Lightweight Multi-person Total Motion Capture Using Sparse Multi-view Cameras

Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu*, Yebin Liu*
Department of Automation and BNRist, Tsinghua University

Abstract

Multi-person total motion capture is extremely challenging when it comes to handle severe occlusions, different reconstruction granularities from body to face and hands, drastically changing observation scales and fast body movements. To overcome these challenges above, we contribute a lightweight total motion capture system for multi-person interactive scenarios using only sparse multi-view cameras. By contributing a novel hand and face bootstrapping algorithm, our method is capable of efficient localization and accurate association of the hands and faces even on severe occluded occasions. We leverage both pose regression and keypoints detection methods and further propose a unified two-stage parametric fitting method for achieving pixel-aligned accuracy. Moreover, for extremely self-occluded poses and close interactions, a novel feedback mechanism is proposed to propagate the pixel-aligned reconstructions into the next frame for more accurate association. Overall, we propose the first light-weight total capture system and achieve fast, robust and accurate multi-person total motion capture performance. The results and experiments show that our method achieves more accurate results than existing methods under sparse-view setups.

1. Introduction

Marker-less motion capture, due to its great potentials for behaviour understanding, sports analysis, human animation, video editing and virtual reality, has been a popular research topic in computer vision and graphics for decades. Within this research field, total motion capture, pioneered by [22] using an extremely dense-view setup (hundreds of cameras), shows impressive results of simultaneous capture of multi-person total interactive behaviours including facial expressions, body and hand poses, and has aroused widespread interest in computer vision community. However, this work [22] suffers from expensive and sophisticated hardware setup and low run-time efficiency.

Recently, to reduce the capture complexity, more and more researches try to perform total motion capture from only a single image or video [41, 55, 13, 45, 35, 64]. By either optimizing a parametric models like SMPL-X [41] and Adam [22] ( [55]) or regressing the model parameters directly from the input images [13], these methods even achieve real-time total motion capture performance for single-person [45, 64]. However, it remains difficult for the monocular methods to handle severe occlusions and challenging poses under multi-person interactive scenarios.

To guarantee both lightweight setups and robust performance, we propose the first lightweight total capture system using only sparse multi-view cameras. However, extending the existing monocular total capture methods to sparse-view multi-person total capture is not trivial. Although the incorporation of multi-view observations may resolve the depth ambiguities for monocular methods, the severe occlusions caused by complex-poses and multi-person interactions will significantly deteriorate the performance for current total capture methods. Specifically, the main challenges include: i) hand/face association across multiple views under drastically changing observation scales and unstable detection results, ii) pixel-aligned fitting between the reconstructed 3D model and the input images, and iii) robust and accurate body association under severe occlusions even for close interactions. To resolve all the challenges above, we propose, as far as we know, the first method to achieve fast, robust and accurate multi-person total motion capture using only light-weight sparse-view cameras.

First of all, compared with the relatively fixed body part scales and satisfactory occlusion-free view point in monocular single-person total capture cases [41, 55, 13, 45, 64],
sparse multi-view setups suffer from hand/face fragments on account of severe occlusions, blurs on hands and even fingers by fast limb movements, and varying hand/face scales across different cameras. Moreover, it remains challenging to associate hands correctly when different hands are located very closely on an image. To resolve these challenges, we propose a novel hand and face bootstrapping algorithm to extract accurate body part features effectively from the sparse and multi-scale images for accurate association. Benefiting from the recent progress in multi-person skeleton pose capture [62], the skeleton-level results are utilized to guide the following object-detection network for more robust and accurate detection. Moreover, we introduce the cross-modality consistency and cross-scale consistency to filter unexpected detection results of fragments caused by occlusions or improper view points.

