Embodied Hands: Modeling and Capturing Hands and Bodies Together **Supplementary Material**

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

ACM Reference Format:
Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together **Supplementary Material**. *ACM Trans. Graph.* 36, 4, Article 245 (July 2017), 2 pages. https://doi.org/10.1145/3130800.3130883

1 MODEL/SCAN MIRRORING

For the creation of the MANO hand model, we first 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. After capturing this data for both right and left hands, we create a single augmented dataset by mirroring the left hand scans to appear as right ones. This approach increases the size of the training data and removes the bias introduced by the handedness of the subjects. In practical terms, it results in a performance improvement as shown in the experiments section. The augmented dataset enables us to train a single consistent hand model for both hands, i.e. we train the right hand model and generate the left one by mirroring. Model components which depend on the global coordinate frame, like the mesh template $\bar{\mathbf{T}}$, the shape blend shapes B_S and the pose blend shapes B_P, require mirroring. The rest of the components (e.g. the blend weights $\mathcal{W}$ and joint regressor $\mathcal{J}$) remain untouched.

We define the sagittal plane in SMPL, x, as our mirroring plane. This entails the following mirroring transformation

$$M = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Mirroring the scan points and the mesh template is trivial, i.e. each scan point p and each vertex v of the template $\bar{\mathbf{T}}$ are mirrored as $p' = pM$ and $v' = vM$. Each of the shape blendshapes $\mathbf{S}_n$ (Eq. (8) in [Loper et al. 2015]) follow a similar procedure

$$B'_S = MB_S = \sum_{n=1}^{|\vec{\beta}|} \beta_n M\mathbf{S}_n \tag{2}$$

$$\mathbf{S}'_n = M\mathbf{S}_n \tag{3}$$

$$\tag{4}$$

since they map a (coordinate frame independent) shape coefficients β_n to vertex displacements.

The pose blend shapes are slightly different, since they are multiplied by hand pose rotations, which depend on the global coordinate frame. Therefore, the output-mirrored pose blend shapes $M\mathbf{P}_n$ should be modified to account for the mirror input transformation required to transform right poses into left ones.

To understand how to achieve this, consider first that a pose is mirrored by applying its rotation in a mirrored coordinate frame. This corresponds to pre- and post- multiplying its corresponding rotation matrix R by the commutative transformation M, $R' = MRM$.

Considering Eq. (9) [Loper et al. 2015], we want to obtain the mirrored pose blendshapes $\mathbf{P}'_n$ such that

$$B'_P \equiv \sum_{n=1}^{|\vec{\beta}|} (R'_n(\vec{\theta}) - R'_n(\vec{\theta}^*))\mathbf{P}'_n \tag{5}$$

$$= MB_P = \sum_{n=1}^{|\vec{\beta}|} (R_n(\vec{\theta}) - R_n(\vec{\theta}^*))M\mathbf{P}_n \tag{6}$$

where R'_n are the scalar elements of the mirrored rotations R'. Therefore, one can obtain the mirrored pose blendshapes $\mathbf{P}'_n$ by applying the rotation un-mirroring transformation (pre- and post-multiplying by M) to each of the 3×3 input blocks in $M\mathbf{P}_n$.

2 EVALUATION - MANO HAND MODEL

In the next figures we present variations of the generalization plot for several comparisons.

Figure 2 shows the performance of our model trained on different datasets. The red curve shows the MANO model trained only on the right hand poses (i.e. no mirroring augmentation). This approach requires to train two separate left and right MANO models, making the models biased by the handedness of the subjects. The blue curve shows the MANO model trained on an augmented dataset, for which left hand poses are mirrored and combined with the right hand ones. This approach leads to a single hand model from a single training procedure (the left model is generated by mirroring the resulting right hand model) with a larger training dataset and no handedness bias. It is our chosen approach for MANO. The plot shows that the latter approach performs favorably, pointing to the need for

*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.

