Supplementary Material

Whose Hands are These? Hand Detection and Hand-Body Association in the Wild

Supreeth Narasimhaswamy[1], Thanh Nguyen[2], Mingzhen Huang[3], Minh Hoai[1,2]
[1]Stony Brook University, [2]VinAI Research [3], University of Buffalo

1. BodyHands Dataset

BodyHands is a new dataset collected to develop and evaluate hand-body association methods. It is a large-scale dataset containing unconstrained images with annotations for hand and body locations and correspondences. Table 1 provides some statistics for the proposed BodyHands dataset.

				# people with annotation for		
	#images	#hands	#people	2 hands	1 hand	0 hand
Train	18861	51915	56047	17402	17111	21534
Test	1629	5983	7048	1642	2699	2707
All	20490	57898	63095	19044	19810	24241

Table 1. **Statistics of the BodyHands dataset**

2. Hand Tracking Evaluation Dataset

To evaluate hand tracking methods in unconstrained conditions, we collect 20 videos from YouTube and manually annotate hand bounding boxes and their trajectories. Specifically, we annotate every 15 frames, and altogether the dataset has 3299 annotated frames, 8,893 hand instances, and 131 hand trajectories. We call this dataset YoutubeHands-20, and this dataset has many videos that contain multiple people interacting in the scene, so tracking hands in such cases is challenging. Note that YoutubeHands-20 has now been expanded by Huang et al. [2] to a larger dataset YoutubeHands containing 200 videos. Fig. 1 shows some representative images from this dataset.

3. Implementation Details

We use Detectron2 [4] to implement the proposed architecture. We set the loss weights λ_1 for the Overlap Estimation Module and λ_2 for the Positional Density Module to be 0.1. We train the network using SGD with an initial learning rate 0.0001 for 20 epochs. We reduce the learning rate by a factor of 10 at 10^{th} and 15^{th} epochs. We train our network on NVIDIA RTX 2080 using a batch size of one.

When conducting the ablation studies for the proposed model, we set the probabilities corresponding to the removed components to be 1, and we do not include the option to match the hand to itself when running the Hungarian Algorithm.

4. Heuristic Method: Hand-Body Association for Hand-Contact Estimation

We consider a simple post-processing heuristic to improve the performance of an off-the-shelf contact estimation network [3] as follows. Given a detected hand H and its corresponding person-contact score s obtained by running the pre-trained hand-contact network of [3], our simple heuristic method will adjust s while leaving the scores of other contact states unchanged. First, we use the hand-body association network developed in this paper to detect hands and obtain the associated human bodies for each detected hand; let $\{(A_i, B_i)\}$ denote the set of hand-body pairs obtained. If H does not overlap with any A_j, we will terminate this process and leave the person-contact score s unchanged. Otherwise, we will associate H with A_j that has the highest IoU with H and subsequently associate H to the body B_j. Second, we use a pre-trained MaskRCNN [1] to detect all people in the image; let $\mathcal{P}$ denote this set. We then associate the body B_j with the person $P_k \in \mathcal{P}$ with the highest IoU with B_j. Third, we consider all detected people in $\mathcal{P}$ different from P_k and determine the overlapping region between them and the hand H. If none of the overlapping regions is larger than 15% of the hand area, we heuristically determine that this hand has a low probability of contact with another person. We then decrease the person-contact score using the formula: $s^{new} = \max(s - 0.5, 0)$. This heuristic improves the average precision for detecting other person-contact from 39.51% to 40.89%. This heuristic is simple, but it is only possible because we have a network that tells us who is the self person among the set of detected people. Note that this heuristic only adjusts the person-contact score.

Sample frames from HandTracking Evaluation Dataset

Figure 1. **Hand tracking evaluation dataset.** Sample frames from the videos used for evaluating hand tracking performance. Each row contains frames from the same video.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision*, 2017. 1

[2] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression, and pose association for hand tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[3] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 1

[4] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1