Supplementary Material: UmeTrack: Unified multi-view end-to-end hand tracking for VR

SHANGCHEN HAN, PO-CHEN WU, YUBO ZHANG, BEIBEI LIU, LINGUANG ZHANG, ZHENG WANG, WEIGUANG SI, PEIZHAO ZHANG, YUJUN CAI, TOMAS HODAN, RANDI CABEZAS, LUAN TRAN, MUZAFFER AKBAY, TSZ-HO YU, CEM KESKIN, and ROBERT WANG, Meta Reality Labs, USA

ACM Reference Format:
Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. 2022. Supplementary Material: UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3550469.3555378

A NETWORK ARCHITECTURE DETAILS

The input shape, output shape, hidden state shape and the layers used for each module are shown in Table 1. The encoder uses the same resnet as [Han et al. 2020] to ensure fair comparisons. The last layer of the encoder is a 1×1 convolution layer for dimensionality reduction purpose. Multi-view fusion uses multiple 1×1 convolutions and ReLU layers. Each 1×1 convolution serves the purpose of feature fusion and dimensionality reduction. The output shape of the multi-view fusion module is the same as the output shape of the encoder. The temporal module is a recurrent neural network with a hidden state using 1×1 convolution and ReLU as the building blocks. Both Regressor-K and Regressor-U are built from residual blocks. The output of Regressor-K contains 20 dimensional joint angles and 21 dimensional root point coordinates. Regressor-U outputs a 1 dimensional hand scale parameter in addition to joint angle and root point outputs.

For root transform prediction, we pre-define 7 points for representing a transformation in the hand local space: $v_H = \{[0,0,0]^T, [1,0,0]^T, [0,1,0]^T, [0,0,1]^T, [1,1,0]^T, [1,0,1]^T, [0,1,1]^T\}$. And the task of a regressor is to predict the location of these points denoted as $\hat{v}$ in the reference camera space. The root transformation can be recovered using Singular Value Decomposition [Sorkine-Hornung and Rabinovich 2016] by solving the following equation:

$$\hat{T}_H = \min_{\hat{T}_H} \sum_i ||\hat{T}_H * v_{H,i} - \hat{v}_i||_2^2 \qquad (1)$$

Table 1. Architecture table

Module	Input	Output	Hidden state	Layers
Encoder	$1 \times 96 \times 96$	$72 \times 6 \times 6$	NA	resnet + Conv11
Multi-view fusion	$144 \times 6 \times 6$	$72 \times 6 \times 6$	NA	(Conv11 + ReLU) ×2 + Conv11
Temporal module	$72 \times 6 \times 6$	$72 \times 6 \times 6$	$18 \times 6 \times 6$	(Conv11 + ReLU) ×2 + Conv11
Skeleton encoder	120	$4 \times 6 \times 6$	NA	linear + reshape
Regressor-K	$76 \times 6 \times 6$	41	NA	residual blocks × 2 + Pool
Regressor-U	$72 \times 6 \times 6$	42	NA	residual blocks × 2 + Pool

REFERENCES

Shangchen Han, Beibei Liu, Randi Cabezas, Christopher Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. 2020. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics* 39 (07 2020). https://doi.org/10.1145/3386569.3392452

Olga Sorkine-Hornung and Michael Rabinovich. 2016. Least-Squares Rigid Motion Using SVD. Technical note.

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9470-3/22/12.
https://doi.org/10.1145/3550469.3555378