OpenNEEDS: A Dataset of Gaze, Head, Hand, and Scene Signals During Exploration in Open-Ended VR Environments

Kara J. Emery
University of Nevada, Reno
Reno, NV, USA
Facebook Reality Labs
Redmond, WA, USA
karaemery@nevada.unr.edu

Marina Zannoli
Facebook Reality Labs
Redmond, WA, USA
Facebook AI
New York City, NY, USA

Lei Xiao
Facebook Reality Labs
Redmond, WA, USA

James Warren
Facebook Reality Labs
Redmond, WA, USA

Sachin S. Talathi
Facebook Reality Labs
Redmond, WA, USA
stalathi@fb.com

ABSTRACT

We present OpenNEEDS, the first large-scale, high frame rate, comprehensive, and open-source dataset of Non-Eye (head, hand, and scene) and Eye (3D gaze vectors) data captured for 44 participants as they freely explored two virtual environments with many potential tasks (i.e., reading, drawing, shooting, object manipulation, etc.). With this dataset, we aim to enable research on the relationship between head, hand, scene, and gaze spatiotemporal statistics and its applications to gaze estimation. To demonstrate the power of OpenNEEDS, we show that gaze estimation models using individual non-eye sensors and an early fusion model combining all non-eye sensors outperform all baseline gaze estimation models considered, suggesting the possibility of considering non-eye sensors in the design of robust eye trackers. We anticipate that this dataset will support research progress in many areas and applications such as gaze estimation and prediction, sensor fusion, human-computer interaction, intent prediction, perceptuo-motor control, and machine learning.

CCS CONCEPTS

• **Human-centered computing** → *User centered design*; • **Computing methodologies** → *Model verification and validation.*

KEYWORDS

datasets, virtual reality, eye tracking, gaze estimation

ACM Reference Format:
Kara J. Emery, Marina Zannoli, Lei Xiao, James Warren, and Sachin S. Talathi. 2021. OpenNEEDS: A Dataset of Gaze, Head, Hand, and Scene Signals During Exploration in Open-Ended VR Environments. 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/10.1145/3448018.3457996

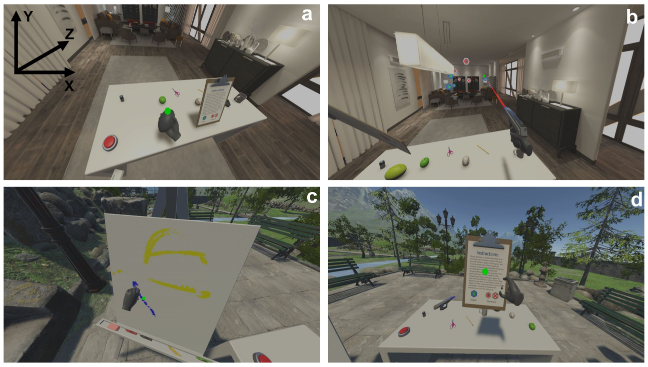

Figure 1: Examples of the indoor and outdoor scenes during open-ended VR gameplay. The 3D gaze vector has been shown in these images for visualization purposes only (green dot). These represent examples of potential tasks: a) object interaction; b) playing a shooting game; c) drawing; d) reading. The axis orientation of the dataset is also displayed in a.

1 INTRODUCTION

Understanding the relationship between head, hand, scene, and gaze information is relevant for many research areas and applications such as eye tracking, multimodal learning, and human-computer interaction. Previous works across a range of fields suggest a relationship between gaze information and hand spatiotemporal statistics [Land and Hayhoe 2001; Li et al. 2013; Ren and Crawford 2009], head motion [Einhäuser et al. 2007; Fang et al. 2015; Hu et al. 2019; Land and Hayhoe 2001; Li et al. 2013; Pelz et al. 2001; Yarbus 1967], and scene content [Cheng et al. 2015; Cornia et al. 2018; Hou et al. 2017; Huang et al. 2015; Itti et al. 1998; Judd et al. 2009; Li and Yu 2015; Rai et al. 2017; Sitzmann et al. 2018; Torralba et al. 2006]. Moreover, recent work has shown that various combinations of eye, head, scene, and hand signals can be leveraged for applications such as gaze estimation [Hu et al. 2019], prediction [Li et al. 2013],

