Fig. 9 shows the mean reconstruction error of each frame using method [12] and ours with 12 views. The errors are smaller by using our method for most frames. More qualitative comparison results are shown in Fig. 7. For closer observation, some regions are enlarged in the corresponding images in Fig. 7. Our method outperforms method [12] especially in some complex motion cases.

We also compare our method with a newest total capture method [39] by using its publicly available code in Fig. 8. It can be seen that our method is able to produce more accurate and realistic 3D human models for multiple closely interacting persons.

Although our method presents good performance on 3D shape and pose estimation for multiple persons with close interactions, there are still some failure cases shown in Fig. 10. Due to complex motion or severe occlusions, our proposed method is hard to obtain the accurate 3D estimation based on wrong 2D estimations in most of views.

4.4. Running time

We conduct our experiments on a desktop with a 16-GB RAM and a NVIDIA GeForce GTX 1080Ti GPU. Take two persons for an example. The 2D pose estimation using OpenPose takes about 0.6*s* per frame. The spatial tracking takes about 1.28*s* for a pair of images, including 1.26*s* for feature extraction using ReID model and 0.02*s* for the matching. The temporal tracking takes about 7.78*s* using the time-consuming

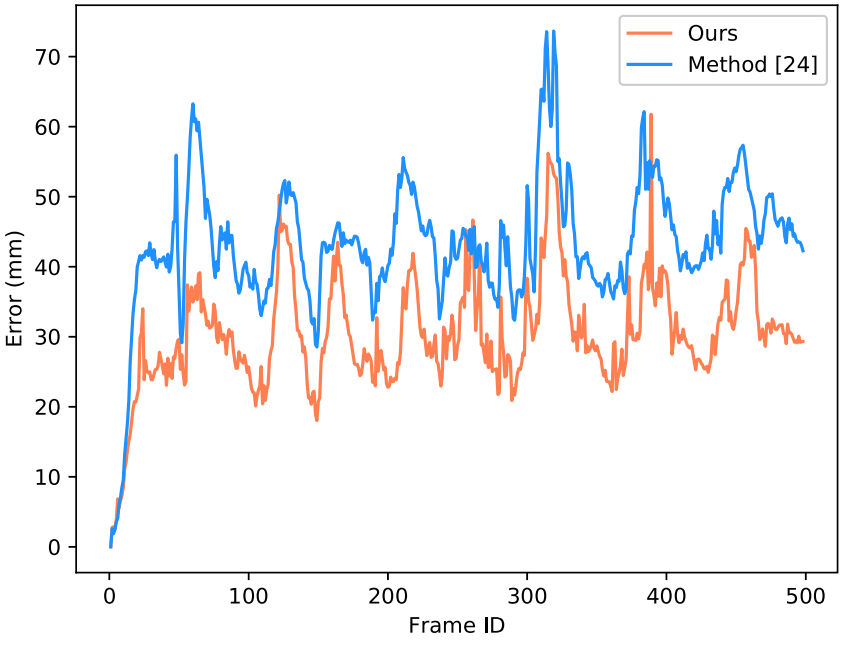

Fig. 9. The errors of 500 frames in *Fight* sequence.

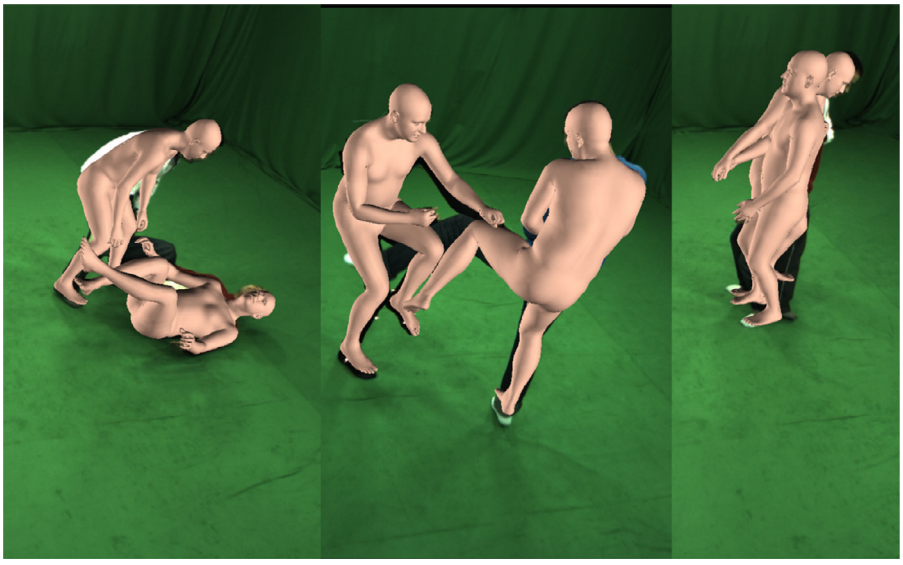

Fig. 10. Examples of failure cases.

