Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies*

Hanbyul Joo[1] Tomas Simon[1,2] Yaser Sheikh[1,2]
[1]Carnegie Mellon University [2]Facebook Reality Labs, Pittsburgh
hanbyulj@cs.cmu.edu, {tomas.simon, yaser.sheikh}@fb.com

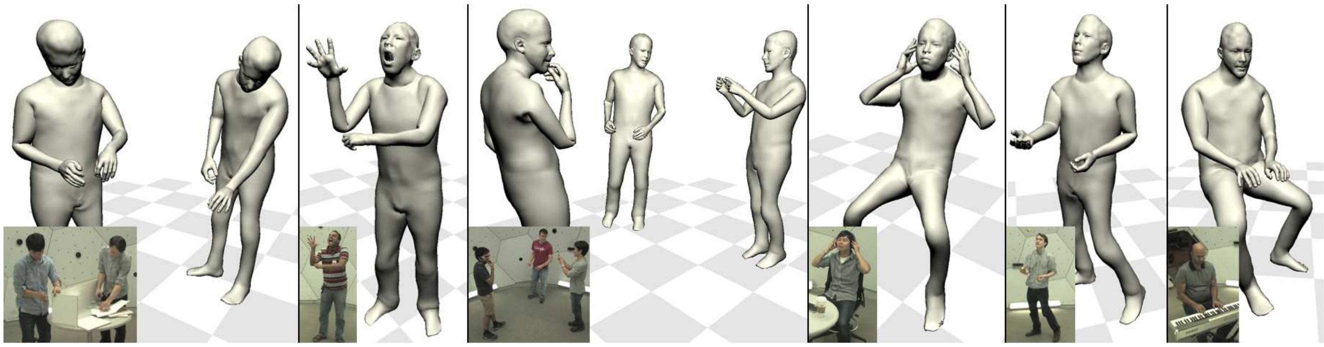


Figure 1: Reconstructing total body motion by Adam. This paper presents a 3D human model capable of concurrently tracking the large-scale posture of the body along with the smaller details of a person's facial expressions and hand gestures.

Abstract

We present a unified deformation model for the markerless capture of human movement at multiple scales, including facial expressions, body motion, and hand gestures. An initial model is generated by locally stitching together models of the individual parts of the human body, which we refer to as "Frank". This model enables the full expression of part movements, including face and hands, by a single seamless model. We capture a dataset of people wearing everyday clothes and optimize the Frank model to create "Adam": a calibrated model that shares the same skeleton hierarchy as the initial model with a simpler parameterization. Finally, we demonstrate the use of these models for total motion tracking in a multiview setup, simultaneously capturing the large-scale body movements and the subtle face and hand motion of a social group of people.

1. Introduction

Social communication is a key function of human motion [9]. We communicate tremendous amounts of information with the subtlest movements. Between a group of interacting individuals, gestures such as a gentle shrug of the shoulders, a quick turn of the head, or an uneasy shifting of weight from foot to foot, all transmit critical information about the attention, emotion, and intention to observers. Notably, these social signals are usually transmitted by the organized motion of the whole body: with facial expressions, hand gestures, and body posture. These rich signals layer upon goal-directed activity in constructing the behavior of humans, and are therefore crucial for the machine perception of human activity.

However, there are no existing systems that can track, without markers, the human body, face, and hands simultaneously. Current markerless motion capture systems focus at a particular scale or on a particular part. Each area has its own preferred capture configuration: (1) torso and limb motions are captured in a sufficiently large working volume where people can freely move [19, 23, 47, 21]; (2) facial motion is captured at close range, mostly frontal, and assuming little global head motion [7, 26, 8, 11, 54]; (3) finger motion is also captured at very close distances from hands, where the hand regions are dominant in the sensor measurements [37, 52, 45, 53]. These configurations make it difficult to concurrently analyze the full spectrum of social signalling.

To overcome this sensing challenge, we present a novel generative body deformation model that has the ability to express the motion of each principal body part. In particu-

*Website: http://www.cs.cmu.edu/~hanbyulj/totalcapture lar, we describe a procedure to build an initial body model, named "Frank", by seamlessly consolidating available part template models [34, 15] into a single skeleton hierarchy. To fit this model to data, we leverage keypoint detection (e.g., faces [20], bodies [58, 16, 36], and hands [44]) in multiple views to obtain 3D keypoints which are robust to multiple people and object interactions. We fit the "Frank" model to a capture of 70 people, and learn a new deformation model, named "Adam", capable of additionally capturing variations of hair and clothing with a simplified parameterization. We present a method to capture the total body motion of multiple people with the 3D deformable model. Finally, we demonstrate the performance of our method on various sequences of social behavior and person-object interactions, where the combination of face, limb, and finger motion emerges naturally.

2. Related Work

Marker-based motion capture systems that track retroreflective markers [2, 59] are the most widely used method to capture human body motion. However, in addition to a laborious process of attaching markers on subjects, these methods still suffer from major limitations including: (1) a necessity of sparsity in marker density for reliable tracking, which limits the spatial resolution of motion measurements [38]; (2) a limitation in automatically handling occluded markers which requires expensive manual clean-up; and (3) markers on the faces, bodies, and hands hinder participants from engaging in natural social interaction. Due to these limitations, capturing the total body motion of interacting people is still a challenging problem even in state-of-the-art motion capture systems [2].

