Gaze Signatures Decode the Onset of Working Memory Encoding

Decoding the Intent to Encode

Candace E. Peacock

Facebook Reality Labs Research, peacock.candace@gmail.com

Brendan David-John

Facebook Reality Labs Research, brendanjohn@ufl.com

Ting Zhang

Facebook Reality Labs Research, tingzhang@fb.com

T. Scott Murdison

Facebook Reality Labs Research, smurdison@fb.com

Matthew J. Boring

Facebook Reality Labs Research, mjb200@pitt.edu

Hrvoje Benko

Facebook Reality Labs Research, benko@fb.com

Tanya R. Jonker*

Facebook Reality Labs Research, tanya.jonker@fb.com

As eye tracking becomes increasingly common in consumer products, these systems have the opportunity to leverage users' gaze behaviors to infer interaction goals and assist users. For example, by inferring that a user is encoding information into working memory (WM), a system could provide supportive tools (e.g., a notepad) to users when they are most likely to need it. The goal of the present work was to explore whether natural gaze dynamics could be used to decode the onset of WM encoding. In an immersive virtual reality task, participants searched for and then encoded objects into WM while their eye movements were recorded. We then computed 61 gaze features and trained a sliding window logistic regression model to decode the onset of WM encoding. The results demonstrated that the model was able to detect WM encoding onsets and that this effect was not simply an artifact of learning when fixations will occur. The findings suggest that natural gaze dynamics can index when users intend to encode information, which could be used to drive adaptive interfaces.

CCS CONCEPTS • **Human-centered computing • HCI theory, concepts, and models**

Additional Keywords and Phrases: cognitive state, eye movements, working memory

ACM Reference Format:

Candace E. Peacock, Brendan David-John, Ting Zhang, T. Scott Murdison, Matthew J. Boring, Hrvoje Benko, and Tanya R. Jonker. 2021. Using gaze to decode the onset of working memory encoding. In Proceedings of the 2021 CHI Workshop on Eye Movements in Cognitive State (CHI '21). Association for Computing Machinery, New York, NY, USA, 4 pages.

1 INTRODUCTION

We move our eyes about three times per second to sample our environment and gather information that is relevant to our current goals [1, 2, 3, 4]. As such, these eye movements might be useful for inferring internal, otherwise unobservable cognitive states, such as the onset of encoding information into working memory (WM) [5, 6, 7]. As eye tracking becomes more common in consumer products, a gaze-based model that infers a user's cognitive state could allow a system to provide adaptive assistance to support their tasks. For example, if a system detects that a user is searching their pantry to construct a mental shopping list, it might launch a list app to help the user record these items and effectively offload the taxing WM task.

Empirical work has demonstrated a link between WM encoding and people's gaze patterns, such as the frequency and distance of saccades between objects in an environment [5, 6, 7]. However, these findings have not been incorporated into predictive models. Furthermore, they link WM encoding to gaze-environment interactions, which requires knowledge of target object location and identity. A model that takes gaze as input on target objects might, however, be impractical. Consumer-grade eye tracking does not track gaze locations with high precision and accuracy for everyone, which undermines the system's ability to accurately identify gaze targets. Furthermore, even if gaze coordinates can be accurately tracked, the system must be able to identify the object at those coordinates, which requires power-hungry cameras and computationally expensive computer vision models. As such, there is benefit in developing models of the intent to encode that do not depend on gaze-environment interactions.

Although there is clear value to developing models that rely on gaze dynamics alone, this is not yet a well-explored area. For example, [8] used aggregate gaze measures (e.g., total fixation counts across time) to discriminate reading from scene viewing. These findings suggest that gaze can be used to decode cognitive state, but they rely on aggregation of gaze data across time, which means they will not be responsive to real-time changes in cognitive state. To develop systems that can react to rapid changes in user goals and drive adaptive interfaces, a model must consume non-aggregate gaze measures (such as saccade amplitudes) in real-time.

Our work addressed two hypotheses. First, we hypothesized that (H1) *gaze dynamics alone can be used to anticipate the onset of encoding*. To address this, we developed a logistic regression model of the intent to encode using gaze data from a virtual WM task (Figure A1). In a follow-up analysis, we hypothesized that (H2) *the predictive model trained to test H1 is sensitive to the onset of WM rather than the onset of fixations*. To ensure that the model captured the gaze dynamics of encoding rather than fixations alone, we performed a stringent test of our model.

