Gaze-informed object sequences for egocentric action recognition using deep learning

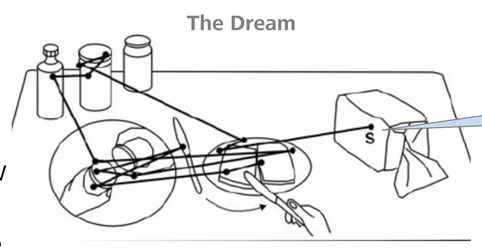
Nora J. Castner[1], Zhengyu Su[2], Siegfried Wahl[1,3]

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Motivation

Understanding *semantics* is a persistent challenge:
Annotating AOIs, for head mounted ET data, is resource intensive.
– AI also has cost to train (time/know how)
→ Focus on pre-trained stuff available.

Open-vocabulary object detectors such as YOLO-World are flexible without retraining [1,2].
→ Like to manual labeling: *You give the model a list of task-fitting objects.*



Scanpath Analysis:
Traditional approach of the scanpath as a sequence of things looked at.

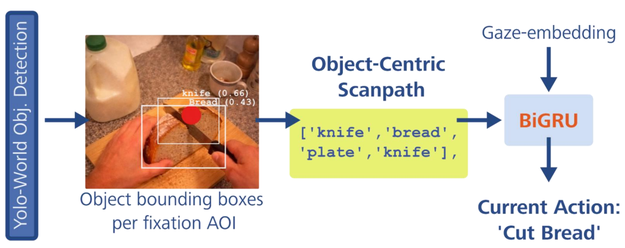
Bread, Plate, Plate, Jam, Jam, Jam, Mustard, ... Sliced Bread, …

Scanpaths + Pre-trained models (SOTA):
– Recurrent architectures (LSTMs, GRUs, BiGRUs) model sequential gaze dynamics & cognitive patterns.
– Recent work integrates scanpaths with PLMs & LLMs [3,5,6,7,8], motivating our use of sequential gaze as input.
– Eye movement-guided segmentation for SAM improves object detection [8].

A *Lazy* Pipeline:
Generates object-level scanpaths by aligning gaze fixations with open-vocabulary object detection
✓ Use freely available models (+)
✓ Fully-automated (+,-)
✓ Evaluated using common scanpath analysis techniques (+)
✓ Ok performance (-)
✓ Source code available (+)

We show that a fully automatic, gaze-informed, open-vocabulary pipeline is technically feasible & carries semantic information, especially for frequent actions.

Pipeline & Preliminary Results



Dataset [4]:
28hrs cooking videos · 86 sessions
32 subjects · SMI Glasses @ 30 Hz
106 action classes (verb + $noun_1$ + $noun_2$)
*Long-tail distribution**

Object Scanpath Generation

a: Define AOI per fixation:
`80 x 80 px` AOI centered on each fixation point.

b: YOLO-World Detection:
Zero-shot open vocab detection. Keep detections whose bounding boxes intersect the AOI (IoU ≥ 0.15).

c: Token Generation:
Winner-take-all: most frequent detected object per fixation → `token`.
No reliable detection → `'x'`.

Scanpath Classification

a: Normalize Scanpath:
Pad (`[PAD]`) or trim scanpath to fixed `max_length` (90th percentile of fixation counts).

b: BiGRU Classification:
- Each token is mapped to a learned embedding vector.
- BiGRU encodes sequence of embeddings.
- Final hidden states passed through a linear layer + softmax to predict action classes.
- **SPECs:** AdamW (learning rate 1×10^{-5} , weight decay 0.01, batch size 4, 40 epochs).

Used predefined training / test split from the original dataset.

