ARL-TR-9821 ● Oct 2023

Estimating User Gaze Depth Perception in Real-Time for Extended Reality Environments

by Nathan Villavicencio and Russell Cohen Hoffing

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9821 ● OCT 2023

Estimating User Gaze Depth Perception in Real-Time for Extended Reality Environments

Nathan Villavicencio
California State University, Fullerton

Russell Cohen Hoffing
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
October 2023	Technical Report	06/05/2023	08/11/2023

4. TITLE AND SUBTITLE
Estimating User Gaze Depth Perception in Real-Time for Extended Reality Environments

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Nathan Villavicencio and Russell Cohen Hoffing

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FD Adelphi, MD 20783	ARL-TR-9821

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID ID: Russell Cohen Hoffing, 0000-0002-3478-8169

14. ABSTRACT
The progression of head-mounted displays (HMDs) to incorporate eye tracking in extended reality settings enables opportunities to develop adaptive HMDs capable of responding to user context and state to provide relevant information faster, which can result in faster training and better decision-making. The purpose of this research is to develop a method that allows for real-time estimation of perceptual depth in extended reality settings using eye-tracking data. In three different experimental conditions (virtual reality, augmented reality, and real), 13 human subjects fixated on targets at four varying depths while eye-tracking data was collected. Here, fixation periods are isolated and segmented in the eye-tracking data and the average inverse eye vergence angle (EVA) and interpupillary distance (IPD) are calculated. Using these two variables as features and the depth as the target, support vector machine (SVM), random forest, and XGBoost models exhibited mean classification accuracy of 50.1%, 48.9%, and 49.0%, respectively. Near/far classification for distances of 0.25 m versus 4.0 m yielded classification accuracies of 61.0% , 85.4%, and 85.6% for SVM, random forest, and XGBoost, respectively. We conclude that user perceptual depth can be estimated using EVA and IPD and adaptive HMDs are feasible using similar models for gaze depth estimation.

15. SUBJECT TERMS
Humans in Complex Systems, virtual reality, augmented reality, head-mounted display, HMD, depth perception

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	16

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Russell Cohen Hoffing	(310) 448-0374

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Head-mounted displays (HMDs) already allow for physical interaction between a user and an extended reality environment via rotational tracking of the head and positional tracking of the body. In more recent HMDs, eye-tracking cameras have been incorporated allowing for more precise physical and even cognitive interaction. In previous research, HMDs with eye-tracking cameras have been shown to be able to estimate the direction and focal depth of a user's gaze. However, little research demonstrates the ability to estimate a user's focal depth in real-time for these extended reality environments. In this research, we propose a method of estimating a user's focal depth in an extended reality environment using current commercial eye-tracking technology.

2. Methodology

The following section outlines the methodology for this research. We begin by outlining the experimental methodology from which our data is sourced. We then discuss the processing of the data, calculation of our features, and machine learning models employed.

2.1 Experimental Methodology

The following experimental methodology was previously performed by members of our research group. While experimental methodology pertinent to this research will be discussed in this section, full details can be found in the original paper (Arefin et al. 2022).

In this experiment, 16 subjects focused on targets at different distances for varying periods of time in real, augmented reality (AR), and virtual reality (VR) environments. As seen in Fig. 1, targets are placed at distances of 0.25, 0.75, 1.5, and 4.0 m. Extended reality environments are presented using the Microsoft Hololens2 and eye tracking is done using Pupil Labs Pupil Core.

