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

$C(S^{(p)}(t))$: Centroid of segment $S^{(p)}(t)$, $M(S^{(p)})$: Area of convex hull of $S^{(p)}(t)$

Temporal consistency of segments
$$\Psi_{TS}(S^{(p)}) = \lambda_{TS} \sum_{t=1}^{T-1} \left( 1 - f_{sim}(S^{(p)}(t), S^{(p)}(t+1)) \right)$$

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

Joint attentionness
$$\Psi_{JA}(J, S^{(1)}, S^{(2)}|G^{(1)}, G^{(2)}) = \sum_{t=1}^{T} \left( \lambda_{JA1} Y(J_t, S^{(1)}(t), S^{(2)}(t), G^{(1)}(t), G^{(2)}(t)) + \lambda_{JA2} Z(J_t) \right)$$

$Y$ measures visual similarity of segments:
$$Y(J_t, S^{(1)}(t), S^{(2)}(t), G^{(1)}(t), G^{(2)}(t)) = J_t \left( 1 - f_{sim}(S^{(1)}(t), S^{(2)}(t)) \right) + (1 - J_t) \alpha \left( G^{(1)}(t), G^{(2)}(t) \right)$$

$\alpha$ computes visual similarities around gaze region like [1]
$$Z(J_t) = \begin{cases} J_t, & \text{magnitude of global motion} > \delta_m \\ 0, & \text{otherwise} \end{cases}$$

Temporal consistency of joint attention
$$\Psi_{TJ}(J) = \lambda_{TJ} \sum_{t=1}^{T-1} |J_t - J_{t+1}|$$


Experiment
Spatial segmentation task

<table>
<thead>
<tr>
<th>Method</th>
<th>P=large</th>
<th>P=small</th>
<th>P=medium</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.633</td>
<td>0.660</td>
<td>0.730</td>
<td>0.735</td>
</tr>
<tr>
<td>Kera et. al. [1]</td>
<td>0.755</td>
<td>0.857</td>
<td>0.977</td>
<td>0.814</td>
</tr>
<tr>
<td>Dus</td>
<td>0.199</td>
<td>0.268</td>
<td>0.464</td>
<td>0.363</td>
</tr>
</tbody>
</table>

Temporal localization task

<table>
<thead>
<tr>
<th>Method</th>
<th>P=large</th>
<th>P=small</th>
<th>P=medium</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.633</td>
<td>0.660</td>
<td>0.730</td>
<td>0.735</td>
</tr>
<tr>
<td>Kera et. al. [1]</td>
<td>0.755</td>
<td>0.857</td>
<td>0.977</td>
<td>0.814</td>
</tr>
<tr>
<td>Dus</td>
<td>0.199</td>
<td>0.268</td>
<td>0.464</td>
<td>0.363</td>
</tr>
</tbody>
</table>

Visualizations
- GT1, GT2
- Baseline1, Baseline2
- Ours

GT1: ground truth of person 1,2
Baseline1: $\Psi_{GO}$ 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