Learning an Efficient Model of Hand Shape Variation from Depth Images

Sameh Khamis[1,2] Jonathan Taylor[2] Jamie Shotton[2] Cem Keskin[2] Shahram Izadi[2] Andrew Fitzgibbon[2]
[1]University of Maryland [2]Microsoft Research

Morphable models of the human body have been a great success story of computer vision and graphics. However, to our knowledge, no morphable model of the human hand has yet been constructed. The hand is in some senses ideal for such modeling: it is normally unclothed, and has huge potential for natural 3D user interfaces. Ballan *et al*. [2] demonstrate that extremely robust hand tracking is possible given a user-specialized hand model, but acquiring the model requires manual rigging and a multi-camera capture setup. Taylor *et al*. [4] demonstrate acquisition of a user-specialized model from a single depth camera, but require long calibration sequences in which all degrees of freedom of the hand have to be exercised.

In this paper, we build a morphable model of hands. Our input, captured from a single depth camera, is a small set of unordered images containing diverse subjects performing varied hand poses, together with a rough initialization pose for each frame. The keys to our approach are twofold.

First, we learn only those aspects of pose and shape that are not explained by a standard rigged model. This reduces the data requirements, but also has the advantage that the output of our system is a standard subdivision surface model driven by a linear blend skinning. This ensures our model can be evaluated extremely efficiently. In contrast, models such as SCAPE [1] and TenBo [3] involve an additional linear solve at test time, which, while readily implementable on GPUs, does represent significant additional computational cost.

Second, we fit the model jointly to all partial scans, rather than first building complete scans per subject and then attempting principal component analysis. As we show in experiments on synthetic and real data, this yields a better model even for unoccluded synthetic data, and a much better model with real scans that contain missing and noisy data.

Figure 1 illustrates how our end-to-end model jointly learns shape and pose parameters. Our shape model follows our intuition that the variation in the shape of a human hand (and skeleton) in a *single pose* is relatively compact and can be described by a low dimensional linear subspace. In particular, given a vector of shape parameters $\beta \in \mathbb{R}^K$, the M vertex positions

$$V(\beta;\mathcal{V}) = \sum_{k=1}^{K} \beta_k V_k \tag{1}$$

of a neutral hand mesh is recovered as a linear combination of K mesh basis matrices $\{V_k\}_{k=1}^K \subseteq \mathbb{R}^{3\times M}$. Likewise the locations

$$L(\beta;\mathcal{L}) = \sum_{k=1}^{K} \beta_k L_k \tag{2}$$

of the B bones of an underlying skeleton is recovered as a linear combination of K bone location basis matrices $\{L_k\}_{k=1}^K \subseteq \mathbb{R}^{3\times B}$.

We explicitly parameterize pose using a vector θ concatenating a set of joint angles, global orientation and translation. Our pose model specifies the articulated deformation that θ invokes on a mesh $V(\beta)$ in a neutral pose using the corresponding skeleton $L(\beta)$ and maps the neutral hand mesh and skeleton to a posed hand mesh $\mathcal{P}(\theta;V(\beta),L(\beta)) \in \mathbb{R}^{3\times M}$, using standard linear blend skinning.

Following [4], we represent the actual surface of our model using a subdivision surface. Given a fixed triangulation of M vertex positions $V \in \mathbb{R}^{3\times M}$, Loop subdivision defines a 'limit surface' $\mathcal{S}(V) \subset \mathbb{R}^3$ by an iterative mesh smoothing procedure, but may also be explicitly parameterized as

$$\mathcal{S}(\mathbf{u};V) : \Omega \times \mathbb{R}^{3\times M} \mapsto \mathbb{R}^3 \,, \tag{3}$$

which is a quartic polynomial mapping from a location $\mathbf{u}$, in an essentially 2D space Ω of surface coordinates, to a point on the 3D subdivision surface.