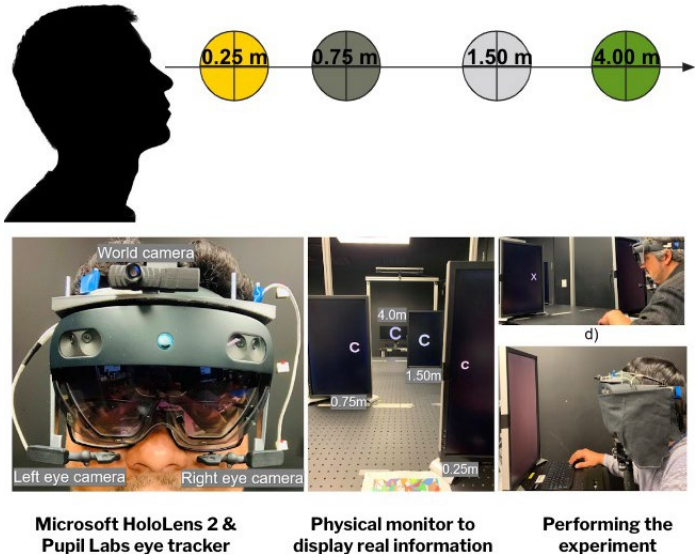

Fig. 1 Top: Diagram outlining the theoretical setup where each subject must focus on a target set at a distance of 0.25, 0.75, 1.5, or 4.0 m for varying periods of time. Bottom: Photos of experimental setup. The left photo shows how the Microsoft Hololens2 and Pupil Labs eye tracker are combined for this experiment. The center photo shows the real experimental setup of targets. The right photos show two different subjects performing the experiment.

2.2 Data Processing

The first processing step was to remove NaNs from the data. Second, samples of data for which the Pupil Labs' pupil confidence score was below 0.75 were discarded. Third, 450 samples following each target switch were discarded to allow time for the subject to switch targets. Lastly, in order to provide stable periods of fixation for which a machine learning model could be trained, the data is segmented into windows of 50 samples and each window is checked for saccades or significant noise. In order to do this, the average x- and y-direction vectors for each window are calculated. If any of the 50 samples in the window deviate from either average direction vector past a threshold of 0.05, the window is rejected. These variables were chosen since they are generally stable except during periods of saccade or noise. This model is solely focused on classifying focal distances during periods of fixation. For this reason, we attempt to omit all periods of saccade and noise from the training data without any attempt to classify one from the other. Using this method an average of 470 (σ = 166) windows were collected for each subject. An example of the extracted windows versus raw data can be seen in Fig. 2.

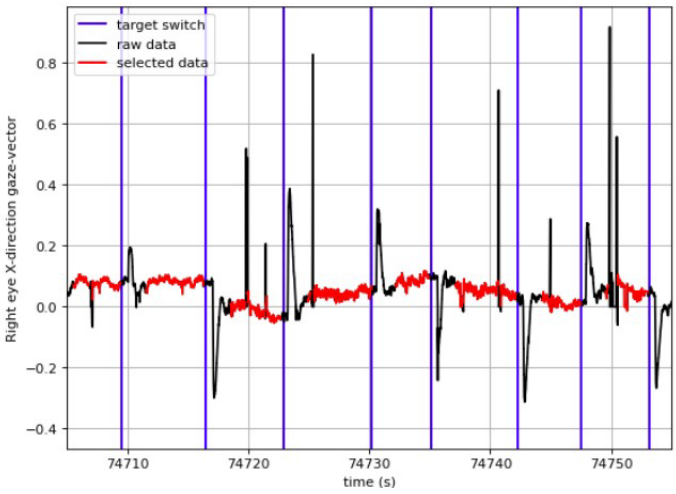

Fig. 2 This graph shows the x-direction gaze vector for selected data (red) vs. the raw data (black). The vertical blue lines represent the time in which the target was officially changed. A grace period where data is not collected following each target switch is imposed to allow the user time to focus on the new target.

For real-time implementation, the third step of removing samples following a target change cannot be implemented since that timing cannot be known in a typical extended reality user experience. However, these target changes are accompanied by notable changes in the x- and/or y-direction vector that would be identified in the final step of processing. The reason for the implementation of the third step is to prevent creating data where the subject is still fixated on the old target but the classification is now for the new target.

2.3 Eye Vergence Angle and Interpupillary Distance Calculation

In order to estimate the focal depth of a user during a given time, we calculate and utilize the eye vergence angle (EVA) and the interpupillary distance (IPD). The EVA represents the angle between the left-eye gaze vector and the right-eye gaze vector while the IPD represents the distance between pupil centers. These two features were chosen because they reflect changes in the ocular motor system that occur when a user switches focal distance.

Using the left-eye gaze vector and right-eye gaze vector, we calculate the EVA using the cosine similarity formula.

$$cos(\theta) = \frac{\vec{L} \cdot \vec{R}}{\|\vec{L}\|\|\vec{R}\|} \tag{1}$$

Pupil Labs provides normalized vectors for the left and right eyes from which the EVA is derived. Pupil Labs also provides the coordinates of the pupil centers for the left and right eyes from which the IPD was calculated.

