Towards gaze-based prediction of the intent to interact in virtual reality

Brendan David-John
brendanjohn@ufl.edu
Facebook Reality Labs Research
Redmond, Washington, USA

Candace E. Peacock
cepeacock@ucdavis.edu
Facebook Reality Labs Research
Redmond, Washington, USA

Ting Zhang
tingzhang@fb.com
Facebook Reality Labs Research
Redmond, Washington, USA

T. Scott Murdison
smurdison@fb.com
Facebook Reality Labs
Redmond, Washington, USA

Hrvoje Benko
benko@fb.com
Facebook Reality Labs Research
Redmond, Washington, USA

Tanya R. Jonker
tanya.jonker@fb.com
Facebook Reality Labs Research
Redmond, Washington, USA

ABSTRACT

With the increasing frequency of eye tracking in consumer products, including head-mounted augmented and virtual reality displays, gaze-based models have the potential to predict user intent and unlock intuitive new interaction schemes. In the present work, we explored whether gaze dynamics can predict when a user intends to interact with the real or digital world, which could be used to develop predictive interfaces for low-effort input. Eye-tracking data were collected from 15 participants performing an item-selection task in virtual reality. Using logistic regression, we demonstrated successful prediction of the onset of item selection. The most prevalent predictive features in the model were gaze velocity, ambient/focal attention, and saccade dynamics, demonstrating that gaze features typically used to characterize visual attention can be applied to model interaction intent. In the future, these types of models can be used to infer user's near-term interaction goals and drive ultra-low-friction predictive interfaces.

CCS CONCEPTS

• **Human-centered computing** → *Mixed / augmented reality*; **HCI theory, concepts and models**.

KEYWORDS

intent prediction, eye tracking, mixed reality, virtual reality, interaction

ACM Reference Format:
Brendan David-John, Candace E. Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ETRA '21: 2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Short Papers), May 25–27, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3448018.3458008

ETRA '21 Short Papers, May 25–27, 2021, Virtual Event, Germany

ACM ISBN 978-1-4503-8345-5/21/05…$15.00
https://doi.org/10.1145/3448018.3458008

1 INTRODUCTION AND RELATED WORK

To enable widespread consumer adoption of augmented reality (AR), virtual reality (VR), and mixed reality technologies, collectively referred to as "XR", these devices must deliver natural experiences that directly benefit users without unnecessarily burdening them. A key component of a natural, effortless user experience is providing input to a system, as the input paradigm allows the user to express their interaction intent to the system. And yet, the predominant interaction models for XR rely on either fully manual input, which has been shown to be physically fatiguing [Hincapié-Ramos et al. 2014; Kang et al. 2020], or voice commands, which are limited in their capability and social acceptability.

XR systems have the opportunity to reduce user effort by anticipating what a user intends to interact with, and then providing them with a "quicklink" to complete the predicted action with less friction [Jonker et al. 2020]. For example, if the XR system was aware that the user was about to interact, it could map an input gesture to the inferred target of interaction and allow the user to complete the action without manually pointing to it, thereby reducing the physical and cognitive burden on the user. The necessity of this approach becomes clear as we move towards all-day wearable AR devices and VR for productivity and work, where long sessions of use will exacerbate fatigue. Adaptive interfaces that accurately predict a user's intent to interact enable adaptive inputs will vastly reduce the amount of physical and cognitive burden on the user, and have the potential to shape the future of XR interaction.

Eye gaze is a promising data stream for predicting user interaction intent. In prior research, eye movements have provided substantial insights into human behavior and cognition, with vast amounts of literature demonstrating the relationship between eye movements and attention [Borji and Itti 2012; Frintrop et al. 2010; Wolfe 2000], cognitive state [Hayhoe and Ballard 2005; Henderson et al. 2013], decision making [Land and Hayhoe 2001; Orquin and Loose 2013], and memory [Ballard et al. 1995; Hollingworth and Luck 2009; Hollingworth et al. 2013; Peacock et al. 2021]. This body of research demonstrates that eye movements characterize internal cognitive states and goals, which lays the groundwork for computational models that can anticipate human behavior and drive adaptive interfaces in real time. As such, we explored the use of gaze data to predict when the user intends to interact with an XR system.