0730-0301/2017/7-ART245
https://doi.org/10.1145/3130800.3130883

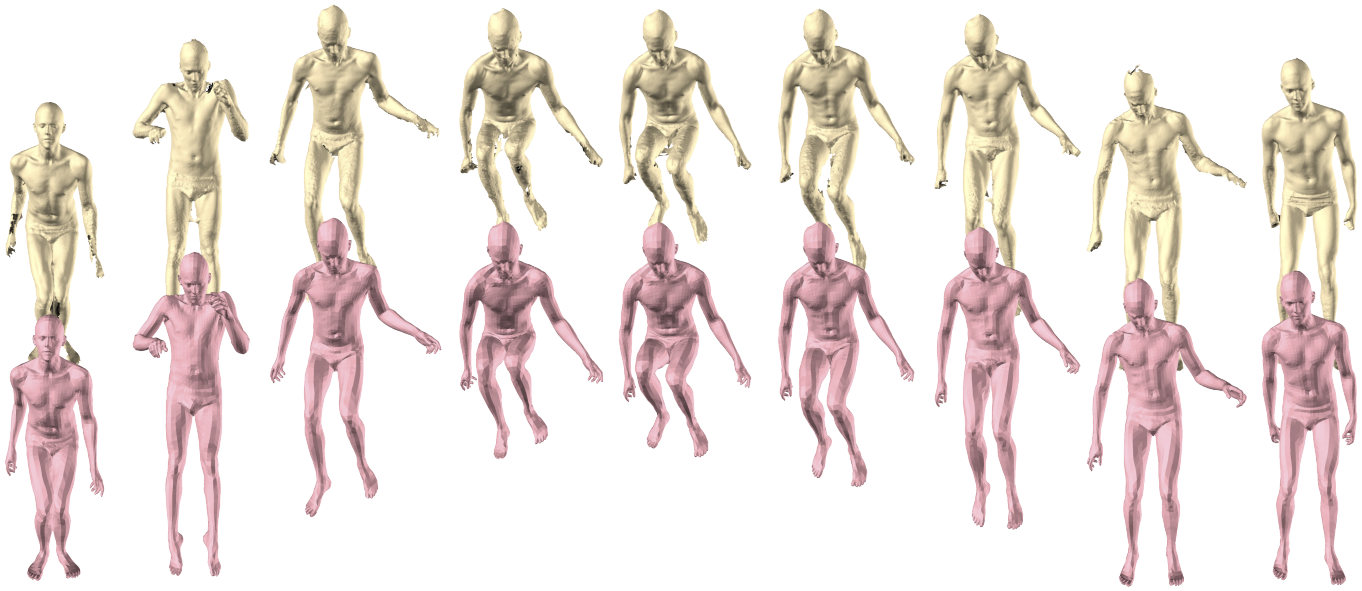

Fig. 1. Sample frames (non-sequential) from a jumping action. Fast motions result in missing hand data in the 3D scans, posing challenges to existing motion capture methods. The 3D scans are shown with *beige* color, while the resulting registration of SMPL+H is shown with *pink*. The proposed model and registration method result in natural motion capture even under fast motions and missing visual data for several frames, e.g. in the hand region.

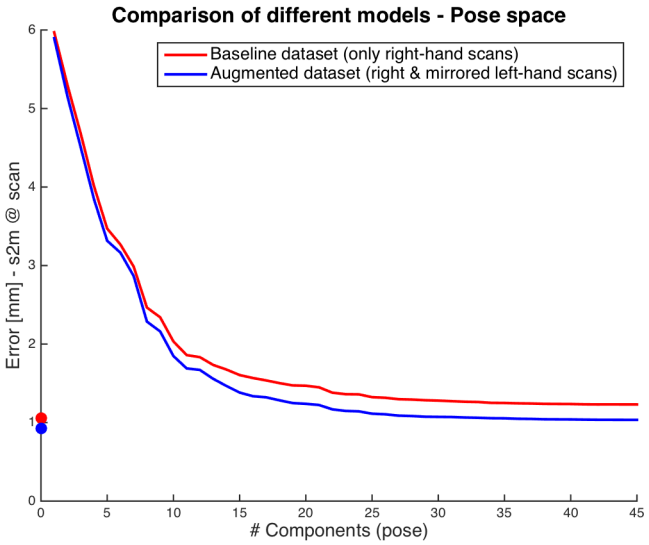

Fig. 2. Generalization plot for the model trained on different datasets. The *baseline dataset* is comprised of only right-hand scans, while the *augmented dataset* is comprised of right- and mirrored left-hand scans. We report the mean scan-to-mesh (s2m) error for a varying number of pose-space components. For $x = 0$ we show the error for the full space. The plot shows that an augmented dataset leads to a lower fitting error of the model, pointing to the need for bigger future training datasets.

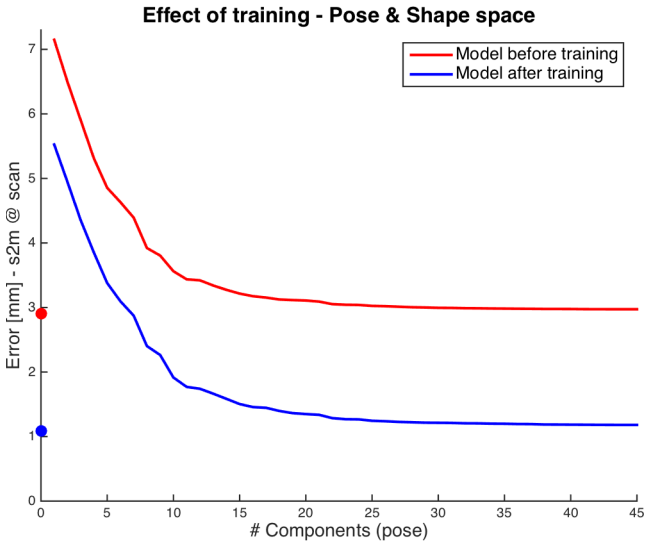

Fig. 3. Generalization plot for the model before and after training. We report the mean scan-to-mesh (s2m) error for a varying number of pose-space components. For $x = 0$ we show the error for the full space.

richer training datasets in the future, potentially using dynamic 4d scans instead of static 3d scans. For the full pose space the former approach has an error of 1.05 mm and the latter 0.93 mm.

Figure 3 shows a comparison of the error before and after training. For the full-space the initial model has an error of 2.90 mm, while the trained model has an error of 1.01 mm.

REFERENCES

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.