Learning video saliency from human gaze using candidate selection

Rudoy, Goldman, Schechtman, Manor

Akanksha Saran
CS381V: Experiment Presentation
Outline

● Description of Gaze Datasets
  - DIEM
  - CRCNS

● Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

● Conclusions
Outline

- **Description of Gaze Datasets**
  - DIEM (Dynamic Images and Eye Movements)
  - CRCNS (Collaborative Research in Computational Neuroscience)

- Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

- Conclusions
DIEM Dataset

- 84 videos captured at 30 fps
- ~50 participants/video
- More than 4500 eye movement traces
- Some videos used with audio data
- Videos on TV news, sports, commercials, movie trailers, wildlife etc.
- Provide gaze information for left and right eye separately for each participant
- X,Y coordinates on the screen, saccade/fixation/blink, pupil dilation
- Eye tracker rate is 1000 Hz
DIEM Dataset Illustration

https://www.youtube.com/watch?v=Q3FgO2_ZuP0

https://www.youtube.com/watch?v=D5K09NPn75c
CRCNS Dataset

- 50 video clips (Itti, 2004; 2005).
- 8 subjects total; 4-6 subjects on each video clip.
- 235 eye movement traces.
- Videos on TV news, sports, commercials, talk shows, Video games (short video snippets combined together)
- (X,Y) at each time point plus additional information when saccades start
- Eye tracker rate is 240 Hz.
- Task: “follow main actors and actions, try to understand overall what happens in each clip. We will ask you a question about main contents. Do not worry about details like specific text messages.”
CRCNS Dataset Illustration

https://www.youtube.com/watch?v=_d1nvM6AI9A  https://www.youtube.com/watch?v=sdq5TV_nKlg
## Properties of the two datasets

<table>
<thead>
<tr>
<th></th>
<th>DIEM</th>
<th>CRCNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single event</td>
<td>Single event videos</td>
<td>Multiple video snippets combined</td>
</tr>
<tr>
<td>videos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4500 gaze</td>
<td>4500 gaze patterns</td>
<td>235 gaze patterns</td>
</tr>
<tr>
<td>patterns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>~50 subjects</td>
<td>~50 subjects per video</td>
<td>~4 subjects per video</td>
</tr>
<tr>
<td>per video</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video frames</td>
<td>Video frames vary in size (1280 x 960)</td>
<td>Fixed size video frame (640 x 480)</td>
</tr>
<tr>
<td>vary in size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1280 x 960)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Quality</td>
<td>High Quality</td>
<td>Low quality</td>
</tr>
<tr>
<td>1000 Hz eye</td>
<td>1000 Hz eye tracker</td>
<td>240 Hz eye tracker</td>
</tr>
<tr>
<td>tracker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some videos</td>
<td>Some videos shown with audio</td>
<td>No audio</td>
</tr>
<tr>
<td>shown with</td>
<td></td>
<td></td>
</tr>
<tr>
<td>audio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Description of Gaze Datasets
  - DIEM
  - CRCNS

- Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

- Conclusions
Outline

- Description of Gaze Datasets
  - DIEM
  - CRCNS

- Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

- Conclusions
Variation in human agreement on fixations (DIEM)

- Per-frame variation in gaze fixations across participants is well bounded for all videos
- Variations for the left and right eye are closely related (as expected)
Variation in human agreement on fixations (DIEM)

- Per-frame variation in gaze fixations across participants is well bounded for all videos
- Variations for the left and right eye are closely related (as expected)
Low variation in human gaze agreement

- close up shots, activity towards center, text

https://www.youtube.com/watch?v=E8PzL6-U1yI
https://www.youtube.com/watch?v=vlEFCc_9y74
High variation in human gaze agreement

- no sound available, not clear what is going on, gives time to examine the room

https://www.youtube.com/watch?v=hzYrz-ixuwc  https://www.youtube.com/watch?v=2j7Gq9tDZ80
Variation in human agreement on fixations (CRCNS)

- Per-frame variation in gaze fixations across participants is well bounded or all videos
- Variations in data is less than DIEM dataset
Variation in human agreement on fixations (CRCNS)

- Per-frame variation in gaze fixations across participants is bound in a small band for all videos
- Variations in data is less than DIEM dataset
Low variation in human fixations (CRCNS)

- Text which limits the variance, motion cues seem to guide subjects

https://www.youtube.com/watch?v=wRKD5InFqs0  https://www.youtube.com/watch?v=mRTKOdQO_Kw
High variation in human fixations (CRCNS)

- less motion allows subjects to focus on different aspects of the scene

https://www.youtube.com/watch?v=5ulk-tJ5YwQ  https://www.youtube.com/watch?v=vnVRrbeElBU
DIEM v/s CRCNS

- Avg standard deviation across participants and across videos
- Normalized with respect to width and height of corresponding frame
- DIEM a more diverse dataset