Although gaze data has several clear uses for XR [Kim et al. 2019; Patney et al. 2016; Sun et al. 2018], it has been underutilized when it comes to providing real-time prediction during interaction [Lengyal et al. 2021]. Some prior investigations have explored how gaze can be used to predict a user's intent to interact, but these focus on the distribution of attention over defined "areas of interest" in the environment. For example, Pfeuffer et al. [2021] produced an interface concept that had elements move between the background and foreground depending on the distribution of gaze between real and virtual content to accentuate the content that is most likely to be of interest. Similarly, others have also defined models that use the distribution of gaze data on specific regions within the environment to identify and highlight task relevant objects [Gebhardt et al. 2019], optimize placement of visual elements [Alghofaili et al. 2019b], visualize potential actions [Karaman and Sezgin 2018], or support cross-device interaction and information sharing [Li et al. 2019]. These examples provided promising results for enhancing interaction through adaptive interfaces, but they all rely on knowledge of the environment and the eye's gaze point in that environment. These systems rely on robust object detection and tracking algorithms to make predictions, which limits their application to mobile XR devices, where the scene is not perfectly known and computational resources are limited. Furthermore, adaptive interfaces that rely on the interaction between gaze and the environment require a well calibrated and highly precise eye-tracker[1], which is not yet feasible in commercial devices nor for all people.

There is a considerable benefit in developing models of the intent to interact that do not depend on the interaction between gaze and the environment, but instead on the dynamics of eye movements. For example, Alghofaili et al. [Alghofaili et al. 2019a] demonstrated that an LSTM model trained only on the angle between gaze and head direction can be used to identify when a user in VR is lost and needs navigation assistance. Similarly, gaze features, such as those based on eye-movement speed, might reveal patterns of slow and fast eye movements that characterize different cognitive processes [Land and Tatler 2009; Land and Hayhoe 2001; Tatler and Vincent 2008], which could then be used to identify patterns that precede important actions in the system (e.g., a selection).

Past work demonstrated that a model trained using gaze dynamics could accurately decode whether a selection was made [Bednarik et al. 2012]. Bednarik et al. used fixation, saccade, and pupil features to estimate when users will click to select a tile to move in a puzzle game, and they achieved good performance with AUC-ROC scores (a metric of model performance; see Sec. 2.4) as high as 0.81 when using all features. However, their prediction model incorporated gaze data up to one fixation *after* the click, which reduces its potential application to a real-time scenario. It is not clear how their model would have performed if it used gaze data prior to selection alone. Nevertheless, their results suggest that it is possible to decode moments of interaction from gaze data alone.

1.1 Model Criteria and Our Approach

An ideal model of the intent to interact should be robust to eye-tracker calibration accuracy, sensor noise, and variation between individuals. Furthermore, it should be able to make predictions in real-time. An ideal model should be able to make real-time prediction of when a user will interact with a system, which is a critical step in deploying intent-to-interact models for XR interfaces, as it will allow the system to provide adaptive interventions at just the right time. As such, in this paper, we explore whether gaze-based models can predict the onset of interaction in XR. To our knowledge, this is the first investigation predicting the temporal onset of interaction intent using nothing more than gaze dynamics leading up to the moment of input. We trained logistic regression models to predict the moment of interaction based on point-and-click events in VR. In our work, we explored two key hypotheses: (H_1) Natural gaze dynamics from eye-tracking can be used to predict the onset of interaction in VR, (H_2) There is a consistent set of gaze features across individuals that reflect eye movements related to interaction.

2 METHOD

We conducted a remote VR study to ensure safe data collection during the COVID-19 pandemic. This section defines the experimental protocol, data processing pipeline (Figure 2, Left), feature extraction approach, models, and metrics.

2.1 Protocol

2.1.1 Participants. Fifteen participants were recruited with informed consent under a protocol approved by the Western Institutional Review Board. Participant age ranged from 18 to 54 and two were left-handed. Participants included five females and ten males, and sampled from a population without extensive experience using virtual reality devices. Participants were screened to have normal or corrected-to-normal vision from contact lenses, as eye glasses interfere with eye-tracking quality. Participants received equipment by mail and interfaced with researchers through video calls to complete the experiment remotely.