Markerless motion capture methods have been explored over the past two decades to achieve the same goal of motion capture systems, but they tend to implicitly admit that their performance is inferior to their marker-based counterpart, advocating their "markerless" nature as the major advantage. Most markerless motion capture methods largely focus on the motion of the torso and limbs. The standard pipeline is based on a multiview camera setup and tracking with a 3D template model [33, 25, 17, 12, 30, 18, 55, 13, 47, 19, 21]. In this approach, motion capture is performed by aligning a 3D template model to the measurements, which can include colors, textures, silhouettes, point clouds, and keypoints. Recent methods exploit a generative deformable body model [4, 34, 40] to express both shape and body variations of humans. Since these body models often assume minimum clothing for subjects, explicit modeling for clothing is needed to capture clothed subjects [63, 39]. Recent advances in 2D keypoint detection [36, 16, 58] make it possible to reliably reconstruct 3D keypoints in a multiview setup, where a 3D model can be fitted [21, 28, 29]. A specific strength of learning-based detectors is that they can provide a "guess" for occluded parts, based on the spatial human body configurations learned from a large-scale 2D pose dataset. Note that we differentiate markerless motion capture approaches, producing motion parameters as output, from multiview performance capture approaches [56, 22] which aim to obtain detailed surface shapes by free-form mesh deformations. With the introduction of commodity depth sensors, single-view depth-based body motion capture also became a popular direction [5, 43]. More recently, a collection of approaches aims to reconstruct 3D skeletons directly from monocular images, either by fitting 2D keypoint detections with a prior on human pose [64, 10] or getting even closer to direct regression methods [65, 35, 51].

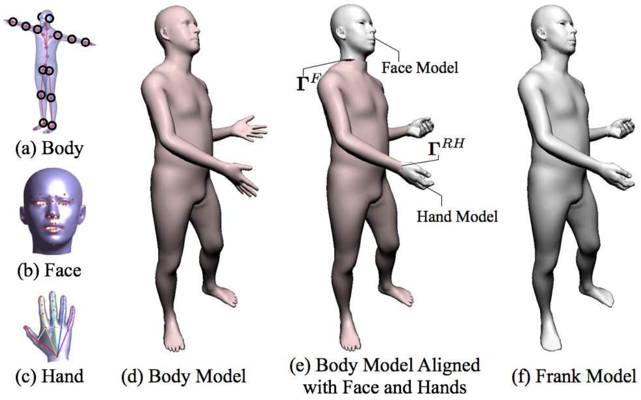
In all earlier work, face and hand motion captures are often considered as separate research domains. Facial scanning and performance capture has been greatly advanced over the last decade. There exist multiview methods showing excellent performance on high-quality facial scanning [7, 26] and facial motion capture [8, 11, 54]. Recently, lightweight systems based on a single camera show compelling performance by leveraging a morphable 3D face model on 2D measurements [24, 20, 32, 50, 15, 14, 60]. Most of these methods are based on a deformable 3D face rig such as the method of Cao et al. [15]. Hand motion capture is mostly led by single depth-sensor based methods [37, 49, 52, 31, 61, 48, 57, 46, 42, 45, 53, 62], with few exceptions based on multi-view systems [6, 46, 41]. Recently, 2D hand keypoint detection and the use of it to obtain 3D hand keypoints in a multiview setup are introduced by Simon et al. [44]. Notably, a generative 3D model that can express body and hands was also introduced by Romero et al. [41].

In contrast, this paper presents the first approach for "total" markerless motion capture of multiple interacting people, producing a parameterized representation that jointly captures the time-varying body pose, hand pose, and facial expressions of each of the interacting participants.

3. Frank Model

The motivation for building the Frank[1] body model is to leverage existing part models: SMPL [34] for the body, FaceWarehouse [15] for the face, and an artist-defined hand rig (shown in Fig. 2). Each of these capture shape and motion details at an appropriate scale for the corresponding part. This choice is not driven merely by the free availability of the component models: note that due to the trade-off between image resolution and field of view of today's 3D scanning systems, scans used to build detailed face models will generally be captured using a different system than that used for the rest of the body. For our model, we merge

[1] Frank is an homage to a certain *Modern Prometheus*.



(a) Body
(b) Face
(c) Hand
(d) Body Model
(e) Body Model Aligned with Face and Hands
(f) Frank Model

Figure 2: Part models and the Frank model. (a) The body model [34]; (b) the face model [15]; and (c) a hand rig. In (a-c), the red dots have corresponding 3D keypoints reconstructed by detectors; (d) Body only model; (e) Face and hand models substitute the corresponding parts of the body model. Alignments are ensured by $\mathbf{\Gamma}$s; and (f) The blending matrix $\mathbf{C}$ is applied to produce a seamless mesh.

all transform bones into a single skeletal hierarchy but keep the native parameterization of each component part to express identity and motion variations. As the final output, the Frank model produces motion parameters capturing the total body motion of humans, and generates a seamless mesh by blending the vertices of the component meshes.

3.1. Stitching Part Models

The Frank model M^U is parameterized by motion parameters $\boldsymbol{\theta}^U$, shape (or identity) parameters $\boldsymbol{\phi}^U$, and a global translation parameter $\mathbf{t}^U$,

$$\mathbf{V}^U = M^U(\boldsymbol{\theta}^U, \boldsymbol{\phi}^U, \mathbf{t}^U), \qquad (1)$$

where $\mathbf{V}^U$ is a seamless mesh expressing the motion and shape of the target subject. The motion and shape parameters of the model are a union of the part models' parameters:

$$\boldsymbol{\theta}^U = \{\boldsymbol{\theta}^B, \boldsymbol{\theta}^F, \boldsymbol{\theta}^{LH}, \boldsymbol{\theta}^{RH}\}, \qquad (2)$$

$$\boldsymbol{\phi}^U = \{\boldsymbol{\phi}^B, \boldsymbol{\phi}^F, \boldsymbol{\phi}^{LH}, \boldsymbol{\phi}^{RH}\}, \qquad (3)$$

where the superscripts represent each part model: B for the body model, F for the face model, LH for the left hand model, and RH for the right hand model. Each of the component part models maps from a subset of the above parameters to a set of vertices, respectively, $\mathbf{V}^B \in \mathbb{R}^{N^B \times 3}$, $\mathbf{V}^F \in \mathbb{R}^{N^F \times 3}$, $\mathbf{V}^{LH} \in \mathbb{R}^{N^H \times 3}$, and $\mathbf{V}^{RH} \in \mathbb{R}^{N^H \times 3}$, where the number of vertices of each mesh part is N^B=6890, N^F=11510, and N^H=2068. The final mesh of the Frank model, $\mathbf{V}^U \in \mathbb{R}^{N^U \times 3}$, is defined by linearly blending them with a matrix $\mathbf{C} \in \mathbb{R}^{N^U \times (N^B + N^F + 2N^H)}$:

