Embodied Hands: Modeling and Capturing Hands and Bodies Together

JAVIER ROMERO*[†], Body Labs Inc.
DIMITRIOS TZIONAS*, Max Planck Institute for Intelligent Systems
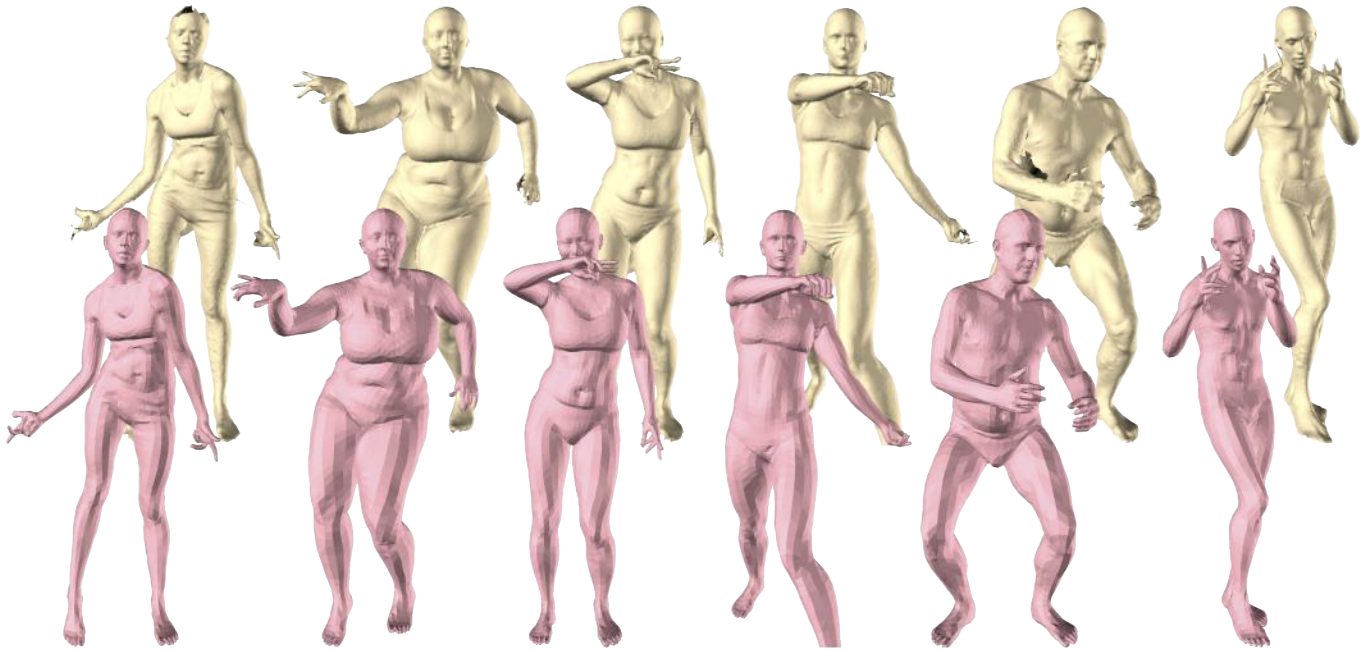MICHAEL J. BLACK, Max Planck Institute for Intelligent Systems

Fig. 1. **Capturing hands and bodies together.** Beige: raw 4D scan. Pink: recovered shape and pose using our model of *hands and bodies interacting together (SMPL+H)*. To deal with noise and missing data, we develop a new model of hands called *MANO* (*hand Model with Articulated and Non-rigid defOrmations*) and combine it with a parameterized 3D body model (SMPL). This combination enables natural performance capture even under severely noisy measurements.

Humans move their hands and bodies together to communicate and solve tasks. Capturing and replicating such coordinated activity is critical for virtual characters that behave realistically. Surprisingly, most methods treat the 3D modeling and tracking of bodies and hands separately. Here we formulate a model of *hands and bodies interacting together* and fit it to full-body 4D sequences. When scanning or capturing the full body in 3D, hands are small and often partially occluded, making their shape and pose hard to recover. To cope with low-resolution, occlusion, and noise, we develop a new model called *MANO* (*hand Model with Articulated and Non-rigid defOrmations*). MANO is learned from around 1000 high-resolution 3D scans of hands of 31 subjects in a wide variety of hand poses. The model is realistic, low-dimensional, captures non-rigid shape changes with pose, is compatible with standard graphics packages, and can fit any human hand. MANO provides a compact mapping from hand poses to pose blend shape corrections and a linear manifold of pose synergies. We attach MANO to a standard parameterized 3D body shape model (SMPL), resulting in a fully articulated body and hand model (SMPL+H). We illustrate SMPL+H by fitting complex, natural, activities of subjects captured with a 4D scanner. The fitting is fully automatic and results in full body models that move naturally with detailed hand motions and a realism not seen before in full body performance capture. The models and data are freely available for research purposes at http://mano.is.tue.mpg.de.

CCS Concepts: • **Computing methodologies** → **Motion capture**;

Additional Key Words and Phrases: Hands, Human body shape, 3D shape, Learning, Performance capture, Motion capture

*Both authors contributed equally to the paper
[†]This research was performed by JR at the MPI for Intelligent Systems.

Authors' email addresses: javier.romero@bodylabs.com; {dtzionas, black}@tue.mpg.de.

1 INTRODUCTION

Bodies and hands are literally inseparable. Yet, despite this, research on modeling bodies and hands has progressed separately. Significant

245:2 • Romero et. al.

advances have been made on learning realistic 3D statistical shape models of full bodies but these models typically have limited, or no, hand articulation. Similarly there is significant work on tracking hands using depth sensors and video sequences but these hands are modeled and tracked in isolation from the body. We argue that the hands and body together are important for communication and that a complete picture of our actions, emotions, and intentions is not possible without the joint analysis of hands and bodies. The growth of interest in virtual and augmented reality has increased this need for characters and avatars that combine realistic bodies and hands in motion. Here we develop a new approach to capture the 4D motion of the hands and body together.

Several factors have led to the separation of hands from bodies. Full body models such as SCAPE [Anguelov et al. 2005] are learned from subjects making a tight fist, while more recent models like SMPL [Loper et al. 2015] assume an open, rigid, hand. Neither looks realistic when animated. Such 3D body models are created from 3D scans of the whole body. At the resolution of most body scanners, the hands are small and the fingers are hard to resolve resulting in noise and "webbing" between the fingers. Additionally, occlusion of the hand by the body and by itself results often in significant missing data. Consequently, to get sufficient resolution and unconstrained movement, most visual hand tracking work focuses on capturing hands alone, possibly together with the forearms, using RGB-D sequences.

We make several contributions, which can be roughly divided into two categories: learning a new model of the hand, and tracking hands and bodies together.

First we collect a new database of detailed hand scans of 31 subjects in up to 51 poses. We capture both left and right hands of men and women, use a wide range of poses, and capture hands interacting with objects.

Second, we use this data to build a statistical hand model similar to the SMPL body model [Loper et al. 2015], which we call *MANO* for *hand Model with Articulated and Non-rigid defOrmations.* Like SMPL, the model factors geometric changes into those inherent to the identity of the subject and those caused by pose. The model is trained to minimize the vertex error in the training set. The pose space uses linear blend skinning for simplicity, with corrective blend shapes that are automatically learned from the scans. MANO is created from the SMPL hand topology, and has analogous components to those in SMPL: a template shape, kinematic tree, shape and pose blend shapes, blend weights and a joint regressor.

Hand articulation is very different, however, from the full body; the hand contains a large number of joints with restricted articulation. Consequently, MANO differs from SMPL in several ways. Like SMPL, MANO uses corrective blend shapes that are a function of pose. Unlike SMPL, following the example of [Mohr and Gleicher 2003] that introduced localized skinning blend weights, we encourage the corrective pose blend shapes to be local by penalizing the dependency of corrections on joints which are far away in geodesic terms. Additionally, the high dimensionality of the full hand space makes fitting to noisy, low-resolution, body scans computationally expensive and prone to local minima. Consequently we reduce the dimensionality of the pose space by computing a linear embedding of the pose parameters from our dataset. We believe this is the first

such analysis based on such high-quality hand data. The resulting MANO model is lightweight, simple to animate, and compatible with existing graphics software.

Third, we combine MANO with the SMPL body model to give a new combined model of *hands and bodies interacting together (SMPL+H).* Here we take the shape space of the hand from the SMPL model, which captures correlations between hand and body morphology. To this we add the MANO joints, kinematic tree, blend weights, and pose blend shapes.

Fourth, we address the problem of capturing full bodies and hands in motion. To that end, we employ a 4D body scanning system that captures full 3D body shape at 60 frames per second. At the resolution of the scanner, the hands can be extremely noisy and low resolution, sometimes disappearing entirely. To recover hand pose, we modify *4Cap*, the temporal mesh registration algorithm used in DYNA [Pons-Moll et al. 2015], to include a simple velocity prior that prevents sudden hand motion in the complete absence of data. With this we fit SMPL+H to full-body 4D sequences to recover the body intrinsic shape and its changing pose, including finger articulation.

To illustrate the models and methods, we fit SMPL+H to a wide variety of complex sequences with fast motions. Figure 1 shows a few examples and many more can be found in the **supplemental video**. In addition to recovering the model parameters, we further allow the vertices of alignments to move to better fit the scan (with a penalty for deviating from the model surface) increasing their realism and level of detail. The resulting models and alignments look much more natural than those using SMPL.

In summary we propose a new model of hand shape and pose that is learned from data, compatible with existing graphics systems, low-dimensional, realistic and compatible with the SMPL body model. By combing MANO with SMPL we are able to jointly capture bodies and hands in motion with high realism and deal with missing and noisy data. This opens up the study of correlated body and hand motion at a level of detail not previously possible. The MANO model and SMPL+H are **available for research purposes** [MANO web] along with the aligned meshes needed to train MANO and the test-datasets for evaluation.

2 RELATED WORK

The body and the hands have often been studied separately. For example, models of neural control have long embraced this divisional approach [Penfield and Boldrey 1937]. Here we argue that modeling and tracking hands and bodies together is important for vision, graphics, and virtual reality.

Hand capture. Capturing the 3D motion of human hands has been studied for decades [Erol et al. 2007; Heap and Hogg 1996] due to its applications in computer graphics, animation, human computer interaction, rehabilitation and robotics to name a few. This interest is increasing [Supančič III et al. 2015; Ye et al. 2013] in light of the recent advancements in consumer RGB-D sensors and virtual reality [Hamer et al. 2009; Leap 2017; Melax et al. 2013; Oberweger et al. 2015; Oikonomidis et al. 2011a, 2012; Romero et al. 2013; Schmidt et al. 2014; Schröder et al. 2014; Sridhar et al. 2015; Tagliasacchi et al. 2015; Tang et al. 2017; Taylor et al. 2016; Tompson et al. 2014; Tzionas et al. 2016; Wang et al. 2013; Xu et al. 2016].