2.1.2 Equipment. Eye and head movements were collected from an HTC Vive Pro Eye HMD. Eye-tracking data was logged at 120 Hz for two participants and 60 Hz for others. The difference in sampling rate was a result of the participant hardware configuration. Prior to the experiment, each participant was walked through a 5-point eye-tracking calibration. The calibration quality was then assessed qualitatively using the 9-point validation available through the SteamVR system dashboard. The distance between validation targets spanned approximately 11 degrees visual angle.[2] Participants had to light up each validation target consistently, achieving a spatial error of no more than 5.5 degrees visual angle[3].

2.1.3 Experiment Design. We designed an item selection task in a virtual pantry. Participants were presented with shelves filled with common food items and were asked to choose items that best fit a provided recipe. Participants had to select at least three items, and at most thirteen. There were six total recipes designed for this task: Fruit Salad, Garden Salad, Smoothie, Protein Shake, Soup, and Stir-fry. Items were selected by pointing the Vive hand controller

[1]Well calibrated and precise refers to the lack of a fixed offset in gaze position after calibration, or drift resulting from slippage.

[2]Visual angle was determined by measuring pixel distances in a 1440×1600 monocular viewport that spans a 110° field of view diagonally.

[3]The calibration accuracy serves as an upper bound on expected gaze error, and as a result our approach accommodates larger offsets in spatial accuracy than typical eye-tracking studies.

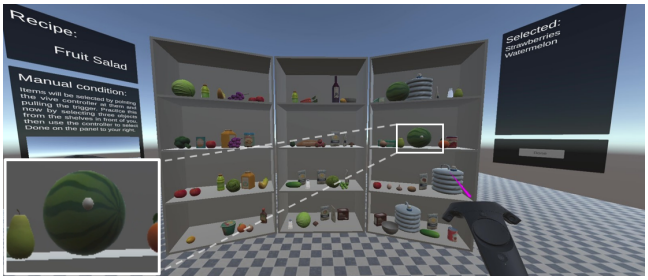

Figure 1: Virtual pantry scene used to select items for the given recipe; (inset) gray sphere indicates intersection with the controller ray.

and pulling the trigger. A gray sphere placed at the intersection between the controller ray and item indicates the item could be selected, (Figure 1, inset). Each recipe was considered one trial. Upon starting the experiment, each participant performed six practice trials with the Fruit Salad recipe. Next, nine blocks of six trials were presented with a random assortment of recipes and item layouts. Each participant encountered the same sequence of recipes and layouts. Before each block, participants were offered a voluntary break of up to five minutes. After each block, participants completed subjective surveys, but these data are not analyzed within the scope of this paper. After completing the experiment a post-session survey was used to capture simulator-sickness scores [Kennedy et al. 1993] and general questions about experience with VR devices. The entire data collection process lasted approximately 30 minutes.

2.2 Pre-processing

Our processing pipeline is visualized in Figure 2. The first step involved transforming the 3D gaze vectors from the eye-in-head frame of reference to an eye-in-world direction using head orientation [Diaz et al. 2013]. Next, we computed angular displacement between consecutive gaze samples, represented as normalized vectors u and v, $\theta = 2 \cdot atan2(\|u - v\|, \|u + v\|)$.[4] Gaze velocity was computed as θ divided by the change in time between gaze samples.

2.2.1 Filtering. Gaze data was filtered to remove noise and unwanted segments before event detection and feature extraction. Data from the practice trials and breaks was discarded prior to analysis, and we remove all gaze samples where gaze velocity exceeds $800°/s$, indicating unfeasibly fast eye movements [Dowiasch et al. 2015]. The removed values were then replaced through interpolation. Finally, a median filter with a width of seven samples was applied to the gaze velocity signal to smooth the signal and account for noise prior to event detection [Pekkanen and Lappi 2017].

2.2.2 Eye Movement Event Detection. I-VT saccade detection was performed on the filtered gaze velocity by identifying consecutive samples that exceeded $70°/s$ [Salvucci and Goldberg 2000]. A minimum duration of 17ms and maximum duration of 200ms was enforced for saccades. I-DT fixation detection was performed by computing dispersion over time windows as the largest angular displacement from the centroid of gaze samples. Time windows where dispersion did not exceed 1° were marked as fixations. A minimum duration of 100ms and maximum duration of 2s was enforced for fixations. An example of fixation and saccade labels are shown in Figure 2.

2.2.3 Ground Truth. To mark the ground truth of input onsets, we used the trigger events from the hand controller. To ensure sufficient samples for training and to allow for temporal variability in features preceding the click, any sliding window that ended within 200 ms of a click event was considered a positive class sample (Fig. 2).

