YouRefIt: Embodied Reference Understanding with Language and Gesture

Yixin Chen[1], Qing Li[1], Deqian Kong[1], Yik Lun Kei[1], Song-Chun Zhu[2,3,4], Tao Gao[1], Yixin Zhu[2,3], Siyuan Huang[1]

[1] University of California, Los Angeles [2] Beijing Institute for General Artificial Intelligence [3] Peking University [4] Tsinghua University

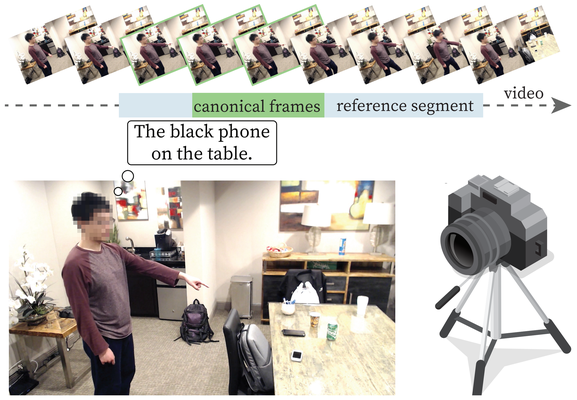
2021 ICCV OCTOBER 11-17 VIRTUAL

Project Page: `https://yixchen.github.io/YouRefIt`

Embodied Reference Understanding (ERU)

We study the machine's understanding of embodied reference: One agent uses both **language and gesture** to refer to an object to another agent in a **shared physical environment**.

Task: Refer to an object in the scene to an imagined person (camera)
Steps:
1. Refer to one object using both pointing gesture and language.
2. After the reference, tap the target object to confirm.
3. Repeat until no more objects.
4. Write down the sentences in the same order as during the recording.
5. Submit both the videos and sentences.



Key Characteristics

- **Multimodal:** People often use both natural language and gestures when referring to an object.
- **Perspective-taking:** Embodied reference requires the awareness that others see things from different viewpoints and the ability to imagine what others see from their perspectives.

Contribution & Discovery

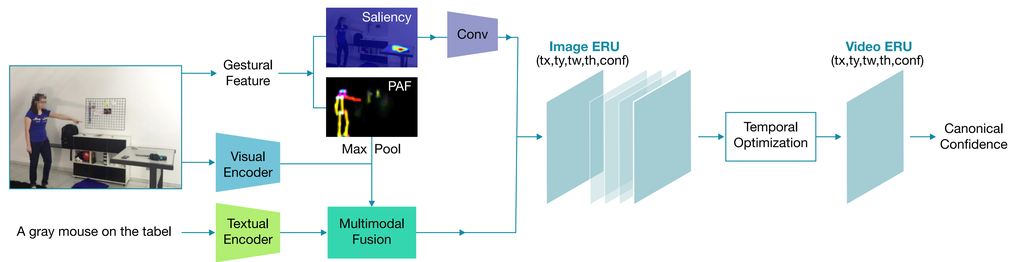
- We crowd-source the first video dataset in physical scenes, **YouRefIt**, to study the reference understanding in an embodied setting.
- We devise two benchmarks, **Image ERU** and **Video ERU**, as the protocols to study and evaluate the embodied reference understanding.
- We propose a **multimodal framework** for ERU tasks with multiple baselines and model variants. The experimental results confirm the significance of the joint understanding of language and gestures in embodied reference.

YouRefIt Dataset

We introduce a new dataset named **YouRefIt**, a video collection of people referring to objects with both natural language and gesture in indoor scenes.

- YouRefIt contains videos crowd-sourced by Amazon Mechanic Turk(AMT), and thus the reference happens in a more natural setting with richer diversity.
- The referrers (human) and the receivers (camera) in YouRefIt share the same physical environment, with both language and gesture allowed for referring to objects.
- YouRefIt includes 432 recorded videos and 4,195 localized reference clips with 395 object categories.
- Each reference process was annotated with segments, canonical frames, bounding boxes of the referred objects, and sentences with semantic parsing.
- Canonical frames are the "keyframes" that the referrer holds the steady pose to indicate what is being referred clearly. Combined with natural language, it is sufficient to use any canonical frame to localize the referred target.

Framework



We devise a novel multimodal framework for ERU task that leverages both the language and gestural cues.