ACM Transactions on Graphics, Vol. 36, No. 6, Article 245. Publication date: November 2017.
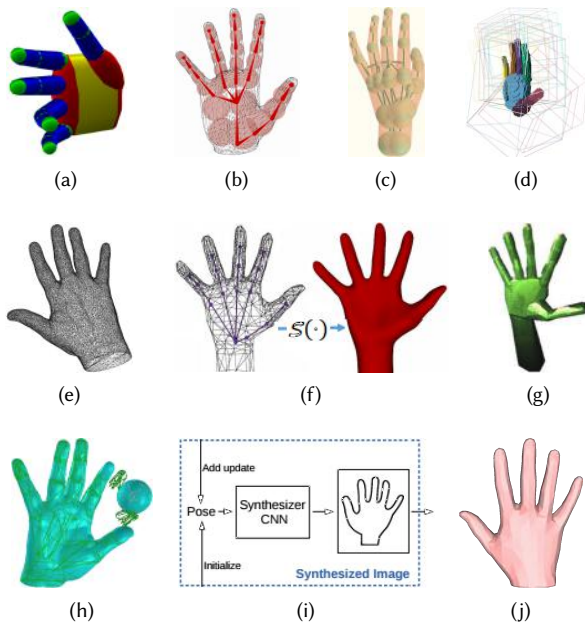
Fig. 2. Hand models. (a) *Primitives* approximation [Oikonomidis et al. 2011a], (b) *Sum-of-Gaussians* model [Sridhar et al. 2013], (c) *Sphere-Meshes* [Tkach et al. 2016] can be thought of as a generalization of the previous models, (d) *Articulated TSDF* for a voxelized shape-primitive hand model [Schmidt et al. 2014], (e) *Triangular Mesh* [Ballan et al. 2012; Tzionas et al. 2016], (f) *Loop Subdivision Surface* of a triangular control mesh [Khamis et al. 2015], (g) *Convex Bodies* for tracking [Melax et al. 2013], (h) *Convex Parts* of a triangular mesh for contact point detection [Tzionas et al. 2016], (i) *Learned Model* using a CNN to synthesize images of a given hand pose [Oberweger et al. 2015], (j) Our proposed hand *MANO* model. Images reproduced from the cited papers or their supplementary videos.

Despite this interest, key problems remain and hands have proven difficult to detect and track. Hands are dexterous, with many degrees of freedom, and the fingers all look similar in color and shape. They also move quickly, resulting in motion blur and making tracking difficult. The 3D shape of the hand leads to self occlusion, which makes hand pose inference ambiguous. The key problem is that *hands are small*; in images or scans of the full body, hands are often difficult to fully resolve.

Even traditional marker-based motion capture has limitations for capturing hands and bodies together. Most capture protocols ignore the hands and most publicly available datasets do not contain any information about finger pose. Hands require small markers which, in a large volume, require many mocap cameras to sufficiently resolve. Markers also limit hand movement, can fall off, and are often occluded. This limits the capture and study of hand and body movement together.

Several methods use additional instrumentation like wearable cameras [Kim et al. 2012], markers [Zhao et al. 2012], data-gloves [Dipietro et al. 2008], or colored-gloves [Wang and Popović 2009] to simplify hand tracking. Such approaches are intrusive, alter the hand shape, and hinder natural hand motion. Furthermore, they need to

be combined and synchronized with other systems to capture the full body and hands together. Recent commercial solutions such as Leap Motion [Leap 2017] require no markers or gloves but only capture the hands without the body.

Most recent research systems use commodity RGB-D cameras [Supančič III et al. 2015; Ye et al. 2013] but these have limited resolution and existing methods only capture the hands and not the full body. Multicamera systems [Ballan et al. 2012; Sridhar et al. 2013; Tzionas et al. 2016] on the other hand scale to large volumes but currently do not enable full body and hand tracking together.

Fitting and tracking. Beginning with Rehg and Kanade [1994], there have been many generative hand models proposed and many methods for fitting them to data. These include many filtering approaches [Bray et al. 2007; MacCormick and Isard 2000; Stenger et al. 2001; Wu et al. 2001], belief-propagation [Hamer et al. 2010, 2009; Sudderth et al. 2004], Particle Swarm Optimization (PSO) [Oikonomidis et al. 2011a], sampling [Oikonomidis et al. 2014], inverse-kinematics [Kim et al. 2012; Wang and Popović 2009], probabilistic line matching [Athitsos and Sclaroff 2003], reduced linear search sub-spaces [Heap and Hogg 1996; Wu et al. 2001], Bayesian filtering with Chamfer matching [Thayananthan et al. 2003], and physics-based methods [Melax et al. 2013]. Schmidt et al. [2014] extend the Signed Distance Function representation to articulated objects, while Qian et al. [2014] combine a gradient based ICP approach with PSO. Sridhar et al. [2013] explore the use of a Sum of Gaussians model for hand tracking in RGB images. None of these approaches address the tracking of full bodies and hands together.

Hand Models. Many hand models have been proposed and are summarized in Fig. 2. A popular approach is to approximate the hand with shape primitives (Fig. 2a) which enable fast evaluation of distances [Oikonomidis et al. 2011a, 2012, 2014; Qian et al. 2014; Rehg and Kanade 1994; Tagliasacchi et al. 2015]. An alternative is the Sum-of-Gaussians model [Sridhar et al. 2015, 2013] (Fig. 2b). The sphere-mesh model in Fig. 2c [Tkach et al. 2016] can be thought of as a generalization of the these models, while models based on shape-primitives are also used as underlying collision models [Oikonomidis et al. 2011b]. Schmidt et al. [2014] voxelize each shape-primitive and compute a Signed Distance Function for the local coordinate frame (Fig. 2d). Melax et al. [2013] use a union of convex bodies for hand tracking, However these approaches only roughly approximate hand shape.

A triangulated mesh with Linear Blend Skinning (LBS) [Lewis et al. 2000] (Fig. 2e) is more realistic and better fits image data [Ballan et al. 2012; Tzionas et al. 2016]. Despite their fixed shape, meshes are useful for computing contact points during interaction [Tzionas and Gall 2015]; this computation can be accelerated by approximating the mesh as an ensemble of convex hulls [Tzionas et al. 2016] (Fig. 2h). A recent alternative [Taylor et al. 2016] models the hand with smooth loop subdivision surfaces [Loop 1987] (Fig. 2f), which facilitate efficient and accurate computation of derivatives. De La Gorce et al. [2011] define a triangulated hand model and introduce scaling terms for each bone, allowing the hand to change shape. They use surface texture and shading to fit this to images in an analysis-by-synthesis approach.

Recently deep learning is providing new options for estimating hands from images and depth maps. Instead of using a single Convolutional Neural Network (CNN) to predict hand joints from the input image as Tompson et al. [2014], Oberweger et al. [2015] train a second CNN to close the loop from output to input and refine the predicted pose. This second CNN (Fig. 2i) is an image synthesizer inspired by Dosovitskiy et al. [2015] that renders a hand given the hand pose predicted by the first CNN.

All above approaches model hands in isolation from the rest of the body, with the exception of [Melax et al. 2013; Taylor et al. 2016] that also model the forearm. However, as it is shown in our experiments section, having very noisy or even completely missing hand and forearm is not uncommon in real full-body data.

Dimensionality Reduction. While hands have many degrees of freedom, a large number of them are not independently controllable and, in natural movements, hand poses are effectively low dimensional. Various glove- [Santello et al. 1998; Todorov and Ghahramani 2004] or marker-based [Schröder et al. 2014] capture methods have been used to study this with the effective dimensionality depending on the task and the capture system. Santello et al. [1998] find that two principal components account for over 80% of the variance in their data. Schröder et al. [2014] find that 3-6 components account for 80 − 90% of the variance. Todorov and Ghahramani [2004] use a different protocol and find the effective dimensionality is about 6.5. While our capture method is different and more detailed, our results are broadly consistent with those of Todorov and Ghahramani [2004].

The practical use of a low dimensional pose space for hands was shown by ElKoura and Singh [2003], who used it in a data-driven approach to generate physically plausible poses for music generation. In addition, Schröder et al. [2014] employ it for efficient hand tracking with noisy RGB-D data. However they train on motion data acquired by an intrusive marker-based approach that alters the shape and pose space, and rely on a fixed shape model constituted by cylindrical shape primitives.

Personalized hand models. Personalizing the hand model to the user improves tracking [Tan et al. 2016]. A personalized template mesh can be reconstructed offline with a multiview-stereo method [Ballan et al. 2012; Tzionas et al. 2016], however this does not scale well with the number of users. De La Gorce et al. [2011] define 56 scaling terms to allow bone lengths to vary but they do not learn how they correlate from data. Taylor et al. [2014] personalize a template hand mesh by adapting its shape to fit the depth data of a user's hand using an off-line calibration step for each user.

Khamis et al. [2015] go further to learn a model of shape variation from scans of 50 people captured with a depth sensor. They fit a linear blend skinned model to partial hand scan data and learn a low-dimensional PCA model of the deviations from this. This is similar to MANO, except that they restrict the pose-dependent deformations to be modeled by LBS; we produce more realistic posed meshes by learning pose-dependent corrective blend shapes [Loper et al. 2015]. With a calibration step [Sharp et al. 2015] their model can be used for realtime tracking [Tan et al. 2016; Taylor et al. 2016].

Body models. The modeling and tracking of bodies has a similar history to hands, paralleling the examples in Fig. 2. Early body models are based on geometric primitives, while more recent ones are learned from data. Early methods capture the statistics of body shape but do not model how the geometry of the body changes with pose [Allen et al. 2003; Seo et al. 2003]. The introduction of the SCAPE model [Anguelov et al. 2005] showed how to learn a model that combines shape and pose variations. SCAPE achieved high realism and spawned many related approaches [Chen et al. 2013; Freifeld and Black 2012; Hasler et al. 2009; Hirshberg et al. 2012; Pons-Moll et al. 2015]. These models represent shape in terms of triangle deformations, which present technical challenges for optimization and fitting. Vertex based models [Allen et al. 2006; Hasler et al. 2010; Loper et al. 2015] are preferable in this respect. None of these models capture hand motion, adopting a fixed posture with either a fist or an open hand.

Here we base our hand model on SMPL [Loper et al. 2015] and replace the hands in SMPL with our new learned MANO model. The original SMPL hands include a small amount of bending and wrist rotation, but this is not sufficient to capture any significant hand pose variation. The new MANO model can be used separately or together with the body (SMPL+H). When used separately, it should be able to replace most of the hand models described above easily since it is built on standard vertex-based modeling methods. When combined with the body, we are able to estimate both hand poses and body movement together. We know of no previous models that attempt to model and capture the full body and detailed hands together.

3 MODEL

Given the difficulty of capturing hands with bodies, we advocate for a two-stage approach in the creation of a dexterous full body model. First, we collect a large number of scans of hands in isolation. These scans are obtained with a scanner configured specifically to capture hands with a fixed wrist position. This allows us to capture the nuances of hand deformation. We then train a hand model using an iterative process of aligning a template to the scans using the model and learning a model from the registered scans. In the second stage, we integrate this hand model with the full body to obtain a single, dexterous and fully articulated body model.

MANO is based on SMPL [Loper et al. 2015] and is compatible with the full body SMPL model. In order to facilitate the integration of the hand models with the full body model, we take the hand vertices from the full body model as a template. The general formulation of a SMPL model M, taken from the original paper [Loper et al. 2015] for completeness, is as follows:

$$M(\vec{\beta}, \vec{\theta}) = W(T_P(\vec{\beta}, \vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W}) \quad (1)$$

$$T_P(\vec{\beta}, \vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \quad (2)$$

where a skinning function W (in our case LBS) is applied to an articulated rigged body mesh with shape T_P, joint locations J defining a kinematic tree, pose $\vec{\theta}$, shape $\vec{\beta}$, and blend weights $\mathcal{W}$.