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/10.1145/3448018.3457996

and classification [Kothari et al. 2020] as well as determining a person's focus of attention in a scene, e.g. [Cheng et al. 2015; Goferman et al. 2012; Itti et al. 1998; Jia and Han 2013; Judd et al. 2009; Kienzle et al. 2007; Koch and Ullman 1987; Liu et al. 2011; Torralba et al. 2006]. Though these findings and approaches together provide evidence for an association between gaze, head, hand, and scene information, there is no dataset nor model that captures the comprehensive relationship between these signals, especially during open-ended exploration (Table 1). Furthermore, the recent success of data-driven machine learning models in capturing the complex associations between multi-modal signals [Ramachandram and Taylor 2017] has driven an increasing demand for datasets of sufficient size and variety to capture and leverage these relationships for gaze estimation and the many other research areas they could support. To address this need, we release OpenNEEDS, a large-scale dataset of sequences of head, hand, scene, and gaze signals sampled at 90Hz as users explore open-ended virtual environments.

The widespread utility of eye tracking technology has created a growing demand for consistent and reliable eye-tracking systems, and there is still a need for new and creative approaches that can enhance the accuracy of eye-tracking data. To demonstrate the utility of OpenNEEDS, we quantify the extent to which these non-eye signals (head, hand, scene) can be leveraged for producing a spatially accurate estimate of gaze. Specifically, we use a machine learning framework to show that gaze estimation models trained on data from each non-eye sensor indivdiually and their complete combination outperform all gaze estimation baselines considered. Overall, we believe that OpenNEEDS has the potential to support a wide range of applications beyond gaze estimation and prediction, such as sensor fusion, intent inference, visual saliency, and human-computer interaction. In addition, we believe that this dataset can inform neuro-, cognitive, and perception science approaches to understanding perceptuo-motor coordination.

2 RECORDING HEAD, HAND, SCENE, AND GAZE DATA

2.1 Data capture

Participants: The dataset was captured from 44 voluntary participants (age (years) = 31.7 (SD 10.5); 20 females; 40 right-handed; ipd (mm) = 62.95 (SD 3.62)). Before taking part in data collection, all participants provided written informed consent for using their data for research and commercial purposes. All participants had normal or corrected-to-normal visual acuity in both eyes. The dataset was anonymized to remove any personally identifiable information.

Apparatus: Participants were fitted with a custom-made prototype virtual reality head-mounted display (VR-HMD) and used Oculus Rift positional sensors and controllers for interacting with the virtual environment. The field-of-view of the VR-HMD was 104°. The HMD was equipped with a custom-made eye-tracker mounted with two synchronized eye-facing infrared cameras. The eye tracker had a median gaze error (p50) of 1.3° following successful calibration. We recorded the head and hand pose at 90 Hz using Oculus Rift's intertial measurement unit (IMU) and Touch controllers respectively. Virtual environments were designed and displayed using the Unity game engine. To produce real-time recordings of the scenes, we captured the pose and position of all interactive objects in the scene and replayed each participants playtime offline to create full-resolution (2560 x 1440 pixels) scene RGBD images and motion vector maps.

Stimuli: Given that gaze distributions have been shown to differ based on the type of scene (e.g., indoor vs. outdoor) and task (e.g., making a sandwich vs. ordering coffee) with which the observer is engaged [Sprague et al. 2015], our scenes were designed to elicit a range of tasks and therefore capture a wide variety of gaze behaviors. We created one indoor and one outdoor scene each with the same interactive content (Figure 1): a table of graspable objects (including a clipboard of instructions and a gun which initiates a shooting game) and a canvas with crayons and an eraser. Together these objects provided the opportunity for many behaviors such as reading, throwing, object-manipulation, drawing, aiming, and shooting, and the indoor and outdoors scenes are primed to capture different gaze distributions [Sprague et al. 2015]. Therefore, we anticipate that this will offer support for more robust and generalizable algorithms utilizing the association between head, hand, scene, and gaze across a variety of contexts.

Procedure: Participants stood in a large room and could move freely within the space. They were asked to freely explore the environments (indoor and outdoor) for up to five minutes each with no further instructions, and were free to quit participation when they wished. A clipboard of written instructions was provided in each environment to encourage their awareness of the types of opportunities for interaction (Figure 1), though no specific tasks were explicitly requested, required, nor tracked throughout the experiment. Eye-tracking calibration was performed at the start of the experiment and between scenes using a custom-built calibration procedure. Each calibration was performed until successful (i.e. p50 of 1.3°), and thus all participants included in the reported dataset met this requirement. Of the 44 total participants, 41 and 43 successfully completed the indoor and outdoor scenes, respectively. The three participants and one participant excluded from the indoor and outdoor scene respectively were left out due to unsuccessful calibration. There were no auditory stimuli. During playtime, we recorded the positions of the interactive objects, the controllers, head pose information, and 3D gaze vectors.