Secondly, using only the pose-regression methods or the key-point detection methods cannot yet guarantee accurate parametric model fitting. Firstly, pose-regression methods [13, 45, 64] are able to reconstruct decent hand gestures in self-occlusion cases, but these one-shot methods cannot guarantee pixel level alignment with 2D joint positions on the image. On the other hand, keypoint-detection methods [41, 55] are capable of providing pixel-aligned geometric features for visible joints, but may need heavy post-processing optimizations, which is quite sensitive to the initialization and usually fails due to self-occlusion. To fully leverage the advantages of both categories and avoid their drawbacks, we propose a new unified two-stage parametric fitting method, in which we leverage the pose-regression result as the initial value to accelerate the convergence for parametric model fitting based on the detected keypoints, and finally achieves pixel-aligned fitting accuracy without losing the efficiency.

Last but not least, for extremely complex poses and close interactions, even 4D association [62] may fail in the body association step, which is an inherent and natural limitation for sparse multi-view setups. To this end, we propose a feedback mechanism in which the reconstructed pixel-aligned human parametric models in the previous frame are propagated into the current frame for enhancing soft visibility information and finally achieve accurate association result. Benefiting from this novel feedback mechanism, our method is able to capture accurate human behaviours even under scenarios with severe occlusions and close interactions.

Our contributions can be concluded as:

- A new hand and face bootstrapping method that involves the body-level skeleton guidance for more accurate body part localization and self-validated consistency scores to filter out the noise of fragmented detection results by unexpected view points or occlusion observations (Sec. 4).
- A new unified two-stage parametric fitting method that fully utilizes both pose-regression and keypoint-detection methods to produce accurate pixel-aligned 3D human models with expressive motion (Sec. 5).
- A new feedback mechanism that propagates the accurate reconstruction into the next frame to further improve the association accuracy especially on the severe occluded occasions (Sec. 6).

2. Related Work

2.1. Total Motion Capture

Total motion capture methods, which aim at markerless multi-scale human behaviour capture (including body motion, facial expressions and hand gestures), have shown great potentials in human 4D reconstruction and high-fidelity neural rendering [42, 49, 28, 63]. As the pioneering method of total motion capture, [22] achieved promising human behaviours capture results under the setup of hundreds of cameras, however, this method relies on the expensive and sophisticated hardware and is therefore hard for applications. On the other end of the spectrum, to achieve lightweight and convenient capture, many works [41, 55, 13, 45, 35, 64] focused on total capture from a monocular setup. Monocular total capture [55] and SMPLify-X [41] optimized parametric human models (SMPL-X [41] and Adam [22]) to fit with the 2D detected keypoints. Choutas et al. [13] directly regressed the parameters of SMPL-X [41] from a single RGB image and refined the captured results of head and hands subsequently. Pose2Pose [35] combined global and local image features for more accurate prediction. FrankMocap [45] regressed parameters of hand and body poses separately and finally integrated two parts into a unified whole body output. Zhou et al. [64] exploited the motion relationship between body and hands to design the network and achieved real-time monocular capture. Overall, although current monocular methods could achieve plausible human total capture performance, they still suffer from depth ambiguity and occlusions.

2.2. Skeleton-based Pose Reconstruction

Single-view 2D and 3D pose estimation methods [54, 43, 19, 16, 9, 12, 26, 58, 1, 34, 23, 39, 33, 60] have been widely explored in recent years, however, they suffer from severe occlusions and ambiguity and cannot produce high-confidence results. To alleviate the occlusion and produce more accurate reconstruction, many works aimed to reconstruct human poses from multi-view input. On the first branch of this direction, some approaches [17, 50, 32, 57, 31, 27, 25, 40] preformed temporal skeleton-based tracking for each frame, but these methods suffer from imperfect initialization and accumulated errors. On another branch, cross-view matching methods associated correspondences (e.g., human instances and keypoints) from different viewpoints and finally reconstructed 3D pose for each performer. Some works utilized 3DPS models to solve 3D joint positions implicitly by skeletal constraints [3, 4] or body part
et al. [21] utilized 2D detection from dense multiple views to vote for possible 3D joint positions. Dong et al. [14] proposed a multi-way matching algorithm to guarantee cycle consistency across all the views. Zhao et al. [62] jointly formulated the temporal tracking and cross-view matching as a 4D association graph and achieved real-time performance. Tu et al. [51] proposed to directly operate in the 3D space while avoiding incorrect decisions in each viewpoint. Lin et al. [30] presented a plane-sweep-based approach to perform multi-view multi-person 3D pose estimation without the explicit cross-view matching. Even though these methods are able to capture 3D human poses using skeletons, they cannot reconstruct full body behaviours, i.e., facial expressions, hand motions, and body surfaces.