Unlike standard LBS models, with a SMPL model the mesh T_P that is posed and skinned, is a function of the pose and shape of the hand. Shape blendshape function B_S allows the base shape, in this case

Fig. 3. Data collection with the 3dMDhand System. The system captures a hand with the wrist resting on a hand stand, and performing different articulated motions. The active stereo system is composed by 5 scanning units, each collecting one color image (top row) with flash-based illumination and two grayscale images (bottom row) using speckle illumination. The result is a high quality, textured scan of the hand (right most image).

of the hand, to vary with identity. The pose blend shape function, B_P, captures deformations of the mesh as a function of the bending of the joints. Traditional LBS models are overly smooth and suffer from "collapse" at the joints. For MANO, we learn corrective blend shapes that correct these artifacts, resulting in more natural-looking finger bending.

Specifically the pose and shape blend shapes are defined as the linear combination of a set of deformations, i.e. vertex offsets:

$$B_P(\vec{\theta}; \mathcal{P}) = \sum_{n=1}^{9K} \left(R_n(\vec{\theta}) - R_n(\vec{\theta}^*) \right) \mathbf{P}_n, \qquad (3)$$

$$B_S(\vec{\beta}; \mathcal{S}) = \sum_{n=1}^{|\vec{\beta}|} \beta_n \mathbf{S}_n. \qquad (4)$$

Here $\mathbf{P}_n \in \mathcal{P}$ are the pose blend shapes and K is the number of parts in the hand model. These are not directly controlled by the hand part rotations but rather by the elements of the part rotation matrices as in SMPL. $R_n(\vec{\theta})$ indexes into the nth element of a vector of concatenated rotation matrix elements and $\vec{\theta}^*$ is the zero pose.

The shape blend shapes are computed from a set of registered hand shapes, normalized to the zero pose, using principal component analysis (PCA) as illustrated in Fig. 4. Consequently here, the β_n are linear coefficients and the vectors $\mathbf{S}_n \in \mathcal{S}$ are principal components in a low-dimensional shape basis that we learn below.

The joint locations, $J(\vec{\beta})$, also depend on shape parameters. These are learned as a sparse linear regression matrix $\mathcal{J}$ from mesh vertices as in SMPL. We refer the reader to [Loper et al. 2015] for more details.

Before we can learn the model parameters $(\mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}, \bar{\mathbf{T}})$, we first need hand scans of many people in many poses and then we need to register our template mesh to these to put them in correspondence. We describe these steps below.

3.1 Hand data

Our hand data was captured with a 3dMDhand System [3dMDhand 2017] (Fig. 3, 3dMD LLC, Atlanta). The system is composed of five scanning units, each containing one color camera with resolution 2448×2048, two grayscale cameras with resolution 1624×1236 and two speckle projectors that provide the illumination for the grayscale images. The scans have a resolution of approximately
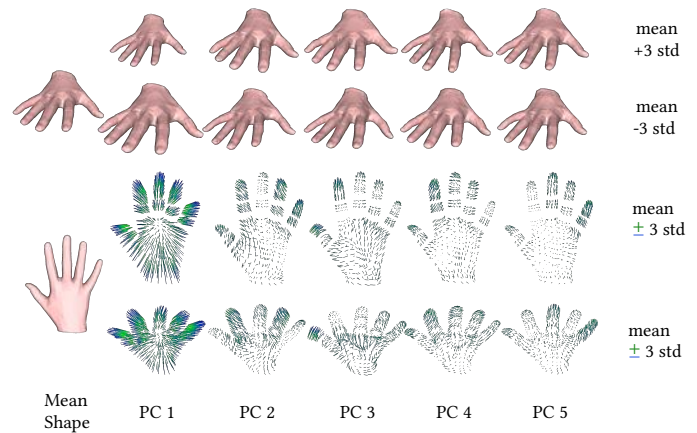

Fig. 4. PCA shape space. Each column depicts the effect of one of the first five *principal components* (PCs) of the learned hand shape space. The effect of each PC is shown by adding ± 3 standard deviations (std) to the mean shape (left-most image), as indicated. See the Supplemental Video.

$50,000$ vertices, include a texture map, and provide an accuracy of 0.2 millimeters root mean square (RMS) error according to the manufacturer.

Using this system, we collected data from both left and right hands of a total of 31 subjects, providing us a total number of 2018 scans. We mirror the left-hand scans to appear as right hands. Mirroring enables us to train a single consistent hand model. We later mirror back the learned right hand model to create a left hand model.

Each subject was captured performing three types of poses: a set of joint exploration poses, the 31 poses from the grasp taxonomy of Feix et al. [2016], and a few mixed poses. One can see a complete set of poses in Fig. 5. Each subject performed a subset of the protocol depending on their availability.

For each scan, we manually remove the geometry corresponding to the arm and handstand. This manual cropping takes approximately 30 seconds per scan. We then automatically segment out the objects in the poses involving object grasps using color information, since they are painted green. More specifically, a vertex is classified as object vertex simply if $G > max(R, 20)$ in 8-bit RGB
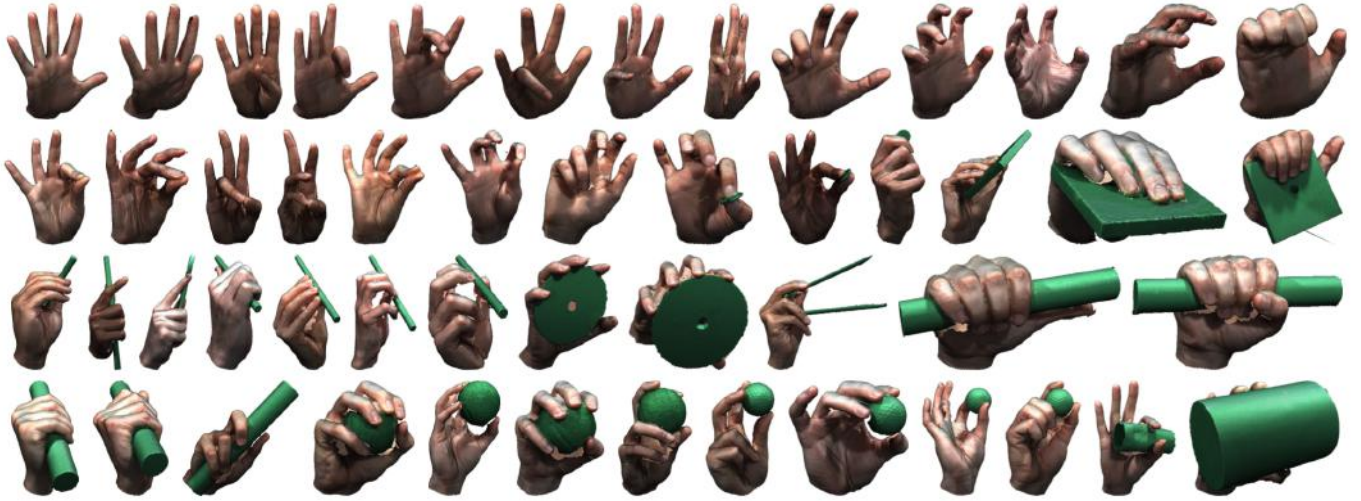
Fig. 5. Hand capture protocol. Each of the 31 subjects performed most of the 51 hand poses shown here. Images here are shown from multiple subjects.

space, where G and R are the values of the green and red channel correspondingly. This leaves the scans of hands grasping objects with significant holes.

3.2 Registration

The next step to building a hand model is to register, or align, a template to all the hand scans, bringing them into correspondence. The process of registering noisy hand scan data is challenging. Mesh registration is a challenging problem in general, but hands are especially difficult due to a high degree of self-similar structures (fingers) and a high level of self and object occlusion. While some previous work uses landmarks to simplify registration, we do not do so because it is impractical when registering a large number of scans. Instead, we bootstrap the creation of the model by manually curating the registrations. Specifically, we start with a crude hand model, use it to register all our scans, manually curate good ones, learn an improved model, and repeat. This gradually improves the alignments and the model.

We treat the registration process as an optimization problem, where we minimize the distance between the scan and a registered mesh, $\mathbf{V}$, with respect to the registration vertex locations, while keeping the registration *likely* according to the model. We do not provide full details here because the approach is similar to previous work [Hirshberg et al. 2012; Loper et al. 2015], with the addition of the manual curation. Specifically we minimize

$$E(\vec{\beta}, \vec{\theta}, \mathbf{V}; \mathbf{S}) = \lambda_g E_g + \lambda_c E_c + \lambda_\theta E_\theta + \lambda_\beta E_\beta \tag{5}$$

$$E_g(\mathbf{V}; \mathbf{S}) = \sum_{s \in \mathbf{S}} \rho(\min_{v \in \mathcal{V}} \|s - v\|) \tag{6}$$

$$E_c(\mathbf{V}, \vec{\beta}, \vec{\theta}) = \sum_i \|D_i(\mathbf{V}) - D_i(M(\vec{\beta}, \vec{\theta}))\|^2 \tag{7}$$

$$E_\theta(\vec{\theta}) = \|\vec{\theta}\|^2 \tag{8}$$

$$E_\beta(\vec{\beta}) = \|\vec{\beta}\|^2. \tag{9}$$

The energy is composed by four main components: data (or geometry) term E_g, coupling term E_c, pose prior E_θ and shape prior E_β. The data term E_g represents the point-to-plane distance between the vertices s on the scan $\mathbf{S}$ and the surface of the registration $\mathcal{V}$ (where calligraphic denotes a continuous surface), robustified with a Geman-McClure error function, ρ, [Geman and McClure 1987].

The coupling term E_c encourages the registered mesh, $\mathbf{V}$, to be similar to the model M, whose parameters $\vec{\beta}, \vec{\theta}$ we also optimize. We define similarity here in terms of the differences between the edges of the model and the registered mesh. The edges are given by the function D, which is simply a linear mapping. By directly optimizing the vertices in the registration we can go beyond the limitations of the model and obtain more faithful registrations, while the coupling term keeps the registration conservatively close to the model.

The shape prior term E_β penalizes the Mahalanobis distance between the optimized shape parameters and the distributions of hand shapes in the CAESAR dataset [Robinette et al. 2002]. Our shape space is orthogonal since it is the result of performing PCA of unposed registrations. We scale its basis vectors according to the square root of the explained variance $\sqrt{cov}$, which requires the coefficients $\vec{\beta}$ to be scaled by its inverse, effectively standardizing them. As a result [Brereton 2015], the Mahalanobis distance is conveniently computed just as the norm of the shape parameters $\vec{\beta}$.

For the pose prior E_θ, we define specific priors for each pose (since they are known given the protocol). We initialize this with a prior that penalizes deviations from the neutral pose and then refine with Gaussian priors for each pose.

The objective in Eq. 5 is highly non-convex. To optimize it, we use dogleg, a quasi-Newton least-squares optimizer [Nocedal and Wright 2006]. The gradients are computed with automatic differentiation with OpenDR [Loper and Black 2014].

We iterate the whole process of registration and model building twice, with a final visual inspection in between to define the

training database of well-registered meshes. After inspection, 1554 registrations of 2018 scans were deemed successful and are included in the released dataset. This inspection takes only approximately 1-2 seconds per scan, as it is a simple binary labeling task. For registration, in the first iteration we use the initial pose prior, while in the second we use the pose specific Gaussian prior, as described above. Registration takes approximately 30 seconds per frame on a 3.7 GHz Quad-Core Intel Xeon E5 computer using 4 threads, where only the closest-point search is multithreaded.

Figure 6 shows a variety of hand scans from one subject and the corresponding aligned meshes. Figure 7 shows a variety of hand shapes from different subjects in the same, relatively flat, pose. These figures give a sense of the level of detail in the aligned meshes.