3 DATA PROCESSING

All recorded signals were referenced to the cyclopean eye (i.e. center of the head between the eyes) making the OpenNEEDS dataset egocentric. The axis orientation is represented in Figure 1a. Each sensor signal was measured at 90Hz for a total of 2,194,865 samples of data. In this section, we summarize the format in which each signal is stored. We also characterize the gaze distributions and fixation biases between scene types and users.

3.1 Sensors

Head: Pose: The head orientation was recorded as a unit quaternion at each frame (specifically X, Y, Z, and W coordinates). The use of quaternions offers singularity-free rotation with few parameters while maintaining correct algebraic operations [LaValle et al. 2014]. *Motion:* We encoded pixel motion in screen space within a range of -1 to 1, representing the offset from the last frame to the

Table 1: Comparing other publicly available datasets in the field of non-eye sensors and eye tracking to OpenNEEDS. No other dataset comprehensively and explicitly captures head, hand, and scene non-eye sensors along with 3D gaze vectors in VR scenes that offer object interaction. However, the existence of other related datasets points to the rich areas of research that could be informed by OpenNEEDS. (Resolution in the table refers that of the scene images).

	Sensors				Other Characteristics				
Dataset	head	hand	scene	gaze	samples	subjs	resolution	FR	task
OpenNEEDS	✓	✓(6-dof)	✓	✓	2,086,507	44	128 x 71	90Hz	VR free task
GTEA Gaze+ [Li et al. 2015]			✓	✓	~700,000	26	1280 x 960	24Hz	meal prep
EGTEA Gaze+ [Li et al. 2018]		✓(masks)	✓	✓	~2,500,000	32	1280 x 960	24Hz	meal prep
Sitzmann et al., 2018	✓		✓	✓	1,980	169	1920 x 1080	120Hz	VR fixed view
SGaze [Hu et al. 2019]	✓		✓	✓	18,000	60	28 x 28	100Hz	VR static scene
GW [Kothari et al. 2020]	✓		✓	✓	~5,800,000	19	1920 x 1080	300Hz	four tasks

current frame. These motion values were stored as an image (128 x 71 pixels) at each frame in the red and green channels in a linear color space at an 8-bit resolution.

Hand: Left- and right-hand controller data were captured for each participant. The hand orientation was recorded as a unit quaternion at each frame (X, Y, Z, and W coordinates). The 3D position of each hand relative to the center of the head was recorded in meters (X, Y, and Z coordinates).

Scene: Color images: The on-screen image presented to the user at each frame was stored as a 3-channel (RGB) image at an 8-bit resolution in sRGB color space. In order to limit data storage to a reasonable size for use, the color images were down-sampled to a pixel resolution of 128 x 71. We chose color over grayscale images given their potential utility for future use cases. *Depth images:* The depth maps were recorded at on-screen resolution and the depth values were the distance in meters from the cyclopean eye along the viewing axis, calculated using the LinearEyeDepth shader function in Unity. For reasonable storage, the depth maps were also down-sampled to a pixel resolution of 128 x 71. *Objects:* The position and quaternion of each interactive object in the scene was recorded for replaying the gameplay offline to store the color, depth, and motion images. The 3D position in meters and unit quaternion over time are included for all 28 of the interactive objects.

Gaze: Ground truth 3D gaze vectors in meters are provided for each frame. Gaze X and Y were limited by the FOV (104°), and gaze Z (depth) limits were [0.3,5] meters. We measured gaze vectors and cornea centers of each participant using a user-calibrated glint-based model [Guestrin and Eizenman 2006].