2.3. 3D Hand Reconstruction

3D hand reconstruction is an essential sub-problem in total capture. Many works [47, 8, 48, 56, 65, 20, 37] that focused on 3D hand pose estimation from a single RGB image have been proposed. Recently, more and more works aimed at recovery of a 3D hand mesh [18, 24, 10] or directly regressing the pose and shape parameters of a parametric hand model (MANO [44]) [2, 6, 61, 66, 11]. However, these methods only focused on single hand reconstruction while ignoring the interaction between hands. Moon et al. [36] proposed InterHand2.6M, a large-scale two hand interaction dataset. Several researchers have explored the problem of pose estimation under two hand interacting situation [38, 52, 29, 59], but the problem of hand pose estimation under multi-person interacting scenario with more hands involved is still unsolved.

3. Overview

3.1. Main Pipeline

As shown in Fig. 2, given multiple synchronous and calibrated RGB videos as input, our pipeline works in a frame-by-frame manner, and outputs a series of parametric human models naturally combining body posture, hand gesture and facial expressions by the following steps:

1. **4D Body Association** (Sec. 3.2): Given multi-view input, we associate the 2D keypoints and triangulate 3D body skeletons using 4D association [62].

2. **Hand and Face Bootstrapping** (Sec. 4): With the body skeletons, we perform hand and face bootstrapping to extract their 2D bounding boxes efficiently and also associate them to different subjects among different views.

3. **Two-stage Parametric Fitting** (Sec. 5): Then we fit parametric human model SMPL-X [41] to these posture, gesture and expression features in a two-stage manner to achieve efficient and accurate pixel-level alignment.

4. **Feedback Mechanism** (Sec. 6): Finally, the tracked human models are propagated into the 4D association step of the next frame to further improve the association accuracy especially on severe occluded occasions.

3.2. 4D Body Association

As a building block of our method, the 4D association [62] contributes a real-time multi-person skeleton tracking framework with sparse multi-view video inputs. By taking the tracked 3D joints from the previous frame and the detected 2D key-points in current frames as graph nodes $D_j$, 4D association algorithm introduces a series of connecting edges: single-view parsing edges $E_P$, cross-view matching edges $E_V$ and temporal tracking edges $E_T$, and finally formulate a unified association graph $G_{4D}$ for optimizing the multi-view body association problem effectively.

4. Hand and Face Bootstrapping

We introduce a hand and face bootstrapping method to (i) extract local body part regions of interest (RoI) and detection from full-body inputs and (ii) eliminate incorrectly associated matches using the proposed non-maximum suppression (NMS) method. Body-level semantic features, hand pose regressions and keypoint detections are integrated into our pipeline. Note that the proposed bootstrapping methods for hand and face are quite similar, but the interactive hand behaviour is much more frequent under practical multi-person scenarios. So in this section, we mainly introduce the hand bootstrapping method which is more representative, and the method for face is similar.