2 METHODOLOGY

2.1 Participants, Apparatus, and Procedure

Thirty-eight participants completed the study. Informed consent was obtained, and protocols were approved by the Western Institutional Review Board. Six participants were excluded as they failed to complete the study due to discomfort or noise disruptions resulting in a sample of N=32 (mean age: 27.7 years, 16 females).

Tasks were performed and gaze data were collected using the HTC Vive headset with Tobii Pro binocular eye tracking (120 Hz). The built-in Tobii 5-point calibration protocol was used.

A practice trial preceded the main trial sequence. In each trial, participants were spawned in one of two virtual apartments and received on-screen instructions to navigate to a specific room using point-and-teleport navigation. An arrow indicated the room's location, and it was visible through the walls to guide navigation (Figure A1). Upon arrival, the navigation arrow and text cue disappeared and either 1) one of the objects in the room became marked with a blue arrow hovering over it, or 2) participants were prompted to navigate to a new room. Participants were to remember the identity of the object marked with the blue arrow. After the initial gaze intersection on the object, the arrow disappeared and either 1) another object became marked in the same room, or 2) participants were prompted to navigate to a new room. At the end of each trial, participants verbally recalled the objects and were then given an optional break. Participants completed 30 trials in the same order. Fifteen trials contained 5 objects and 15 contained 9 objects for a total of 210 encoded objects.

2.2 Feature Computation and Sliding Window Framework

Encoding onsets were defined as the fixation after the first gaze ray intersection with either the arrow or object (whichever occurred first). The time-series gaze data collected in one trial were then trimmed into 5 or 9 clips per trial based upon the number of encoded objects. Each clip began with the appearance of the blue arrow and ended with the onset of encoding. Two exclusion criteria were applied: 1) Clips corresponding to forgotten items were discarded (14.43% data loss) as it was impossible to know whether incorrect trials were due to forgetting or inattentiveness; 2) To account for tracker error, we used a strict criterion: only clips in which encoding onsets were less than 500ms after the first gaze intersection were used (M = 17.40% data loss; range = 2.13% to 34.52%).

Gaze data were transformed by a rotation matrix to correct for head orientation [9]. The identification by velocity threshold was used for event detection [10]. Gaze velocity was computed as the angular distance between samples divided by the change in time. A saccade was detected if gaze velocity was greater than 70°/s for 12 to 300ms [9, 11]. A fixation was detected if gaze velocity was less than 20°/s for 50 to 1500ms [9]. We then computed 61 gaze features which included gaze velocity, 58 fixation/saccade features [11], the k-coefficient, [12] and dispersion (see A.2 for full feature set). Features derived from fixations/saccades have missing values when no events are detected and were therefore linearly interpolated since logistic regression cannot handle missing data (see next section). Each gaze feature was also z-scored within-participant. To augment the number of true data points, data occurring 20ms prior to the encoding onset was marked as a true class.

To create input samples for model training, a sliding window of N ms (later determined through hyper-parameter search) with a step size of 1 sample (~8.33ms; how many samples to move forward in time) was used on each feature. The class label (encoding or not encoding) was determined by the class of the last timestamp in a window. Features from fixation/saccade events may contain duplicate values. To increase the computational efficiency and reduce collinearity, we downsampled by averaging every 5 samples (~42ms) within each window.

Logistic regression models were used to detect the intent to encode because they are interpretable and lightweight. Due to highly imbalanced data with 99.1% more null samples, the class weights of the models were set to be inversely proportional to the number of samples for each class. Models were trained within-participant. Each participant's data were split into 90% training and 10% test sets. A stratified 10-fold cross-validation with three repeats was used for hyperparameter tuning and feature selection on the training set.

The area under the precision-recall curve (AUC-PR) was used for model evaluation, which is well-suited to imbalanced data. A shortcoming of AUC-PR is that the baseline value is derived from the chance rate of true examples, which varies by individual and window size, making it difficult to compare model performance. To create a standardized chance rate for each individual, we resampled the training and testing data to have a fixed percentage of true classes (0.9%) based on the average true class percentage across individuals.