3.2 Fixation biases and inter-subject agreement

To explore whether OpenNEEDS captures typical gaze behavior, and therefore assess its robustness for gaze estimation and gaze prediction models in particular, we analyzed the gaze distributions for each scene type and specifically determined whether previously reported fixation biases exist in our dataset [Hu et al. 2019; Judd et al. 2009; Nuthmann and Henderson 2010; Sitzmann et al. 2018]. Following the procedure in [Nuthmann and Henderson 2010], we show that our dataset exhibits a center bias with a slight downward vertical shift and a broader distribution along the vertical dimension (Figure 2a). We found that the screen center and mean position of all fixations [Cornia et al. 2018; Hu et al. 2019; Li et al. 2013; Nakashima et al. 2015] were appropriate baseline metrics for our dataset given that approximately 20% of all fixations lie within 5° eccentricity of the screen center and 50% of all fixations lie within 10° of the fixation mean respectively (Figure 2b).

Quantifying the agreement in gaze behavior across users is crucial to understand the robustness of our data and thus its utility for applications with success contingent on user generalizability. We assessed inter-observer congruency by means of a receiver operating characteristic (ROC) curve [Le Meur and Baccino 2013; Torralba et al. 2006], following a previously reported procedure [Sitzmann et al. 2018]. For each user, we determined the extent to which their gaze behavior agrees with the average gaze behavior of all other users. A fast convergence of the ROC curve to its maximum value of 1 (i.e. a high AUC) is indicative of high agreement between the gaze behavior of the i^{th} user and all other users. Figure 3 shows an ROC curve for each of the 43 and 41 users that participated in the outdoor and indoor scenes, respectively, and indicates the high average AUC for both scenes, suggesting a strong similarity in gaze behaviors across users.

4 GAZE ESTIMATION

The widespread utility of eye-tracking technology has created a growing demand for consistent and reliable eye-tracking systems. Though there have been significant improvements in eye tracking, there is still a need for new and creative approaches that can enhance the accuracy of eye-tracking data, specifically for performance in noisy environments (e.g., strong environmental lighting, facial occlusions). Previous studies suggest that non-eye signals are useful for gaze estimation by incorporating head pose/motion into visual saliency models to boost their performance [Nakashima et al. 2015; Sitzmann et al. 2018], using a combination of head motion, hand position, and gaze temporal dynamics for gaze prediction in egocentric video [Li et al. 2013], and combining head motion and a coarse visual saliency map for gaze estimation in VR [Hu et al. 2019]. Though these previous approaches provide support for the idea that head, hand, and scene signals are useful for estimating gaze, they have not yet addressed how these signals individually and in complete combination contribute to gaze estimation. Furthermore, the datasets on which these models were built capture gaze behaviors in constrained situations, such as only one type of task [Li et al. 2013] or in VR environments without the potential for object interaction [Hu et al. 2019]. Thus, using OpenNEEDS, we can better characterize the contribution of non-eye sensors to gaze estimation in more naturalistic conditions. As such, in this section we report on a baseline experiment using a machine learning framework to quantify the contribution to gaze estimation of

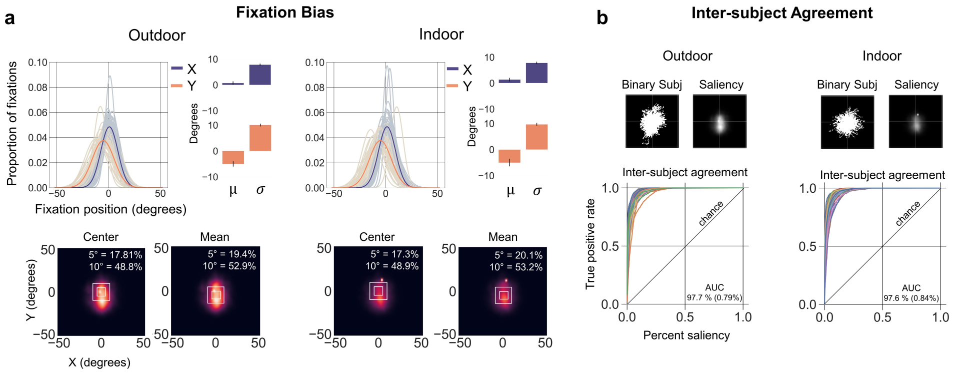

Figure 2: Fixation biases and inter-subject agreement in gaze behavior in OpenNEEDS. a) We found that both center and mean biases are present in OpenNEEDS, as windows centered around these 2D points account for a large percentage of all fixations. b ROC curves are calculated by assessing the quantity of fixations for a given subject (e.g. Binar Subj) that are accounted by for the *n*% most salient reasons of the average *ground truth saliency map* (e.g. Saliency). A fast convergence to 1 (i.e. high AUC) is indicative of high agreement between a given subject and all others. The ROC curve for each subject, and the average and standard deviation of the AUC of the ROC curves across users are shown in the lower right-hand quadrant.

each individual non-eye sensor (head, hand, and scene) and the full combination thereof via an early sensor fusion approach [Poria et al. 2017].

