Understanding Embodied Reference with Touch-Line Transformer

Yang Li✉, Xiaoxue Chen, Hao Zhao✉, Jiangtao Gong, Guyue Zhou, Federico Rossano, Yixin Zhu✉

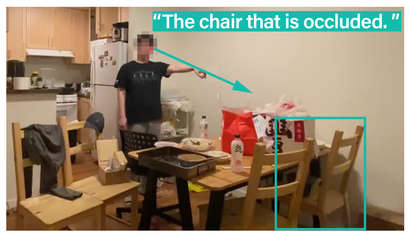
ICLR | 2023

Robots interact better with humans by learning pointing gestures that originated from touch

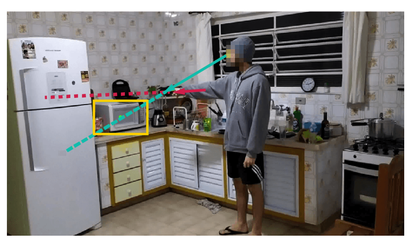
Many robots have difficulty understanding what humans are referring to because most modern learning algorithms do not effectively utilize both gestural information and language information at the same time. However, in order to properly understand human intents, it is critical to consider both gestures and language. We propose to let robots learn pointing gestures that originated from touch to improve their ability in interacting with humans.

Utilize both gestural and language information

To accurately locate the referent in complex scenes, it is necessary to consider both gestural and language information. Take the scenario depicted in the following figure as an example, a man is referring to the chair in the green box. Without considering his pointing gesture, his language cannot uniquely refer to the chair because multiple chairs are present in this context. Conversely, without considering his language, one cannot distinguish the intended referent “the chair” from other nearby objects with only nonverbal expressions.



To effectively model pointing gestures, we propose to use the accurate virtual touch line instead of the elbow-wrist line. As shown in the figure below, people often incorrectly use the red elbow-wrist line to interpret pointing gestures and locate referents (Herbort & Kunde, 2018). However, the object the man is pointing at is the microwave on the green virtual touch line instead of the refrigerator on the red elbow-wrist line. A recent psychology study (O’Madagain et al., 2019) corroborates observations above and suggests that pointing gestures are originated from touch.



Experiment results

Our approach outperforms prior state-of-the-art methods by 16.4%, 23.0%, and 25.0% under the IoU threshold of 0.25, 0.50, and 0.75, respectively.

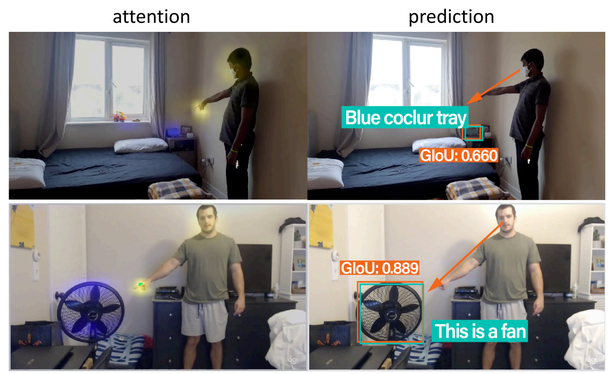
	IoU=.25	IoU=.50	IoU=.75
FAOA (Yang et al., 2019)	44.5	30.4	8.5
ReSC (Yang et al., 2020)	49.2	34.9	10.5
YouRefIT PAF-only (Chen et al., 2021)	52.6	37.6	12.7
YouRefIt Full (Chen et al., 2021)	54.7	40.5	14.0
Ours (Inpainting)	59.1 (+4.4)	51.3 (+10.8)	32.4 (+18.4)
Ours (No explicit gestural key points)	64.9 (+10.2)	57.4 (+16.9)	37.2 (+23.2)
Ours (EWL)	69.5 (+14.8)	60.7 (+20.2)	35.5 (+21.5)
Ours (VTL)	71.1 (+16.4)	63.5 (+23.0)	39.0 (+25.0)
Human	94.2	85.8	53.3

Our model performs better when learning virtual touch lines (VTLs) than when learning elbow-wrist lines (EWLs).

IoU	None	EWL	VTL
0.25	64.9	69.5 (+4.6)	71.1 (+6.2)
0.50	57.4	60.7 (+3.3)	63.5 (+6.1)
0.75	37.2	35.5 (-1.7)	39.0 (+1.8)

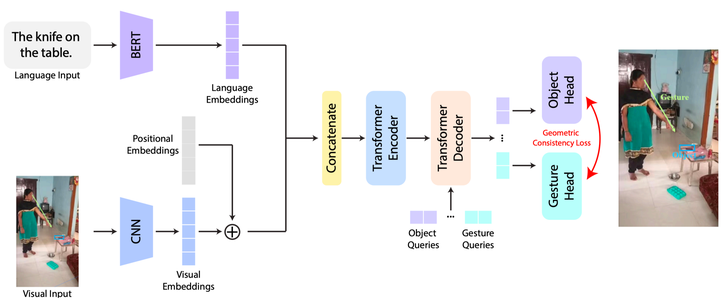
As shown in the figures below, EWLs are unreliable for predicting referent locations because they frequently do not pass through the referents (objects in green boxes). In contrast, the elbow-wrist lines are very good indicators for referents.

Our visualizations of attention weights indicate that our model successfully learns gestural features that boost performance. We use yellow for gesture keypoint queries and blue for matched object queries.



Model Architecture

Our framework consists of a multi-modal encoder, a Transformer decoder, and prediction heads. Language and visual inputs are first encoded by the text encoder and visual encoder to obtain language and visual embeddings, respectively. Next, these embeddings are concatenated and fed into the Transformer encoder to learn multimodal representations. The Transformer decoder and prediction heads output the predicted bounding box and VTL/ EWL. A geometric consistency loss is integrated to encourage the use of gestural signals.



References

Oliver Herbort and Wilfried Kunde. How to point and to interpret pointing gestures? instructions can reduce pointer–observer misunderstandings. Psychological Research, 82(2):395–406, 2018.

Cathal O’Madagain, Gregor Kachel, and Brent Strickland. The origin of pointing: Evidence for the touch hypothesis. Science Advances, 5(7):eaav2558, 2019.