$$\mathbf{V}^U = \mathbf{C} \begin{bmatrix} \left(\mathbf{V}^B\right)^T & \left(\mathbf{V}^F\right)^T & \left(\mathbf{V}^{LH}\right)^T & \left(\mathbf{V}^{RH}\right)^T \end{bmatrix}^T, \qquad (4)$$

where T denotes the transpose of a matrix. Note that $\mathbf{V}^U$ has fewer vertices than the sum of part models because there are redundant parts in the body model (e.g., face and hands of the body model). In particular, our final mesh has N^U=18540 vertices. Fig. 2 (e) shows the part models that are aligned, and (f) shows the final mesh topology of the Frank model after applying the the blending matrix $\mathbf{C}$ at the mean shape in the rest pose. The blending matrix $\mathbf{C}$ is a very sparse matrix; most rows have a single column set to one with zeros elsewhere and simply copy the vertex locations from the corresponding part models with minimal interpolation at the seams.

In the Frank model, all parts are rigidly linked by a single skeletal hierarchy, which is crucial as an output of motion capture. This unification is achieved by substituting the hands and face branches of the SMPL body skeleton with the corresponding skeletal hierarchies of the detailed part models. All parameters of the Frank model are jointly optimized for motion tracking and identity fitting. The parameterization of each of the part models is detailed in the following sections.

3.2. Body Model

For the body, we use the SMPL model [34] with minor modifications. In this section, we summarize the salient aspects of the model in our notation. The body model, M^B, is defined as follows,

$$\mathbf{V}^B = M^B(\boldsymbol{\theta}^B, \boldsymbol{\phi}^B, \boldsymbol{t}^B), \qquad (5)$$

with $\mathbf{V}^B = \{\mathbf{v}_i^B\}_{i=1}^{N^B}$. The model uses a template mesh of N^B=6890 vertices, where we denote the i-th vertex as $\mathbf{v}_i^B \in \mathbb{R}^3$. The vertices of this template mesh are first displaced by a set of blendshapes describing the *identity* or body shape. Given the vertices in the rest pose, the posed mesh vertices are obtained by linear blend skinning (LBS) using transformation matrices $\mathbf{T}_j^B \in$ SE(3) for each of the J joints,

$$\mathbf{v}_i^B = \mathbf{I}_{3\times 4} \cdot \sum_{j=1}^{J^B} w_{i,j}^B \mathbf{T}_j^B \begin{pmatrix} \mathbf{v}_i^{B0} + \sum_{k=1}^{K_b} \mathbf{b}_i^k \phi_k^B \\ 1 \end{pmatrix}, \qquad (6)$$

where $\mathbf{b}_i^k \in \mathbb{R}^3$ is the i-th vertex of the k-th blendshape, ϕ_k^B is the k-th shape coefficient in $\boldsymbol{\phi}^B \in \mathbb{R}^{K_b}$ with K_b=10 the number of identity body shape coefficients, and $\mathbf{v}_i^{B0}$ is the i-th vertex of the mean shape. The transformation matrices $\mathbf{T}_j^B$ encode the transform for each joint j from the rest pose to the posed mesh in world coordinates, which is constructed by traversing the skeleton hierarchy from the root joint with pose parameter $\boldsymbol{\theta}^B$ (see [34]). The j-th pose parameter θ_j^B is the angle-axis representation of the relative rotation of joint j with respect to its parent joints. $w_{i,j}^B$ is the weight with which transform $\mathbf{T}_j^B$ affects vertex i, with $\sum_{j=1}^{J^B} w_{i,j}^B{=}1$ and $\mathbf{I}_{3\times 4}$ is the 3×4 truncated identity matrix to transform from homogeneous coordinates to a 3 dimensional vector. We use $J^B{=}21$ with $\boldsymbol{\theta}^B \in \mathbb{R}^{21\times 3}$, ignoring the last joint of each hand of the SMPL model. For simplicity, we do not use the pose-dependent blendshapes[2].

3.3. Face Model

As a face model, we build a generative PCA model from the FaceWarehouse dataset [15]. Specifically, the face part model, M^F, is defined as follows,

$$\mathbf{V}^F = M^F(\boldsymbol{\theta}^F, \boldsymbol{\phi}^F, \mathbf{T}^F), \tag{7}$$

with $\mathbf{V}^F = \{\mathbf{v}_i^F\}_{i=1}^{N^F}$, where the i-th vertex is $\mathbf{v}_i^F \in \mathbb{R}^3$, and $N^F{=}11510$. The vertices are represented by combining shape and expression subspaces:

$$\hat{\mathbf{v}}_i^F = \mathbf{v}_i^{F0} + \sum_{k=1}^{K_f} \mathbf{f}_i^k \phi_k^F + \sum_{s=1}^{K_e} \mathbf{e}_i^s \theta_s^F \tag{8}$$

where, as before, $\mathbf{v}_i^{F0}$ denotes i-th vertex of the mean shape, and ϕ_k^F and θ_s^F are the k-th face identity (shape) and s-th facial expression (pose) parameters respectively. Here, $\mathbf{f}_i^k \in \mathbb{R}^3$ is the i-th vertex of the k-th identity blendshape ($K_f = 150$), and $\mathbf{e}_i^s \in \mathbb{R}^3$ is the i-th vertex of the s-th expression blendshape ($K_e = 200$).