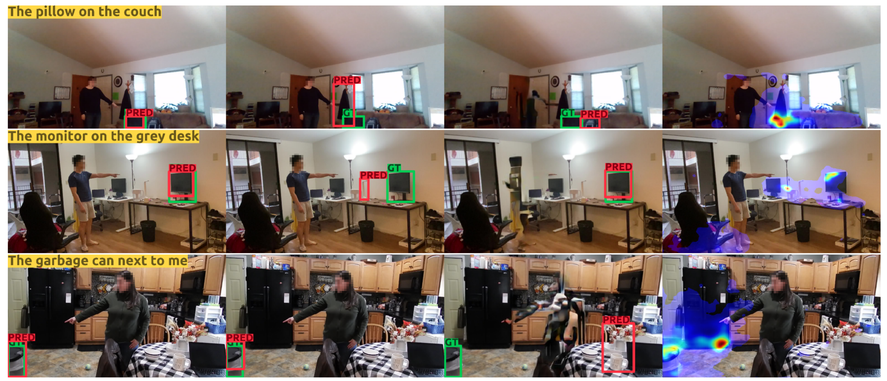
Image ERU

Given the canonical frame and the sentence from an embodied reference instance, Image ERU aims at locating the referred object in the image through both the human language and gestural cues.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Language-only												
MAttNet$_{\text{pretrain}}$	14.2	2.3	4.1	34.7	12.2	2.4	3.8	29.2	9.1	1.0	2.2	23.1
FAOA$_{\text{pretrain}}$	15.9	2.1	9.5	34.4	11.7	1.0	5.4	27.3	5.1	0.0	0.0	14.1
FAOA$_{\text{inpaint}}$	23.4	14.2	23.6	32.1	16.4	9.0	17.9	22.5	4.1	1.4	4.7	6.2
ReSC$_{\text{pretrain}}$	20.8	3.5	17.5	40.0	16.3	0.5	14.8	36.7	7.6	0.0	4.3	17.5
ReSC$_{\text{inpaint}}$	34.3	20.3	38.9	44.0	25.7	8.1	32.4	36.5	9.1	1.1	10.1	16.0
Gesture-only												
RPN+Pointing$_{15}$	15.3	10.5	16.9	18.3	10.2	7.2	12.4	11.0	6.5	3.8	9.1	6.6
RPN+Pointing$_{30}$	14.7	10.8	17.0	16.4	9.8	7.4	12.4	9.8	6.5	3.8	8.9	6.8
RPN+Saliency[27]	27.9	29.4	34.7	20.3	20.1	**21.1**	26.8	13.2	12.2	**10.3**	**17.9**	8.6
Ours$_{\text{no_lang}}$	41.4	29.9	48.3	46.3	30.6	17.4	37.0	37.4	10.8	1.7	13.9	16.6
Language + Gesture												
FAOA[59]	44.5	30.6	48.6	54.1	30.4	15.8	36.2	39.3	8.5	1.4	9.6	14.4
ReSC[58]	49.2	32.3	54.7	60.1	34.9	14.1	42.5	47.7	10.5	0.2	10.6	20.1
Ours$_{\text{PAF_only}}$	52.6	35.9	60.5	61.4	37.6	14.6	49.1	49.1	12.7	1.0	16.5	20.5
Ours$_{\text{Full}}$	**54.7**	**38.5**	**64.1**	**61.6**	**40.5**	16.3	**54.4**	**51.1**	**14.0**	1.2	17.2	**23.3**
Human	94.2±0.2	93.7±0.0	92.3±1.3	96.3±1.7	85.8±1.4	81.0±2.2	86.7±1.9	89.4±1.7	53.3±4.9	33.9±7.1	55.9±6.4	68.1±3.0

Image ERU

Our results reveal that gestural cues are as critical as language cues in resolving ambiguities and overloaded semantics.



(a) Ours$_{\text{Full}}$ (b) Ours$_{\text{no_lang}}$ (c) ReSC$_{\text{inpaint}}$ (d) Saliency Map

Video ERU

Given a referring expression and a video clip that captures the whole dynamics of a reference action with consecutive body movement, Video ERU aims at recognizing the canonical frames and estimate the referred target at the same time.

Model	IoU=0.25				IoU=0.5				IoU=0.75			
	all	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*	*all*	*small*	*medium*	*large*
Frame-based	**55.2**	42.3	**58.9**	**64.8**	**41.7**	**22.7**	53.4	**48.8**	16.9	1.6	21.8	**27.0**
Transformer	52.3	40.2	55.6	58.3	38.8	21.2	54.1	47.1	13.9	1.5	20.8	22.7
ConvLSTM	54.8	**43.1**	57.5	60.0	39.3	22.5	**54.8**	46.7	**17.3**	**1.8**	**24.3**	25.5
Ours$_{\text{Full}}$	54.7	38.5	64.1	61.6	40.5	16.3	54.4	51.1	14.0	1.2	17.2	23.3

Result indicates that the canonical frames can provide sufficient language and gestural cues for clear reference, however temporal information can improve the performance of canonical frame detection.