DeepMatching feature extraction method (7.75*s*). The shape and pose optimization takes about 90s per frame for two persons.

5. Conclusion

In this paper, we present a fully automatic method to jointly capture body, feet, face, and hands of multiple persons involving close interactions. To overcome challenges caused by multiple closely interacting persons, we design a spatio-temporal tracker that uses discriminative appearance descriptors and pose similarities to get multi-person associations in spatial and temporal domain. For more accurate reconstruction, we estimate 3D shape and pose of each person by fitting a SMPL-X model to multi-view videos using spatio-temporal constraints. Experimental results shows that our method outperforms state-of-the-art methods quantitatively and qualitatively.

In future work, we will try to speed up our method. Furthermore, we will also try to recover the geometry details on the shape with shape-from-shading approaches, and generate a full texture map for a more vivid virtual character.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the Tianjin Research Program of Application Foundation and Advanced Technology under Grant 18JCYBJC19200, and the National Natural Science Foundation of China (Grant 61827805 and Grant 61861166002).

References

[1] Y.-W. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, et al., Towards fully mobile 3D face, body, and environment capture using only head-worn cameras, IEEE Trans. Vis. Comput. Graph. 24 (11) (2018) 2993–3004.

[2] T. Waltemate, D. Gall, D. Roth, M. Botsch, M.E. Latoschik, The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response, IEEE Trans. Vis. Comput. Graph. 24 (4) (2018) 1643–1652.

[3] P. Caserman, A. Garcia-Agundez, S. Goebel, A survey of full-body motion reconstruction in immersive virtual reality applications, IEEE Trans. Vis. Comput. Graph. (2019).

[4] A. Kanazawa, M.J. Black, D.W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7122–7131.

[5] G. Pavlakos, L. Zhu, X. Zhou, K. Daniilidis, Learning to estimate 3D human pose and shape from a single color image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 459–468.

[6] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, B. Schiele, Neural body fitting: Unifying deep learning and model-based human pose and shape estimation, in: Proceedings of the International Conference on 3D Vision (3DV), IEEE, 2018, pp. 484–494.

[7] P. Yao, Z. Fang, F. Wu, Y. Feng, J. Li, Densebody: directly regressing dense 3D human pose and shape from a single color image, arXiv preprint arXiv:1903.10153(2019).

[8] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graph. (TOG) 34 (6) (2015) 248.

[9] C. Wu, C. Stoll, L. Valgaerts, C. Theobalt, On-set performance capture of multiple actors with a stereo camera, ACM Trans. Graph. (TOG) 32 (6) (2013) 161.

[10] A. Mustafa, H. Kim, J.-Y. Guillemaut, A. Hilton, General dynamic scene reconstruction from multiple view video, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 900–908.

[11] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, C. Theobalt, Markerless motion capture of interacting characters using multi-view image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2720–2735.

[12] K. Li, N. Jiao, Y. Liu, Y. Wang, J. Yang, Shape and pose estimation for closely interacting persons using multi-view images, in: Computer Graphics Forum, 37, Wiley Online Library, 2018, pp. 361–371.

[13] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: realtime multi-person 2D pose estimation using part affinity fields, arXiv preprint arXiv:1812.08008(2018).

[15] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P.V. Gehler, B. Schiele, DeepCut: joint subset partition and labeling for multi person pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4929–4937.

[16] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, DeeperCut: a deeper, stronger, and faster multi-person pose estimation model, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 34–50.

[17] M. Kocabas, S. Karagoz, E. Akbas, MultiPoseNet: fast multi-person pose estimation using pose residual network, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 417–433.

[18] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, RMPE: regional multi-person pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.

[19] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4903–4911.

[20] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[22] C. Kim, F. Li, A. Ciptadi, J.M. Rehg, Multiple hypothesis tracking revisited, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4696–4704.

