Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data – Supplemental Document –

Yuxiao Zhou[1] Marc Habermann[2,3] Weipeng Xu[2,3] Ikhsanul Habibie[2,3] Christian Theobalt[2,3] Feng Xu[1]

[1]BNRist and School of Software, Tsinghua University, [2]Max Planck Institute for Informatics, [3]Saarland Informatics Campus

In the following, we provide more quantitative evaluations and comparisons to previous methods (Sec. 1). Further, we show more qualitative results of our method on sequences from the internet and publicly available datasets (Sec. 2). Finally, we provide more technical details about our method (Sec. 3).

1. Additional Quantitative Evaluation

In the Dexter+Object (DO) [7] and EgoDexter (ED) [4] datasets only ground truth finger tips were annotated. Existing works adopted two methods for alignment before evaluation: [8, 1, 3] aligned the predictions with ground truth depth, and [2, 9, 6, 10] aligned the centroid of predicted and ground truth finger tips. In the main document, we report our results with centroid alignment. In Tab. 1, we report our results with both two alignment approaches, and compare with other works. We demonstrate that our approach is superior to others under both alignment methods.

We further implement a optimization-based IK solver to compare with the proposed IKNet. Following [5, 2], we use the main coefficients of pose PCA bases to represent the hand pose to make it physically plausible. The shape representation remains the same as for the IKNet. More specifically, the pose is represented by $\theta \in \mathbb{R}^{12}$, and the shape $\beta \in \mathbb{R}^{10}$. We define the energy function as the Euclidean distance between corresponding joints of the input and the posed hand model, and use Levenberg–Marquardt algorithm to solve for θ and β:

$$\theta, \beta = \arg\min_{\theta,\beta} \sum_{i=1} ||X_i - J_i(\theta, \beta)||_2,$$

where X_i is the input location of joint i, and $J_i(\theta, \beta)$ is joint i's location in the MANO [5] hand model with shape β and pose θ. The IK solver is evaluated on the same datasets as IKNet with DetNet's prediction as input. As shown in Tab. 2, the IK solver has a lower accuracy than our proposed IKNet. The main limitations of the IK solver are: first, the energy function is highly non-convex which lets the solver converge to local optima; second, without a complex pose prior, the solver is not robust to noise and errors in the input. In contrast, our IKNet naturally avoids the convexity issue as at test time there is only a single feed-forward pass. Further, it is able to handle inaccurate inputs thanks to the learned prior and the 3DPosData.

We also perform an ablation study on the delta maps of the DetNet. As shown in Tab. 3, the delta maps significantly improve the result. This is because the delta maps provide additional information about the relative positions of neighboring joints.

2. Qualitative Results

Our architecture can also perform motion capture on in-the-wild sequences by combining 2D and 3D predictions. More specifically, given the hand bounding box of the first frame, the model first estimates 2D and 3D joint locations from the corresponding image crops, and then updates the bounding box such that the 2D predictions lie in the center of the new bounding box, which is used for cropping the next frame. Thus, our system can predict hand pose from in-the-wild sequences fully automatically. In Fig. 1, we show our results on several in-the-wild sequences obtained from the internet. Note that our method can generalize to these unseen images and is also robust to occlusions, challenging poses and fast motions. In Fig. 2, we present more qualitative results on the DO and ED dataset with severe object and self occlusions. Again our results look plausible even for such challenging scenarios.

3. Pose Representation in the MANO Model

In the original MANO [5] model, the pose parameters $\bar{\theta} \in \mathbb{R}^{16\times3}$ represent the rotations of 16 joints. More specifically, $\bar{\theta}_j$ of joint j represents the rotation of the bone between j and its children $C(j)$ in the kinematic tree. Thus, the pose parameters $\bar{\theta}$ do not include 5 finger tips that do not have child joint. Under this representation, if one joint has more than one child joint, the child joints have to share the same rotation, e.g. the wrist joint. In our architecture,

Centroid Alignment				
Method	DO		ED	
	AUC	PCK	AUC	PCK
Ours	**.948**	**.816**	**.811**	**.611**
Zhang et al. [9]	.825	.600	-	-
Boukhayma et al. [2]	.763	.489	.674	.383
Z&B [10]	.573	.200	-	-
Spurr et al. [6]	.511	.220	-	-
Depth Alignment				
Ours	**.946**	**.808**	**.773**	**.572**
Xiang et al. [8]	.912	.741	-	-
Baek et al. [1]	.650	.359	-	-
Mueller et al. [3]	.482	.240	-	-

Table 1. Comparison with other methods with different alignment methods on DO and ED. The PCK is evaluated with an error threshold of $20mm$. Our approach achieves higher accuracy on all datasets for both types of alignment.