3.3 Hand Model

Given this set of curated registrations, the goal is to learn the parameters of a SMPL-style hand model so that it fits the registrations. We start with the same strategy as used for the body in [Loper et al. 2015]. The model parameters, namely $(\mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}, \bar{T})$, are optimized iteratively keeping the rest constant. As in SMPL, personalized templates $\hat{T}_i$ are used for optimizing the pose related components $(\mathcal{P}, \mathcal{W})$, and its linear decomposition performed with PCA populates the shape space $\mathcal{S}$ and template $\bar{T}$. However, we modify a number of components in SMPL to take into account differences between the hands and the body.

MANO contains 15 joints plus the global orientation. Unlike most joints in the body, many hand articulations are anatomically restricted to one degree of freedom, while our model considers them as ball joints for simplicity. Hand pose is therefore effectively overparameterized. Since most of the parameters in a SMPL model belong to the pose-dependent blend shapes B_P (which grows linearly with number of joints), the regularization of B_P for a highly articulated object like the hand is important to avoid overfitting. One of the effects of this overfitting is the non-locality of pose-dependent deformations obtained when a model is trained under the procedure described in [Loper et al. 2015].

Consequently, we reformulate the model to have a stronger regularization that promotes pose-dependent blend shapes B_P that are local to the joints that influence them. Since pose-dependent blend shapes map pose rotation elements to vertex displacements, a natural way to achieve this is by penalizing the dependency of vertex displacements on joints whose rotation centers are far away.

More concretely, we replace the constant cost λ_P associated with all pose blend shape elements in SMPL (Eq. 14 [Loper et al. 2015]) by a cost that depends on the distance between the input joint and the output vertex

$$\Lambda_P(i, j) = \mathcal{J}_j D_{\text{geo}}(t_i, \bar{T}) \tag{10}$$

where $D_{\text{geo}} \in \mathbb{R}^N$ is the geodesic distance between a particular vertex t_i and the rest of the vertices in the template mesh $\bar{T}$, and $\mathcal{J}_j \in \mathbb{R}^N$ denotes the joint regressor matrix for joint j. Λ_P establishes a cost for each pair of input joint and output vertex. Since each input joint spans 9 scalar inputs corresponding to its rotation matrix, and each output vertex corresponds to 3 scalars, each element in Λ_P is expanded into a 3×9 block before computing the product with

$\mathcal{P}$. The geodesic distance to a weighted average of points is not defined; therefore, we replace it by the weighted average of the geodesic distances. This new regularization scheme results in more local and regularized pose-dependent deformations, as illustrated in the **supplemental video**.

Second, the zero pose of our model represents a flat hand (as in the CAESAR dataset [Robinette et al. 2002], for compatibility with SMPL), which is far from the mean pose in our dataset (see Fig. 10). The deformations entailed by the difference between the zero and mean poses result in unnatural templates due to the optimization order of the model. More specifically, the initial optimization of the personalized templates, $\hat{T}_i$, is performed using a model without pose blend shapes, B_P, which are optimized afterwards. The templates then absorb the pose-dependent corrections between the zero and mean pose, resulting in unnaturally extruded knuckles (see top row in Fig. 8). For this reason the template optimization should not take into account registrations with extreme poses until the pose dependent deformations work reasonably. We express this by weighting the registrations used in the optimization of $\hat{T}_i$ according to how much we believe they can be potentially affected by the pose-dependent blend shapes, namely

$$w_i = \left[\sum_{n=1}^{9K} \| R_n(\vec{\theta}) - R_n(\vec{\theta}^*) \|_F^2 \right]^{-2} \tag{11}$$

where R_n represents the nth element of the concatenation of rotation matrices corresponding to the pose axis angles $\vec{\theta}$ or the rest pose $\vec{\theta}^*$ and $\| \cdot \|_F$ denotes the Frobenius norm. We choose to use the Frobenius norm of the difference of rotation axes since it has a clear Euclidean interpretation, while it is also rotation invariant and has no problems with periodicity, as opposed to other representations like the difference of angles. The use of double squaring is a simple heuristic to heavily penalize deviations from the rest pose. The weights of Equation (11) are only applied in the first iteration of $\hat{T}_i$ optimization, when the pose-dependent blend shapes are zero, and essentially penalize deviations from the rest pose that activate strong pose blend shapes.

Third, we exploit the symmetry of left and right hands by creating a single right hand model from right and (mirrored) left data. We then mirror the resulting model to obtain the left hand model. This allows us to virtually double the amount of training data, which helps limit overfitting. Details of the mirroring are included in the **supplementary material**

Finally, another difference with respect to SMPL is that, after the initial set of registrations, we do not use the CAESAR data for our shape space. We instead compute our shape space using the neutral poses from our pose subjects. Figure 7 shows some examples. The reason for this is that CAESAR registrations have some systematic bias due to the combination of the pose, occlusions, and technology used to create the dataset.

With these changes in the training procedure, we now train MANO following [Loper et al. 2015]. This step takes approximately 42 hours on a single 3.7 GHz Quad-Core Intel Xeon E5 computer using 8 threads. Learned components of our model, like the pose-dependent corrective blend shapes are critical to correct the errors of

Fig. 6. Examples of the captured poses for a single subject. The *top* row shows some of the captured *scans*, while the *bottom* row shows the corresponding mesh *alignments* with *pink* color.

Fig. 7. Examples of the captured *"flat-hand"* pose for various subjects. The *top* row shows some of the captured *scans*, while the *bottom* row shows the corresponding mesh *alignments* with *pink* color.

Fig. 8. Three different views (columns) of the average pose-subject templates $\hat{T}_i$ for a model optimized with constant (first row) and vertex-joint-dependent (second row) regularization. Pay attention to the bumps in the knuckles in the first row, which are attenuated considerably in the second.

linear blend skinning and to capture natural bending of the fingers; see Fig. 9.

Hand pose embedding. In order to make the model practical for the purpose of scan registration, we will try to expose a set of parameters that efficiently explain the most common hand poses. Since research in neuroscience has shown that most hand pose variance lies on low-dimensional manifolds [Santello et al. 1998], we will parameterize each hand posture by a set of coefficients that map a low-dimensional manifold. The only requirement for this manifold is to be differentiable in order to apply the chain rule within our optimization framework. Striving for simplicity, we choose a linear mapping obtained with PCA on the poses in our dataset, in the axis angle representation (after mapping all axis angles to the range $(-\pi, \pi)$). The principal components (PCs) of our pose space can be seen in Fig. 10. For high-quality data, high accuracy can be achieved using many PCs. For noisy data though, one needs fewer

PCs in a trade-off between accuracy (many PCs) and robustness (few PCs). For our data, we observed that 6 components per hand suffice to model most common hand poses, capturing approximately 81% of the variance in the training poses (see Experiments).

3.4 SMPL+H: Model integration

Unlike previous hand models, our goal is to obtain an integrated dexterous full body model that can be registered to full body scans. We therefore need to integrate the right and left hand models with a full SMPL body model. The basic integration is straightforward: the components of the model related to the finger articulation (blend weights, pose-dependent deformations and joint-regressors) are taken from the trained MANO models, while the components related to the rest of the body are taken from the full body SMPL model, including the wrist. In order to capture the correlations between body and hand shape, we use the shape space from the full body model [Loper et al. 2015; Robinette et al. 2002] to determine the shape of the hands.

Similarly to the SMPL model, this body+hands model is fully compatible with animation packages and game engines, and its simplicity makes it fast and simple to use. The full articulation of the hands comes at a cost though: the dimensionality of the pose space has increased more than twofold. This is rather counterproductive, since most of the pose degrees of freedom are devoted to a small area of the body in which typically data is either noisy or missing. Consequently, rather than modeling the full dimensionality of the hand pose space, we use a 6-dimensional linear embedding described above for each hand. Then, the kinematic structure of SMPL has 66 degrees of freedom without the hands and SMPL+H then has 78.
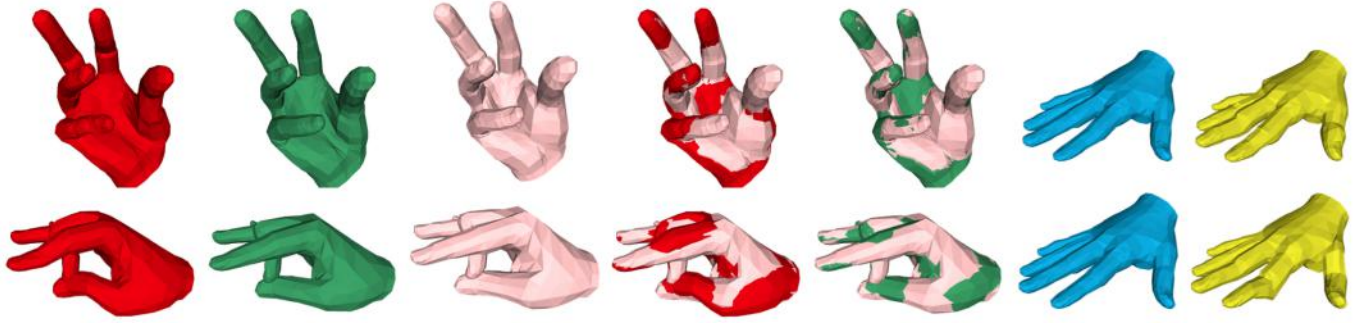
Fig. 9. **The effect of pose blendshapes**. From left to right, model without pose-dependent blendshapes in red, model with pose-dependent blendshapes in green, registration in pink, two overlays, personalized template in cyan, personalized template plus pose-dependent blendshapes in yellow. Pose blend shapes prepare the hand template for the subsequent articulation, to correct for artifacts of the latter. Therefore the deformations of the template in yellow correct the artifacts in red (best viewed on screen) and result in a good model after posing, shown with green color. Each row shows two different poses; the first one shows a large correction on the bending of the pinky finger, while the second one shows the correction of the articulation in the middle finger.
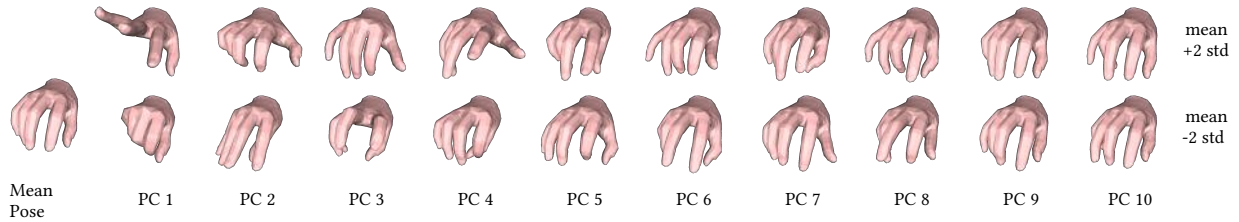

Fig. 10. **PCA pose space.** The left-most image depicts the *mean pose*, while the rest of the columns depict the effect of the first ten *principal components* (PCs) of the pose space. The effect of each PC is shown by adding ±2 standard deviations (std) to the mean pose, as indicated.

4 CAPTURING BODY AND HAND MOTION

To evaluate SMPL+H we use 4D sequences of bodies and hands in motion; i.e. temporal sequences of 3D scans. We fit the proposed model to the scans, bringing all of them into correspondence.

4.1 Camera system

For this purpose we use a 3dMDbody.u [3dMDbody 2017] multi-camera active stereo system capturing 3D scans at 60 frames per second (fps). The camera system consists of 22 pairs of stereo cameras, 22 color cameras, 34 speckle projectors and arrays of white-light LED panels. Here we do not use the color cameras. For each frame t the system outputs a 3D scan S^t computed from the stereo cameras and explaining a projected texture pattern.