Finally, a transformation $\mathbf{T}^F$ brings the face vertices into world coordinates. To ensure that the face vertices transform in accordance to the rest of the body, we assume that the mean face $\mathbf{v}_i^{F0}$ is aligned with the body mean shape as shown in Fig. 2, which is manually done in building the model. This way, we can apply the transformation of the body model's head joint $\mathbf{T}_{j=F}^B(\boldsymbol{\theta}^B)$ as a global transformation for the face model in Eq. 9. However, to keep the face in alignment with the body, an additional transform matrix $\boldsymbol{\Gamma}^F \in$ SE(3) is required to compensate for displacements in the root location of the face joint due to body shape changes in Eq. 6.

Finally, each face vertex position is given by:

$$\mathbf{v}_i^F = \mathbf{I}_{3\times 4} \cdot \mathbf{T}_{j=F}^B \cdot \boldsymbol{\Gamma}^F \begin{pmatrix} \hat{\mathbf{v}}_i^F \\ 1 \end{pmatrix}, \tag{9}$$

where the transform $\boldsymbol{\Gamma}^F$, which is directly determined by the body shape parameters $\boldsymbol{\phi}^B$, aligns the face model with the body model.

3.4. Hand Model

We use an artist-rigged hand mesh. Our hand model has $J^H{=}16$ joints and the mesh is again deformed via linear blend skinning. The hand model has a fixed shape, but we introduce scaling parameters for each bone to allow for different finger sizes. The transform for the j-th joint is parameterized by the Euler angle rotation with respect to its parent, $\boldsymbol{\theta}_j^H \in \mathbb{R}^3$, and an additional anisotropic scaling factor along each axis, $\boldsymbol{\phi}_j^H \in \mathbb{R}^3$. Specifically, the linear transform for the j-th joint in the bone's local reference frame becomes $\mathrm{eul}(\boldsymbol{\theta}_j^H)\cdot\mathrm{diag}(\boldsymbol{s}_j^H)$, where $\mathrm{eul}(\boldsymbol{\theta}_j^H)$ converts from an Euler angle representation to a 3×3 rotation matrix and $\mathrm{diag}(\boldsymbol{\phi}_j^H)$ is the 3×3 diagonal matrix with the X,Y,Z scaling factors ϕ_j^H on the diagonal. The vertices of the hand in world coordinates are given by LBS with weights $w_{i,j}^H$:

$$\mathbf{v}_i^H = \mathbf{I}_{3\times 4} \cdot \mathbf{T}_{j=H}^B \cdot \boldsymbol{\Gamma}^H \cdot \sum_{j=1}^{J} w_{i,j}^H \mathbf{T}_j^H \begin{pmatrix} \mathbf{v}_i^{H0} \\ 1 \end{pmatrix}. \tag{10}$$

where $\mathbf{v}_i^{H0}$ denotes i-th vertex of the mean shape , $\mathbf{T}_j^H$ is each bone's composed transform (with all parents in the hierarchy), $\mathbf{T}_{j=H}^B \in$ SE(3) is the transformation of the corresponding hand joint in the body model, and $\boldsymbol{\Gamma}^H$ is the transformation that aligns the hand model to the body model. As with the face, this transform depends on the shape parameters of the body model.

4. Motion Capture with Frank

We fit the Frank model to data to capture the total body motion, including the major limbs, the face, and fingers. Our motion capture method relies heavily on fitting mesh correspondences to 3D keypoints, which are obtained by triangulation of 2D keypoint detections across multiple camera views. To capture shape information we also use point clouds generated by multiview stereo reconstructions. Model fitting is performed by an optimization framework to minimize distances between corresponded model joints and surface points and 3D keypoint detections, and iterative closest point (ICP) to the 3D point cloud. Note that more details are provided in the supplementary material.

4.1. 3D Measurements

We incorporate two types of measurements in our framework as shown in Fig. 3: (1) corresponded 3D keypoints, which map to known joints or surface points on the mesh models (see Fig. 2), and (2) uncorresponded 3D points from multiview stereo reconstruction, which we match using ICP.

3D Body, Face, and Hand Keypoints: We use the OpenPose detector [27] in each available view, which produces 2D keypoints on the body with the method of Cao et al. [16], and hand and face keypoints using the method of

[2] For our target sequences, the modeling error between the SMPL model [34] and the 3D surface measurements is dominated by clothing artifacts, which the pose-blendshapes were not trained on.

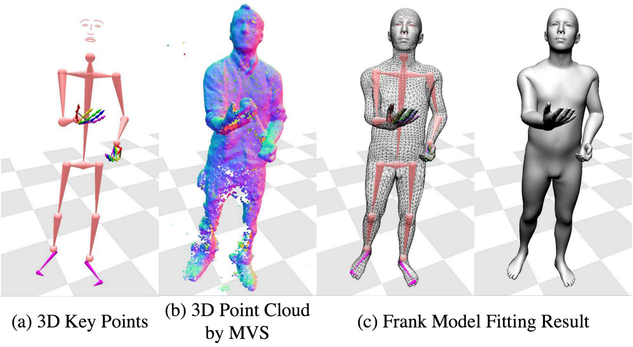


Figure 3: Fitting Frank: The optimization takes, as input, (a) 3D keypoints, and (b) point clouds, and produces (c) a fitted skeleton and mesh as output.

Simon et al. [44]. 3D body skeletons are obtained from the 2D detections using the method of [29], which uses known camera calibration parameters for reconstruction. The 3D hand keypoints are obtained by triangulating 2D hand pose detections, following the method of [44], and similarly for the facial keypoints. Note that subsets of 3D keypoints can be entirely missing if there are not enough 2D detections for triangulation, which can happen in challenging scenes with inter-occlusions or motion blur.

3D Feet Keypoints: An important cue missing from the OpenPose detector is keypoints on the feet. For motion capture, this is an essential feature to accurately determine the orientation of the feet. We therefore train a keypoint detector for the tip of the big toe, the tip of the little toe, and the ball of the foot. We annotate these 3 keypoints per foot in each of around 5000 person instances of the COCO dataset, and use the architecture of Wei et al. [58] with a bounding box around the feet determined by the 3D body detections. We also apply multiview bootstrapping in the Panoptic Studio to improve the quality, as described by Simon et al. [44].