<table>
<thead>
<tr>
<th></th>
<th>DIEM (left eye)</th>
<th>DIEM (right eye)</th>
<th>CRCNS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1748</td>
<td>0.1863</td>
<td>0.1294</td>
</tr>
</tbody>
</table>
Outline

● Description of Gaze Datasets
  - DIEM
  - CRCNS

● Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - **Gaze Patterns over time**
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

● Conclusions
Gaze patterns over time (DIEM)

- Gaze pattern for a subject with *moderate variation* in fixations over time
- Fixations localize in certain regions over the entire frame
Gaze patterns over time

- Gaze pattern for a subject with **largest variation** in fixations over time
- Fixations localize in certain regions over the entire frame

Frames: 720 x 576
Gaze patterns over time

- Gaze pattern for a subject with **smallest variation** in fixations over time
- Fixations localize in certain regions over the entire frame
- Candidate regions form a valid hypothesis to model video saliency
Outline

- Description of Gaze Datasets
  - DIEM
  - CRCNS

- Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with plausible Candidate Regions
  - Correlation between pupil dilation and fixations

- Conclusions
Gaze fixation overlap with plausible Regions (Hit Rate for DIEM dataset)

- Overlap with **per-frame face detections** (every 10 frames)
- Overlap with high magnitude **optical flow** regions (every 15 frames)
- Overlap with **per-frame static saliency** (every 10 frames)

<table>
<thead>
<tr>
<th></th>
<th>Faces</th>
<th>Optical Flow</th>
<th>Static saliency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hit Rate</strong></td>
<td>30.62 %</td>
<td><strong>49.25 %</strong></td>
<td>37.02 %</td>
</tr>
</tbody>
</table>
Gaze Hits with Faces

- Not detecting the other face helps reasoning about most of the ground truth fixations
Gaze Misses with Faces

- Motion cue dominates
Gaze Misses with Faces

- Text reading over a few frames
Gaze Misses with Faces

- Frontal face detector does not detect the side view
Gaze Hits with Optical Flow

- Includes a large region with insignificant motion
- High recall
Gaze Hits with Optical Flow

- Brightness constancy constraint violated
- Entire object falsely detected as having motion
- High recall
Gaze Hits with Optical Flow

- Likely frames from a scene-cut detector
- Almost every pixel in the frame has significant motion
Gaze Misses with Optical Flow

- Center of the frame accounts for most ground truth fixations
Gaze Misses with Optical Flow

- Insignificant motion

Flow thresholded image
Gaze Hits with Static Saliency

- Static saliency can extract out text in the center of the image
- The subject could be in the process of reading the text
Gaze Hits with Static Saliency

- Redundant information from face detector and static saliency
Gaze Hits with Static Saliency

- Almost all ground truth fixations accounted for
Gaze Misses with Static Saliency

- None of face detector, optical flow or static saliency accounts for the ground truth fixations here
Gaze Misses with Static Saliency

- Motion cue dominates
Gaze fixation overlap with plausible Regions

- Optical flow can reason for about 50% of the ground truth gaze data
- Frontal face detector fails to detect faces in all scenarios
- Static saliency (GBVS) can reason about text in center of image frames
- Multiple cues can reason about the same ground truth gaze point
- Static cues not sufficient to model all gaze fixations,
- Scope for modeling transitions dynamically between frames
- Scope for other cues to be used
Outline

- Description of Gaze Datasets
  - DIEM
  - CRCNS

- Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

- Conclusions
Correlation between pupil dilation and event tags

- Each frame is labeled with an event tag by the eye tracking device
- Types of event tags - Fixation, Saccade, Blink
- Right eye (0.47), Left eye (0.35)
Correlation between pupil dilation and fixations

- Each frame is labeled with an event tag by the eye tracking device
- Only frames with the ‘fixation’ event tag considered
- Right Eye (0.48), Left Eye (0.31)
Correlation between pupil dilation and fixations

- Pupil dilation has a weakly positive correlation with gaze fixation
- In general, higher dilation in pupil indicates “interest”
- Pupil dilation not used to model video saliency in the paper by Rudoy et al.
- The right eye has a consistently higher correlation with gaze fixation versus the left eye (measurement bias of the tracker?)
- Not very reliable
Outline

● Description of Gaze Datasets
  - DIEM
  - CRCNS

● Analysis of Human Gaze Datasets for Videos
  - Variation in human agreement on fixations
  - Gaze Patterns over time
  - Ground Truth overlap with Candidate Regions
  - Correlation between pupil dilation and fixations

● Conclusions
Conclusions

- Gaze fixation across participants have tight bounds of variations
  - Candidate regions form a valid hypothesis to model video saliency
- Fixations localize in certain regions over the entire frame
- Static cues not sufficient to model all gaze fixations
  - Scope for modeling transitions dynamically between frames
- Pupil dilation and human gaze fixations are weakly positively correlated
- Written text forms another crucial candidate region
Thank you!