Classification performance:
Subsets of the most freq. action classes

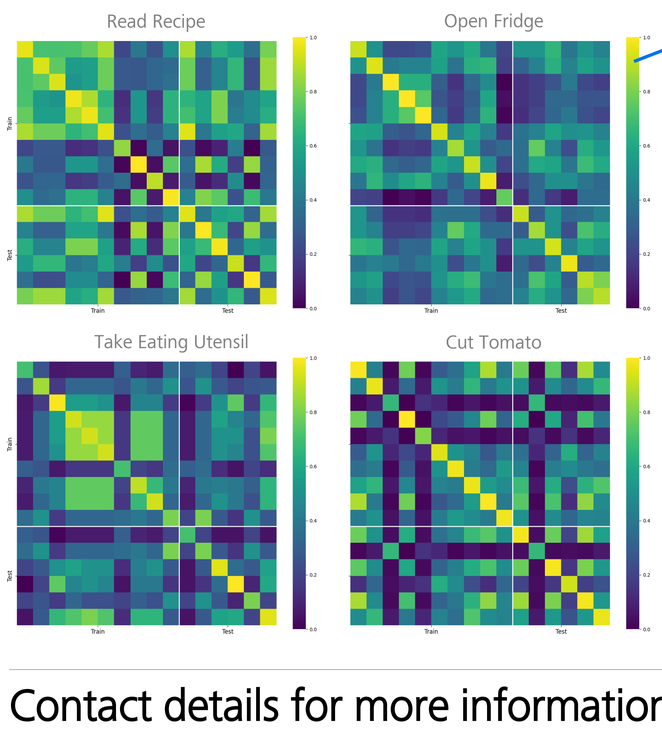
Subset	Mean Acc	Mean Class Acc
2 Classes	76.44 ± 1.16	78.67 ± 1.06
3 Classes	59.37 ± 1.38	54.64 ± 1.98
4 Classes	53.49 ± 1.24	42.74 ± 1.27
All (106*)	8.58 ± 1.3	2.14 ± 0.67

Subsets are formed by the most frequent actions in EGTEA Gaze+ (e.g., ***Read recipe***, ***Open fridge***, ***Take eating utensil***, ***Cut tomato***).
We report mean over 3 runs.

Above-chance performance without any manual correction. * Baseline accuracy is ≈ 5% (chance level).

Accuracy decreased predictably with increasing class count, but minority classes are underlearned.

Key Findings



NW Similarity of the scanpath sequences for the top 4 classes.

✓ **Proof of Concept:** Gaze-informed token scanpaths capture action semantics.
✓ YOLO-World is viable for automated AOI-Labeling.
✓ YOLO-World + BiGRU is flexible; components can be swapped or improved independently.
✓ Fully-automated pipeline too strict. Manual intervention to complement it.

Outlook
- Compare *gaze-conditioned vs. gaze-free* YOLO-World object sequences to quantify the added value of gaze.
- Benchmark against *manually coded AOIs* & established scanpath classifiers on the same dataset.
- Extend to AR/XR settings for intent recognition & assistive guidance.

GitHub

Forschungsinstitut für Augenheilkunde Universitätsklinikum Tübingen
Zeiss Vision Science Lab

References:

[1] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16901-16911).

[2] YOLO-World Model - Ultralytics YOLO Docs

[3] Duo Yang and Nora Hollenstein. 2023. PLM-AS: Pre-trained language models aug mented with scanpaths for sentiment classification. *In Proceedings of the Northern Lights Deep Learning Workshop*, Vol. 4.

[4] Li, Y., Liu, M., & Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 619-635).

[5] Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 2:Short Papers). 217–224.

[6] Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2024. Seeing eye to AI: human alignment via gaze-based response rewards for large language models. *arXiv preprint* arXiv:2410.01532 (2024).

[7] Kiegeland, S., Reich, D. R., Cotterell, R., Jäger, L. A., & Wilcox, E. (2024, July). The pupil becomes the master: Eye-tracking feedback for tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*.

[8] Catarina Moreira, Jeffrey Cockburn, and Monica S Castelhano. 2025. A Framework for Leveraging LLMs for Scene Analysis and Cognitive Processing. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 8, 2 (2025), 1–18.

Contact details for more information:

nora.castner@zeiss.com

Affiliations: [1]Carl Zeiss Vision International GmbH, Aalen, Germany, [2] Trinity College, Dublin, Ireland, [3]Institute for Ophthalmic Research, University of Tübingen, Germany

ZEISS Vision Science Lab

ZEISS VISION CARE

ZEISS

Seeing beyond