Since the predictive window size may vary by gaze feature, we computed the AUC-PR for a set of window sizes (5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120) ranging from 42ms to 1000ms and chose the window size with the largest AUC-PR for each feature. Recursive feature selection was then applied to identify useful gaze features. Individual features were added in a random order to the model. Features were retained if they increased the AUC-PR and were dropped otherwise after the addition of a feature. Features were concatenated with their optimal window size if they aligned at the same endpoint. Features were relatively consistent across participants: 56 out of 61 features were selected for 50% or more participants. The top three consistent features were fixation detection, gaze velocity, and the angular displacement between the current and previous saccade landing points.

3 RESULTS

H1 was first tested. Here, our model performed above chance ($M = 0.29$, $SD = 0.21$, chance = 0.009) on unseen test data as per a one-sample t-test ($t(31) = 7.34$, $p < 0.001$), demonstrating that our model detected encoding onsets without knowledge of the environment. We next tested whether the model trained to test H1 had simply learned fixation onsets (H2). If this was the case, then the H1 model should confuse null classes preceding fixations with true classes preceding encoding fixations, thereby resulting in chance performance. To match our selection for true classes, we filtered null classes from the test set to have the same onsets: Null classes were included if the last sample in a window containing no fixation was followed by the onset of a fixation. These filtered test cases were then resampled to match the standardized chance rate (0.009) (see also A.3.1). Overall, the H1 model performed significantly above chance ($M = 0.26$, $SD = 0.23$, $t(31) = 6.11$, $p < 0.001$), suggesting that the model did not learn to detect fixation onsets (Figure 1). Finally, there was no difference between the H1 and H2 results as per a paired t-test, suggesting that any differences in performance were not great enough to detect statistically ($t(31) = 1.04$, $p = 0.31$). Overall, the results showed that the model decoded the intent to encode without knowledge of the environment and that this was not due to the model learning fixation onsets. We also report the results using AUC-ROC for interpretability (A.3.2) and discuss the between-subject variation in performance (A.3.3).

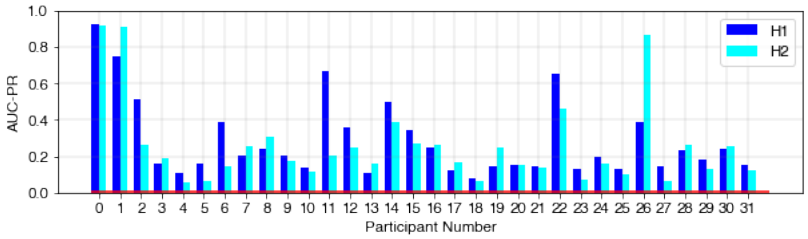

Figure 1: The blue (cyan) bars represent the AUC-PR for H1 (H2). The red line is chance (0.009).

4 DISCUSSION

The present work tested whether gaze dynamics can decode the onset of WM encoding without knowledge of the environment. Overall, our model performed above chance, demonstrating that gaze dynamics do reflect the intent to encode (H1). Importantly, these models were performant using consumer-grade eye-tracking devices, suggesting potential for integration into working systems. Furthermore, a control analysis verified that the model did not simply learn to detect fixation onsets (H2), suggesting that the model's gaze features are sensitive to the intent to encode beyond general fixation onsets.

Prior empirical work has relied on gaze-environment interactions to make inferences about encoding [5, 6, 7]. Using knowledge of the environment to validate internal goals may be problematic in consumer settings, though, as eye-tracking technologies are inaccurate and computer vision is computationally expensive. To provide a novel solution, we used gaze dynamics alone to successfully decode encoding intent.

Although we validated the intent to encode model in a single setting, future work should examine if these models generalize to different tasks and environments. Furthermore, logistic regression does not capture relative changes in the time-series signal; therefore, time-series modeling might improve performance of the models and provide more theoretical insights about the relationship between gaze and encoding intent. Finally, because high-performing participants retained more features than low-performing participants (Figure A2), it may be the case that the chosen features did not capture the full range of gaze behaviors across people. Future work may wish to explore whether additional features are sensitive to encoding onsets.

In conclusion, the present study leveraged gaze features to decode the intent to encode. We provided a framework using gaze dynamics that did not rely on knowledge of the environment to decode cognitive state before the onset of encoding. Although gaze features aggregated across time can differentiate tasks, such as scene viewing versus reading [8], our work demonstrates the potential to detect the *onsets* of new cognitive states in real-time. This has great potential utility for adaptive interfaces as these onsets can be used to launch new assistive tools in response to real-time changes in gaze behavior.