2.4 Machine Learning Models

Due to the nature of our data, we looked to implement algorithms capable of supervised classification. For each 50-sample window, the average EVA and IPD are calculated as features. The classification task then becomes classifying the fixation distance for the entire 50-sample window given the average EVA and IPD. We ultimately decided to implement support vector machine (SVM), random forest classifier, and an XGBoost classifier. All hyperparameters were kept at default values with the exception of number of trees in the random forest. This value was increased to 200 after which no improvement in performance is seen. A train/test split of 80/20 for each subject was used to assess the performance of each model. In order to simulate real conditions, the training and testing set was split chronologically (rather than being shuffled) to prevent windows from the same gaze period being in both the training and testing set. To assess classification accuracy, the machine learning model uses the average EVA (radians) and IPD of an accepted window to predict target distance. Because the grace period after each target transition is presented, we know each window only has one ground truth distance. Here, accuracy is reflective of predictions across all windows from the testing set for each participant.

3. Results

Figure 3 shows an example distribution of inverse EVA versus IPD for a single subject where the hue of each point represents the ground truth distance. While there is a lot of overlap, an overall trend can be seen where the inverse EVA (y-axis) increases as the target distance increases. As is the case with most subjects, the IPD (x-axis) does not appear to provide as much insight for classification. The IPD appears to gradually increase as distance increases. However, the overlap between each target prevents IPD from being a useful predictor in this distribution.

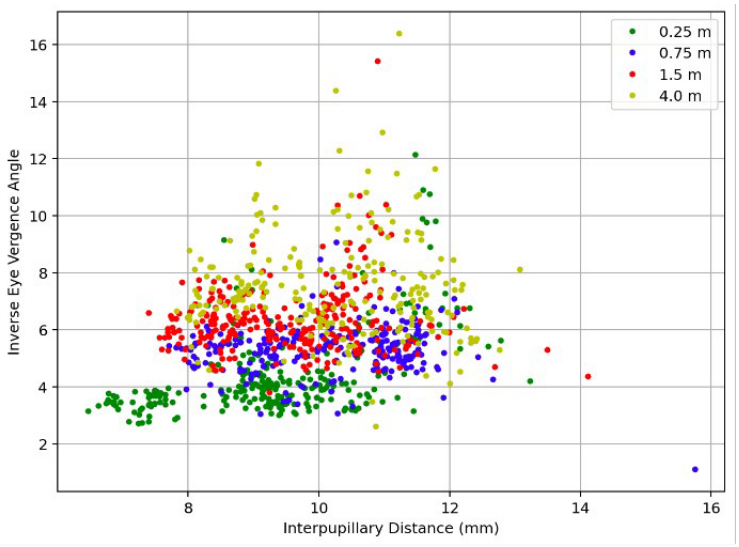

Fig. 3 The plot shows an example of the distribution of inverse EVA vs. IPD for a single subject. Each point represents the average IPD and inverse EVA for a given window where the color represents the ground truth location of the target during this window.

The performance of each machine learning model can be seen in Table 1. SVM, random forest, and XGBoost had classification accuracies of 50.1% ± 13.7, 48.9% ± 19.3, and 49.0% ± 13.6, respectively. All three models have similar mean accuracies. However, random forest has a larger standard deviation, meaning its performance may be more sensitive to the specific subject.

Table 1 Average classification accuracy and standard deviation of each machine learning classification model. All models were trained and tested on each subject individually.

Machine learning model	Mean classification accuracy
SVM	50.1 (σ = 13.7)
random forest	48.9 (σ = 19.3)
XGBoost	49.0 (σ = 13.6)

In addition to the four targets, a near/far classification with targets only at 0.25 and 4.0 m was attempted using the same machine learning model parameters and data (omitting the other two distances). From Fig. 4, it can be seen that all models significantly improve at 50 samples per window. However, the SVM model showed the least improvement with a mean accuracy of 61.0% for a window size of 50 samples. Random forest and XGBoost have similar performances with 50-sample window mean accuracies of 85.4% and 85.6%, respectively. The effect of window size was investigated here as well, showing small improvements as window size is increased past 50 samples per window.