The camera pairs are positioned in a way to cover multiple overlapping view-points and allow a wide range of full-body motions such as arm movements, running on the spot, throwing or jumping. All cameras synergistically capture full bodies and no camera is dedicated to the capturing of hands. This enables the use of holistic models like the proposed one, however it poses additional challenges for hands due to low resolution for the hand region, the big size of the projected speckle pattern compared to finger width, and the high velocity of finger motion. These factors result in large occluded regions without observed points during several frames. The above suggest the the need for strong hand models that are combined with full body models.

4.2 Sequences

Using the camera system described in Section 4.1 we capture a new set of challenging 4D sequences with a variety of subjects and actions performed. The captured sequences are separated in 3 groups. In the first group we capture 11 sequences of 10 subjects, 5 male and 5 female, of varying size and shape, each performing an unconstrained improvised action. For these sequences we perform a comparison between the proposed model and SMPL. The next group of 28 sequences are captured for a single subject performing actions designed around hand motion including basketball, finger counting or keyboard typing, among others. The final set of 2 sequences captures 2 professional female actors performing improvised movements that express fear. The duration of the sequences (in frames) is 189 to 1125 for the first group, 204 to 890 for the second and more than 3500 for the third one. In total our dataset has 41 sequences with a total of 27, 156 frames and corresponding 3D scans. All subjects are dressed with identical minimal clothing, namely tight fitting swimwear bottoms for both men and women and a sports bra top for women. All subjects gave written informed consent to participate in this study.

4.3 Mesh registration

The goal of full body registration is to bring all temporal 3D scans into correspondence by registering a single 3D template for body and

245:10 • Romero et. al.

hands to all of them. The template $\bar{\mathbf{T}}$ is a 3D watertight triangulated mesh comprised of 6890 vertices and $13,776$ triangles.

Similar to Section 3.2, registration is cast as a minimization problem. We try to minimize the discrepancy between the model and the data by minimizing the distance between the scan and the registration with respect to the registration vertex locations.

For registration we use a two stage approach, similar to [Loper et al. 2015; Pons-Moll et al. 2015]. We first compute a specific template $\bar{\mathbf{T}}^j$ for each subject, j, using a subset of the scans, typically the first 50 frames where the subject is mostly static. We then use $\bar{\mathbf{T}}^j$ to align the proposed model to all the scans (including the subset of the first step). By adopting the subject-specific templates the optimization of the second stage becomes significantly faster. To keep local optimization tractable, optimization for each frame is initialized with the pose of the previous frame in both stages. In the absence of a good initialization for the first frame, we introduce an extra initial stage for that frame in which a stronger pose prior is used.

4.3.1 Subject-specific template. For the first frame, where initialization is far away from the optimum, we exploit the fact that subjects were instructed to start with an approximate "A-pose" by following a conservative approach with strong pose priors. For this frame we solve only for a model fit by optimizing pose and shape

$$E_0(\vec{\beta}, \vec{\theta}; \mathbf{S}, \lambda_a) = \lambda_a \lambda_{\bar{g}} E_{\bar{g}} + \frac{\lambda_{\theta_b}}{\lambda_a} E_{\theta_b} + \frac{\lambda_{\theta_h}}{\lambda_a} E_{\theta_h} + \lambda_\beta E_\beta \quad (12)$$

where

$$E_{\bar{g}}(\vec{\beta}, \vec{\theta}; \mathbf{S}) = E_{g^*}(\vec{\beta}, \vec{\theta}; \mathbf{S}) + 0.1 E_{g'}(\vec{\beta}, \vec{\theta}; \mathbf{S}) \quad (13)$$

$$E_{g^*}(\vec{\beta}, \vec{\theta}; \mathbf{S}) = \sum_{s \in \mathbf{S}} \rho(\min_{m \in \mathcal{M}(\vec{\beta}, \vec{\theta})} \|s - m\|) \quad (14)$$

$$E_{g'}(\vec{\beta}, \vec{\theta}; \mathbf{S}) = \sum_{m \in \mathcal{M}(\vec{\beta}, \vec{\theta})} \rho(\min_{s \in C} \|s - m\|) \quad (15)$$

$$E_{\theta_b}(\vec{\theta}_b) = \|\vec{\theta}_b\|^2 \quad (16)$$

where $\mathcal{M}(\vec{\beta}, \vec{\theta})$ refers to the surface entailed by the model vertices $M(\vec{\beta}, \vec{\theta})$, C to the surface entailed by the scan vertices $\mathbf{S}$, and the *data* term $E_{\bar{g}}$ includes a *scan-to-mesh* energy E_{g^*} similar to Equation (6) and a *mesh-to-scan* $E_{g'}$. We found empirically that 20 shape parameters $\beta_i \in \vec{\beta}$ suffice to effectively capture shape identity.

While in general E_{g^*} suffices for fitting, the penalization of mismatching model vertices imposed by $E_{g'}$ avoids problems like excessively long hands in the templates extracted in this stage. In that respect, $E_{\bar{g}}$ can be seen as an approximation of the Hausdorff distance, as argued in [Tagliasacchi et al. 2016; Tkach et al. 2016]; replacing the *max* operators in the Hausdorff formulation by *sum* makes it differentiable and adds extra robustness to spurious scan data.

Furthermore, E_β is a *shape* term similar to Equation (9), E_{θ_b} is a pose prior similar to Equation (8). This prior is applied for the *body* pose parameters $\vec{\theta}_b$ up to (and including) the wrist, excluding the fingers, i.e. $\vec{\theta}_h$. We empirically found that a Gaussian pose prior is not suitable for hands since in case of disappearing hand data in scans it traps the hand pose in a local minimum around the mean

pose, even when data reappears. Since hand pose deviations are not Gaussian, we found Gaussian Mixture Models (GMM) to be more expressive. Inspired by [Bogo et al. 2016; Olson and Agarwal 2013], we employ a GMM by approximating the sum in the GMM by a max operator

$$E_{\theta_h}(\vec{\theta}_h) \equiv -\log \sum \left(g_j \mathcal{N}(\vec{\theta}_h \vec{\mu}_{\theta_h, j}, \Sigma_{\theta_h, j})\right) \quad (17)$$

$$\approx -\log \max_j \left(g_j \mathcal{N}(\vec{\theta}_h \vec{\mu}_{\theta_h, j}, \Sigma_{\theta_h, j})\right) \quad (18)$$

$$= \min_j \left(-\log \left(g_j \mathcal{N}(\vec{\theta}_h \vec{\mu}_{\theta_h, j}, \Sigma_{\theta_h, j})\right)\right) \quad (19)$$

where g_j are the mixture model weights of N Gaussians, and $\vec{\mu}_{\theta_h, j}$, $\Sigma_{\theta_h, j}$ are the mean and covariance matrices of each component of the mixture. The GMM is computed separately for left and right but based on the same data, namely all the left and right poses from the MANO dataset, mirrored as required. We empirically found $N = 10$ to work well. The objective Function (12) is used in an annealing framework based on the annealing weight λ_a. At the beginning of the process, where the initial solution is far away from the desired optimum, optimization can be easily trapped in a local minimum. We therefore start with a smaller weight $\lambda_a = 100$ to have a higher prior term and a lower data term. After a rough alignment with this weight we repeat registration with a higher weight $\lambda_a = 500$ to allow for better fitting with a weaker prior. The objective Function (12) is used only for the first frame, as a first step for rough initialization of the following stage.

Next we optimize shape, pose, and a full coupled aligned mesh, $\mathbf{V}$, by minimizing

$$E(\vec{\beta}, \vec{\theta}, \mathbf{V}; \mathbf{S}, \lambda_a) = \lambda_a \lambda_g E_g(\mathbf{V}; \mathbf{S}) + \frac{\lambda_{\theta_b}}{\lambda_a} E_c + \frac{\lambda_{\theta_h}}{\lambda_a} E_{\theta_h} + \lambda_\beta E_\beta + \lambda_r E_r \quad (20)$$

where

$$E_r(\vec{\theta}) = \sum_k \|\theta_k - \theta'_k\|^2 \quad (21)$$

and where E_g and E_c are defined by Equations (6) and (7), respectively. Recall that E_c is a "coupling" term that regularizes the aligned mesh to be close to the model.

We drop the body pose prior E_{θ_b} to allow more freedom during fitting. However, we keep the hand pose prior E_{θ_h} since hand data can be very noisy, even disappearing completely for several frames. For this reason the data term is simplified to E_g, removing the sum over model points in Equation (15) which is particularly sensitive to holes in the scan. The term E_β is defined in Equation (9). As a form of temporal pose smoothness against jittery motion, we introduce a zero-velocity regularizer E_r that penalizes large changes between the current θ and the previous θ' pose parameters during optimization, especially in the complete absence of data.

After aligning all frames in the initial subsequence, we then unpose all of these aligned meshes similar to [Loper et al. 2015] and manually select a subset through visual inspection, to filter out unposing artifacts. We then average these unposed alignments to obtain the subject-specific template $\bar{\mathbf{T}}^j$.

4.3.2 Subject-specific sequence alignment. After first acquiring a subject-specific template $\bar{\mathbf{T}}^j$ we avoid optimizing over $\vec{\beta}$ in the

ACM Transactions on Graphics, Vol. 36, No. 6, Article 245. Publication date: November 2017.

second stage, to gain significant optimization speed up and robustness. We therefore perform alignment similar to Section 4.3.1 but we drop the shape term E_β . Therefore the first frame is initialized with model-based optimization

$$E_0(\vec{\theta}; \mathbf{S}, \lambda_a) = \lambda_a \lambda_{\bar{g}} E_g(\vec{\beta}, \vec{\theta}; \mathbf{S}) + \frac{\lambda_{\theta_b}}{\lambda_a} E_{\theta_b} + \frac{\lambda_{\theta_h}}{\lambda_a} E_{\theta_h} \qquad (22)$$

followed by coupled alignment

$$E(\vec{\theta}, \mathbf{V}; \mathbf{S}, \lambda_a) = \lambda_a \lambda_g E_g(\mathbf{V}; \mathbf{S}) + \frac{\lambda_c}{\lambda_a} E_c + \frac{\lambda_{\theta_h}}{\lambda_a} E_{\theta_h} + \lambda_r E_r. \qquad (23)$$

Subsequent frames are optimized with coupled alignment only. All the terms are defined as in Section 4.3.1 By adopting the two stage approach of first acquiring a subject-specific template and then tracking with subject-specific registration, we reduce the dimensionality of the latter by 20 parameters and, most importantly, make the registrations more accurate and robust.

4.3.3 Optimization. The objective functions (12), (20), (22), (23) are minimized using a gradient-based dogleg minimization [Nocedal and Wright 2006]. Gradients are computed with automatic differentiation using OpenDR [Loper and Black 2014]. The optimization takes approximately 4 minutes per frame on a 3, 7 GHz Quad-Core Intel Xeon E5 computer using 4 threads, where only the search of closest-points is multithreaded.

Fig. 11. Representative scans of the test-dataset created to evaluate the MANO model. The dataset contains 50 scans of 6 subjects performing single- and double-finger articulation, as well as a more coordinated movement of all fingers.

5 EXPERIMENTS

In the following we present quantitative and qualitative evaluation of MANO and qualitative evaluation of the SMPL+H model.

