Reconstructing Hand Shape and Appearance for Accurate Tracking from Monocular Video

Pratik Kalshetti
pmkalshetti@iitb.ac.in
Indian Institute of Technology Bombay
Mumbai, India

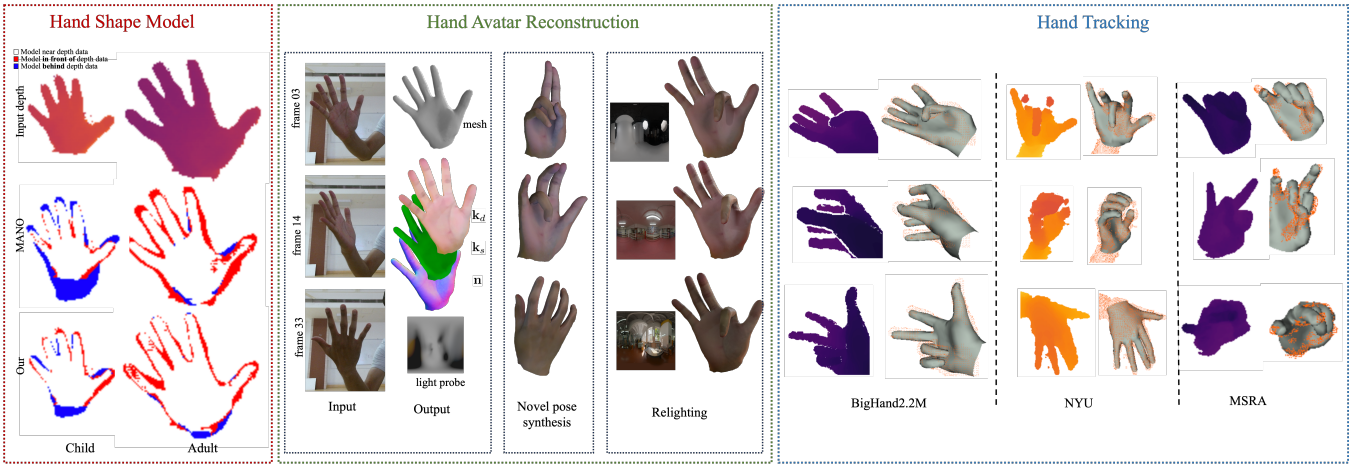

Figure 1: Our contributions include 1) Hand shape model: We introduce a new hand shape model that can represent various hand sizes compared to the popular hand model, MANO [Romero et al. 2017]. 2) Hand avatar reconstruction: Given a monocular RGB video of a user's hand, we reconstruct the user's hand geometry and appearance along with environment lighting. Our reconstructed avatar can be posed and rendered under novel lighting. 3) Hand tracking: We present a tracking framework that registers our proposed hand model to a sequence of depth frames and achieves competitive accuracy on various datasets.

ABSTRACT

A virtual animatable hand avatar capable of representing a user's hand shape and appearance, and tracking the articulated motion is essential for an immersive experience in AR/VR. Recent approaches use implicit representations to capture geometry and appearance combined with neural rendering. However, they fail to generalize to unseen shapes, don't handle lighting leading to baked-in illumination and self-shadows, and cannot capture complex poses. In this thesis, we 1) introduce a novel hand shape model that augments a data-driven shape model and adapt its local scale to represent unseen hand shapes, 2) propose a method to reconstruct a detailed hand avatar from monocular RGB video captured under real-world environment lighting by jointly optimizing shape, appearance, and lighting parameters using a realistic shading model in a differentiable rendering framework incorporating Monte Carlo path tracing, and 3) present a robust hand tracking framework that accurately registers our hand model to monocular depth data utilizing a modified skinning function with blend shapes. Our evaluation demonstrates that our approach outperforms existing hand shape and appearance reconstruction methods on all commonly used metrics. Further, our tracking framework improves over existing generative and discriminative hand pose estimation methods.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; *Tracking*.