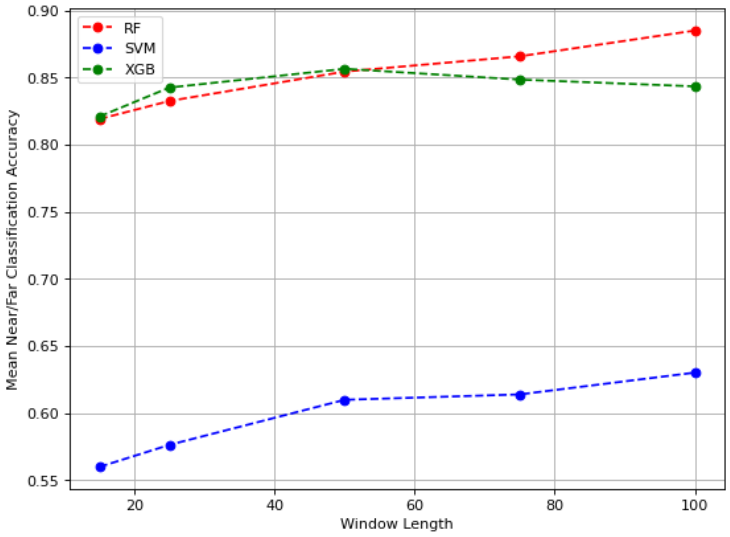

Fig. 4 Plot of window length vs. mean near/far classification accuracy

4. Discussion

The main takeaway from these results is that using IPD and inverse EVA as predictors allows for a significant increase in classification performance when compared to chance (25%). Overall, the performance of each machine learning model yielded comparable results. While SVM did slightly outperform random forest and XGBoost in this data set, more subjects would be needed to know for certain which algorithm works best overall. Random forest has a higher classification standard deviation than SVM and XGBoost, suggesting that it can perform extremely well for some subjects and poorly for others.

We consider the hyperparameters and window sizes to be points of consideration for use in future models. Besides random forest where the number of trees were set intentionally (Section 2, Methodology), the other hyperparameters in these models were set to default. Tuning these hyperparameters provides an avenue for increasing the mean classification accuracy in the future. Window sizes were kept to 50 samples (192 ms) to better reflect the duration of fixations and saccades for a user in an unrestricted environment. Increasing the window sizes does result in a small increase in accuracy for this data set, but may be too large for the natural fixation lengths of free gaze and ultimately result in worse classification/response.

The distributions of both EVAs and IPDs were found to vary notably from subject to subject, preventing the creation of a single model based on multiple subjects. Physical differences such as distance between the eyes appear to be correlated, but other factors may also exist.

Future work should involve the development of an application that can implement this predictive algorithm to yoke the display of a virtual object to the user's estimated focal depth.

5. Conclusion

The goal of this project was to demonstrate the feasibility of adaptive HMDs by creating machine learning models capable of classifying focal distance using the eye-tracking data available from current commercial off-the-shelf HMDs. To train these models, a previous data set where subjects fixated on a target set at distances of 0.25, 0.75, 1.5, and 4.0 m in real, AR, and VR environments was used. Using SVM, random forest, and XGBoost classification accuracies of 50.1% ± 13.7, 48.9% ± 19.3, and 49.0% ± 13.6 were achieved, respectively. All three models perform similarly and were able to significantly increase the classification performance when compared to the chance classification rate of 25%. Near/far classification for distances of 0.25 m versus 4.0 m yielded classification accuracies of 61.0% , 85.4%, and 85.6% for SVM, random forest, and XGBoost, respectively. Random forest and XGBoost show superior performance compared to SVM for this near/far classification. From this we conclude that adaptive HMDs that respond to user depth of focus are feasible.

6. References

Arefin MS, Swan JE II, Cohen Hoffing R, Thurman SM. Tracking perceptual depth changes with eye vergence and inter pupillary distance in a virtual reality environment. J Vision. 2022;22(14):3838. doi: 10.1167/jov.22.14.3838.

List of Symbols, Abbreviations, and Acronyms

AR	augmented reality
ARL	Army Research Laboratory
DEVCOM	US Army Combat Capabilities Development Command
EVA	eye vergence angle
HMD	head-mounted display
IPD	interpupillary distance
NaN	Not a Number
SVM	support vector machine
VR	virtual reality

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB
1 (PDF)	DEVCOM ARL FCDD RLA FD R COHEN HOFFING