Specifically, given sparse multi-view image inputs at frame $t$, we firstly leverage the 4D association algorithm (Sec. 3.2) to get the associated 2D body keypoints in each view and the triangulated 3D body skeletons. Secondly, we indicate preliminary screened RoIs $\{RoI_{a}^\alpha\}$ through body skeleton semantic information, then a lightweight object-detection network is utilized for further localizing tight and reliable RoIs $\{RoI_{b}^\beta\}$ in these initiatory screened areas to boost the key point detection and parametric regression performance of hands. However, there may exist several RoI $\beta$ corresponding to different hands in a single RoI $\alpha$, due to close interactions and bad view directions as shown in Fig. 4 (b). This will lead to severe ambiguity in the later hand association step. To eliminate these ambiguous RoIs, we propose a double-check non-maximum suppression (NMS) method to guarantee both cross-modality (between key point detection and parametric regression of hands) and cross-scale (between body reconstruction and hand reconstruction) consistency. Next, we will introduce the 2D hand localization and association in detail.

4.1. 2D Hand Localization

We conduct 2D hand localization in a coarse-to-fine manner: first generating initial bounding box for each hand according to the reconstructed 3D body skeleton and semantic information, and then refine the initial bounding
where the guidance of the reconstructed body skeletons:

\[
\{\text{RoI}_{k,\alpha}\} = \{\text{Rect}(o_p^c, r_p^c)|z_p^c = 1, p = 1, 2, ..., P\},
\]

where \(O_p\) and \(R\) is sphere center and radius. \(O_p\) could be simply extrapolated from the 3D position of wrist and elbow, and \(R\) is a constant parameter defined in terms of realistic physical scale. \(o_p^c\) and \(r_p^c\) are the projected circle center and radius in view \(c\), respectively. \(f_c\) is the focal length of camera \(c\), \(P_c(\cdot)\) is perspective projection function, and \(d_c(\cdot)\) is the distance to camera’s image plane. An indicated variable \(z_p^c\) is introduced for whether the wrist joint of person \(p\) has been assigned a 2D keypoint detection in view \(c\).

Since current hand keypoints detection and regression methods still rely on tight and accurate bounding boxes for achieving good performance, we further refine the initial bounding boxes using the iterative hand detector [53]. Note that this strategy helps us filter out the inconsequential areas efficiently at the coarse level, thus reducing unnecessary computation and accelerating the hand localization process.

For generating the initial bounding box for a hand, we leverage the reconstructed 3D body skeleton to interpolate the center of hand and construct a 3D bounding sphere with constant radius to handle size variations of hands on 2D images caused by perspective projection. We then generate the initial 2D bounding box according to the projected center and radius of the 3D bounding ball in each view. Specifically, we estimate \(\{\text{RoI}_{k,\alpha}\}\) of person \(k\) in each view \(c\) under the guidance of the reconstructed body skeletons:

\[
\text{RoI}_{k,\alpha} = \{(\text{Rect}(o_p^c, r_p^c)|z_p^c = 1, p = 1, 2, ..., P)\},
\]

where \(o_p^c\) and \(r_p^c\) are the projected circle center and radius in view \(c\), respectively. \(f_c\) is the focal length of camera \(c\), \(P_c(\cdot)\) is perspective projection function, and \(d_c(\cdot)\) is the distance to camera’s image plane. An indicated variable \(z_p^c\) is introduced for whether the wrist joint of person \(p\) has been assigned a 2D keypoint detection in view \(c\).

Since current hand keypoints detection and regression methods still rely on tight and accurate bounding boxes for achieving good performance, we further refine the initial bounding boxes using the iterative hand detector [53]. As shown in Fig. 3, we utilize that one-pass hand detection network to further extract more precise RoIs \(\text{RoI}_{k,\beta}\). We demonstrate that our two-step localization method outperforms the light-weight detector (e.g., 100DOH [46] used in FrankMocap [45]) on both the speed and accuracy (Fig. 11).

4.2. Hand Association

Since 2D body joints have been associated and 3D body skeletons of different subjects have been triangulated in previous steps, in this section, we mainly focusing on how to assign correct bounding boxes of hands to the 3D wrist joints in each view.

