Decoding Attention from Gaze: A Benchmark Dataset and End-to-End Models

Karan Uppal [1], Jaeah Kim [2], Shashank Singh [3]

[1] Indian Institute of Technology Kharagpur, [2] Carnegie Mellon University, [3] Max Planck Institute for Intelligent Systems

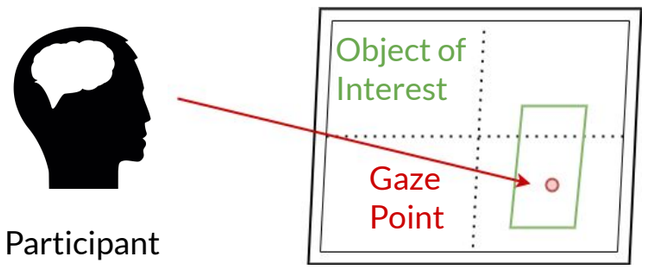
Innovation and Global Cooperation for Sustainability
12-14 March, 2024

1. Problem Motivation

Eye-tracking has the potential to provide rich behavioural data about human cognition, having applications in a variety of fields. However, analysing these rich data can be challenging. This paper studies using computer vision tools for "attention decoding" with the aim to automate attention decoding and provide a publicly available dataset, Multiple Object Eye-Tracking (MOET) dataset.

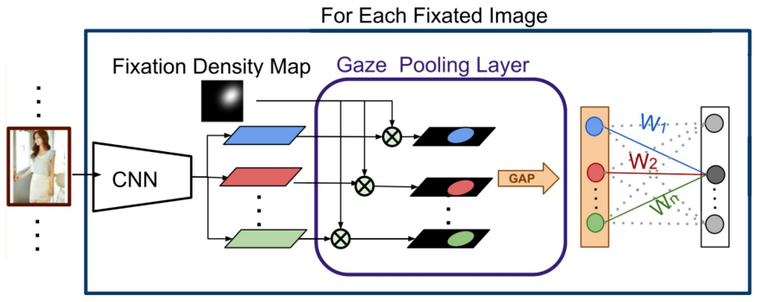
Attention Decoding

decoding a participant's latent attentional state from their measured gaze behavior



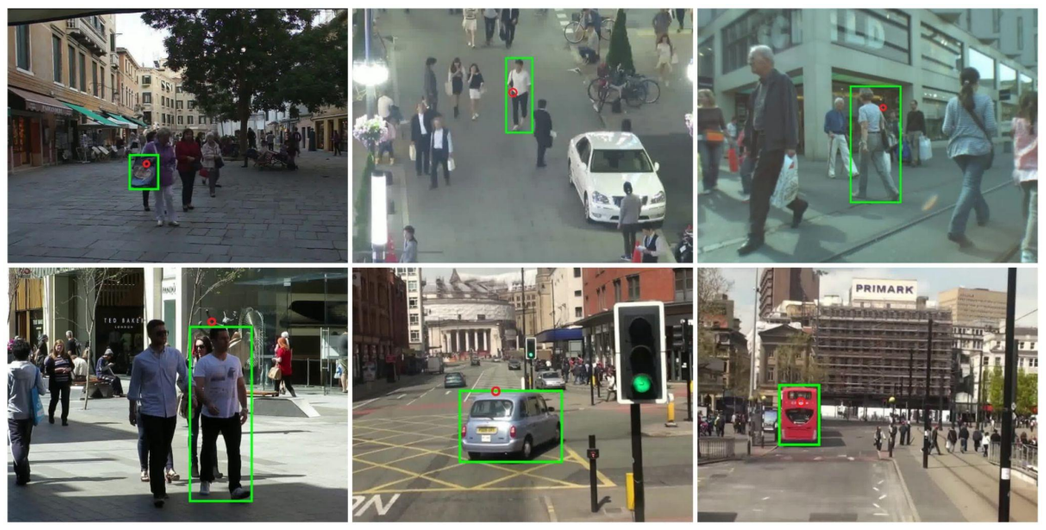
2. Gaze Density Map

Taking inspiration from Deep Gaze Pooling[2], we modeled the gaze as a Gaussian, centred at the coordinates of the gaze. The whole gaze sequence is aggregated into a gaze density map which can be interpreted as a temporal localized feature importance map.



3. MOET Dataset

This is an extension of the Multiple Object Tracking 2016 (MOT16)[1] benchmark dataset with eye-tracking data obtained from 16 participants viewing the videos while tracking distinct target objects, which are annotated with labels and bounding boxes.