A major contribution of this work is showing how to jointly learn all the model parameters from a set of noisy depth images of users' hands. We assume we have a diverse set of S different subjects. For each subject s, we have F_s depth frames of the user performing various hand articulations. In each frame f, a set of N_{sf} data points $\{\mathbf{x}_{sfn}\}_{n=1}^{N_{sf}} \subset \mathbb{R}^3$ is extracted. We cast our objective as the problem of minimizing an energy function that measures how well the posed surface explains the data points. The data term is

$$E_{\text{data}} = \sum_s \sum_f \sum_n \min_{\mathbf{u}\in\Omega} \rho\left(\|\mathbf{x}_{sfn} - \mathcal{S}(\mathbf{u};\mathcal{P}(\theta_{sf};V(\beta^s),L(\beta^s)))\| \right) \tag{4}$$

where $\rho(\cdot)$ corresponds to a robust kernel applied to the point position error. The apparently complex 'closest point on a subdivision surface' term is optimized efficiently using the lifting trick used in [4]. The data term is added to regularization terms that encode priors about the basis representation $\mathcal{V} = \{V_k\}$ and $\mathcal{L} = \{L_k\}$, the shape parameters $\mathcal{B} = \{\beta^s\}$, the pose parameters $\Theta = \{\theta_{sf}\}$, and the the skinning weights. We can then learn the model parameters by jointly minimizing the energy over all parameters simultaneously.

Our experiments demonstrate that the learned model is robust to noise in the training data. We also show that, at test time, the learned shape basis is able to generalize to data from unseen subjects in unseen poses and outperform a strong sensible baseline. Finally, we expect to accurately fit a personalized model from as few as one or two frames, a clear advantage over other approaches.

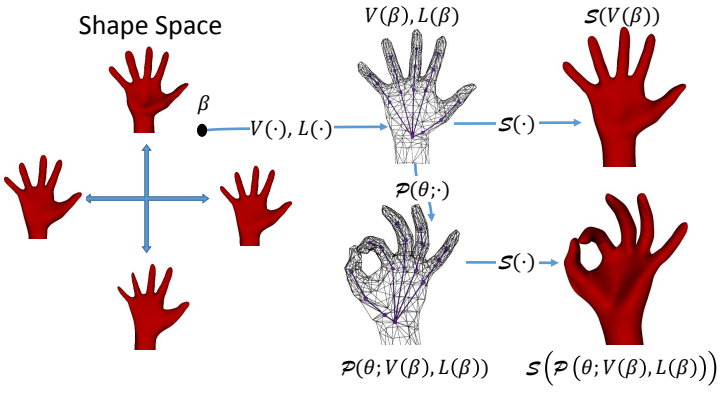

Figure 1: Our deformable surface model takes into account both pose (via an animation-ready kinematic model) and shape (in a shape space). A set of shape parameters $\beta \in \mathbb{R}^K$ in shape space (left) specifies (upper center) a neutral mesh $V(\beta) \in \mathbb{R}^{3\times M}$ and skeleton parameters $L(\beta) \in \mathbb{R}^{3\times B}$. A set of joint angles θ deforms the mesh to obtain a specific posed mesh $\mathcal{P}(\theta;V(\beta),L(\beta)) \in \mathbb{R}^{3\times M}$ (bottom left) using the linear blend skinning function $\mathcal{P}(\cdot)$. A subdivision surface function $\mathcal{S}(\cdot)$ maps these meshes to smooth 3D surfaces (right column). Simultaneously optimizing the parameters on the full pipeline from joint angles to 3D shape gives the parameters that best relate the end-to-end model to sparse and noisy real data.

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. *ACM Trans. Graphics*, 24(3), 2005.

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

[3] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *Proc. CVPR*, June 2013.

[4] J. Taylor, R. Stebbing, V. Ramakrishna, C. Keskin, J. Shotton, S. Izadi, A. Hertzmann, and A. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *Proc. CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.