2.3 Feature Extraction

We explored a set of 61 total features, including gaze velocity, dispersion, event detection labels, low-level eye movement features derived from events [George and Routray 2016], and the K coefficient to discern between focal and ambient behavior [Krejtz et al. 2016]. A full list of the features is included in the Supplementary Material.

Gaze velocity and dispersion provide continuous signals that represent how fast gaze is moving and how spread out gaze points are over a period of time, respectively. The dispersion algorithm requires a time parameter that indicates the amount of gaze data to be included in the computation. We set this time parameter to one second to match the maximum possible duration of model input (Sec. 2.3.2). Low-level features were extracted from each fixation and saccade event [George and Routray 2016]. The feature set included fixation duration, saccade amplitude, saccade peak velocity and statistical measures such as skewness and kurtosis applied to gaze samples from each event. To represent these features as a continuous time-series, we set the value for each gaze sample as the feature value from the most recent fixation or saccade event, i.e., each was carried forward in time until the next detected event.

The K coefficient [Krejtz et al. 2016] signal is unique within our feature set as it is influenced by both fixation and saccade events, measuring ambient/focal eye movement behavior. Ambient eye movements describe exploratory eye movements that are used to visually sample the scene, whereas focal describes deliberate eye movements used to foveate something of interest. The coefficient is defined as the difference between z-scores of fixation duration and saccade amplitude, with positive values indicating focal behavior from long fixations and short saccades, and negative values ambient behavior from short fixations and large saccades. This measure is typically computed over a long sequence of eye movements; however, we considered a shorter window of five seconds preceding each gaze sample. We expected participants to quickly shift search and selection behavior during our task and using longer time windows would reduce the temporal resolution of the feature.

2.3.1 Temporal Resample. Temporal resampling was applied to account for irregular gaze data sampling and to account for two participants with 120Hz data. Linear interpolation was used to sample each feature uniformly in time at 60Hz.

2.3.2 Sliding Windows. Model input was defined as sliding windows of the time-series data. Sliding windows have three parameters: window size, step size, and binning factor. Window size defined the duration of a predictive window used for model input, while step size determined how many samples to move forward in time when generating the sliding windows. The binning factor is used to

[4]This equation is more numerically stable for small values of θ than the typical arccosine of the dot product method.

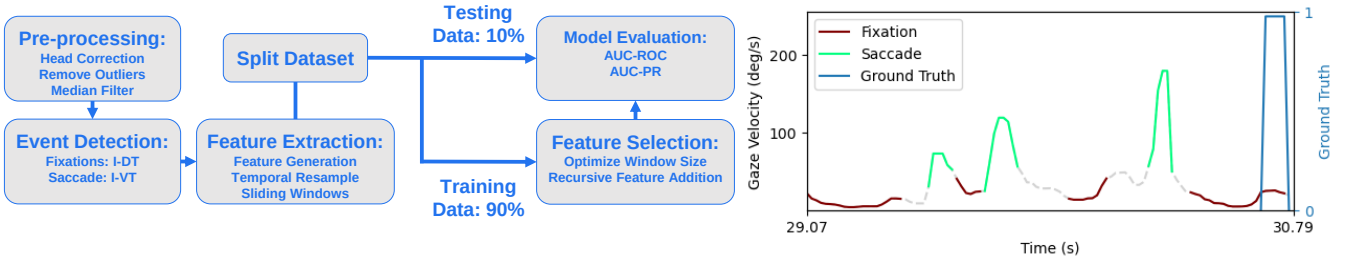

Figure 2: Pipeline used to detect events, extract features, perform feature selection, and evaluate model performance. Right: An example of gaze velocity signal preceding a click event. Fixations and saccades are indicated by dark red and light green colors respectively.

reduce the number of model inputs while still capturing temporal patterns by averaging sequential feature values in time. For example, a 60 sample window that spans one second can be compressed into a 30 sample window with a binning factor of two by averaging feature values from the first two samples, the third and fourth sample, and so on. Reducing the number of model inputs reduces the degrees of freedom for the model, lowering computation time during model training. We used a step size of one and a binning factor of five based on model performance during initial analyses. We identified an optimal window size for each gaze feature per individual using cross-validation of a logistic regression model trained using only that feature. We evaluated window sizes of 10, 20, 30, 40, 50, and 60 samples, spanning 167ms to 1000ms. The window size with the highest model performance was selected as optimal for that feature, and stored per individual for use in feature selection and model evaluation.