We leverage classical non-maximum suppression (NMS) [5] algorithm but proposed two novel consistency scores to effectively filter out ambiguous RoIs. Specifically, cross-modality consistency score \(\zeta^c\) and cross-scale consistency score \(\xi^c\) are proposed to judge which match will be finally retained. In practice, hands usually come close or even overlapped in some side views, and interactions among individuals will lead to more ambiguities. Specifically, considering the case that \(\text{RoI}_{k_1,\alpha} \cap \text{RoI}_{k_2,\alpha} \neq \emptyset, k_1 \neq k_2\), as illustrated in Fig. 4, \(\text{RoI}_{k_1,\beta}\) (dark blue) and \(\text{RoI}_{k_3,\beta}\) (red) share the same sub-region, which results in association ambiguities. Inspired by traditional NMS algorithms in object-detection pipeline, we come up with a self-validated association algorithm to filter out redundant \(\text{RoI}_{\beta}\) and re-
Figure 4. Illustration of hand association algorithm. (a) Body skeletons obtained by [62]. (b) Association ambiguity may happen when $RoI_{k1,\alpha} \cap RoI_{k2,\alpha} \neq \emptyset, 1 \neq k_2$. Blue and red dotted line rectangles are $RoI_{k1,\alpha}$ and $RoI_{k2,\alpha}$, respectively. Then 3 tight bounding boxes $RoI_{k1,\alpha}$, $RoI_{k2,\alpha}$, $RoI_{k3,\alpha}$ (dark blue) and $RoI_{k3,\beta}$ (red) are further extracted from the two initial rectangles. We can observe that the right hand lies in the overlapping area of $RoI_{k1,\alpha}$ and $RoI_{k2,\alpha}$, leading to redundant proposals and confusing partition. (c) and (d) show that pose-regression network is specific for one chirality input. (e) is the result of heatmap-based detector which is invariant with chirality. (f) shows that after our double-check NMS procedure, the correct distributed $RoI_{k1,\alpha}$ are retained and the false one $RoI_{k1,\beta}$ is discarded.

 Maintaining the correct match. Firstly, for each view $c$, we locate all redundant RoIs by calculating Intersection of Union (IoU) from total individuals’ hands proposals. Secondly, we calculate cross-modality consistency score $\zeta_c$ and cross-scale consistency score $\xi_c$ for each RoI proposal.

The first metric, cross-modality consistency score $\zeta_c$, is used to penalize the inconsistency between different detection modalities. As shown in Fig. 4 (c), (d) and (e), the heatmap-based feature is invariant with respect to flipping translation, but pose-regression network needs right chirality assurance for achieving reasonable results. This divergence can help us to distinguish the left or right association ambiguities. Denote $J_h = 21$ as the hand joint number, $S_{h,\text{reg}}^b \in \mathbb{R}^{2 \times J_h}$ as 2D hand joint positions from pose-regression network, $S_{h,\text{dect}}^b \in \mathbb{R}^{2 \times J_h}$ as the output of keypoint-detection network, and $w, h$ as the size of $RoI_{\beta}$. Then $\zeta$ is formulated as

$$\zeta = \frac{1}{J_h} \sum_{j=1}^{J_h} \max(0, 1 - \frac{2\|S_{\text{reg},j}^b - S_{\text{dect},j}^b\|_2}{\sqrt{w^2 + h^2}}). \quad (2)$$

On the other hand, the second metric, cross-scale consistency score $\xi$, is formulated to punish unreasonable wrist misalignment between the local hand estimator and global body estimator. Denote $S^h_{\text{wrt}} \in \mathbb{R}^2$ as the associated 2D wrist position from full-body detections, $S^h_{\text{dect},jw}$ as wrist joint position by local hand keypoints detector. Finally, $\xi$ is defined as

$$\xi = \max(0, 1 - \frac{2\|S^h_{\text{wrt}} - S^h_{\text{dect},jw}\|_2}{\sqrt{w^2 + h^2}}). \quad (3)$$

Finally, we sum up these two scores as confidence measurements to apply the NMS algorithm to reserve the one with the highest score. We demonstrate that our double-check NMS method helps to improve the association accuracy in confusing situations.