KEYWORDS

hand shape and appearance reconstruction, hand pose estimation

ACM Reference Format:
Pratik Kalshetti. 2023. Reconstructing Hand Shape and Appearance for Accurate Tracking from Monocular Video. In *SIGGRAPH Asia 2023 Doctoral Consortium (SA Doctoral Consortium '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3623053.3623371

SA Doctoral Consortium '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0392-8/23/12…$15.00
https://doi.org/10.1145/3623053.3623371

1 INTRODUCTION

A user-specific digital hand avatar is essential to realize a realistic, immersive experience in virtual reality (VR) and augmented reality (AR) applications. In our context, an avatar is represented by

shape (e.g., triangle mesh) and appearance (e.g., spatially-varying material). The realistic appearance of the user's hand texture plays a vital role in providing a sense of embodiment, whereas accurate modeling of the user's hand shape and tracking is essential for precise interaction with virtual objects. We focus on the problem of automatically generating such a personalized hand avatar from a monocular RGB video and then animating it to resemble the user's hand pose in a monocular depth video.

There has been a lot of research in estimating the shape and pose of humans [Jiang et al. 2023] and faces [Grassal et al. 2022]. However, human hands have unique challenges, like large self-occlusion and substantial pose variation compared to full-body and face. Further, hands play a much more critical role in an interactive experience, so developing methods tailored for hands is essential.

State-of-the-art hand pose estimation methods use a data-driven hand shape model (e.g., MANO [Romero et al. 2017]) to estimate the pose of the hand from images [Boukhayma et al. 2019]. However, this data-driven hand model cannot adapt to unseen hand shapes with significant deviations from the training set and thus adversely affects tracking. We tackle this problem by introducing a new shape model *adaptive MANO (aMANO)* that augments MANO's shape space with local scale adaptation [Kalshetti and Chaudhuri 2022b]. This local scale adaptation enables capturing users with substantially different hand sizes than those covered by the original MANO shape space. Specifically, we use a set of local scale parameters that scale each of the bones in the hand model and a modified skinning function to handle this local scale adaptation. Additionally, to capture fine-level details, we subdivide the mesh and introduce vertex offsets from the template mesh along the normal.

To capture the appearance of human hands, state-of-the-art approaches broadly fall into learning a PCA-based texture basis [Li et al. 2022; Qian et al. 2020] from a large set of hand scans or directly estimating digital avatars from monocular video [Chen et al. 2023; Karunratanakul et al. 2023]. However, none of these approaches disentangle lighting from the intrinsic appearance of the user's hand and thus fail to account for self-occlusion. We propose a method to reconstruct a hand avatar from a monocular RGB video of a user's hand acquired under real-world environment lighting, within minutes. Our method jointly optimizes the parametric shape, material, and lighting using a realistic shading model in a differentiable rendering framework. Specifically, we use a differentiable deferred renderer [Hasselgren et al. 2022], which computes direct illumination using Monte Carlo path tracing. This enables us to generate realistic shading (c.f. HandAvatar [Chen et al. 2023]) under more general real-world environments (c.f. HARP [Karunratanakul et al. 2023]).

Finally, we present an accurate and robust tracking framework that registers our proposed shape model to a sequence of monocular depth images. Our articulated registration method embeds a blend-shape model with the modified skinning function into an energy-minimization formulation and fits the observed depth data to obtain accurate tracking. We also reparameterize the pose at each joint and enforce joint angle limits and PCA pose prior for robustness. These ideas allow us to achieve competitive tracking accuracy compared to state-of-the-art discriminative [Huang et al. 2020; Wan et al. 2018] and generative [Tagliasacchi et al. 2015; Tkach et al. 2017] hand tracking methods.

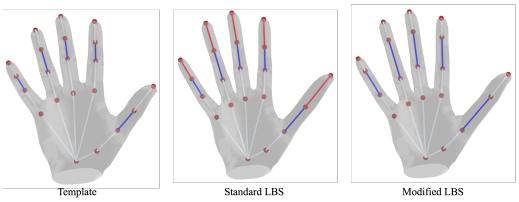

Figure 2: In standard LBS, if we scale the middle phalange (blue), the distal phalange (red) is also scaled. The modified LBS with our local scale parameters enables precise local control and avoids these undesired scaling artifacts.

We summarize our primary contributions below.