5.1 MANO Evaluation

Compactness. Figure 12 plots the compactness of the MANO pose and shape space, left and right, respectively. These plots show the variance of the training-set captured by a varying number of components. The plot for the pose space shows that 6, 10 and 15 components explain 81%, 90% and 95% correspondingly of the full space. Although more components lead to a lower error, they also lead to over-fitting to noise and higher computational demands. Therefore, a larger number of components is suitable for clean hand-only datasets while full-body data benefits for the extra regularization provided by a small number of components.

Generalization. To evaluate the generalization capabilities of the pose space in MANO we create a test-dataset of 50 right hand scans as shown in Figure 11, in which scans with severe occlusions have been removed. The subjects in this dataset are not part of the data

used for model training. Figure 13 shows the generalization plot for the pose space, illustrating how the trained model generalizes to unseen poses. The plot depicts the mean and standard deviation of the scan-to-mesh (s2m) error in millimeters (mm) for each number of components for the low-D space. The plot shows that MANO generalizes nicely and the error decreases monotonically for an increasing number of components for the low-D pose space. As the number of components increases, the curve of the low-D space approximates asymptotically the minimum error of 0.93 mm achieved by the full space.

To study the generalization of the shape space, the 6 subjects of the test-dataset in Figure 11 are insufficient. Therefore, we perform a leave-one-out evaluation of the shapes in the *training set.* We report the mean absolute error (mabs) on the template vertices of each test-subject. Figure 14 depicts the generalization plot for the shape space, that shows how the model generalizes to unseen shapes. Similarly to Figure 13 for the pose space, this plot for the shape space decreases monotonically for an increasing number of components.

Training dataset size. We then study the performance of our model trained on different datasets. First we examine the effect of including mirrored poses in the dataset by comparing a model trained only on the right hand poses, and another trained on the dataset augmented with the mirrored left ones. While the former approach could potentially exploit the differences between hands by creating two specific MANO models for left and right hands, the latter approach exploits more training data for both hands by fusing the datasets. For the full-dimensional pose space, the former approach has an error of 1.05 mm and the latter 0.93 mm. Both approaches were evaluated in the independent test-dataset, as shown in Figure 11. Since this test dataset contains bigger variance for the pose space, we choose to use a fixed personalized template for each subject and optimize only over pose. For brevity we include in **supplementary material** a generalization plot that compares the two approaches for a varying number of components for the pose space. As previously described, the chosen model for MANO is the one that exploits both right and mirrored left data, due to its lower error and more coherent training procedure.

Effect of pose blend shapes. Regarding the effect of pose blend shapes, we fit the test-dataset scans with the pose blend shapes activated and deactivated, using again a fixed personalized template for each subject. The latter is essentially an LBS approach performing similarly to methods like [Ballan et al. 2012; Taylor et al. 2016; Tzionas et al. 2016]. The error for the full pose space is 0.93 mm for activated blend shapes, while for deactivated 1.3 mm. This numerical gain is in accordance with the qualitative results shown in Fig. 9. We thus conclude that the pose blend shapes facilitate not only improved visual realism, but also better fitting.

Effect of parameter learning. One remaining question concerns the overall effect of learning the model parameters $(\mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}, \bar{\mathbf{T}})$ on the accuracy of fitting new data. Since we want to study the overall effect of learning all parameters, for this experiment we change the evaluation protocol, and we optimize over both pose and shape during fitting. For the independent data in Figure 11, an untrained model reports for the full pose-space an s2m error of 2.90
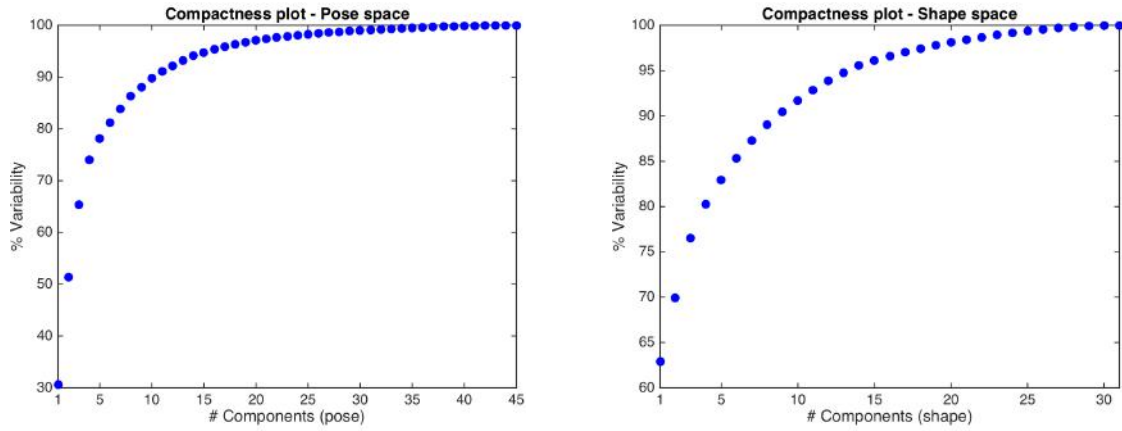
Fig. 12. Compactness plots for the pose (left) and shape (right) space. For the former, 6 and 15 components explain 81% and 95% correspondingly of the space.
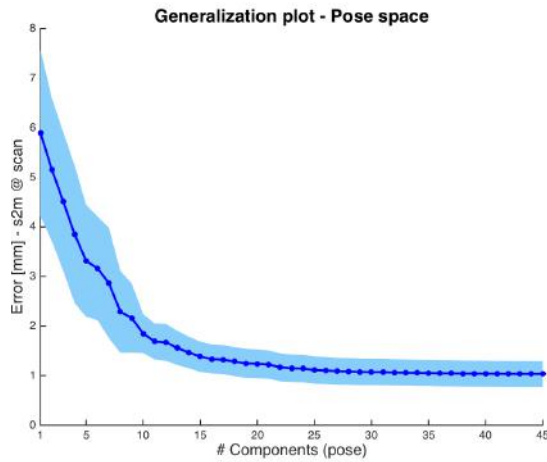

Fig. 13. Generalization plot for the pose space. We report the mean scan-to-mesh (s2m) error for a varying number of components for the low-D pose space. For each number of components we depict the mean and standard deviation of the error. The error for the full space is 0.93 mm.
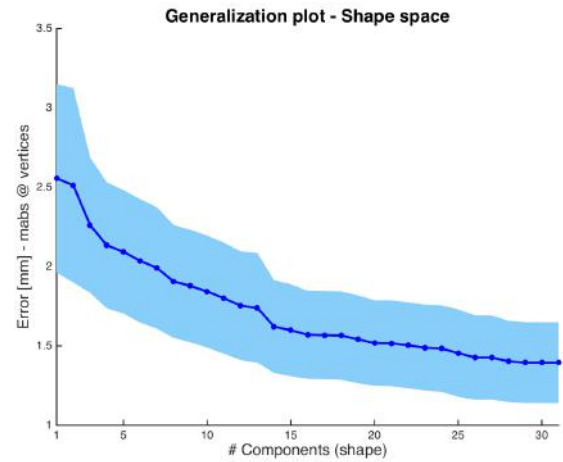
Fig. 14. Generalization plot for the shape space. For benchmarking we follow a leave-one-out approach on the training subjects. We report the mean absolute error (mabs) on the vertices of the template of each test-subject, for a varying number of components for the low-D shape space. For each number of components we depict the mean and standard deviation of the error.

mm, while a model with learned parameters is more than 60% better, with an error of 1.01 mm. For brevity we include in **supplementary material** another generalization plot that compares a varying number of pose components.

5.2 SMPL+H Evaluation

Our test sequences for SMPL+H are captured with a scanner configured to accomodate full bodies and not only hands. Therefore the hands are often extremely noisy as shown in the scans (top row) of Figures 1, 15, 17. Please note that, as shown in Figure 17 (top), it is not uncommon that the whole hand or even the arm completely disappears in the input scans for several consecutive frames. The above challenges call for a strong low-D model as described in Section 3.1

for MANO and Section 3.4 for SMPL+H. Based on the compactness plot for the pose space shown in Figure 12, we use 6 components for the evaluation in this section as a trade-off between accuracy and robustness.

As shown in Figure 15, the holistic approach of SMPL+H improves significantly the visual realism compared with SMPL. The figure shows that SMPL (Fig. 15 green) always keeps flat open hands, which often results in a lack of realism and expressiveness. In contrast, SMPL+H (Fig. 15 pink) is able to accurately capture expressive hand poses, significantly increasing realism compared with SMPL.

This joint modeling proves robustness against severe conditions that traditionally challenge performance capture methods. Figure
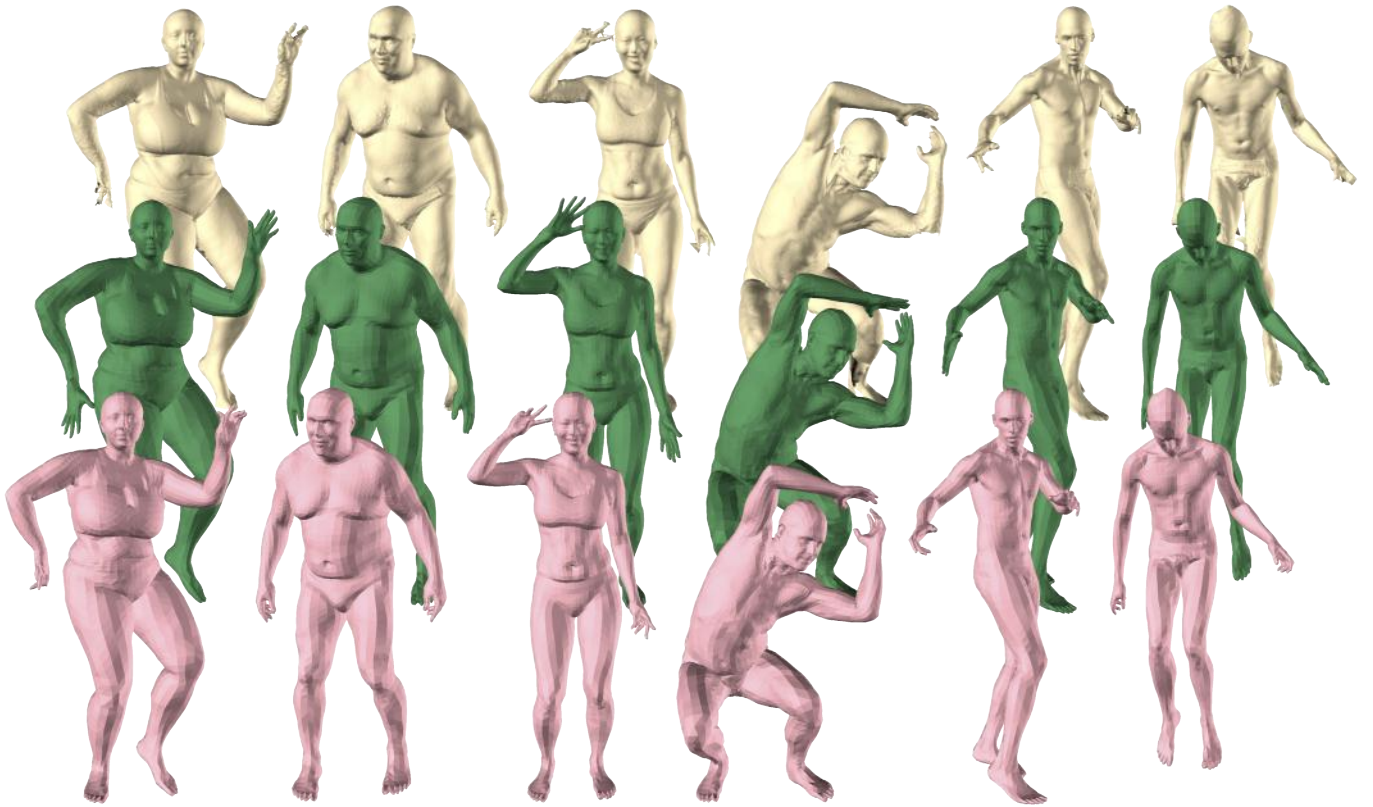
Fig. 15. Example 3D scans (*white*) from our 4D sequences and corresponding registrations of SMPL+H (*pink*). A small discrepancy between (white) scans and corresponding (pink) registrations suggests an expressive model and accurate registration method. Our holistic approach results in natural motion capture even under challenging conditions. We compare SMPL+H (*pink*) with SMPL [Loper et al. 2015] (*green*). The latter is representative of existing state-of-the-art methods without joint modeling of body and hands, and always results in open flat hands. On the contrary, SMPL+H performs significantly more realistically, capturing accurate and expressive hand poses.