5. Two-stage Parametric Fitting

We observe that previous methods usually utilize either parametric pose regression [45] or heatmap-based keypoints [41] for total motion capture. However, they do have their own limitations. On one hand, although pose-regression networks can produce plausible results even under occlusions, they can not guarantee accurate 2D alignment with the input image. On the other hand, heatmap-based networks provide accurate 2D detections for visible joints, but they are still suffer from depth ambiguities and are susceptible to local minima during optimization. In this paper, we unify them together in a two-stage parametric fitting scheme, which contains local initialization and total optimization, to boost the total motion capture performance.

Local Initialization To accelerate convergence and prevent optimization deviation, it is essential to initialize the motion of each body part to a reasonable status. Specifically, for hand pose initialization, we pick a decent initial value from the semantic pose-regression gestures according to the hand association score $\zeta$ (Eqn. 2) and $\xi$ (Eqn. 3). Besides, for body/head pose initialization, we solve the SMPL-X body pose by minimizing the following energy function directly to guarantee more accurate initialization:

$$E_{\text{body}} = \lambda_{b3d}E_{b3d} + \lambda_{pr1}E_{pr1} + \lambda_{b2d}E_{b2d} \quad (4)$$

Here, $E_{b3d}$ is the distance from parametric model’s joints to the corresponding reconstructed 3D body skeletons. $E_{pr1}$ and $E_{b2d}$ are used to regularize natural pose and shape optimization as in SMPLify-X [41].

Total Optimization In this stage, we leverage the accurate 2D hands keypoints and faces landmarkers to further optimize the initial SMPL-X model for accurate total capture:

$$E_{\text{total}} = E_{\text{data}} + E_{\text{reg}}, \quad E_{\text{data}} = \lambda_{b3d}E_{b3d} + \lambda_{b2d}E_{b2d} + \lambda_{f2d}E_{f2d},$$

$$E_{\text{reg}} = \lambda_{pr1}E_{pr1} + \lambda_{b2d}E_{b2d} + \lambda_{h,h}E_{h,h} + \lambda_{f2d}E_{f2d},$$

where $E_{b2d}$ and $E_{f2d}$ are 2D data terms to minimize the distances between the 2D projections of the SMPL-X joints.
and the detected 2D keypoints in all the valid viewpoints. $E_{\theta,h}$ and $E_z$ are L-2 norm to keep the optimized gestures and expressions within reasonable ranges. Note that we can additionally leverage the consistency scores $\zeta$ and $\xi$ (Eqn. 2 and 3) to balance the detection results in different views, so $E_{h2d} = \sum_i \sum_c \frac{\zeta + \xi}{2} \cdot e_{h2d}$, where $c$ is the view index.

6. Feedback Mechanism

Last but not least, for severe occlusions and close interactions, we put forward a feedback mechanism to boost the tracking performance of the association algorithm in return. On one hand, detailed limb detector contributes to extremity reconstruction with higher precision, which are leveraged to refine the body skeleton results. On the other hand, we re-render the human model to each view for the next frame to extend the tracking edges $E_T$ of $G_{4D}$ with additional visibility information. As shown in Fig. 6, we obtain the initial segmentation by rendering the optimized parametric models back to input images. Meanwhile, in order to enhance the robustness with body movements, we implement distance transformation to smooth the boundary of the rendered mask.