3D Point Clouds: We use the commercial software RealityCapture [1] to obtain 3D point clouds from the multiview images, with associated point normals.

4.2. Objective Function

We initially fit every frame in the sequence independently. For clarity, we drop the time index from the notation and describe the process for a single frame, which optimizes the following cost function:

$$E\left(\boldsymbol{\theta}^U, \boldsymbol{\phi}^U, \boldsymbol{t}^U\right) = E_{\text{keypoints}} + E_{\text{icp}} + E_{\text{seam}} + E_{\text{prior}} \quad (11)$$

We use Levenberg-Marquardt with the Ceres Solver library [3] with multiple stages to avoid local minima. See the supplementary material for the details.

Anatomical Keypoint Cost: The term $E_{\text{keypoints}}$ matches 3D keypoint detections, which are in direct corresponce to our mesh models. This term includes joints (or end effectors) in the body and hands, and also contains points corresponding to the surface of the mesh (e.g., facial keypoints and the tips of fingers and toes). Both of these types of correspondence are expressed as combinations of vertices via a regression matrix $\mathbf{J} \in \mathbb{R}^{C \times N^U}$, where C denotes the number of correspondences and N^U is the number of vertices in the model. Let $\mathcal{D}$ denote the set of available detections in a particular frame. The cost is then:

$$E_{\text{keypoints}} = \lambda_{\text{keypoints}} \sum_{i \in \mathcal{D}} ||\mathbf{J}_i \mathbf{V} - \mathbf{y}_i^T||^2, \quad (12)$$

where $\mathbf{J}_i$ indexes a row in the correspondence regression matrix and represents an interpolated position using a small number of vertices, and $\mathbf{y}_i \in \mathbb{R}^{3 \times 1}$ is the 3D detection. The $\lambda_{\text{keypoints}}$ is a relative weight for this term.

ICP Cost: The 3D point cloud measurements are not a priori in correspondence with the model meshes. We therefore establish their correspondence to the mesh using Iterative Closest Point (ICP) during each solver iteration. We find the closest 3D point in the point cloud to each of the mesh vertices, and compute the point-to-plane residual, i.e., the distance along the normal direction,

$$E_{\text{icp}} = \lambda_{\text{icp}} \sum_{\mathbf{v}_j \in \mathbf{V}^U} \mathbf{n}(\mathbf{x}_{j^*})^T (\mathbf{x}_{j^*} - \mathbf{v}_j), \quad (13)$$

where $\mathbf{x}_{i^*}$ is the closest 3D point to j-th vertex $\mathbf{v}_j$, $\mathbf{n}(\cdot) \in \mathbb{R}^3$ represents the point's normal, and λ_{icp} is a relative weight for this term.

Seam Constraints: The part models composing the Frank model are rigidly linked by the skeletal hierarchy. However, the independent surface parameterizations of each of the part models may introduce discontinuities at the boundary between parts (e.g., a fat arm with a thin wrist). To avoid this artifact, we encourage the vertices around the seam parts to be close by penalizing differences between the last two rings of vertices around the seam of each part, and the corresponding closest point in the body model in the rest pose expressed as barycentric coordinates.

Prior Cost: Depending on the number of measurements available in a particular frame, the set of parameters of M^U may not be determined uniquely (e.g., the width of the fingers). More importantly, the 3D point clouds are noisy and cannot be well explained by the model due to hair and clothing, which are not captured by the SMPL and FaceWarehouse meshes, resulting in erroneous correspondences during ICP. Additionally, the joint locations of the models are not necessarily consistent with the annotation criteria used to train the 2D detectors. We are therefore forced to set priors over model parameters to avoid the model from overfitting to these sources of noise, $E_{\text{prior}} = E^F_{\text{prior}} + E^B_{\text{prior}} + E^H_{\text{prior}}$. The prior for each part is defined by corresponding shape and pose priors, for which we use zero-mean standard nor-

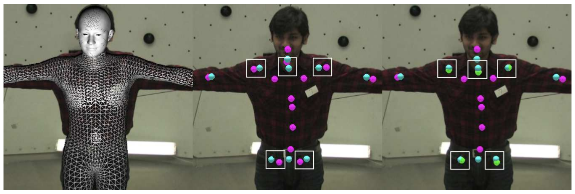


Figure 4: Regressing detection target 3D positions. (Left) The template model is aligned with target object; (Mid.) The torso joints of the template model (magenta) have discrepancy from the joint definitions of 3D keypoint detection (cyan); (Right) The newly regressed target locations (green) are more consistent with 3D keypoint detections.

mal priors for each parameter except for scaling factors, which are encouraged to be close to 1.

5. Creating Adam

We derive a new model, Adam, enabling total body motion capture with a simpler parameterization than the part-based Frank model. In particular, this new model has a single joint hierarchy and a common parameterization for all shape degrees of freedom, tying together the face, hand, and body shapes and avoiding the need for separate part parameterizations or seam constraints. To build the model, it is necessary to align the reconstructed meshes with all body parts (face, body, and hands) of diverse subjects where the model can learn the variations. To do this, we leverage our Frank model and apply it on a dataset of 70 subjects where each of them performs a short range of motion in a multi-view camera system. We select 5 frames for each person in different poses, resulting in 350 meshes, and reconstruct them with our Frank model, producing aligned meshes with joint locations to build Adam. Because we derive the model from clothed people, the blendshapes explain variations of clothing at a coarse level.