2.4 Metrics

Model performance for prediction is typically measured using the area-under-the-curve (AUC) of the Receiver Operator Characteristic (ROC) curve [Bradley 1997]. The ROC curve is constructed to model true positive rate as a function of false positives at different threshold values. Larger values indicate better predictive performance of the model, and all results are compared to a baseline value of 0.5 that represents a no skill classifier that performs classification by guessing. In addition, we computed the AUC of the Precision-Recall (PR) curve, which is a better metric for highly imbalanced data [Davis and Goadrich 2006; Tatbul et al. 2018]. The AUC-PR metric is more sensitive to a large number of null classes that are mis-classified as false positives. One shortcoming of AUC-PR is that the baseline value is derived from the chance rate of positive examples, which will vary by individual and sliding window parameters, making it difficult to compare model performance directly between individuals and models. To create a standardized rate of chance for each individual, we resampled our data to have a fixed percentage of positive classes that was equal to the average across individuals (4.5%).

2.5 Feature Selection

A recursive feature addition process was used to identify the top features for each individual. This process started with an empty set and iterated over each feature in a random order for inclusion. During each iteration, the features were used to train a logistic regression model with ten-fold cross validation and three repeats, which resulted in an averaged AUC-PR score. When an added feature improved the AUC-PR, the feature was retained. George and Routray [2016] employed a similar procedure, but with recursive feature elimination[5]. For each individual, a different order of features was considered, and the models were trained using the optimal window size for each feature as described in Section 2.3.2. Individuals retained 20 features on average, with a minimum of 12 and maximum of 26 features.

2.6 Model

We explored whether logistic regression models, commonly used for gaze data [Costela and Castro-Torres 2020; Gingerich and Conati 2015], could predict the intent to interact using sliding windows.

3 RESULTS

H_1: Natural gaze dynamics from eye-tracking can be used to predict the onset of interaction in VR. To explore whether natural gaze dynamics could successfully predict the onset of interaction, we assessed model performance against chance. For the presented analysis all models were trained with the features listed in Table 1 and used the optimal window sizes determined for each participant. Across our sample of participants, we found a maximum AUC-PR of 0.194 and minimum of 0.065 (Table 2), demonstrating that the trained models for all participants produced above-chance performance on test data (chance=0.045). A similar pattern was observed for AUC-ROC: we found a maximum score of 0.92 and minimum of 0.717. Figure 3 illustrates the mean PR and ROC curves (black line) along with the curves for each participant (light gray line). Average model performance was significantly greater than chance for both AUC-PR (Mean=.121, t(14)=7.2,p<.0001) and AUC-ROC (Mean=.767, t(14)=18.36,p<.0001) using a one sample t-test.

H_2: There is a consistent set of gaze features across individuals that reflect eye movements related to interaction. To provide insight into the signal within eye-tracking data that was driving model performance, we identified the features that were selected by more than half of the participants using the recursive feature selection protocol. This resulted in a set of twelve features, which are described in Table 1.

[5]We first evaluated the elimination approach and found that the retained set depended heavily on the random order in which they were considered.

Table 1: Features selected for model evaluation and the number of participants in which they were retained.

Feature	Count	Feature	Count	Feature	Count
Fixation Detection	13 (87%)	Std. Dev. of Vert. Gaze during Saccade	9 (60%)	Saccade Duration	8 (53%)
Gaze Vel.	12 (80%)	Kurtosis of Vel. during Saccade	9 (60%)	K Coefficient	8 (53%)
Average Vel. during Fixation	10 (67%)	Skew of Vel. during Saccade	9 (60%)	Std. Dev. of Vel. during Saccade	8 (53%)
Skew of Horiz. Accel. during Saccade	10 (67%)	Skew of Horiz. Vel. during Saccade	9 (60%)	Ang. Distance from Prev. Saccade	8 (53%)