17 shows that SMPL+H performs naturally even under severe missing data, fast motions or measurement noise, wich are especially prevalent for the case of hands.

Finally, it is worth noting that our method is not perfect. Two typical failure cases are shown in Figure 16: scans of people interacting with objects, and hand poses which are not well covered by the low-dimension pose space.

Extensive results can be seen in the **supplementary video**.

6 CONCLUSIONS

We propose MANO, a new model of the human hand that is learned from examples, is low-dimensional, is easy to pose and fit to data, and is compatible with graphic engines because it is built on linear blend skinning with blend shapes. The realism of the shape model and its simple formulation makes it appealing for many applications.

We combine MANO with a model of the body to produce SMPL+H, the first model of human body shape that is learned from examples and includes full hand articulation. The expressiveness and robustness of this model is demonstrated by fitting it to 4D scan

Fig. 16. Typical failure cases for hands. Each pair shows the scan (white) and the corresponding registration of SMPL+H (pink). Left: At the moment hand-object interaction is not explicitly modeled, thus wrong correspondences are established between the hands of the model and the object scan data. Right: Unusual hand poses may fall outside the low-dimensional pose space.
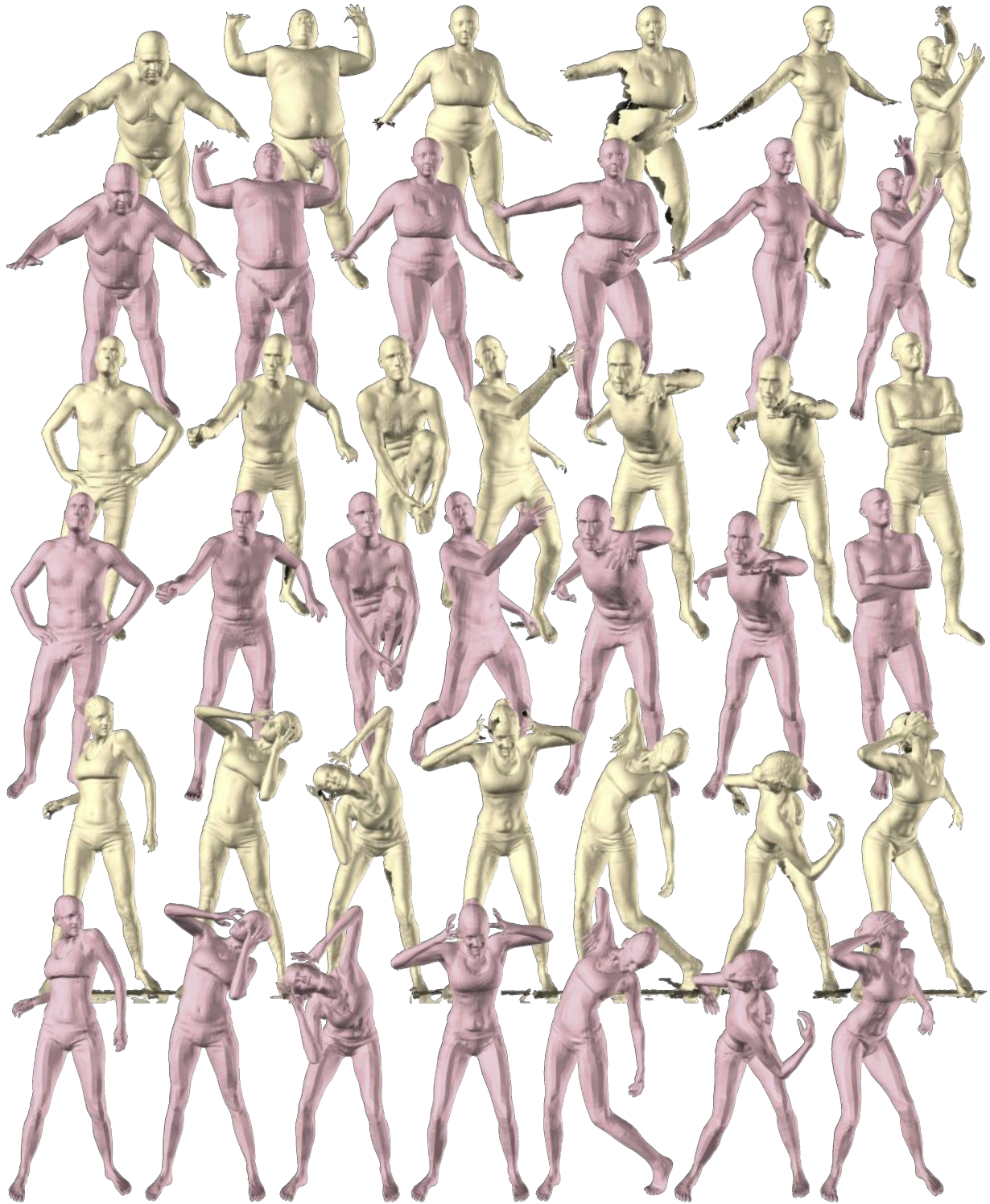
Fig. 17. Example 3D scans (*white*) from our 4D sequences and corresponding registrations of SMPL+H (*pink*). A small discrepancy between (white) scans and corresponding (pink) registrations suggests an expressive model and accurate registration method. Our holistic approach results in natural motion capture even under challenging conditions, like the illustrated cases of severe missing data due to fast motion, occlusion, finger-webbing or measurement noise. Please note that it is not uncommon that the whole hand or even the arm completely disappear for several frames, posing increased challenges.

data of people performing natural movements. We show that the optimization is stable even when the raw data is noisy, low resolution, and partially missing. By capturing the hand and body together we extract avatars that appear natural in their movements.

6.1 Discussion

Model: While here we build on top of SMPL [Loper et al. 2015], our dataset, design choices, and insights can be applied in training related models of the hand using representations like SCAPE [Anguelov et al. 2005].

Training dataset: The quantitative evaluation of MANO in Section 5.1 shows that our model benefits from augmenting the dataset through mirroring. This points to the need for richer training datasets in the future, potentially using dynamic 4D scan sequences instead of static 3D scans.

Real-time applications: Blend shapes encode vertex offsets (per-vertex displacement), as described in Section 3, thus their memory footprint is comparable to the footprint of the mesh. Furthermore, they are linear as described with Equations (3) and (4), thus their runtime is comparable to LBS. As a result our model with linear blend shapes is both memory and runtime efficient, therefore it is suitable also for real-time applications.

Low-dimensional pose space efficiency: Computational efficiency is further aided by the faster optimization in the low-dimensional pose space. However a low-dimensional space can not model the full space by definition. When subjects use their hands in a natural manner, it explains their poses well, while complex unnatural poses may fall outside this space. In the latter case the low-dimensional pose space can be used as an computationally efficient initialization for a subsequent optimization in the full pose space. Alternatively, we plan to train non-linear latent space models using deep learning but this will require more training data.

Failure cases: Our SPML+H sequences (Sec. 4.2) include failure cases with unusual hand poses that fall outside the low-dimensional pose space, as described above. Example cases are included in our **supplementary video**. Furthermore, we capture a sequence that includes hand-object interaction (a person picks up a heavy ball, see Figure 16), a typical failure case since at the moment we do not explicitly model hand-object interaction. However such cases can be tackled in future work inspired by works like [Ballan et al. 2012; Bambach et al. 2015; Oikonomidis et al. 2011b; Rogez et al. 2015; Sridhar et al. 2016; Tzionas et al. 2016; Tzionas and Gall 2015]. Please note that all failure cases are included in our dataset [MANO web] to spur future work.

Industrial applications: Our approach is relevant for industrial applications in several different ways. First, while our capture environment is complex, the ability to capture fully body and hand performance without markers may make this appropriate for high-end applications. Second, the body model with hand articulation is directly usable by animators as it is designed to be compatible with existing animation systems. We believe this is the first learned model of bodies and hands for which this is true. Third, this provides a practical step towards the simultaneous capture of hands

and bodies, which will enable animators to learn how body pose and hand pose are correlated; this is useful both as reference material and for training machine learning models relating the two.

6.2 Limitations and Future Work

There are several avenues for future work. We showed tracking results with SMPL+H that involved small amounts of self contact but we currently do not explicitly reason about this. A major limitation of previous 3D body shape tracking systems is that they do not allow self contact, while many natural motions and poses require it. We argue that a good hand model is valuable towards this direction, as we should be better able to fit it to scans where it is in contact with other body parts.

Contact with other objects is also important. Our hand training set includes such cases but we remove the objects and do not explicitly reason about hand-object contact during fitting. Of course, reasoning about such interaction with objects requires one to also reason about the object's surface. This is beyond our scope here but the MANO model should help facilitate research in this direction. The extension with interaction-specific blend shapes could be useful to model non-rigid skin deformation during manipulation.

We believe that high-quality data such as our dataset [MANO web] is crucial to build a strong model like MANO. However, after creating a strong model with such costly data, MANO is valuable for fitting lower quality data of other modalities like RGB-D or RGB, or for training methods to deal with such data.

The same applies for the combination of body and hands. Here we fit SMPL+H to 4D capture data. It would be interesting to fit the model to lower-quality RGB-D sequences. Since body models like SMPL have been fit to such data [Bogo et al. 2015] with high-quality results, doing so for the body and hands together may be possible. Again the challenge will be estimating the fingers from low-resolution data.

To address that challenge, we plan to use MANO and SMPL+H for a very different purpose – to generate realistic training data for deep learning, a topic gaining increasing research focus [Simon et al. 2017; Varol et al. 2017]. The MANO training set includes high-quality texture maps so we can generate hands with realistic textures, pose them, light them, and then put them in scenes with varying backgrounds. We then plan to train deep networks to estimate the shape and pose of the hand from images.

Finally we plan to use SMPL+H registrations to train models capturing how hand motion is correlated with body motion. Existing approaches [Jörg et al. 2012] are based on synthetic data from animations. Inspired by them, we would like to capture hand and body correlations based on real markerless mocap data, while using realistic pose and shape spaces. Existing mocap datasets [CMU 2000; Ionescu et al. 2014] do not contain hand information and SMPL+H could fill this gap. Given sufficient SMPL+H registrations, learning a mapping from body motions to hand poses would be possible. This would help towards more realistic and robust motion capture, and towards enriching existing animations or mocap sequences by adding correlated and plausible hand poses.

ACKNOWLEDGMENTS

REFERENCES

2000. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu. (2000).

