Deep Future Gaze: Gaze Anticipation On Egocentric Videos Using Adversarial Networks

Mengmi Zhang (NUS, A*STAR), KT Ma (A*STAR), Joc-Hwee Lim (A*STAR), Qi Zhao (UMN), Jiashi Feng (NUS)

**Novel problem**

- To anticipate where people will look in the next few seconds from a single ego-centric video frame
- Potential novel applications
  - Reduce system's response time by pre-processing:
    - Pre-render virtual environment and objects
    - Pre-fetch online information. E.g. AR tour guide, contextual advertisements
  - Reduce user's reaction time with proactive feedback. E.g. fall prevention, driving, object search, navigation, task assistance.

**Challenges**

- Non-linear interactions of features:
  - Task: search, object manipulation etc.
  - Foreground objects: hands, objects of interests, etc.
  - Background motion: head motion
- Gaze in egocentric videos:
  - Eye-tracking errors
  - Gaze, hands, objects of interests outside Field Of View
  - Humans: distractions, hesitations, individual quirks

**Egocentric eye-tracking dataset**

- Object Search Task (OST) dataset
  - Home Environment: 2 bedrooms, living room, washroom, kitchen, study room, balcony
  - 57 videos, 15 mins each
  - Navigation, putting/taking objects, reading, writing, searching

**Proposed architecture**

- Generate future frames, then perform saliency prediction
  - Future Frame Generation Module
  - Temporal Saliency Prediction Module (GP)

**Design factors**

- Future semantic and motion information help gaze anticipation
- Additional feedback to generator leads to more realistic future semantic and motion estimation
- Dual-streams 3D ConvNet to separate foreground and background contributes to better frame generation
- Gaze anticipation depends on both temporal and spatial information

**Example**

- Ground Truth
- Anticipated Gaze
- Future Frames

**Quantitative comparisons**

- AUC
- AAE

**Future works**

- Head/body movements anticipation
- Improve generalizability
- Personalization
- Novel applications

https://github.com/Mengmi/deepfuturegaze_gan

Additional feedback to generator leads to more realistic future semantic and motion estimation

Dual-streams 3D ConvNet to separate foreground and background contributes to better frame generation

Gaze anticipation depends on both temporal and spatial information

<table>
<thead>
<tr>
<th>AAE</th>
<th>DFG(Ours)</th>
<th>SALICON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time t+16</td>
<td>11.7</td>
<td>18.4</td>
</tr>
<tr>
<td>Time t+32</td>
<td>11.8</td>
<td>16.6</td>
</tr>
</tbody>
</table>