5.1. Fitting Clothes and Hair

The Frank model captures the shape variability of human bodies and faces, but does not account for clothing or hair, since it keeps the original model space of part models ([34] and [15]). To learn a new set of linear blendshapes that better capture the rough geometry of clothed people and also roughly model hair, we need the meshes to match the geometry of the source data more accurately. For this purpose, we deform the meshes outside of the shape-space along each the normal direction of each vertex. For each vertex $\mathbf{v}_i$ in the Frank model, the deformed mesh vertex $\tilde{\mathbf{v}}_i$ is represented as:

$$\tilde{\mathbf{v}}_i = \mathbf{v}_i + \mathbf{n}(\mathbf{v}_i)\delta_i, \tag{14}$$

where $\delta_i \in \mathbb{R}$ is a scalar displacement meant to compensate for the discrepancy between the Frank model vertices and the 3D point cloud, along the normal direction at each vertex. We pose the problem as a linear system,

$$\begin{pmatrix} \mathbf{N}^T \\ (\mathbf{WLN})^T \end{pmatrix} \Delta = \begin{pmatrix} (\mathbf{P} - \mathbf{V}^U)^T \\ \mathbf{0} \end{pmatrix}, \tag{15}$$

where $\Delta \in \mathbb{R}^{N^U}$ contains the stacked per-vertex displacements, $\mathbf{V}^U$ are the vertices in the Frank model, $\mathbf{P} \in \mathbb{R}^{N^U \times 3}$ are corresponding point cloud points, $\mathbf{N} \in \mathbb{R}^{N^U \times 3}$ contains the mesh vertex normals, and $\mathbf{L} \in \mathbb{R}^{N^U \times N^U}$ is the Laplace-Beltrami operator to regularize the deformation. We also use a diagonal weight matrix $\mathbf{W} \in \mathbb{R}^{N^U \times N^U}$ to avoid large deformations where the 3D point cloud has lower resolution than the original mesh, such as details in the face and hands.

5.2. Detection Target Regression

There exists an important discrepancy between the joint locations of the LBS model (i.e., the 3D centers of rotation for bone deformation) and the location of the keypoint detections (which come from manually annotated guesses of where the anatomical joints are in 2D images). This is shown in Fig. 4. This difference has the effect of pulling the model towards a bad fit even while achieving a low keypoint cost, $E_{\text{keypoints}}$, especially for shoulders and hips. We alleviate this problem by computing a new regression function, $\hat{\mathbf{J}}^A \in \mathbb{R}^{J^A \times N^U}$, which relates the vertices in the body model to the expected location of 3D keypoint detections. However, to be able to learn these regressors, we require instances of the fitted model vertices as well as the 3D keypoint detections.

Therefore, we first fit the Frank model (with additional shape variations) using the original joint locations as detection targets, and obtain aligned meshes across all subjects. Based on these outputs, we can build the regression matrix using the locations of 3D keypoint measurements as targets instead of Frank model's joint locations. Similar to the joint regression in SMPL [34], we first select a subset of vertices in the proximity of each detection target, and estimate a fixed, sparse linear combination of these vertices that approximates the location of the 3D keypoint across all fitted meshes. This optimization is posed as an L1-regularized least-squares problem with non-negative constraints, where we additionally impose that the vertex weights sum to one, resulting in an interpolation.

The results are shown in Fig. 4. Note that this new regressor is used only for the optimization in Eq. (12), whereas the original joint regressor from SMPL [34], $\mathbf{J}^A$, is used for LBS. However, we also add rows to the joint regression matrix to account for the additional finger joints, which we solve for in the same way. The resulting matrix is $\mathbf{J}^A \in \mathbb{R}^{J^A \times N^U}$ where N^U is the number of vertices of Adam (the same as Frank) and $J^A = 61$ is the number of joints in Adam model including 21 body joints and 20 finger joints (including 5 finger tips) for each hand.

Table 1: Accuracy of Silhouettes from different models

	SMPL[34]	Frank	Frank ICP	Adam ICP
Mean	84.79%	85.91%	87.68%	87.74%
Std.	4.55	4.57	4.53	4.18

5.3. Building the Shape Deformation Space

After model fitting with Δ displacement, we warp each frame's surface to the rest pose, applying the inverse of the LBS transform. With the fitted surfaces warped to this canonical pose, we do PCA analysis to build a joint linear shape space that captures shape variations across the entire body. As in Section 3.3, we separate the expression basis for the face and retain the expression basis from the FaceWarehouse model, as our MVS point clouds are of too low resolution to fit facial expressions.

The Adam model is parameterized as:

$$M^A(\boldsymbol{\theta}^A, \boldsymbol{\phi}^A, \boldsymbol{t}^A) = \mathbf{V}^A \tag{16}$$

with $\mathbf{V}^A = \{\mathbf{v}_i^A\}_{i=1}^{N^A}$ and N^A=18540 which is equal to the vertices in Frank, N^U. As in SMPL, the vertices of this template mesh are first displaced by a set of blendshapes in the rest pose, $\hat{\mathbf{v}}_i^A = \mathbf{v}_i^{A0} + \sum_{k=1}^{K_A} \mathbf{s}_i^k \phi_k^A$, where $\mathbf{s}_i^k \in \mathbb{R}^3$ is the i-th vertex of the k-th blendshape, ϕ_k^A is the k-th shape coefficients of $\boldsymbol{\phi}^A \in \mathbb{R}^{K_b}$, and $K_A = 40$ is the number of identity coefficients, $\mathbf{v}^{A0}$ is the mean shape and $\mathbf{v}_i^{A0}$ is its i-th vertex. Note that these blendshapes now capture variation across the face, hands, and body. These are then posed using LBS as in Eq. (6) after obtaining joint locations by the joint regressor matrix $\mathbf{J}^A$.

5.4. Tracking with Adam

The cost function to capture total body motion using Adam is similar to Eqn. 11 without the seam term:

$$E\big(\boldsymbol{\theta}^A, \boldsymbol{\phi}^A, \boldsymbol{t}^A\big) = E_{\text{keypoints}} + E_{\text{icp}} + E_{\text{prior}}. \tag{17}$$

However, Adam is much more amenable to optimization than Frank: it has a single set of unified shape and pose parameters for all parts, and does not require seam constraints between disparate models.