- We introduce a new hand shape model that augments the widely used MANO hand model, with local scale adaptation to capture unseen hand shapes with substantially different sizes. (Sec. 2) [Kalshetti and Chaudhuri 2022a,b]
- We propose a method to reconstruct a hand avatar from monocular RGB video of the hand captured under real-world environment lighting. Our method jointly optimizes hand shape and appearance in a differentiable rendering framework in minutes. (Sec. 3) [Kalshetti and Chaudhuri 2024]
- We present an accurate and robust tracking framework to register our proposed hand model to a sequence of depth images using articulated registration employing a modified skinning function with pose blend-shapes. (Sec. 4) [Kalshetti and Chaudhuri 2019, 2022b]

2 HAND SHAPE MODEL

Our parametric mesh model is based on the parametric human hand model, MANO [Romero et al. 2017], which is described by a function of pose $\theta \in \mathbb{R}^{45}$ (capturing local rotation at each of the 15 joints) and shape $\beta \in \mathbb{R}^{10}$ (coefficients for the PCA shape blend shapes) returning $N = 778$ vertices and $F = 1538$ faces. It is learned from 2018 scans of only 31 subjects (primarily adults). As a result, these PCA shape blend-shapes cannot capture hand shapes with significant deviations from training data.

We introduce a new hand shape model, *adaptive MANO (aMANO)*, that augments MANO's shape space with local scale adaptation [Kalshetti and Chaudhuri 2022b]. We incorporate this local scale adaptation into the standard LBS by anisotropically scaling the bone j in the reference frame using a local scale parameter, ϕ_j, as

$$v_i' = \sum_{j=1}^{n_b} W_{b_{ij}} \left\{ a_j' + R_j \left(W_{e_{ij}} s_j + (-a_j + v_i) \right) \right\} \tag{1}$$

where $s_j = (\phi_j - 1)(b_j - a_j)$ and W_e is the endpoint weight that captures the influence of a vertex relative to a bone's endpoints that is essential to avoid undesired scaling artifacts [Kalshetti and Chaudhuri 2022b] as shown in Fig. 2.

Additionally, to capture fine-level details, we subdivide the mesh and introduce vertex offsets $\mathbf{D} \in \mathbb{R}^{3N}$ from the template mesh. Our hand model is differentiable with respect to the model parameters, $\phi, \beta, \mathbf{D}, \theta$, thus enabling gradient backpropagation.

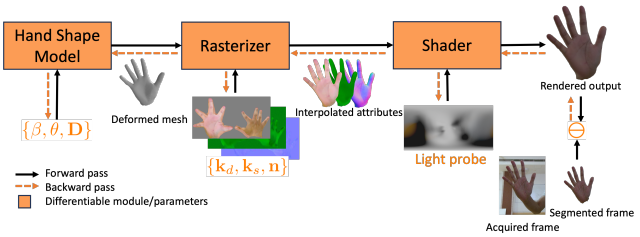

Figure 3: Hand avatar reconstruction pipeline.

3 HAND AVATAR RECONSTRUCTION

Given a monocular RGB video of a user's hand captured under real-world environment lighting, our goal is to reconstruct a personalized hand avatar of the user. We represent the hand avatar's geometry using our proposed hand shape model and its appearance using Disney's principled BRDF [McAuley et al. 2012]. We also assume that the hand is illuminated under an unknown environment lighting, which we model using a HDR environment probe.

Our method jointly optimizes the parametric mesh, material, and lighting using a realistic shading model in a differentiable rendering framework [Hasselgren et al. 2022] incorporating Monte Carlo path tracing. We first rasterize the scene into geometry buffers (G-buffer), including the surface intersection point and normal and interpolated material parameters for each pixel. Given the G-buffer, we perform the shading pass to determine the outgoing radiance. This requires us to calculate visibility, which is evaluated by tracing a shadow ray in the incoming light direction. This formulation allows us to handle self-shadows caused by self-occlusion among fingers. The rasterizer and shading are differentiable and thus allow gradients to be backpropagated to lighting, material, and geometry parameters. We use an analysis-by-synthesis approach such that the rendered optimized avatar under the optimized lighting matches the input video frames (see Fig. 3).

4 HAND TRACKING

Our hand tracking framework registers our hand shape model onto a sequence of depth frames. The registration energy is written as a weighted sum of several terms We refer the reader to Kalshetti

$E_{\text{data}_{3D}}$	the model explains the depth point cloud
$E_{\text{data}_{2D}}$	the model lies inside the observed sensor silhouette
E_{bound}	angle at each joint should respect kinematic bounds
E_{pca}	hand pose lies in a low-dimensional manifold
E_{int}	fingers cannot inter-penetrate
E_{reinit}	model's fingertips are close to detected fingertips
E_{shape}	avoid drifting from human hand shape
E_{temp}	avoid jittery tracking

and Chaudhuri [2022b] for details about these terms. We linearize each term and solve using Levenberg-Marquardt. Further, we use a discrete optimization over the 3D correspondences at each iteration: we sample a new set of barycenters and update the correspondences if any new correspondences are closer than the previous ones. This aids in faster convergence and allows a surprisingly small number of barycenters to be used at each iteration (see Fig 4).

