Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics

Evonne Ng
UC Berkeley

Hanbyul Joo
Facebook AI Research

Shiry Ginosar
UC Berkeley

Trevor Darrell
UC Berkeley

1 Results video

The supplementary video shows sequences of various individuals in different conversational settings from both in-the-wild, and Panoptic Studio settings.

The results show that our model, trained on monologue videos, generalizes to people not seen in the training set, and to other conversational domains such as the multi-person setting in the Panoptic Studio. From body input alone, our model realistically depicts when the speaker's fingers are flexed as they emphasize a point, when the fingers are curl up during a pause, and even when the fingers are pointing at a screen.

We demonstrate that our body-prior approach can be used to augment missing hands for body-only mocap data and for existing 3D pose estimation methods that predict the body pose only, by feeding in those existing body pose annotations as input to our model. Furthermore, we demonstrate that our extended network, which uses an additional image observation, produces accurate hand pose estimates of the *seen* hand despite fast gestural movements.

Consistent with the quantitative results provided in the main paper, compared against different baselines, we notice a significant difference between **MTC** hand pose estimates, and hands inferred by **Ours w/** B. While our method is able to capture the motion blurred hands of a lecturer swinging their hands upwards, **MTC** fails to do so. Furthermore, whenever the hands are close together, inducing a great deal of self-occlusion from the fingers, **MTC** fails while **Ours w/** B continues to predict reasonable results from observing the body pose alone. For all frames, our method always produces reasonable, well-formed hand pose estimates, while **MTC** produces much more jittery and often ill-formed hand poses.

In the hand pose estimation scenario, we show a failure case where **Ours w/** B produces an incorrect, albeit reasonable, hand pose sequence. In this situation, there are multiple possible hand pose sequences that may correspond to the given body pose dynamic. The additional appearance-based cue allows our model to snap to the correct hand poses seen within the video. We show **Ours w/ B+I** can capture more distinct semantic hand gestures, accurately capturing a raised index finger as the speaker says, and indicates, "one". Again, **MTC** produces jittery and ill-formed hands in comparison.

2 Person-specific models discussion

While we focus on inferring 3D hands within the domain of conversational gestures across various individuals, we acknowledge that gestures can be individual-specific. Hence, while there exists common gestural patterns that can be learned and applied across individuals, we get further improvements by training a speaker-specific model, given ample data on the single speaker exists.

We train separate models for each individual in the training set. We use the same network architecture as described in the main paper. For each person-specific model, we also only test on videos of that given individual. We compare the person-specific models **Indiv. w/** B and **Indiv. w/** $B + I$ against

Preprint. Under review.

	Indiv. w/ B	Ours w/ B	Indiv. w/ $B+I$	Ours w/ $B+I$	Median	NN
Error (mm)	2.03	2.21	1.79	2.05	3.29	3.05

Table 1: Avg. joint errors (in mm; approximated by 30 cm avg. shoulder length; lower is better) on a hold-out test set from the in-the-wild dataset. While person-specific models outperform models trained across individuals in a broader conversational domain, the improvements are slight.

Video ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Votes for ours (%)	75	73	57	50	52	68	36	48	54	48	52	71	41	75	36	44	77	36	44

Table 2: Perceptual user study on in-the-wild videos demonstrates that our hands synthesized from *body-only input* is competitive against the image-based method, MTC. We show the percentage of votes our method gets on each of the 19 video pairs shown in the evaluation.

our person-agnostic models **Ours w/** B and **Ours w/** $B+I$ (trained and tested across multiple people all at once) in Table 1. The errors are calculated as the euclidean distance between the predictions and the extracted psudo-ground truth MTC annotations. We report errors averaged over all sequences from a hold-out test set with videos of each individual from the in-the-wild dataset. Note that while we evaluate against psudo-ground truth MTC annotations here, all numbers reported in the main paper are evaluated against ground truth from the Panoptic Studio.

Table 1 demonstrates that we achieve slight improvement gains from training and testing in a person-specific setting as opposed to across multiple individuals. However, the margin separating **Ours w/** B and **Ours w/** $B+I$, from their respective person-specific model is much lower than those separating ours from all other baselines. This supports the idea that while gesture dynamics can be person-specific, there exists common conversational body dynamics that generalizes across individuals.

3 In-the-wild dataset composition

Since the psudo-ground truth annotations extracted via MTC can be especially noisy on low-resolution images of hands, we filter out low resolution videos from our in-the-wild dataset during training. We thus train on 4 different speakers, whose videos are consistently of high resolution, for which MTC annotations are relatively reliable. We ensure this subset still consists of a good diversity of speakers from various settings, discussing a wide range of topics. Hence, this subset was sufficient to train our model, which generalizes to novel individuals.

4 Perceptual evaluation details

In the main paper, we provide high-level analysis for the perceptual evaluation. To give more fine-grain analysis of our results, Table 2 shows a breakdown of the total percentage of votes our method receives (over **MTC**) for each of the 19 video pairs included in the evaluation. Our hands synthesized from the body prior alone **Ours w/** B, receives greater than 50% of the votes 11 out of the 19 times. Furthermore, in videos where **MTC** does outperform, 50-36% of the evaluators still vote for ours. This means the preference for **MTC** over ours is not obvious. Hence, our synthesized hands from the body prior only is perceptually competitive against current SOTA image-based hand estimation methods.

5 Architecture details

As described in the main paper, our fully convolutional encoder-decoder network consists of a body encoder, a UNet, and a hand decoder. For the UNet architecture, the input temporal extent is T and we temporally down sample such that $T' = T/2$ as the bottleneck size. We find that any $T' < T/2$ results in hand sequences that are overly-smooth, and any $T' \geq T/2$ led to perceptually indistinguishable results.

Furthermore, we train with an adversarial discriminator, for which we use a series of 1D convolutional layers. We train the discriminator every three epochs. Experiments show that training the discriminator more frequently led to no noticeable improvements, while training less frequently led to overly smooth outputs.