Supplementary Materials for Intelligent Carpet: Inferring 3D Human Pose from Tactile Signals

Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma,
Tomás Palacios, Antonio Torralba, Wojciech Matusik
Massachusetts Institute of Technology
{yiyueluo, liyunzhu, mfoshey, wanshou, pratyuss, tpalacios, torralba, wojciech}@mit.edu

1. Dataset

Our synchronized tactile and visual data were captured with 10 volunteers with 2 rounds of 15 different actions, including walking, squatting, waist turning, arm-waving, twisting, standing on toes, side-walking, lunging, lying, rolling, sit-ups, push-ups, bending, sitting, and stepping, where each action takes about 1 minute (800 frames). No additional instructions were provided, and the volunteers performed the actions in their own ways at random positions and orientations. We used one round of the recorded data for training and split the other for validation and testing. To evaluate how well the model generalizes to unseen individuals and activities, we used the dataset of 7 volunteers or 12 tasks for training and split the held-out data (unseen people or tasks) for testing.

2. Accuracy of 3D pose label

We agree with the reviewers that adding additional cameras can improve the quality of the ground truth human poses. To evaluate the quality of our 3D labels, we manually label 100 pairs of randomly-sampled visual frames and triangulating them into 3D keypoints. Among the 200 frames, the probability of a keypoint missed from our vision system is 2.4 %. For all detected keypoints, our optimization pipeline slightly improves the overall quality of the 3D label, with the mean absolute error (compared to the manually labeled ground truth) drops from 6.5 $\pm$ 3.2 cm to 6.4 $\pm$ 3.2 cm. For the missing keypoints, our optimization pipeline generates a reasonable estimation with a mean absolute error of 7.7 $\pm$ 2.9 cm, which is comparable to the detected keypoints. As demonstrated in the supplementary video, our optimized 3D labels are reasonably accurate and consistent with minor jitterings, which we believe is sufficient for our system.

3. Ablation studies

We performed additional ablation studies on our 3D pose estimation model by quantitatively evaluating the effectiveness of the 3D indexing volume (Layer #8 in Figure 4) and the link length loss. We also performed an ablation study on the repeating tensor along the last dimension (between Layer #7 and #8) by comparing the performance of our model to a model where the tactile feature maps were expanded to 3D by zero-padding along the last dimension. As demonstrated in Table 1, our model obtains the best performance.

	Our model	W/o 3D indexing volume	Zero-padding expansion	W/o link length loss $\mathcal{L}^{\text{link}}$
L2 (cm)	**14.0 $\pm$ 11.4**	19.8 $\pm$ 12.0	66.5 $\pm$ 29.7	15.8 $\pm$ 11.5

Table 1. Ablation studies on different modeling decisions.