6. Results

We perform total motion capture using our two models, Frank and Adam, on various challenging sequences. For experiments, we use the dataset captured in the CMU Panoptic Studio [28, 29]. We use 140 VGA cameras to reconstruct 3D body keypoints, 480 VGA cameras for feet, and 31 HD cameras for faces and hands keypoints, and 3D point clouds. We compare the fits produced by our models with the simplified[3] SMPL model [34].

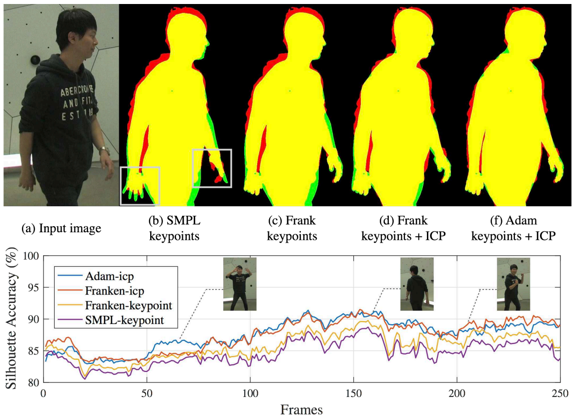


Figure 5: (Top) The silhouette from different methods overlayed with ground-truth. The ground truth is drawn in the red channel and the rendered silhouette masks from each model are drawn in the green channel. Thus, the correctly overlapped region is shown as yellow color. (Bottom) Silhouette accuracy compared to the ground truth silhouette.

6.1. Quantitative Evaluation

We evaluate how well each model can match a moving person by measuring overlap with the ground truth silhouette across 5 different viewpoints for a 10 second sequence. To obtain the ground truth silhouette, we run a background subtraction algorithm using a Gaussian model for the color of each pixel, with post-processing by morphological transforms to remove noise. As an evaluation metric, we compute the percentage of overlap compared to the union between the GT silhouettes and the rendered forground masks after fitting each model. Here, we compare the fitting results of 3 different models: SMPL, our Frank, and our Adam models. The results are shown in Fig. 5 and Table 1. We first compare accuracy between SMPL and Frank model by using only 3D keypoints as measurement cues. The major source of improvement of Frank over SMPL is in the articulated hand model (by construction, the body is almost identical). Including ICP term as cues provides better accuracy. Finally in the comparison between our two models, they show almost similar performance. Ideally we expect Adam to outperform Frank because it has more expressive power for hair and clothing, and it shows better performance for certain body shapes (frame 50-75 in Fig. 5). However, Adam sometimes produces artifacts showing lower accuracy: it tends to generate thinner legs, mainly due to poor 3D point cloud reconstructions in the training data[4]. However, Adam is simpler for total body motion capture and has potential to be improved once a large dataset is available with a more optimized capture setup.

[3] In all our comparison, we disabled the pose-dependent blendshapes of SMPL, and thus here SMPL model means the body part of Frank.

[4] Due to dark clothing combined with fewer camera views of the legs.

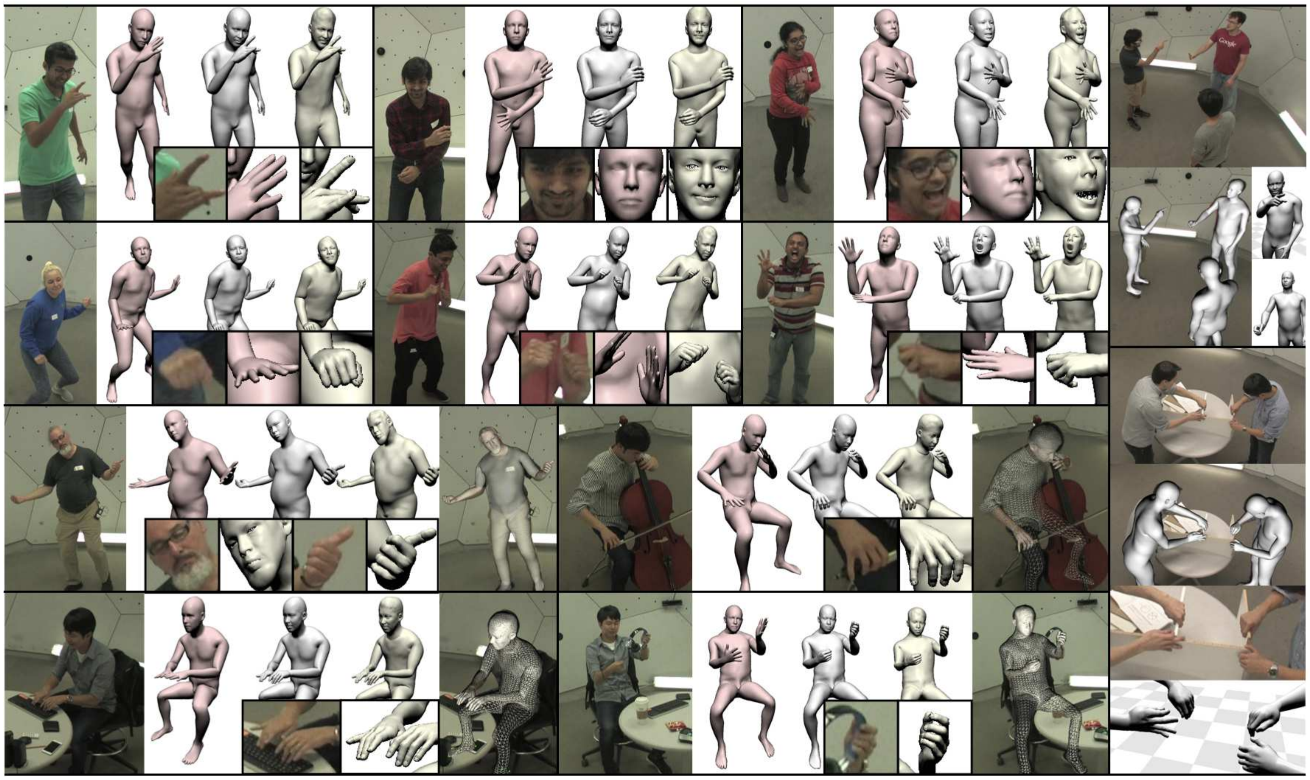


Figure 6: Total body reconstruction results on various human body motions. For each example scene, the fitting results from three different models are shown by different colors (pink for SMPL [34], silver for Frank, and gold for Adam).