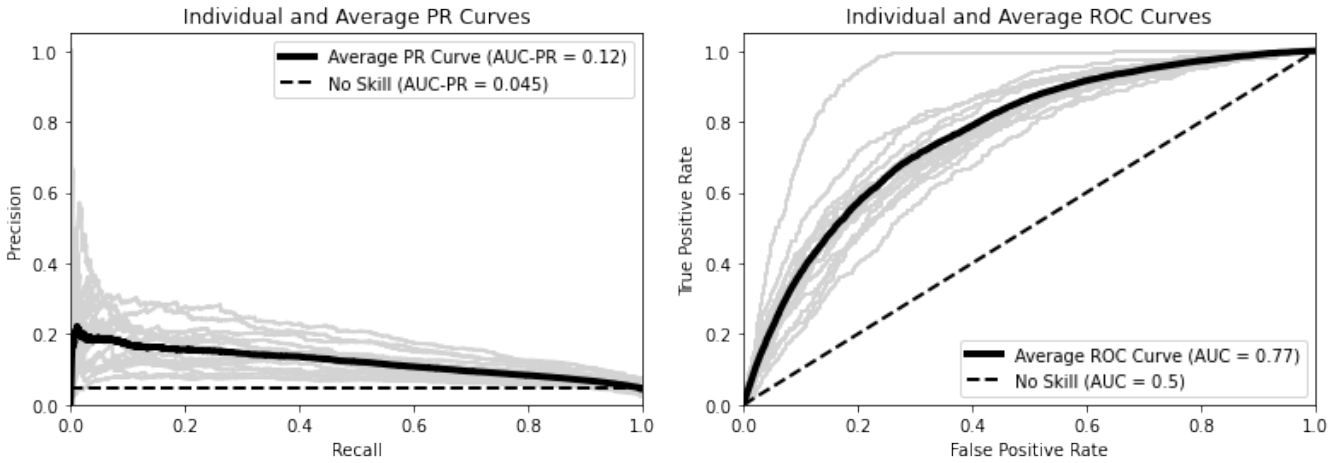

Figure 3: PR (left) and ROC (right) performance curves illustrating the average curve (black line), individual curves (gray lines).

4 DISCUSSION

We investigated two key hypotheses in this work: (H_1) Natural gaze dynamics from eye-tracking can be used to predict the onset of interaction in VR, (H_2) There is a consistent set of gaze features across individuals that reflect eye movements related to interaction. We validated H_1 by demonstrating that our models produced above-chance prediction of the intent to interact in VR across two different metrics. These results suggest that we were able to predict moments of interaction using the dynamics of gaze alone, even independently of any knowledge of the eye's location in the environment or the identity of the gazed-upon object. In fact, AUC-ROC scores from our model were comparable to that of Bednarik et al. [2012] even though our model only used eye-tracking data that preceded the selection event, while earlier demonstrations used eye events that both preceded and followed the selection. Please see the Supplementary Material for a direct comparison to [Bednarik et al. 2012] with additional performance metrics.

We found that H_2 was partially supported by identifying the top gaze features for modeling across participants. Twelve features were selected by over half of the time, however they were not common to all participants. Typically, during search or during orienting to a target, people tend to produce an initial large eye movement proportional to the eye's distance from the target, followed by one or more smaller corrective saccades that bring the visual target into the fovea and allow the extraction of that target's visual features [Krejtz et al. 2017]. The number and amplitude of these corrective eye movements directly depend on the initial distance of the saccade target due to a systematic hypometric bias [Lisi et al. 2019], which is exemplified by an "undershooting" of the larger initiated saccade. For lower amplitude saccades indicative of ambient orienting, this behavior is less pronounced due to a smaller initial absolute error. These known patterns of eye movements are captured by the K coefficient, which is a measure of ambient/focal attention, as well as specific statistical features for saccadic eye movements, such as the skew. This result corroborates past findings from eye movement literature that gaze can be used to predict interaction [Hayhoe and Ballard 2014; Hayhoe and Matthis 2018], and that there are patterns of eye-movements that characterize particular phases of visual search, orientation, and selection.

Limitations and Future Work. Although our results are limited to our specific task and to VR, the proposed modeling framework can be applied to gaze data and corresponding head rotations collected in XR as it does not depend on information about the environment or offsets in spatial accuracy less than 5.5 degrees visual angle. Furthermore, given the alignment between our top predictive features and the substantial literature on gaze behaviours during visual search and interaction, we expect the modeling framework to be generalizable to novel tasks. Our results are limited to offline analysis, however commercial eye trackers are capable of real-time event detection that would support the proposed modeling framework. In future work, these types of modeling effort might be improved upon through the use of time-series models, including recurrent neural networks and variants such as long short-term memory models. While logistic regression models are interpretable and light-weight, we see the potential to incorporate deep networks to boost performance. Deep learning approaches have already been applied to enhance mid-air interactions and reduce user fatigue [Cheema et al. 2020]. In future work trained models can be deployed as part of an adaptive interface to better understand how well intent-to-interact models perform in practice.