3dMDbody 2017. 3dMDbody. http://www.3dmd.com/3dmd-systems/4d/3dmddynamic-body. (2017). Accessed: 01/09/2017.

3dMDhand 2017. 3dMDhand. http://www.3dmd.com/3dmdhandfoot-system. (2017). Accessed: 01/09/2017.

Brett Allen, Brian Curless, and Zoran Popović. 2003. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 22, 3 (2003), 587–594.

Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. 2006. Learning a Correlated Model of Identity and Pose-dependent Body Shape Variation for Real-time Synthesis. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '06)*. Eurographics Association, 147–156.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 408–416.

Vassilis Athitsos and Stan Sclaroff. 2003. Estimating 3D Hand Pose from a Cluttered Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 432–439.

Luca Ballan, Aparna Taneja, Juergen Gall, Luc Van Gool, and Marc Pollefeys. 2012. Motion Capture of Hands in Action using Discriminative Salient Points. In *European Conference on Computer Vision (ECCV)*. 640–653.

Sven Bambach, Stefan Lee, David Crandall, and Chen Yu. 2015. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. In *International Conference on Computer Vision (ICCV)*. 1949–1957.

Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. In *International Conference on Computer Vision (ICCV)*. 2300–2308.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision – ECCV 2016*.

Matthieu Bray, Esther Koller-Meier, and Luc Van Gool. 2007. Smart Particle Filtering for High-Dimensional Tracking. *Computer Vision and Image Understanding (CVIU)* 106, 1 (2007), 116–129.

Richard G. Brereton. 2015. The Mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics* 29, 3 (2015), 143–145.

Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. 2013. Tensor-Based Human Body Modeling. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 105–112.

Martin De La Gorce, David J. Fleet, and Nikos Paragios. 2011. Model-Based 3D Hand Pose Estimation from Monocular Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 33, 9 (2011), 1793–1805.

Laura Dipietro, Angelo M Sabatini, and Paolo Dario. 2008. A survey of glove-based systems and their applications. *Transactions on Systems, Man, and Cybernetics* 38, 4 (2008), 461–482.

Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. 2015. Learning to generate chairs with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1538–1546.

George ElKoura and Karan Singh. 2003. Handrix: animating the human hand. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 110–119.

Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding (CVIU)* 108, 1-2 (2007), 52–73.

Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. 2016. The GRASP Taxonomy of Human Grasp Types. *IEEE Transactions on Human-Machine Systems* 46, 1 (2016), 66–77.

Oren Freifeld and Michael J. Black. 2012. Lie Bodies: A Manifold Representation of 3D Human Shape. In *European Conf. on Computer Vision (ECCV) (Part I, LNCS 7572)*, A. Fitzgibbon et al. (Eds.) (Ed.). Springer-Verlag, 1–14.

Stuart Geman and Donald E. McClure. 1987. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, Vol. 52.

Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. 2010. An Object-Dependent Hand Pose Prior from Sparse Training Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 671–678.

Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. 2009. Tracking a hand manipulating an object. In *International Conference on Computer Vision (ICCV)*. 1475–1482.

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. A statistical model of human pose and body shape. *Computer Graphics Forum* 28, 2 (2009), 337–346.

Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn, and Hans-Peter Seidel. 2010. Learning Skeletons for Shape and Pose. In *Proceedings of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D '10)*. ACM, New York, NY, USA, 23–30.

Tony Heap and David Hogg. 1996. Towards 3D hand tracking using a deformable model. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 140–145.

David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. 2012. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conf. on Computer Vision (ECCV)*. 242–255.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

Sophie Jörg, Jessica Hodgins, and Alla Safonova. 2012. Data-driven finger motion synthesis for gesturing characters. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 189.

Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. 2015. Learning an Efficient Model of Hand Shape Variation From Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2540–2548.

David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: Freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *ACM Symposium on User Interface Software and Technology (UIST)*. 167–176.

Leap 2017. Leap Motion. https://www.leapmotion.com. (2017). Accessed: 01/09/2017.

John P. Lewis, Matt Cordner, and Nickson Fong. 2000. Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-driven Deformation. In *ACM Transactions on Graphics (SIGGRAPH)*. 165–172.

Charles Loop. 1987. *Smooth Subdivision Surfaces Based on Triangles.* Department of Mathematics, University of Utah.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Matthew M. Loper and Michael J. Black. 2014. OpenDR: An Approximate Differentiable Renderer. In *Computer Vision - ECCV 2014 (Lecture Notes in Computer Science)*, Vol. 8695. Springer, 154–169.

John MacCormick and Michael Isard. 2000. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision (ECCV)*. 3–19.

MANO. web. http://mano.is.tue.mpg.de. (web). Accessed: 2017-09-01.

Stan Melax, Leonid Keselman, and Sterling Orsten. 2013. Dynamics Based 3D Skeletal Hand Tracking. In *Graphics Interface*. 63–70.

Alex Mohr and Michael Gleicher. 2003. Building efficient, accurate character skins from examples. *ACM Transactions on Graphics (TOG)* 22, 3 (2003), 562–568.

Jorge Nocedal and Stephen Wright. 2006. *Numerical optimization.* Springer Science.

Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. 2015. Training a Feedback Loop for Hand Pose Estimation. In *International Conference on Computer Vision (ICCV)*. 3316–3324.

Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2011a. Efficient model-based 3D tracking of hand articulations using Kinect. In *British Machine Vision Conference (BMVC)*. 101.1–101.11.

Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2011b. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *International Conference on Computer Vision (ICCV)*. 2088–2095.

Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2012. Tracking the articulated motion of two strongly interacting hands.. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1862–1869.

Iason Oikonomidis, Manolis I.A. Lourakis, and Antonis A. Argyros. 2014. Evolutionary Quasi-random Search for Hand Articulations Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3422–3429.

Edwin Olson and Pratik Agarwal. 2013. Inference on Networks of Mixtures for Robust Robot Mapping. *International Journal of Robotics Research (IJRR)* 32, 7 (2013), 826–840.

Wilder Penfield and Edwin Boldrey. 1937. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60, 4 (1937), 389.

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. 2015. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 34, 4 (July 2015), 120:1–120:14.

Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and Robust Hand Tracking from Depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1106–1113.

James M. Rehg and Takeo Kanade. 1994. Visual tracking of high dof articulated structures: an application to human hand tracking. In *European Conference on Computer Vision (ECCV)*. 35–46.

Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. 2002. *Civilian American and European Surface Anthropometry Resource (CAESAR) Final Report.* Technical Report AFRL-HE-WP-TR-2002-0169. US Air Force Research Laboratory.

Gregory Rogez, James S. Supančič III, and Deva Ramanan. 2015. First-Person Pose Recognition Using Egocentric Workspaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4325–4333.

Javier Romero, Hedvig Kjellström, Carl H. Ek, and Danica Kragic. 2013. Non-parametric hand pose estimation with object context. *Image and Vision Computing* 31, 8 (2013), 555 – 564.

Marco Santello, Martha Flanders, and John F. Soechting. 1998. Postural Hand Synergies for Tool Use. *Journal of Neuroscience* 18, 23 (1998), 10105–10115.

Tanner Schmidt, Richard Newcombe, and Dieter Fox. 2014. DART: Dense Articulated Real-Time Tracking. In *Robotics: Science and Systems (RSS)*.

Matthias Schröder, Jonathan Maycock, Helge Ritter, and Mario Botsch. 2014. Real-time hand tracking using synergistic inverse kinematics. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. 5447–5454.

Hyewon Seo, Frederic Cordier, and Nadia Magnenat-Thalmann. 2003. Synthesizing Animatable Body Models with Parameterized Shape Modifications. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '03)*. 120–125.

Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. 2015. Accurate, Robust, and Flexible Real-time Hand Tracking. In *ACM Conference on Human Factors in Computing Systems (CHI)*. 3633–3642.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *CVPR* (2017).

Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. 2015. Fast and Robust Hand Tracking Using Detection-Guided Optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3221.

Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. 2016. Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In *European Conference on Computer Vision (ECCV)*. 294–310.

Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. 2013. Interactive Markerless Articulated Hand Motion Tracking using RGB and Depth Data. In *International Conference on Computer Vision (ICCV)*. 2456–2463.

Björn Stenger, Paulo R.S. Mendonça, and Roberto Cipolla. 2001. Model-based 3D tracking of an articulated hand. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 310–315.

Erik B. Sudderth, Michael I. Mandel, William T. Freeman, and Alan S. Willsky. 2004. Visual Hand Tracking Using Nonparametric Belief Propagation. In *Workshop on Generative Model Based Vision*. 189–189.

James S. Supančič III, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. 2015. Depth-based hand pose estimation: data, methods, and challenges. In *International Conference on Computer Vision (ICCV)*. 1868–1876.

Andrea Tagliasacchi, Sofien Bouaziz, Mark Pauly, and Hao Li. 2016. Modern techniques and applications for real-time non-rigid registration. In *SIGGRAPH ASIA 2016 Courses*.

Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. 2015. Robust Articulated-ICP for Real-Time Hand Tracking. *Eurographics Symposium on Geometry Processing (SGP)* 34, 5 (2015), 101–114.

David Joseph Tan, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, and Jamie Shotton. 2016. Fits Like a Glove: Rapid and Reliable Hand Shape Personalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5610–5619.

Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. 2017. Latent Regression Forest: Structured Estimation of 3D Hand Poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 7 (July 2017), 1374–1387.

Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Toby Sharp, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. 2016. Efficient and Precise Interactive Hand Tracking through Joint, Continuous Optimization of Pose and Correspondences. *ACM Transactions on Graphics (SIGGRAPH)* 35, 4 (2016), 143:1–143:12.

Jonathan Taylor, Richard Stebbing, Varun Ramakrishna, Cem Keskin, Jamie Shotton, Shahram Izadi, Aaron Hertzmann, and Andrew Fitzgibbon. 2014. User-Specific Hand Modeling from Monocular Depth Sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 644–651.

Arasanathan Thayananthan, Björn Stenger, Philip H.S. Torr, and Roberto Cipolla. 2003. Shape context and chamfer matching in cluttered scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 127–133.

Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. 2016. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (TOG)* 35, 6 (2016).

Emo Todorov and Zoubin Ghahramani. 2004. Analysis of the synergies underlying complex hand manipulation. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 2. 4637–4640.

Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM Transactions on Graphics (TOG)* 33, 5 (2014), 169:1–169:10.

Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. 2016. Capturing Hands in Action using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision (IJCV)* 118, 2 (2016), 172–193.

Dimitrios Tzionas and Juergen Gall. 2015. 3D Object Reconstruction from Hand-Object Interactions. In *International Conference on Computer Vision (ICCV)*. 729–737.

Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In *CVPR*.

Robert Y. Wang and Jovan Popović. 2009. Real-time Hand-tracking with a Color Glove. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 63:1–63:8.

Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. 2013. Video-based hand manipulation capture through composite motion control. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 43:1–43:14.

Ying Wu, John Y. Lin, and Thomas S. Huang. 2001. Capturing natural hand articulation. In *International Conference on Computer Vision (ICCV)*. 426–432.

Chi Xu, Ashwin Nanjappa, Xiaowei Zhang, and Li Cheng. 2016. Estimate Hand Poses Efficiently from Single Depth Images. *International Journal of Computer Vision (IJCV)* 116, 1 (2016), 21–45.

Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. 2013. A Survey on Human Motion Analysis from Depth Data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. 149–187.

Wenping Zhao, Jinxiang Chai, and Ying-Qing Xu. 2012. Combining marker-based mocap and RGB-D camera for acquiring high-fidelity hand motion data. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 33–42.