4.1 Feature engineering

To prepare OpenNEEDS for training gaze estimation models, we discarded all frames with missing gaze vectors and outlying ground truth gaze vectors (i.e. with greater than 99 percentile scores), reducing the dataset to 1,933,551 frames. To control for bias to users or scenes, we created a data subset that required an equal number of frames from each user and each scene type. This resulted in our final dataset of 31 different users each with 2.5 minutes of game play from each scene for a total of 837,000 frames. Given that the goal of this paper is to release OpenNEEDS, we only intend to report a baseline experiment that demonstrates this dataset's utility, particularly in the context of gaze estimation models. Thus, we did not perform an exhaustive analysis of the optimal gaze estimation model that could be designed using OpenNEEDS, nor did we train our model on the comprehensive set of signals available, but rather we report a baseline model trained on a representative subset of signals including one from each category: head, hand, and scene. The signal subset used as the dataset for our experiment was as follows:

Head: We included head orientation as input to our gaze estimation model. The head orientation was represented by the originally recorded unit quaternion at each frame.

Hand: The left and right hand orientations and positions were included as input to our gaze estimation model. Both hand orientations were represented by their originally recorded unit quaternions at each frame. Both hand positions were represented by their stored 3D coordinates in meters relative to the cyclopean eye.

Scene: Rather than including the raw RGB images presented on screen at each frame, we included their visual saliency maps (i.e. a value within the range (0,1) that represents the expected density of eye fixations for each pixel given the content of the scene). To create these saliency maps, we processed each of the original images (down-sampled to 64x36 pixels) using a pre-trained, state-of-the-art SAM-ResNet saliency predictor [Cornia et al. 2018]. We further down-sampled each saliency map to 32x18 pixels for training.

Annotations: We transformed the original ground truth 3D gaze vectors to 2D gaze angles, (θ, ϕ), for training and testing.

4.2 Sensor Models

We characterize the task of gaze estimation as a supervised regression problem. We used gradient boosting regression trees (GBRT), which are a powerful algorithm for supervised regression problems that produce an estimate by combining predictions across many learners (trees) to create a powerful "committee" of weighted votes. *Gradient boosting* is a particular type of boosting strategy that iteratively includes the trees with predictions closest to the maximal descent direction (the negative gradient) to help prevent overfitting to the training data [Hastie et al. 2009]. Thus, a GBRT is an additive model of the following form: $F_m(x) = F_{m-1}(x) + h_m(x)$, where F is the GBRT model, and h_m are the basis functions modeled as small regression trees of fixed size. With each boosting iteration, a new tree is added to the GBRT model, F, and the weights at each iteration are computed by the following equation: $w_m = argmin_w \frac{1}{N}\Sigma_i^N L(y_i, F(\mathbf{X}_i, w))$, where L is the squared error loss function in our case. We trained this model to estimate the 2D gaze angle, $\mathbf{Y} = (\theta, \phi)$, as a function of the input features, $\mathbf{X}$. For a more complete description of GBRTs, see [Chen and Guestrin 2016; Friedman 2001; Hastie et al. 2009; James et al. 2013].

Comparison of the performance of each of the GBRT models to the baselines was accomplished using k-fold (here 5-fold) cross validation, resulting in a 80% (training data subset)/20% (test data subsets) split. We instituted three different methods for creating the data subsets for this procedure: In the *Random* method, original data were shuffled into five equally-sized subsets. In the *Subject-stratified* method, each of the five subsets was comprised of an equal representation of each subject. In the *Subject-independent* method, each of the five subsets included data from a unique set of users. We trained a separate GBRT model for each non-eye sensor i.e., head, hand, scene, and one model combining all these sensors with equal weighting. A separate GBRT model was trained for each of the five-fold cross-validation methods, i.e., random, subject-stratified, subject-independent, for a total of 12 GBRT models. For each model, we applied an optimization procedure to decide values for the following GBRT-hyperparameters: learning rate, number of trees, tree depth, child weight, minimum loss reduction required for partitioning γ, and the L1 regularization term on the weights α. The optimal GBRT-hyperparameters chosen for each model are shown in Table S2.