Table 2: Area under the curve metrics from within-subjects evaluation. AUC-PR and AUC-ROC have a baseline score of 0.045 and 0.5 respectively. Bold values indicate the top performing model for the corresponding metric.

Participant	AUC-PR	AUC-ROC	Participant	AUC-PR	AUC-ROC	Participant	AUC-PR	AUC-ROC
P001	0.150	0.717	P008	**0.194**	0.754	P014	0.083	0.721
P002	0.065	0.688	P009	0.110	**0.922**	P015	0.084	0.732
P004	0.128	0.793	P010	0.141	0.748	P016	0.107	0.763
P005	0.072	0.737	P011	0.118	0.796			
P006	0.171	0.744	P012	0.092	0.757			
P007	0.186	0.822	P013	0.107	0.807	**Mean (SD)**	0.121 (0.04)	0.767 (0.06)

5 CONCLUSION

Models that predict a user's intent to interact from eye movements can be applied to drive a predictive XR interface. Our results suggested that our modeling framework can predict moments of interaction with a logistic regression model that is interpretable and practical for real-time deployment. Thus, predicting a user's intent to interact enables an adaptive XR interface that has potential to provide users with easy-to-use, minimally fatiguing XR interactions for all-day use.

ACKNOWLEDGMENTS

The authors would like to thank Vivien Francis for development of the data collection environment, along with Taylor Bunge, Chip Connor, Nour Shoora, and Joseph Montgomery for implementing the remote data collection infrastructure. The authors would also like to thank Mei Gao and Anna Yu for valuable feedback in the design of questionnaire prompts.

REFERENCES

Rawan Alghofaili, Yasuhito Sawahata, Haikun Huang, Hsueh-Cheng Wang, Takaaki Shiratori, and Lap-Fai Yu. 2019a. Lost in style: Gaze-driven adaptive aid for vr navigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

Rawan Alghofaili, Michael S Solah, Haikun Huang, Yasuhito Sawahata, Marc Pomplun, and Lap-Fai Yu. 2019b. Optimizing visual element placement via visual attention analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 464–473.

Dana H Ballard, Mary M Hayhoe, and Jeff B Pelz. 1995. Memory representations in natural tasks. *Journal of cognitive neuroscience* 7, 1 (1995), 66–80.

Roman Bednarik, Hana Vrzakova, and Michal Hradis. 2012. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the symposium on eye tracking research and applications*. 83–90.

Ali Borji and Laurent Itti. 2012. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 185–207.

Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

Noshaba Cheema, Laura A Frey-Law, Kourosh Naderi, Jaakko Lehtinen, Philipp Slusallek, and Perttu Hämäläinen. 2020. Predicting mid-air interaction movements and fatigue using deep reinforcement learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

Francisco M Costela and José J Castro-Torres. 2020. Risk prediction model using eye movements during simulated driving with logistic regressions and neural networks. *Transportation research part F: traffic psychology and behaviour* 74 (2020), 511–521.

Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.

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 (2013), 5–5.

Stefan Dowiasch, Svenja Marx, Wolfgang Einhäuser, and Frank Bremmer. 2015. Effects of aging on eye movements in the real world. *Frontiers in human neuroscience* 9 (2015), 46.

Simone Frintrop, Erich Rome, and Henrik I Christensen. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)* 7, 1 (2010), 1–39.

Christoph Gebhardt, Brian Hecox, Bas van Opheusden, Daniel Wigdor, James Hillis, Otmar Hilliges, and Hrvoje Benko. 2019. Learning Cooperative Personalized Policies from Gaze Data. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 197–208.

Anjith George and Aurobinda Routray. 2016. A score level fusion method for eye movement biometrics. *Pattern Recognition Letters* 82 (2016), 207–215.

Matthew Gingerich and Cristina Conati. 2015. Constructing models of user and task characteristics from eye gaze data for user-adaptive information highlighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9, 4 (2005), 188–194.

Mary Hayhoe and Dana Ballard. 2014. Modeling task control of eye movements. *Current Biology* 24, 13 (2014), R622–R628.

