東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

JST CREST Project
COLLECTIVE VISUAL SENSING

Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Videos

Yifei Huang, Minjie Cai, Hiroshi Kera, Ryo Yonetani, Keita Higuchi and Yoichi Sato
The University of Tokyo

ICCV 17
International Conference on Computer Vision 2017

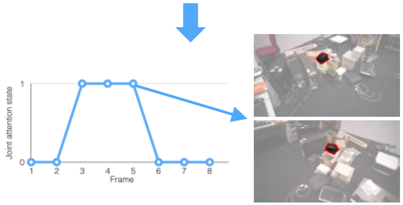
Goal

Discovering objects of joint attention using multiple first-person videos (FPVs) with points of gaze (PoG) data

Task

- **Temporally localize** time intervals of joint attention
- **Spatially segment** the object of joint attention

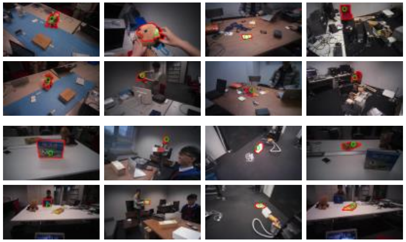
Input: multiple FPVs with PoG data



Output: Joint attention states and object segmentation

Dataset

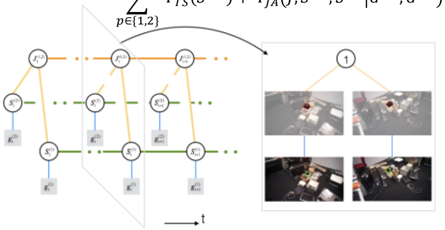
- 24 pairs of egocentric videos with gaze data (20 ~ 60 secs)
- 5 different environments, 20+ different objects
- Annotation of joint attention period & object segments



Problem Formulation

Given gaze position ***G***, we aim to infer joint attention state ***J*** and segment the object of joint attention (***S***), by minimizing the objective function:

$$\Psi(S^{(1)},S^{(2)}|G^{(1)},G^{(2)}) = \sum_{p\in\{1,2\}} \Psi_{GO}(S^{(p)}|G^{(p)}) + \sum_{p\in\{1,2\}} \Psi_{TS}(S^{(p)}) + \Psi_{JA}(J,S^{(1)},S^{(2)}|G^{(1)},G^{(2)}) + \Psi_{TJ}(J)$$



Gaze proximity and objectness

$$\Psi_{GO}(S^{(p)}|G^{(p)}) = \sum_{t=1}^{T}\left(\lambda_{GO1}\frac{\left\|C(s_t^{(p)})-\boldsymbol{g}_t^{(p)}\right\|_2}{\left|s_t^{(p)}\right|^{\frac{1}{2}}} + \lambda_{GO2}\left(1-\frac{\left|s_t^{(p)}\right|}{\left|H(s_t^{(p)})\right|}\right)\right),$$

$C\left(s_t^{(p)}\right)$:Centroid of segment $s_t^{(p)}$, $\left|H\left(s_t^{(p)}\right)\right|$: Area of convex hull of $s_t^{(p)}$

Temporal consistency of segments

$$\Psi_{TS}(S^{(p)}) = \lambda_{TS}\sum_{t=1}^{T-1}(1-f_{sim}(s_t^{(p)},s_{t+1}^{(p)}))$$

f_{sim} : cosine similarity of features extracted from segments

Joint attentionness

$$\Psi_{JA}(J,S^{(1)},S^{(2)}|G^{(1)},G^{(2)}) = \sum_{t=1}^{T}(\lambda_{JA1}Y\left(j_t,s_t^{(1)},s_t^{(2)},\boldsymbol{g}_t^{(1)},\boldsymbol{g}_t^{(2)}\right) + \lambda_{JA2}Z(j_t))$$

Y measures visual similarity of segments:

$$Y\left(j_t,s_t^{(1)},s_t^{(2)},\boldsymbol{g}_t^{(1)},\boldsymbol{g}_t^{(2)}\right) = j_t\left(1-f_{sim}\left(s_t^{(1)},s_t^{(2)}\right)\right) + (1-j_t)\alpha\left(\boldsymbol{g}_t^{(1)},\boldsymbol{g}_t^{(2)}\right)$$

α computes visual similarities around gaze region like [1]

$$Z(j_t) = \begin{cases} j_t\ , magnitude\ of\ global\ motion > \delta_m \\ 0, otherwise \end{cases}$$

Temporal consistency of joint attention

$$\Psi_{TJ}(J) = \lambda_{TJ}\sum_{t=1}^{T-1}|j_t - j_{t+1}|$$

[1] Kera et. al. CVPRW2016

Experiment

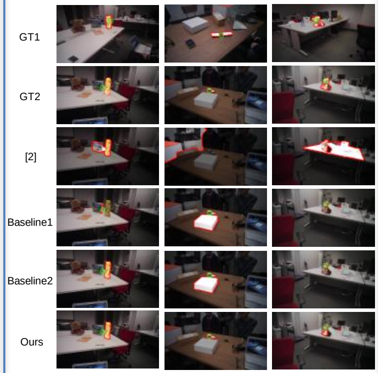
Spatial segmentation task

Method	FtF-large	FtF-small	SbS-large	SbS-small	Avg.
ObMiC [2]	0.287	0.212	0.065	0.336	0.225
Baseline1	0.552	0.599	0.681	0.691	0.631
Baseline2	0.611	0.629	0.723	0.726	0.672
Ours	**0.633**	**0.660**	**0.730**	**0.735**	**0.690**

Temporal localization task

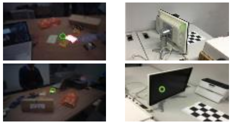
Method	FtF-large (%)		FtF-small (%)		SbS-large (%)		SbS-small (%)		Avg. (%)
	P	R	P	R	P	R	P	R	F1 score
Kera *et al.* [1]	74.5	89.7	69.7	**93.8**	72.9	**96.5**	67.1	83.4	79.0
Ours	**91.9**	**92.8**	**84.7**	86.5	**94.3**	92.6	**79.7**	**98.7**	**89.3**

Visualizations



- GT1,2: ground truth of person 1,2
- Baseline1: Ψ_{GO} only, Baseline2: $\Psi_{GO} + \Psi_{TS}$
- [2]: Fu et. al. CVPR2014

Failure cases



- Different objects with similar appearance
- Same object with different appearances

Future work

- Use predicted gaze instead of eye tracker
- Use 3D geometric relation between FPVs