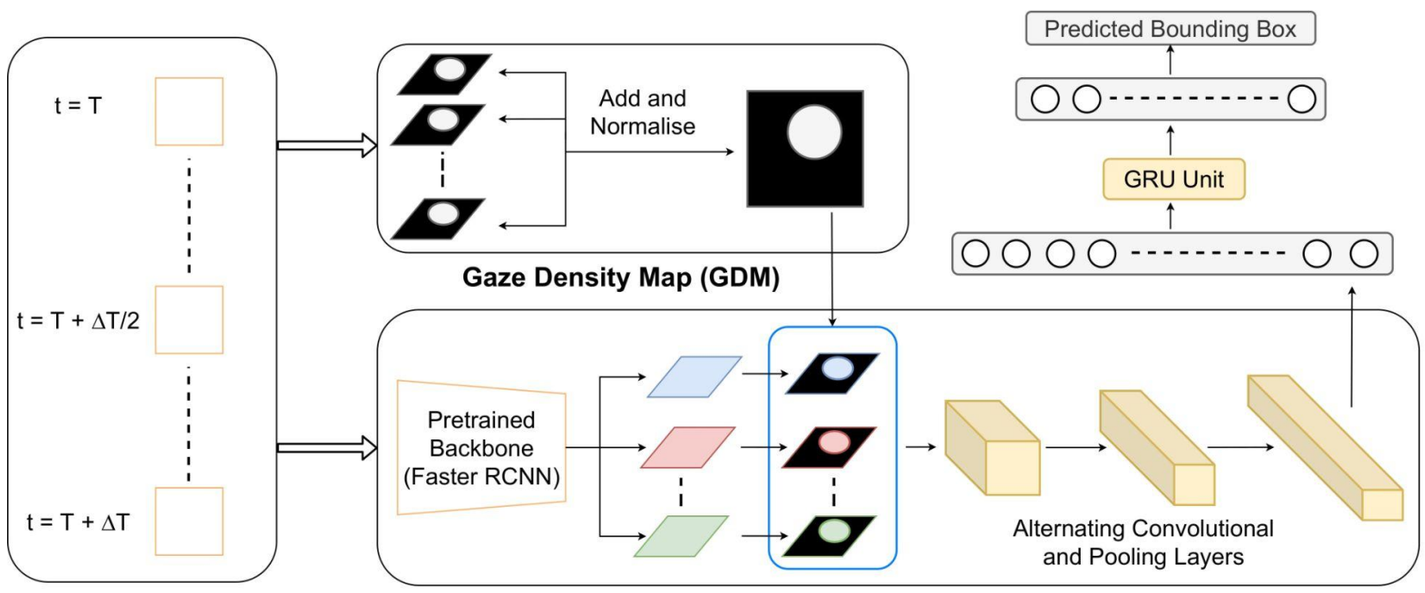


Example frames from stimulus videos in the experiment. The target object is indicated by a green bounding box and the gaze point is indicated by a red circle

Containing 179536 annotated frames from a total of 14 videos featuring a variety of scenarios, the dataset is preprocessed (imputation of missing data, omitting transition frames, omitting missing gaze data frames, and verifying overt condition) to ensure adequate data quality. The most important step is verifying the over condition since our basis assumption is that the participants gaze directed at the object to which they are attending. This results in a final dataset of 105532 frames.

4. End-to-end Attention Decoding Architectures

The cumulative gaze data is used to construct the gaze density map. Each frame is passed through the architecture shown here. In order to incorporate data across multiple consecutive frames, we use a GRU unit, followed by a fully connected layer to predict the bounding box of the object of interest. We refer to this model as the GRU model. In another variant of the same, we replace the GRU unit with a fully connected layer. We refer to this model as the Feedforward Model.



5. Results and Conclusion

The end-to-end models we considered were generally outperformed by two-stage methods combining off-the-shelf object detectors with heuristic rules to select a bounding box from those output by the object detector. At the same time, heuristic methods performed far worse than the optimal "oracle" predictor, suggesting that there may be substantial room for improvement in attention decoding methods.

Approach	Algorithm	Mean IoU (Std. Dev.)
Completely Rule-Based	Fixed Box Baseline	0.244 (0.060)
Object Detector (OD) + Rule	OD Baseline	0.495 (0.112)
	OD Mod	0.546 (0.089)
	OD Oracle	0.717 (0.077)
End-to-End (Ours)	Feedforward Model	0.443 (0.116)
	GRU Model	0.406 (0.153)

References:

[1] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.
[2] Sattar, H., et. al. (2020). Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. Neurocomputing, 387, 369-382.