Method	DO		ED		STB	
	AUC	PCK	AUC	PCK	AUC	PCK
DetNet + IKNet	**.948**	**.816**	**.811**	**.610**	**.898**	**.732**
DetNet + IK Solver	.724	.432	.798	.566	.595	.274
DetNet	.923	.734	.804	.601	.891	.710

Table 2. Comparison between IKNet and a classical IK solver. The PCK is evaluated with an error threshold of $20mm$. While IKNet improves the accuracy, the IK solver lowers it due to the lack of complex prior knowledge.

Method	DO		ED		STB	
	AUC	PCK	AUC	PCK	AUC	PCK
DetNet	**.923**	**.734**	**.804**	**.601**	**.891**	**.710**
DetNet w/o D	.847	.595	.772	.546	.820	.628

Table 3. Evaluation of delta maps D in DetNet. The PCK is evaluated with an error threshold of $20mm$. Delta maps improve the result significantly by providing intermediate supervision on neighboring joints.

we altered the original representation to $\theta \in \mathbb{R}^{21\times 3}$ with additional 5 finger tip joints. More specifically, except for the root wrist joint, rotation θ_j for joint j refers to the rotation of the bone between j and its parent joint $P(j)$. This formulation enables that child joints with the same parent can have different poses, resulting in more varied poses, which can be learned from other datasets, e.g. our constructed 3DPos-Data.

Figure 1. Our results on sequences from internet. The bounding box is tracked by our 2D predictions. Note that our approach can generalize well to these images and accurately tracks the hand motion even though the sequence contains fast motions and challenging poses.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3d hand shape and pose from images in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *The IEEE International Conference on Com-*

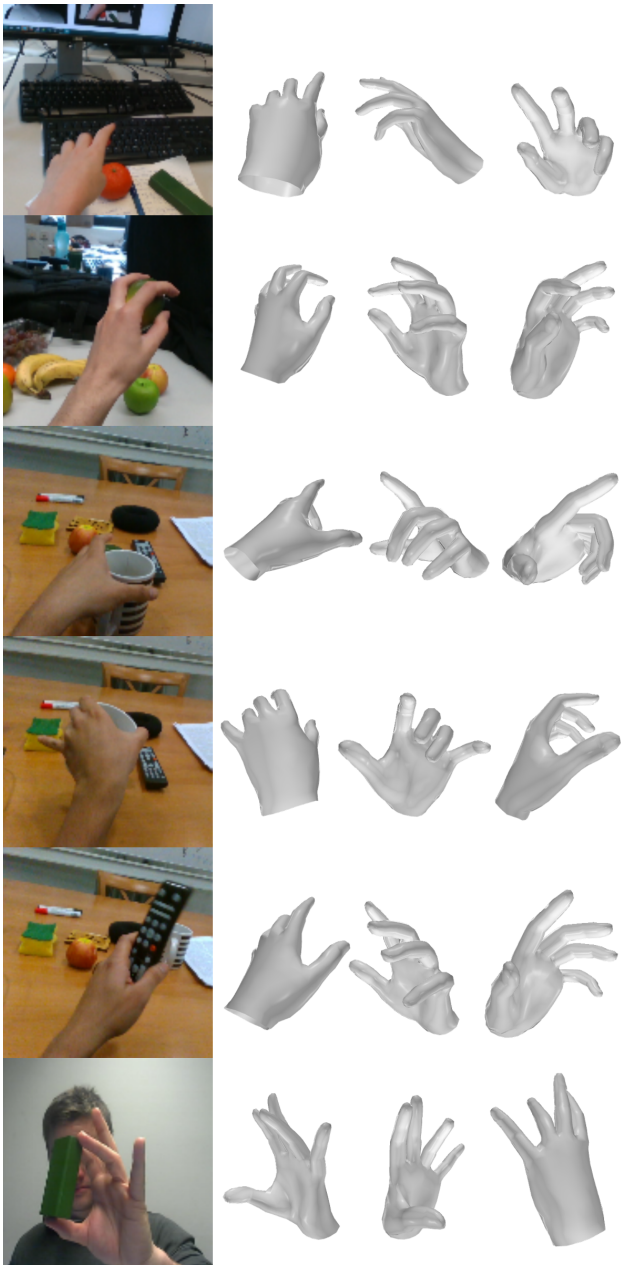

Figure 2. Our results on sequences from the DO and ED dataset. Note that our method is robust to self occlusions and hand object interactions.

puter Vision (ICCV), 2017.

[5] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, Nov. 2017.

[6] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *The European Conference on Computer Vision (ECCV)*, 2016.

[8] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[10] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.