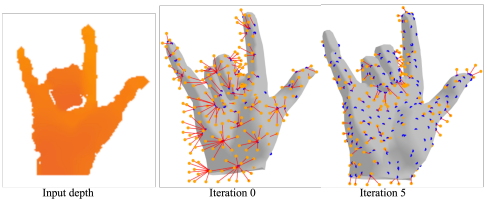

Figure 4: Fast convergence in our tracking framework where the red lines denote the correspondence between barycenters (blue) on the mesh and the point cloud (orange).

Table 1: Quantitative comparison of our proposed aMANO hand model with the popular MANO hand model on our captured dataset containing users across various demographics.

Demographics	E_{3D} (in mm)↓		E_{2D} (in pixels)↓	
	MANO	aMANO	MANO	aMANO
Children	5.4	**4.5**	1.491	**0.758**
Adult (female)	5.8	**5.5**	0.765	**0.523**
Adult (male)	5.9	**5.4**	0.471	**0.363**

Table 2: Quantitative evaluation of our hand avatar reconstruction on InterHand2.6M 30fps dataset [Moon et al. 2020].

Method	PSNR↑	SSIM↑	LPIPS↓
HARP [Karunratanakul et al. 2023]	16.157	0.866	0.167
HandAvatar [Chen et al. 2023]	29.423	0.914	0.088
Ours	**31.179**	**0.936**	**0.061**

5 EXPERIMENTAL RESULTS

We evaluate the capability of our proposed aMANO hand shape model over the state-of-the-art MANO hand model on our captured dataset in Table 1 We observe that MANO fails to capture hand sizes that are significantly far from its training data (e.g., children), whereas our proposed aMANO model gracefully adapts to hands of different sizes (see Fig. 1 (left)).

We also evaluate our hand avatar reconstruction method in Table 2 and observe that our method outperforms existing methods on all commonly used metrics. We also observe that our method accurately captures fine geometric details, as seen in the shading images in Fig. 5, whereas these details are missing in the shading images of HandAvatar. We further evaluate our rendering result with recent monocular methods using volume rendering (HumanNeRF [Weng et al. 2022]) and surface rendering (SelfRecon [Jiang et al. 2022]) in Fig. 6. Unlike our fast test-time optimization (minutes), these methods require long training times (hours).

Finally, we demonstrate the capability of our tracking framework by comparing it with state-of-the-art hand pose estimation methods on NYU and MSRA datasets in Table 3. Our method produces accurate and robust mesh registration on various datasets in complex poses (see Fig. 1 (right)).

Ethical Considerations. To protect users' fingerprint privacy, all RGB images in the paper are down-sampled before the acquisition.

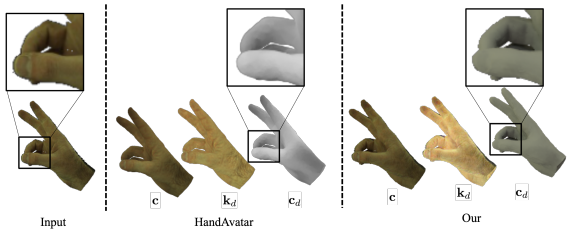

Figure 5: Compared to the local-pair occupancy field of HandAvatar [Chen et al. 2023], our method models shadow rays to produce more accurate self-shadows.

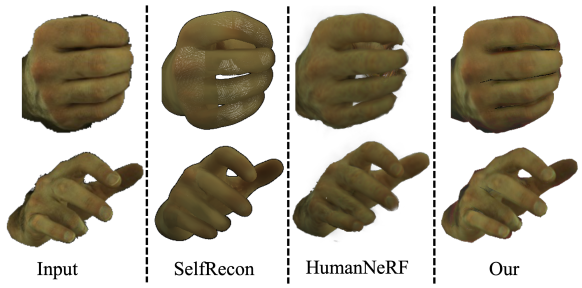

Figure 6: Our method reconstructs high-fidelity texture details compared to the surface rendering method of SelfRecon [Jiang et al. 2022] and correctly captures illumination compared to the volume rendering method of HumanNeRF [Weng et al. 2022].