Evaluation Metric: We report the gaze estimation results in terms of the spatial accuracy of the gaze estimate for each model. We defined spatial accuracy as the angular error in degrees between the estimated and ground truth 2D gaze angles.

Individual Sensor Models: We trained an individual GBRT model for each non-eye sensor (head, hand, and scene). To quantify the independent contribution of each sensor to the gaze estimate, we performed five-fold cross-validation on each non-eye sensor model using each of the three different data grouping methodologies described above. We compared the performance of the individual sensor models to the baseline models to determine whether each sensor is useful for gaze estimation. Across all cross-validation methods, each individual sensor model (i.e., head, hand, and scene) had a lower average angular error than each of the three baseline models (i.e., random, center, and mean) confirming that these non-eye sensor can be leveraged for gaze estimation.

Early Sensor Fusion Model: For a baseline assessment of the combined utility of the measured non-eye sensors for gaze estimation, we trained an early fusion model by concatenating the multimodal features (head, hand, and scene) into a single vector of input [Poria et al. 2017]. Here, we did not apply dimensionality reduction techniques to the features, nor did we explore more complex sensor fusion approaches (i.e. intermediate and late [Atrey et al. 2010; Khaleghi et al. 2009; Ramachandram and Taylor 2017]). Our goal was to provide a baseline estimate of the combined utility of the OpenNEEDS non-eye sensors for gaze estimation. Further investigation into advanced sensor fusion methods (e.g. intermediate and late fusion) will allow us and others to fully exploit the complementary nature of the sensor modalities.

4.3 Baseline Models

The baseline models are defined based on fixation tendencies revealed in previous research [Judd et al. 2009; Nuthmann and Henderson 2010] and in our dataset (see Section 3.2). We cross-validated the performance of each baseline model for each of the three five-fold grouping methods, i.e. random, subject-stratified, and subject-independent.

Random: This model estimates the 2D gaze angles directly from a random normal distribution ($\mathcal{N}(0,1)$) truncated by the field-of-view of the HMD. Based on gaze distributions reported previously, e.g. [Judd et al. 2009; Li et al. 2013; Nuthmann and Henderson 2010], we found a standard normal distribution to be a reasonable approximation of expected fixation patterns.

Center: This study (Figure 2) and many others have reported a center bias for fixations [Clarke and Tatler 2014; Judd et al. 2009; Nuthmann and Henderson 2010], and this bias has been exploited to enhance the accuracy of visual saliency models, e.g. [Cornia et al. 2018; Kruthiventi et al. 2017; Kümmerer et al. 2015; Nakashima et al. 2015; Vig et al. 2014]. Our center baseline model estimates the 2D gaze angle as the center point (i.e. $Y(0,0)$) for each frame.

Mean: Given that the mean fixation point compared to the center point accounted for a slightly higher percentages of all total fixations in our dataset (see Figure 2a), we also used the mean fixation as a baseline for gaze estimation. The mean fixation was also used as a baseline comparison for a previous VR gaze estimation model [Hu et al. 2019]. Our mean baseline model estimates the 2D gaze angle as the mean gaze vector of the training set.

4.4 Results

Figure 3 and Table 2 show that the gaze estimation model for each individual non-eye sensor (head, hand, and scene) and their complete combination outperform each of the baseline models across all five-fold cross-validation methods. The combined early fusion model outperforms all other models when the training and test data are subject-independent, suggesting that this model is more generalizable to novel users and thus more relevant for typical gaze estimation applications. In general, the substantial improvement in accuracy across models when all users are represented in the training and test data (random and subject-stratified cross-validation methods) suggests user calibration, for individual differences in features such as head-eye latency, handedness, arm length, etc., could improve the performance of the subject-independent models, shifting their accuracy closer to that of the other methods. The subject-stratified hand GBRT model achieves the lowest error across all models, reflecting the strong relationship between hand and gaze position in OpenNEEDS (Figure S1 and Table S1). That the hand GBRT model outperforms the combined model in this condition suggests that our early fusion method might not be the optimal multimodal design to optimally leverage the non-eye sensor signals for gaze estimation. To begin to assess this, we tripled the number of trees for the GBRT early fusion model and found that the angular error improved to 2.75°, however, still worse than that of the hand GBRT model (2.46°). This finding suggests the need for designing more complex multimodal models to capture the full complementary power of the non-eye sensor signals for gaze estimation and how they can best be combined with eye data to augment traditional eye-tracking approaches. We intend and encourage others to improve upon our reported results and to define further research areas and applications for this dataset.

