Towards efficient calibration for webcam eye-tracking in online experiments

Shreshth Saxena
Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
shreshth.saxena@ae.mpg.de

Elke B. Lange
Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
elke.lange@gmail.com

Lauren K. Fink
Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany; Max Planck NYU Center for Language, Music, and Emotion, Frankfurt/M., Germany, New York, USA
lauren.fink@ae.mpg.de

ABSTRACT

Calibration is performed in eye-tracking studies to map raw model outputs to gaze-points on the screen and improve accuracy of gaze predictions. Calibration parameters, such as user-screen distance, camera intrinsic properties, and position of the screen with respect to the camera can be easily calculated in controlled offline setups, however, their estimation is non-trivial in unrestricted, online, experimental settings. Here, we propose the application of deep learning models for eye-tracking in online experiments, providing suitable strategies to estimate calibration parameters and perform personal gaze calibration. Focusing on fixation accuracy, we compare results with respect to calibration frequency, the time point of calibration during data collection (beginning, middle, end), and calibration procedure (fixation-point or smooth pursuit-based). Calibration using fixation and smooth pursuit tasks, pooled over three collection time-points, resulted in the best fixation accuracy. By combining device calibration, gaze calibration, and the best-performing deep-learning model, we achieve an accuracy of 2.58^0—a considerable improvement over reported accuracies in previous online eye-tracking studies.

CCS CONCEPTS

• **Applied computing**;

KEYWORDS

webcam eye-tracking, gaze calibration, online experiments, deep learning

ACM Reference Format:
Shreshth Saxena, Elke B. Lange, and Lauren K. Fink. 2022. Towards efficient calibration for webcam eye-tracking in online experiments. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22), June 08–11, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3517031.3529645

ETRA '22, June 08–11, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9252-5/22/06.
https://doi.org/10.1145/3517031.3529645

1 INTRODUCTION

Eye-tracking is a widely used experimental method across disciplines spanning psychology, neuroscience, industrial engineering, marketing/advertising, sport sciences, computer science, etc. [Duchowski, 2002]. Over the last few decades, the method has seen a significant surge in its application, along with huge advancements in the technology applied for tracking eye movements, from earlier galvanometric methods using electrodes placed around the eye [Mowrer et al., 1935], to modern computer vision methods that capture eye movements from eye/face images [Cheng et al., 2021]. A recent trend of webcam-based eye-tracking has contributed substantially to making eye-tracking more accessible and affordable.

The possibility of recording eye movements from any webcam enabled device allows researchers to sample global populations and replicate results cross-culturally, in an economical and time-efficient manner. Webcam-based gaze prediction methods have seen great improvements recently with the application of deep learning [Zhang et al., 2015; Krafka et al., 2016; Fischer et al., 2018; Park et al., 2019]. Webcam eye-tracking has also been applied as a real-time, web-browser-based solution for analyzing web browsing behavior [Papoutsaki et al., 2016; XLabsGaze, 2016], collecting large-scale visual saliency data [Xu et al., 2015], or running online studies [Semmelmann and Weigelt, 2018]. While the application of webcam eye-tracking is increasing at a rapid pace, a comprehensive review of methods, best practices, and data analyses from these low resolution eye-trackers is yet to be done. Such research is particularly relevant for the reliable and replicable application of webcam eye-tracking in online studies. Reviews and comparisons for lab-based eye-trackers [Carter and Luke, 2020; Ehinger et al., 2019] exist in abundance; however, they are not directly applicable for webcam-based methods due to differences in the gaze-tracking algorithms (model-based vs appearance based), sampling frequency (100–1000 Hz vs 15-30 Hz), and added noise in unrestricted webcam recordings. Moreover, rigorous comparisons of different calibration procedures and strategies are highly useful in designing eye-tracking studies and are not readily available for emerging low-resolution, webcam-based eye-tracking methods.

With a less restrictive set up, one critical component of webcam eye-tracking during online experiments is the calibration routine. Calibration is required to project model gaze predictions to 2D gaze points on the screen. This mapping relies on setup-based parameters, such as user-screen distance, camera intrinsic parameters and the monitor-camera pose relationship, which could be estimated

by intricate procedures like Rodrigues et al. [2010]'s mirror-based method for screen calibration and OpenCV's checkboard-pattern method for camera calibration [Bradski, 2000]. Such calculations are typically done by the experimenter while setting up an in-lab apparatus; however, they would require excessive customization of the procedure, and participant co-operation, to be implemented in online setups. We, therefore, suggest alternatives to estimate these parameters in online studies (see sections 2.5.1 and 2.5.2). In addition, we compare different gaze calibration strategies, based on calibration data size, collection time (beginning, middle or end of the study), and calibration task (fix-point and smooth-pursuit). Analysis of these strategies reveal interesting insights that one should consider when designing online eye tracking studies.