For a given 2D keypoint detection candidate $c$, we use the same denotation $z^k(c)$ in [62] to refer the possibility to connect that candidate to person $k$. Benefiting from our feedback module, the tracking edges in 4D association [62] (Sec. 3.2) are extended with visibility priors. We define the enhanced tracking edges $\tilde{z}^k(c)$ as:

$$\tilde{z}^k(c) = \frac{\tau^k(c)}{\sum_{i=1}^{K} \tau^i(c)} z^k(c),$$

where $\tau^i(c) \in [0,1]$ means the continuous occupancy of person $i$. As shown in Fig. 6 (d), $\tau^i(c)$ is negative interrelated to the distance to its binary mask, $\tau^k(c) = 1$ refers to the fully contained situation. As a result, our feed-back mechanism enhances skeleton tracking performance and reduces jittering in projection coincidence cases.

7. Results

In Fig. 7, we demonstrate example results by our system. With the sparse multi-view setup, our method produces expressive human parametric models under multi-person interactive scenarios.

7.1. Implementation Details

Our light-weight total capture system are implemented with 6 synchronized RGB cameras (resolution $2048 \times 2048$) on a single PC (i5-6600K CPU, NVIDIA RTX 3090 GPU). We use Openpose [9] as our body pose estimator, SRHandNet [53] as our hand instance detector and keypoint detector. We leverage the hand pose-regression network of Frankmocap [45] gesture regressor. FaceAlignment [7] are used for face keypoints extraction. Besides, we accelerate all neural network inference by implementing half-precision arithmetic on NVIDIA TensorRT platform. The CNN performance is shown in Table. 1. Besides, our body association backbone takes nearly 10ms to recover human skeletons, the limb localization and association method is fast enough to be neglected. Our parametric fitting workflow takes 150 ms for stage one and 350 ms for stage two (20 Gauss-Newton iterations and parallel for each person). On the whole, our system run-time depends on the captured individual number and view number. Empirically, our pipeline runs about 1 fps for 2 person with 6 views, and the processing speed slows down to 0.3 fps for 7 persons with 8 views. As for hyper-parameters, sphere radius $R$ in hand localization is set to 0.15$m$, and association NMS threshold is 0.5. In two-stage parametric fitting, we set $\lambda_{h3d} = 10$, $\lambda_{h2d} = 0.0001$, $\lambda_{f2d} = 0.0003$, $\lambda_{pri} = \lambda_{\theta,h} = 0.01$, and $\lambda_{\beta} = \lambda_{\varepsilon} = 0.01$.

<table>
<thead>
<tr>
<th>Network</th>
<th>Input</th>
<th>Batchsize</th>
<th>Speed(FPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Openpose [9]</td>
<td>368 $\times$ 368</td>
<td>6</td>
<td>43.1</td>
</tr>
<tr>
<td>FaceAlignment [7]</td>
<td>256 $\times$ 256</td>
<td>4</td>
<td>109.5</td>
</tr>
<tr>
<td>SRHandNet [53]</td>
<td>256 $\times$ 256</td>
<td>8</td>
<td>50.0</td>
</tr>
<tr>
<td>HandHMR [45]</td>
<td>224 $\times$ 224</td>
<td>8</td>
<td>202.1</td>
</tr>
</tbody>
</table>

Table 1. Inference speed of the CNN networks used in our system.
Figure 7. Results by our system. From the left to right are input reference images, parametric model alignment, facial and hand alignment and 3D visualization from a novel view, respectively. (a) Results of the hand-object-interaction case from our captured data using 6 views, (b) results of a multi-person-interaction scenario using 6 views, (c) results on CMU dataset [22] using 8 views.

7.2. Comparison

Since our method is the first to enable lightweight total capture from sparse multi-view, we compared our method with SOTA single view method FrankMocap [45] in Fig. 11 and ground truth from Total Capture [22] in Fig. 9. What’s more, MPJPE (Mean Per Joint Position Error) are provided for Total Capture dataset Tab. 2.

7.3. Evaluation: Hand Bootstrapping

We compare our hand bootstrapping method with SOTA monocular total capture method Frankmocap [45]. To ensure fairness as much as possible, we reduce our system to 2 close front view camera. Fig. 11 (a) shows the reconstruction failure of FrankMocap [45] caused by mixing up left and right hands to the same region proposal. Thanks to the proposed NMS method in hand association, our method can robustly reconstruct more accurate hands in Fig. 11 (b).