REFERENCES

[1] Alfred L Yarbus. 1967. Eye movements during perception of complex objects. In Eye Movements and Vision. Springer, Boston, MA, 171–211. DOI:https://doi.org/10.1007/978-1-4899-5379-7_8

[2] G T Buswell. 1935. How people look at pictures: a study of the psychology and perception in art. Oxford, England.

[3] Michael F Land and Mary M Hayhoe. 2001. In what ways to eye movements contribute to everyday activities? Vision Research 41, 25–26 (2001), 3559–3565. DOI:https://doi.org/10.1016/S0042-6989(01)00102-X

[4] John M. Henderson. 2003. Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7, 11 (2003), 498–504. DOI:https://doi.org/10.1016/j.tics.2003.09.006

[5] Dejan Draschkow, Melvin Kallmayer, and Anna C. Nobre. 2020. When Natural Behavior Engages Working Memory. Current Biology (December

2020). DOI:https://doi.org/10.1016/j.cub.2020.11.013

[6] Jason A. Droll and Mary M. Hayhoe. 2007. Trade-offs between gaze and working memory use. Journal of Experimental Psychology: Human Perception and Performance 33, 6 (2007), 1352–1365. DOI:https://doi.org/10.1037/0096-1523.33.6.1352

[7] Dana Ballard, Mary Hayhoe, and Jeff Pelz. 1995. Memory Representations in Natural Tasks. Journal of Cognitive Neuroscience 7, (December 1995), 66–80. DOI:https://doi.org/10.1162/jocn.1995.7.1.66

[8] John M. Henderson, Svetlana V. Shinkareva, Jing Wang, Steven G. Luke, and Jenn Olejarczyk. 2013. Predicting Cognitive State from Eye Movements. PLOS ONE 8, 5 (May 2013), e64937. DOI:https://doi.org/10.1371/journal.pone.0064937

[9] Gabriel Diaz, Joseph Cooper, Dmitry Kit, and Mary Hayhoe. 2013. Real-time recording and classification of eye movements in an immersive virtual environment. Journal of Vision 13, 12 (October 2013), 5–5. DOI:https://doi.org/10.1167/13.12.5

[10] Dario D Salvucci and Joseph H Goldberg. Identifying Fixations and Saccades in Eye-Tracking Protocols. 8.

[11] Anjith George and Aurobinda Routray. 2016. A Score-level Fusion Method for Eye Movement Biometrics. Pattern Recognition Letters 82, (October 2016), 207–215. DOI:https://doi.org/10.1016/j.patrec.2015.11.020

[12] Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning Ambient/Focal Attention with Coefficient K. ACM Trans. Appl. Percept. 13, 3 (May 2016), 1–20. DOI:https://doi.org/10.1145/2896452

APPENDICES

A.1 Task Figure

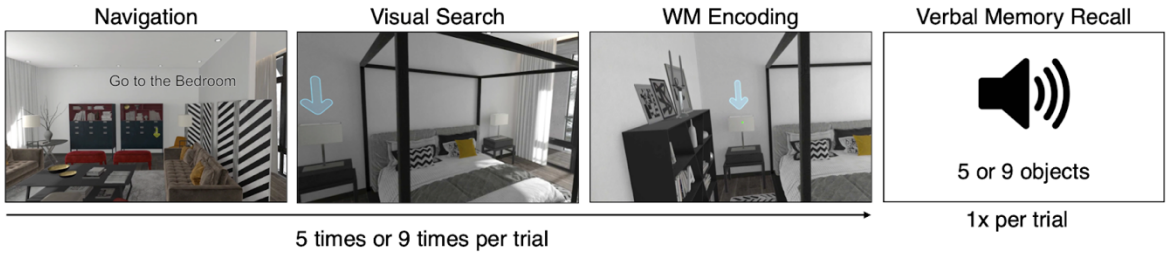

Figure A1: Task figure.

A.2 Feature Set

The following features were computed. M3S2K refers to the computation of mean, median, maximum, standard deviation, skewness, and kurtosis.