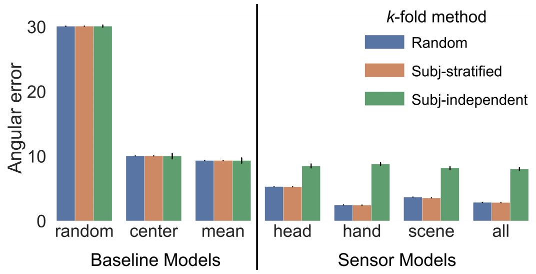

Figure 3: The spatial accuracy (i.e. angular error) in the gaze estimate for baseline and sensor GBRT models (error bars indicate 1 SEM). Each non-eye sensor alone and an early fusion model outperform baseline metrics.

Table 2: Baseline and sensor model performances for gaze estimation reported in terms of error between predicted and ground truth 2D gaze angles. The average and standard deviation (in parentheses) of the angular error (calculated by five-fold cross-validation as the mean and standard deviation of the prediction error of the test set for each fold) is provided for each model.

	Baseline Models			Sensor Models			
k-fold method	random	center	mean	head	hand	scene	all
Random	30.11 (0.025)	10.07 (0.004)	9.36 (0.006)	5.32 (0.007)	2.49 (0.013)	3.70 (0.012)	2.89 (0.014)
Subj-stratified	30.12 (0.022)	10.07 (0.015)	9.36 (0.008)	5.30 (0.019)	2.46 (0.016)	3.58 (0.010)	2.86 (0.007)
Subj-independent	30.12 (0.210)	10.04 (0.575)	9.35 (0.527)	8.51 (0.389)	8.81 (0.327)	8.21 (0.279)	8.05 (0.257)

5 CONCLUSION

We presented OpenNEEDS, a publicly available dataset capturing head, hand, scene, and gaze signals as participants freely explored interactive open-ended virtual environments. We demonstrated that the non-eye signals of OpenNEEDS are informative for estimating gaze, and we anticipate that this dataset will inform a variety of future research endeavors and applications such as eye tracking, sensor fusion, and human-computer, intent prediction, perceptuo-motor control, and machine learning. The dataset is available for download at https://www.dropbox.com/work/OpenNEEDS_2020 upon request (please email either of the corresponding authors with your name and organization for permission).

ACKNOWLEDGMENTS

The authors wish to thank Matthew Chapman for his significant help with data capture and Gusti Lulu Fatima for her substantial contribution to data collection. This work was supported by Facebook Reality Labs Research.

REFERENCES

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16, 6 (Nov 2010), 345–379. https://doi.org/10.1007/s00530-010-0182-0

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794. https://doi.org/10.1145/2939672.2939785

M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. Hu. 2015. Global Contrast Based Salient Region Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 569–582. https://doi.org/10.1109/TPAMI.2014.2345401

Alasdair D.F. Clarke and Benjamin W. Tatler. 2014. Deriving an appropriate baseline for describing fixation behaviour. *Vision Research* 102 (Sep 2014), 41–51. https://doi.org/10.1016/j.visres.2014.06.016

M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. 2018. Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model. *IEEE Transactions on Image Processing* 27, 10 (2018), 5142–5154. https://doi.org/10.1109/TIP.2018.2851672

Wolfgang Einhäuser, Frank Schumann, Stanislavs Bardins, Klaus Bartl, Guido Böning, Erich Schneider, and Peter König. 2007. Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems* 18, 3 (Jan 2007), 267–297. https://doi.org/10.1080/09548980701671094

Yu Fang, Ryoichi Nakashima, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Eye-Head Coordination for Visual Cognitive Processing. *PLOS ONE* 10, 3 (Mar 2015), e0121035. https://doi.org/10.1371/journal.pone.0121035

Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. http://www.jstor.org/stable/2699986

S. Goferman, L. Zelnik-Manor, and A. Tal. 2012. Context-Aware Saliency Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 10 (2012), 1915–1926. https://doi.org/10.1109/TPAMI.2011.272

E. D. Guestrin and M. Eizenman. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering* 53, 6 (2006), 1124–1133. https://doi.org/10.1109/TBME.2005.863952

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction* (2nd ed ed.). Springer.

Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. 2017. Deeply Supervised Salient Object Detection With Short Connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. 2019. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (2019), 2002–2010. https://doi.org/10.1109/TVCG.2019.2899187

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259. https://doi.org/10.1109/34.730558

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics, Vol. 103. Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Yangqing Jia and Mei Han. 2013. Category-Independent Object-Level Saliency Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

T. Judd, K. Ehinger, F. Durand, and A. Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

B. Khaleghi, S. N. Razavi, A. Khamis, F. O. Karray, and M. Kamel. 2009. Multisensor data fusion: Antecedents and directions. In *2009 3rd International Conference on Signals, Circuits and Systems (SCS)*. 1–6. https://doi.org/10.1109/ICSCS.2009.5412296

Wolf Kienzle, Felix A. Wichmann, Matthias Franz, and Bernhard Schölkopf. 2007. A Nonparametric Approach to Bottom-Up Visual Saliency. In *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press, 689–696. https://proceedings.neurips.cc/paper/2006/file/a2d10d355cdebc879e4fc6ecc6f63dd7-Paper.pdf

Christof Koch and Shimon Ullman. 1987. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*. Springer Netherlands, 115–141. https://doi.org/10.1007/978-94-009-3833-5_5

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B. Pelz, and Gabriel J. Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (Dec 2020), 2539. https://doi.org/10.1038/s41598-020-59251-5

S. S. S. Kruthiventi, K. Ayush, and R. V. Babu. 2017. DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. *IEEE Transactions on Image Processing* 26, 9 (2017), 4446–4456. https://doi.org/10.1109/TIP.2017.2710620

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. 2015. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. arXiv:1411.1045 [cs.CV]

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

S. M. LaValle, A. Yershova, M. Katsev, and M. Antonov. 2014. Head tracking for the Oculus Rift. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 187–194. https://doi.org/10.1109/ICRA.2014.6906608

Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods* 45, 1 (Mar 2013), 251–266. https://doi.org/10.3758/s13428-012-0226-9

Guanbin Li and Yizhou Yu. 2015. Visual Saliency Based on Multiscale Deep Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yin Li, Alireza Fathi, and James M. Rehg. 2013. Learning to Predict Gaze in Egocentric Video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Yin Li, Miao Liu, and James M. Rehg. 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)*.

Yin Li, Zhefan Ye, and James M. Rehg. 2015. Delving Into Egocentric Actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. 2011. Learning to Detect a Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 2 (2011), 353–367. https://doi.org/10.1109/TPAMI.2010.70

Ryoichi Nakashima, Yu Fang, Yasuhiro Hatori, Akinori Hiratani, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. 2015. Saliency-based gaze prediction based on head direction. *Vision Research* 117 (Dec 2015), 59–66. https://doi.org/10.1016/j.visres.2015.10.001

A. Nuthmann and J. M. Henderson. 2010. Object-based attentional selection in scene viewing. *Journal of Vision* 10, 8 (Jul 2010), 20–20. https://doi.org/10.1167/10.8.20

Jeff Pelz, Mary Hayhoe, and Russ Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139, 3 (Aug 2001), 266–277. https://doi.org/10.1007/s002210100745

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (Sep 2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A Dataset of Head and Eye Movements for 360 Degree Images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 205–210. https://doi.org/10.1145/3083187.3083218

D. Ramachandram and G. W. Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108. https://doi.org/10.1109/MSP.2017.2738401

L. Ren and J. D. Crawford. 2009. Coordinate transformations for hand-guided saccades. *Experimental Brain Research* 195, 3 (May 2009), 455–465. https://doi.org/10.1007/s00221-009-1811-8

V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. 2018. Saliency in VR: How Do People Explore Virtual Environments? *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1633–1642. https://doi.org/10.1109/TVCG.2018.2793599

William W. Sprague, Emily A. Cooper, Ivana Tošić, and Martin S. Banks. 2015. Stereopsis is adaptive for the natural environment. *Science Advances* 1, 4 (2015). https://doi.org/10.1126/sciadv.1400254 arXiv:https://advances.sciencemag.org/content/1/4/e1400254.full.pdf

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review* 113, 4 (Oct 2006), 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Eleonora Vig, Michael Dorr, and David Cox. 2014. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alfred L. Yarbus. 1967. *Eye Movements and Vision*. Springer US. https://doi.org/10.1007/978-1-4899-5379-7