7.4. Evaluation: Two-stage Parametric Fitting

We conduct ablation study of two-stage fitting metric on CMU dataset [55] and demonstrate that our method makes different modality detectors benefit from each other. On the one hand, as shown in Fig. 8 (a)(d), we perform our two-stage fitting algorithm with only pose regression results, namely we leverage orthogonal projected joints from pose-regression network to take over heatmap-based 2D correspondences in stage two. Misalignment artifacts are shown in detail as pose-regression detector could not guarantee pixel-aligned accuracy. One the other hand, keypoints-detection-only results are shown in Fig. 8 (b)(e). Without pose regression network to initialize hand poses with reasonable gestures, the optimization is easy to fall into a local minimum.

7.5. Evaluation: Feedback Mechanism

We evaluate our feedback module in Shelf dataset [3]. As shown in Fig. 10(a), the left elbow of the salmon person is distributed to the background green person without feed-
Figure 8. Qualitative evaluation of two-stage parametric fitting. (a) and (d) are the results with parametric-regression-only metric (blue). (b) and (e) are the results with keypoints-detection-only metric (salmon). (c) and (f) are the results with our two-stage fitting strategy combing both metrics (pink). Meanwhile, we visualize pose-regression network outputs and heatmap-based network outputs in right of (a) and (b), respectively. Red dots in right of (d) (e) (f) refer to the ground-truth 3D hand annotations.

Figure 9. Comparison of our sparse-view method(8 view used) to ground truth from Total Capture Dataset[22]. The mesh refers to our reconstructed parametric model (SMPL-X), and the red keypoints are the ground truth from the Total Capture dataset.

Figure 10. Qualitative evaluation of feedback mechanism. (a) shows the original association results of [62]. (b) are our reconstructed model of last frame. (c) shows that our feedback mechanism boosts body association performance with visibility prior.

Table 2. Quantitative evaluation on Total Capture dataset. We calculate the MPJPE of body&head joints on a video segment (750 frames) which involves both large variation movements and meticulous hand gestures (noted that hand annotations are few for challenge gestures) with 5 cameras for a comprehensive evaluation.

<table>
<thead>
<tr>
<th>Type</th>
<th>Body</th>
<th>Head</th>
<th>LHand</th>
<th>RHand</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPJPE(mm)</td>
<td>33.4</td>
<td>21.7</td>
<td>22.6</td>
<td>19.3</td>
</tr>
</tbody>
</table>

Table 3. Ablation study of feedback mechanism on Shelf dataset. Numbers are percentage of correct parts(PCP).

<table>
<thead>
<tr>
<th>Shelf</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o feedback</td>
<td>99.0</td>
<td>96.2</td>
<td>97.6</td>
<td>97.6</td>
</tr>
<tr>
<td>w/ feedback</td>
<td>99.5</td>
<td>97.0</td>
<td>97.8</td>
<td>98.1</td>
</tr>
</tbody>
</table>

8. Discussion

Conclusion In this paper, we propose, as far as we know, the first multi-person total motion capture framework with only a sparse multi-view setup. Based on the proposed hand and face bootstrapping, two-stage parametric fitting and feedback mechanism, our method can enable lightweight, fast, robust and accurate capture of the body pose, hand gesture and facial expression of each character even under the scenarios with severe occlusions and close interactions.

Limitation and Future Work We can mainly recover the facial expression by the jaw joint, and cannot reconstruct subtle facial expressions due to the low-resolution facial image input, which we leave for future research.

Acknowledgements This work is supported by the National Key Research and Development Program of China No.2018YFB2100500 and the NSFC No.62125107, No.61827805, No.62171255, and the Shuimu Tsinghua Scholarship.