[23] S. Tang, B. Andres, M. Andriluka, B. Schiele, Multi-person tracking by multicut and deep matching, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 100–111.
[24] U. Iqbal, A. Milan, J. Gall, Posetrack: joint multi-person pose estimation and tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2011–2020.
[25] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, B. Schiele, Arttrack: articulated multi-person tracking in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6457–6465.
[26] J. Dong, W. Jiang, Q. Huang, H. Bao, X. Zhou, Fast and robust multi-person 3D pose estimation from multiple views, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7792–7801.
[27] X. Zhou, M. Zhu, K. Daniilidis, Multi-image matching via fast alternating minimization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4032–4040.
[28] Y. Xiu, J. Li, H. Wang, Y. Fang, C. Lu, Pose flow: efficient online pose tracking, arXiv preprint arXiv:1802.00977(2018).
[29] J. Gall, B. Rosenhahn, T. Brox, H.-P. Seidel, Optimization and filtering for human motion capture, Int. J. Comput. Vis. 87 (1–2) (2010) 75.
[30] A. Yao, J. Gall, L.V. Gool, R. Urtasun, Learning probabilistic non-linear latent variable models for tracking complex activities, in: Advances in Neural Information Processing Systems, 2011, pp. 1359–1367.
[31] E. Simo-Serra, A. Quattoni, C. Torras, F. Moreno-Noguer, A joint model for 2D and 3D pose estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3634–3641.
[32] R. Roberts, J. Lewis, K. Anjyo, J. Seo, Y. Seol, Optimal and Interactive Keyframe Selection for Motion Capture, in: SIGGRAPH Asia 2018 Technical Briefs, 2018, pp. 1–4.
[33] M.Z. Lifkooee, C. Liu, Y. Liang, Y. Zhu, X. Li, Real-time avatar pose transfer and motion generation using locally encoded laplacian offsets, J. Comput. Sci. Technol. 34 (2) (2019) 256–271.
[34] A. Toshev, C. Szegedy, DeepPose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.
[35] S. Park, J. Hwang, N. Kwak, 3D human pose estimation using convolutional neural networks with 2D pose information, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 156–169.
[36] H. Joo, T. Simon, Y. Sheikh, Total capture: a 3D deformation model for tracking faces, hands, and bodies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8320–8329.
[37] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, M.J. Black, Keep it SMPL: automatic estimation of 3D human pose and shape from a single image, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 561–578.
[38] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P.V. Gehler, J. Romero, I. Akhter, M.J. Black, Towards accurate marker-less human shape and pose estimation over time, in: Proceedings of the International Conference on 3D Vision (3DV), IEEE, 2017, pp. 421–430.
[39] D. Xiang, H. Joo, Y. Sheikh, Monocular total capture: posing face, body, and hands in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10965–10974.
[40] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, C. Schmid, BodyNet: volumetric inference of 3D human body shapes, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 20–36.
[41] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2D human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[42] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al., Panoptic studio: a massively multiview system for social interaction capture, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2017) 190–204.
[43] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2013) 1325–1339.
[44] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A. Osman, D. Tzionas, M.J. Black, Expressive body capture: 3D hands, face, and body from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10975–10985.
[45] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.
[46] S. Geman, Statistical methods for tomographic image reconstruction, Bull. Int. Stat. Inst 4 (1987) 5–21.
[47] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, DeepFlow: Large displacement optical flow with deep matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1385–1392.
[48] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, ACM Trans. Graph. (TOG) 36 (6) (2017) 194.
[49] J. Romero, D. Tzionas, M.J. Black, Embodied hands: modeling and capturing hands and bodies together, ACM Trans. Graph. (TOG) 36 (6) (2017) 245.
[50] M. Teschner, S. Kimmerle, B. Heidelberger, G. Zachmann, L. Raghupathi, A. Fuhrmann, M.-P. Cani, F. Faure, N. Magnenat-Thalmann, W. Strasser, et al., Collision detection for deformable objects, in: Computer Graphics Forum, 24, Wiley Online Library, 2005, pp. 61–81.
[51] L. Ballan, A. Taneja, J. Gall, L. Van Gool, M. Pollefeys, Motion capture of hands in action using discriminative salient points, in: Proceedings of the European Conference on Computer Vision, Springer, 2012, pp. 640–653.
[52] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, J. Gall, Capturing hands in action using discriminative salient points and physics simulation, Int. J. Comput. Vis. 118 (2) (2016) 172–193.
[53] J. Nocedal, S.J. Wright, Nonlinear equations, Numerical Optimization (2006) 270–302.
[54] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, C. Theobalt, Markerless motion capture of interacting characters using multi-view image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1249–1256.