- Gaze velocity
- Dispersion: 500ms time window; computed as the maximum angular distance between centroid of samples in time window and each sample
- Fixation detection: sample-level categorical feature indicating whether a sample was a fixation or not
- Saccade detection: sample-level categorical feature indicating whether a sample was a saccade or not.
- K-coefficient: 1000ms time window; computed from how exploratory (ambient) or directed (focal) gaze was [12]
- Features derived from [11]
    - Fixation duration
    - Standard deviation of horizontal gaze position during fixation
    - Standard deviation of vertical gaze position during fixation
    - Path length of gaze samples during fixation
    - Angular displacement between current and previous fixation centroids
    - Angular displacement between fixation centroid and last sample of previous fixation

- Horizontal skewness of gaze samples during fixation
- Vertical skewness of gaze samples during fixation
- Horizontal kurtosis of gaze samples during fixation
- Vertical kurtosis of gaze samples during fixation
- Dispersion of gaze samples during fixation
- Average velocity of gaze samples during fixation
- Saccade duration
- Dispersion of gaze samples during saccade
- M3S2K of gaze velocity during saccade
- M3S2K of gaze acceleration during saccade
- Standard deviation of horizontal gaze position during saccade
- Standard deviation of vertical gaze position during saccade
- Path length of gaze samples during saccade
- Angular displacement between current and previous saccade landing points
- Angular displacement between current and previous saccade centroids
- Saccadic ratio: peak velocity / saccade duration
- Saccade amplitude
- M3S2K of horizontal gaze velocity during saccade
- M2S2K of vertical gaze velocity during saccade
- M3S2K of horizontal gaze acceleration during saccade
- M3S2K of vertical gaze acceleration during saccade

A.3 Supplementary Results

A.3.1 H2 Results Using Average Chance From H2

Because the H2 chance was greater than H1 (due to using filtered/fewer null cases preceding fixations), we also computed the H2 results using the average chance from H2 (8%) to resample the data. The results were still significant ($M = 0.63$, $SD = 0.19$; $t(31) = 16.14$, $p < 0.001$), suggesting that the model was not predicting above chance on the null classes preceding fixations due to our resampling method.

A.3.2 Results using AUC-ROC

We also computed the results using the area under the receiver operating characteristic curve (AUC-ROC) since this is more common and interpretable than AUC-PR. Overall, the results were unchanged. When testing H1, the results showed that the model performed above chance ($M = 0.96$; $SD = 0.02$, chance = 0.5) on the unseen test data as per a one-sample t-test ($t(31) = 113.82$, $p < 0.001$) suggesting that gaze features can decode the intent to encode without the use of environmental features. The H2 results were also unchanged ($M = 0.95$; $SD = 0.04$, chance = 0.5; $t(31) = 59.64$, $p < 0.001$), suggesting that the model was not simply detecting when fixations would occur. Overall, the results using AUC-ROC showed that the models performed exceedingly well at decoding the intent to encode.

A.3.3 High- Vs. Low-Performing Participants

Some participant models performed better than others. One reason for this could be that there were more features predictive of encoding in high-performing participants relative to the low-performing participants. Indeed, a

supplementary analysis showed that test AUC-PRs were significantly correlated to the number of features retained per participant (Pearson r = 0.37, p = 0.04), suggesting that there are likely individual differences in which features are predictive of the intent to encode.

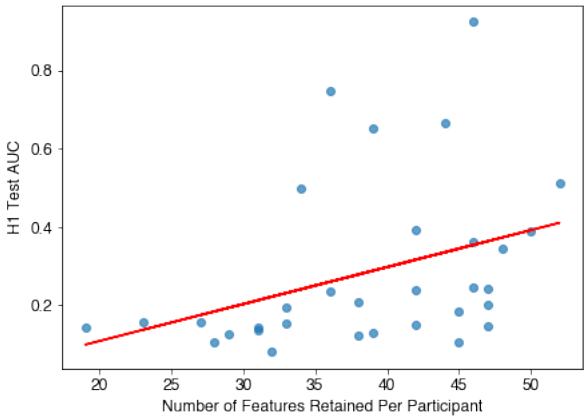

Figure A2: The relationship between test AUCs and the number of features retained for each participant. Each blue circle corresponds to a participant. The red line corresponds to the Pearson correlation between test AUC-PRs and the number of features retained.