Mary M Hayhoe and Jonathan Samir Matthis. 2018. Control of gaze in natural environments: effects of rewards and costs, uncertainty and memory in target selection. *Interface focus* 8, 4 (2018), 20180009.

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 (2013), e64937.

Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1063–1072.

Andrew Hollingworth and Steven J Luck. 2009. The role of visual working memory in establishing object correspondence across saccades. *Journal of Vision* 9, 8 (2009), 414–414.

Andrew Hollingworth, Michi Matsukura, and Steven J Luck. 2013. Visual working memory modulates low-level saccade target selection: Evidence from rapidly generated saccades in the global effect paradigm. *Journal of Vision* 13, 13 (2013), 4–4.

Tanya R. Jonker, Ruta Desai, Kevin Carlberg, James Hillis, Sean Keller, and Hrvoje Benko. 2020. The Role of AI in Mixed and Augmented Reality Interactions. In *CHI2020 ai4hci Workshop Proceedings*. ACM.

Hyo Jeong Kang, Jung-hye Shin, and Kevin Ponto. 2020. A comparative analysis of 3d user interaction: how to move virtual objects in mixed reality. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 275–284.

Çağla Çığ Karaman and Tevfik Metin Sezgin. 2018. Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty. *International Journal of Human-Computer Studies* 111 (2018), 78–91.

Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 3, 3 (1993), 203–220.

Jonghyun Kim, Youngmo Jeong, Michael Stengel, Kaan Akşit, Rachel Albert, Ben Boudaoud, Trey Greer, Joohwan Kim, Ward Lopes, Zander Majercik, et al. 2019. Foveated AR: dynamically-foveated augmented reality display. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–15.

Krzysztof Krejtz, Arzu Çöltekin, Andrew T Duchowski, and Anna Niedzielska. 2017. Using coefficient K to distinguish ambient/focal visual attention during map viewing. *Journal of eye movement research* 10, 2 (2017), 1–13.

Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning ambient/focal attention with coefficient K. *ACM Transactions on Applied Perception (TAP)* 13, 3 (2016), 1–20.

Michael Land and Benjamin Tatler. 2009. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press.

Michael F Land and Mary Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision research* 41, 25-26 (2001), 3559–3565.

Gabor Lengyal, Kevin Carlberg, Majed Samad, and Tanya R. Jonker. 2021. Predicting visual attention using the hidden structure in eye-gaze dynamics. In *CHI2021 Eye Movements as an Interface to Cognitive State (EMICS) Workshop Proceedings*. ACM.

Zhen Li, Michelle Annett, Ken Hinckley, and Daniel Wigdor. 2019. Smac: A simplified model of attention and capture in multi-device desk-centric environments. *Proceedings of the ACM on Human-Computer Interaction* 3, EICS (2019), 1–47.

Matteo Lisi, Joshua A Solomon, and Michael J Morgan. 2019. Gain control of saccadic eye movements is probabilistic. *Proceedings of the National Academy of Sciences*

116, 32 (2019), 16137–16142.

Jacob L Orquin and Simone Mueller Loose. 2013. Attention and choice: A review on eye movements in decision making. *Acta psychologica* 144, 1 (2013), 190–206.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 179.

Candace E. Peacock, Brendan David-John, Ting Zhang, T. Scott Murdison, Matthew J. Boring, Hrvoje Benko, and Tanya R. Jonker. 2021. Gaze Signatures Decode the Onset of Working Memory Encoding. In *CHI2021 Eye Movements as an Interface to Cognitive State (EMICS) Workshop Proceedings*. ACM.

Jami Pekkanen and Otto Lappi. 2017. A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific reports* 7, 1 (2017), 1–13.

Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics* (2021).

Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

Qi Sun, Anjul Patney, Li-Yi Wei, Omer Shapira, Jingwan Lu, Paul Asente, Suwen Zhu, Morgan McGuire, David Luebke, and Arie Kaufman. 2018. Towards virtual reality infinite walking: dynamic saccadic redirection. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and recall for time series. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1924–1934.

Benjamin W Tatler and Benjamin T Vincent. 2008. Systematic tendencies in scene viewing. *Journal of Eye Movement Research* 2, 2 (2008).

Jeremy M Wolfe. 2000. Visual attention. *Seeing* (2000), 335–386.