Table 3: Quantitative evaluation of our tracking framework on NYU [Tompson et al. 2014] and MSRA [Sun et al. 2015].

Dataset	Method	E_{3D} (in mm)↓	E_{2D} (in pixels)↓
NYU	Huang et al. [2020]	10.1	0.242
	Our	**6.7**	**0.227**
MSRA	Wan et al. [2018]	6.8	0.819
	Our	**5.2**	**0.475**

6 CONCLUSION

We introduce an intuitive and mathematically robust extension to existing hand shape models to accommodate users with different hand sizes, including that of children, which was not possible until now. We propose a method to reconstruct the shape and appearance of a user's hand from a monocular RGB video under real-world environment lighting, within minutes. In contrast to existing methods where illumination is baked into the appearance, our method disentangles the intrinsic properties of the underlying appearance and environment lighting, leading to realistic self-shadows. We also present a state-of-the-art hand tracking framework that registers our hand model to depth images and improves over existing generative and discriminative hand pose estimation methods. Our work takes an important step to achieve an immersive experience in AR/VR systems by providing a sense of embodiment by capturing the shape and appearance of the user's hand and improving the accuracy and robustness of hand tracking.

Limitations and future work. Our method assumes that the hand is segmented in the input images and requires a reasonably good initialization, both of which suffer when the hand is occluded. Interesting future directions would be to extend our avatar model to capture pose-dependent muscle bulges and wrinkles to increase realism, and simulate contact forces during interaction with objects.

ACKNOWLEDGMENTS

I thank my PhD advisor, Parag Chaudhuri, for introducing the problem and supporting me throughout this journey.

REFERENCES

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 2019. 3d hand shape and pose from images in the wild. In *CVPR*.

Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. 2023. Hand Avatar: Free-Pose Hand Animation and Rendering From Monocular Video. In *CVPR*.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *CVPR*.

Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. 2022. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. In *NeurIPS*.

Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. 2020. AWR: Adaptive Weighting Regression for 3D Hand Pose Estimation. In *AAAI*.

Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *CVPR*.

Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2023. InstantAvatar: Learning Avatars From Monocular Video in 60 Seconds. In *CVPR*.

Pratik Kalshetti and Parag Chaudhuri. 2019. Unsupervised Incremental Learning for Hand Shape and Pose Estimation. In *ACM SIGGRAPH 2019 Posters*.

Pratik Kalshetti and Parag Chaudhuri. 2022a. Local Scale Adaptation for Augmenting Hand Shape Models. In *ACM SIGGRAPH 2022 Posters*.

Pratik Kalshetti and Parag Chaudhuri. 2022b. Local Scale Adaptation to Hand Shape Model for Accurate and Robust Hand Tracking. *Computer Graphics Forum* 41, 8 (2022), 219–229.

Pratik Kalshetti and Parag Chaudhuri. 2024. Intrinsic Hand Avatar: Illumination-aware Hand Appearance and Shape Reconstruction from Monocular RGB Video. In *WACV*. (to appear).

Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. 2023. HARP: Personalized Hand Reconstruction From a Monocular RGB Video. In *CVPR*.

Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. 2022. NIMBLE: A Non-Rigid Hand Model with Bones and Muscles. *ACM TOG* 41, 4, Article 120 (2022).

Stephen McAuley, Stephen Hill, Naty Hoffman, Yoshiharu Gotanda, Brian Smits, Brent Burley, and Adam Martinez. 2012. Practical Physically-Based Shading in Film and Game Production. In *ACM SIGGRAPH 2012 Courses*.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *ECCV*.

Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. 2020. HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *ECCV*.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM TOG* 36, 6 (2017), 245:1–245:17.

Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. 2015. Cascaded hand pose regression. In *CVPR*.

Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. 2015. Robust articulated-icp for real-time hand tracking. In *Computer Graphics Forum*.

Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon. 2017. Online generative model personalization for hand tracking. *ACM TOG* 36, 6 (2017), 1–11.

Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM TOG* 33 (2014).

C. Wan, T. Probst, L. Gool, and A. Yao. 2018. Dense 3D Regression for Hand Pose Estimation. In *CVPR*.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *CVPR*.