6.2. Qualitative Results

We run our method on sequences where face and hand motions naturally occur. The sequences include short range of motion for 70 people used to build Adam, social interactions of multiple people, a furniture building sequence with dexterous hand motions, musical performances (cello and guitars), and commonly observable daily motions such as typing. Most of these sequences are rarely demonstrated in previous markerless motion capture methods since capturing subtle details is key to achieve realism. Example results are shown in Figure 6 but are best seen in the accompanying videos. Here, we also qualitatively compare our models (in silver color for Frank, and gold for Adam) with SMPL (without pose-blendshapes, in pink) [34]. Note that total body motion capture based on our models produces more realism by capturing subtle details from the hands and faces.

7. Discussion

We present the first markerless method to capture total body motion including facial expression, body motion from torso and limbs, and hand gestures at a distance. To achieve this result, we present two types of models, Frank and Adam, which can express motion in each of the parts. Our reconstruction results show compelling and realistic results, even when using only sparse 3D keypoint detections to drive the models. As a current limitation of our system, Adam lacks expressive power in surface details due to the limited number of subjects in training. However, the major value of Adam model over Frank lies in its simpler representation to capture total body motion, which can be useful for other applications.

There are two interesting points our paper raises. First, markerless hand motion capture, often considered too challenging compared to body and face captures, shows better localization quality in our results. Body joints are located inside the body and are hard to localize for clothed subjects, and the accuracy of face reconstruction greatly decreases once the face is not facing any camera. However, hands are often bare and the hand keypoint detector [44] provides guessed measurements with high confidence even in self-occlusions, which can be fused in multiple views. Second, our results show a potential that markerless motion capture can eventually outperform its marker-based counterpart. Marker-based methods strongly suffer from occlusions, making it hard to capture both body and hands together, while our method can still exploit measurements for occluded parts by learning-based keypoint detectors.

References

[1] Realitycapture software. www.capturingreality.com/.

[2] Vicon motion systems. www.vicon.com.

[3] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ToG*, 2005.

[5] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Consumer Depth Cameras for Computer Vision*. Springer, 2013.

[6] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012.

[7] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In *TOG*, 2010.

[8] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. Sumner, and M. Gross. High-quality passive facial performance capture using anchor frames. In *TOG*, 2011.

[9] R. Birdwhistell. Kinesics and context: Essays on body motion communication. In *University of Pennsylvania Press, Philadelphia.*, 1970.

[10] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *CoRR*, 2016.

[11] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. In *TOG*, 2010.

[12] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. In *IJCV*, 2004.

[13] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region and motion-based 3D tracking of rigid and articulated objects. In *TPAMI*, 2010.

[14] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. In *TOG*, 2015.

[15] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. In *TVCG*, 2014.

[16] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[17] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part i: Theory and algorithms. In *IJCV*, 2005.

[18] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi. Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation. In *IJCV*, 2010.

[19] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH*, 2008.

[20] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. Intraface. In *FG*, 2015.

[21] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015.

[22] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. In *CVPR*, 2008.

[23] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*. IEEE, 2009.

[24] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. In *TOG*, 2013.

[25] D. Gavrila and L. Davis. Tracking of humans in action: A 3-D model-based approach. In *ARPA Image Understanding Workshop*, 1996.

[26] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. In *TOG*, 2011.

[27] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, H. Joo, and Y. Sheikh. Openpose. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[28] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[29] 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. In *TPAMI*, 2017.

[30] R. Kehl and L. V. Gool. Markerless tracking of complex human motions from multiple views. In *CVIU*, 2006.

[31] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012.

[32] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. In *TOG*, 2013.

[33] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. In *TPAMI*, 2013.

[34] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *TOG*, 2015.

[35] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single RGB camera. In *TOG*, 2017.

[36] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[37] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, 2012.

[38] S. I. Park and J. K. Hodgins. Capturing and animating skin deformation in human motion. In *TOG*, 2006.

[39] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *TOG*, 2017.

[40] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. In *TOG*, 2015.

[41] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. In *TOG*, 2017.

[42] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *CHI*, 2015.

[43] J. Shotton, A. Fitzgibbon, M. Cook, and T. Sharp. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[44] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[45] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015.

[46] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *ICCV*, 2013.

[47] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, 2011.

[48] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015.

[49] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3D articulated hand posture. In *CVPR*, 2014.

[50] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[51] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2017.

[52] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[53] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. In *IJCV*, 2016.

[54] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. In *TOG*, 2012.

[55] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *TOG*, 2008.

[56] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. In *SIGGRAPH*, 2009.

[57] C. Wan, A. Yao, and L. Van Gool. Direction matters: hand pose estimation from local surface normals. In *arXiv preprint arXiv:1604.02657*, 2016.

[58] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[59] H. Woltring. New possibilities for human motion studies by real-time light spot position measurement. In *Biotelemetry*, 1973.

[60] C. Wu, D. Bradley, M. Gross, and T. Beeler. An anatomically-constrained local deformation model for monocular face capture. In *TOG*, 2016.

[61] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *ICCV*, 2013.

[62] Q. Ye, S. Yuan, and T.-K. Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *ECCV*, 2016.

[63] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017.

[64] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *CVPR*, 2015.

[65] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshop on Geometry Meets Deep Learning*, 2016.