DeePoint: Visual Pointing Recognition and Direction Estimation Supplementary Material

Shu Nakamura[*] Yasutomo Kawanishi[†] Shohei Nobuhara[*] Ko Nishino[*,†]
[*]Graduate School of Informatics, Kyoto University [†]RIKEN
snakamura@vision.ist.i.kyoto-u.ac.jp, yasutomo.kawanishi@riken.jp, {nob, kon}@i.kyoto-u.ac.jp

In this supplementary material, we show the results of two ablation studies, each of which demonstrates the necessity of key components of DeePoint.

A. Contribution of Temporal Encoder

Our model is composed of two transformer encoders, namely Joint Encoder and Temporal Encoder, cascaded one after another. We evaluate the contribution of Temporal Encoder by ablating Temporal Encoder (*DP w/o TE*) and comparing it with the full DeePoint (*DP*). In *DP w/o TE*, the output of Joint Encoder is concatenated and fed into an MLP instead of being processed by Temporal Encoder. The hidden layer sizes of the MLP is set to $(2880, 960, 960, 192)$, to make the number of parameters roughly the same as that of Temporal Encoder (The number of parameters of the MLP is 3.8 million, while that of Temporal Encoder is 3.1 million).

Table 1 shows the results of this ablation comparison. Although Temporal Encoder requires less parameters than the MLP, they perform better in terms of both precision/recall and angular error, which demonstrates that Temporal Encoder explicitly and efficiently incorporates temporal coordination.

Model	Angular error (↓)	Prec./Rec. (↑)
DP (Ours)	$\mathbf{14.05}^\circ$	$\mathbf{0.625}/\mathbf{0.838}$
DP w/o TE	14.36°	$0.610/0.796$

Table 1. Comparison between DeePoint and a variant with Temporal Encoder replaced with an MLP (*DP w/o TE*). *DP* performs better in both angular error and precision/recall of pointing recognition than *DP w/o TE*, with a smaller number of parameters. Modeling temporal movements and their coordination is essential for pointing recognition, which is successfully and efficiently achieved with Temporal Encoder in DeePoint.

B. Contribution of Body Parts

In DeePoint, the appearance, movements, and spatial coordination of body parts are encoded by Joint Encoder. It opportunistically uses as many body parts as detected by pose estimation including the hand, head, elbow, and shoulder joints. We evaluate the importance of encoding these body parts by ablating them. This is achieved by allowing Joint Encoder to have access to only limited detected keypoints by hiding other keypoints with masked attention. In *DP-Hand&Head*, the model has access to the keypoints that correspond to the head and the hands, that is, the nose, left/right eyes, left/right ears, and left/right hands. In *DP-Hand*, it can only use the keypoints of the left/right hands.

Model	Angular error (↓)	Prec./Rec. (↑)
DP (Ours)	$\mathbf{14.05}^\circ$	$\mathbf{0.625}/\mathbf{0.838}$
DP-Hand&Head	15.12°	$0.613/0.813$
DP-Hand	17.32°	$0.601/0.797$

Table 2. Ablation of body parts. DeePoint with access to all body parts performs the best in both F-measure and mean angular error. Encoding the appearance, movements, and spatio-temporal coordination of all joints is important for accurate pointing recognition and 3D direction estimation.

Table 2 shows the results of this ablation study of body parts. The results show that the encoded tokens from the hand and head are not enough to recognize pointing and estimate its 3D directions. Having access to the tokens of the head or other body parts and joints contributes to improving the accuracy of estimation. These results clearly show that Joint Encoder successfully leverages these spatial body configurations encoded in the arrangement of body parts and also eloquently show that pointing is a full-body gesture.