Our online experiment collected webcam recordings of participants performing a battery of standard eye-tracking tasks, coupled with multiple iterations of calibration trials. The recordings were later processed offline through the gaze estimation models to estimate gaze predictions. This approach allows a modular and flexible comparison of multiple models and calibration strategies on the same input (sequential video frames). Moreover, by reducing the computation load of real-time gaze prediction, we aim to tackle the inconsistent temporal resolution experienced in previous online eye-tracking studies [Semmelmann and Weigelt, 2018]. We applied webcam-based eye-tracking, utilizing state-of-the-art deep learning methods [Zhang et al., 2015; Park et al., 2019; Zhang et al., 2020] for appearance-based gaze estimation. These methods predict eye gaze vectors from the recorded video frames, which are then calibrated to gaze points on the 2D stimulus-presentation plane. Since these deep learning models were pre-trained on a curated database, our calibration procedure also aids in correcting for personal appearance of participants and physical setup assumptions. Below, we report the first of a series of planned analyses [Saxena et al., 2021] to evaluate the application of deep learning for online eye-tracking, across a battery of tasks and dependent measures.

2 EXPERIMENTAL METHOD

2.1 Pre-registration

The study procedure was pre-registered prior to any human observation, and prior to the full collection of data, together with an addendum specifying updates on defining measurement noise to optimize recordings [Saxena et al., 2021]. All ethical approvals, participation criteria, and general procedures can be found in detail there.

2.2 Participants

A total of 72 participants completed the online study. Data for 31 participants had to be excluded based on fps and face detection rates (see Saxena et al. [2021] for a detailed report on the exclusion criteria). The final data consisted of 41 participants (12 male and 29 female participants), aged 20 to 33 years (M = 26, SD = 3). A majority of the participants reported having normal vision (30 of 41) and identified as students (34 of 41). Participants were not allowed to wear glasses while performing the experiment; use of contact lenses was allowed.

2.3 Experiment Platform

The online experiment platform LabVanced [Finger et al., 2017] was used to present the stimuli and record participant responses and webcam videos. The experiment was performed in full-screen mode. All tasks were designed in an arbitrary coordinate system of frame units where 1 visual degree = 54.05 fu (frame units). To allow compatibility with different screen sizes, device calibration (see section 2.5.1) was performed before starting the study and stimulus presentation was fixed to a presentation frame size of 29.6° x 16.65° (1600 x 900 fu), centered on the screen. Participation was therefore only possible with a laptop screen meeting that size criterion. Time events were recorded as UNIX timestamps and later mapped to video frame numbers using constant fps (frames per second) calculated from video parameters.

2.4 Gaze Tracking Models

We selected three deep learning methods that have each reported state-of-the-art results on gaze-tracking datasets collected in-the-wild: 1) the MPIIGaze [Zhang et al., 2015] model that was among the first attempts to tackle appearance-based gaze estimation with ConvNets 2) the FAZE model by Park et al. [2019] that proposed meta-learning to train an adaptable gaze estimator and 3) the ETHXGaze [Zhang et al., 2020] model that was trained on a high-quality, large-scale dataset collected under extreme head pose and gaze variation. Open-sourced Pytorch implementations[1] [2] [3] of all three models were used with customized inference and analysis scripts to adapt to the format of data collected from the online experiment. Image pre-processing steps, (histogram equalisation, normalisation, scaling, etc.), facial keypoints detection, and data normalization as proposed in [Sugano et al., 2014; Zhang et al., 2018], were applied as defaults. During inference, we assumed a single face in all video frames and discarded frames where no faces were detected. The FAZE model applies on a few-shot fine-tuning procedure, in addition to gaze-calibration. To ensure direct comparison with the other two models, and to keep consistent inference steps for all three models, the baseline pre-trained model was used without fine-tuning. While this might affect the accuracy of final gaze predictions, it should not interfere with judging the effect of calibration strategy, which is the focus of interest in this study.

2.5 Tasks

2.5.1 Device Calibration. Device calibration ensures a fixed size presentation frame over varying screen sizes of different participants and allows to robustly estimate the participant's distance from the screen. This distance is utilized by the gaze prediction models to normalize face images [Sugano et al., 2014; Zhang et al., 2018]. Screen size parameters and participant-screen distance are also required to project the predicted gaze vectors (pitch and yaw) from the deep learning models to 2D points and to calculate visual angles.

For screen-size calibration (Fig. 1A, left), participants were required to place a standard-sized ID card (85.60 × 53.98 mm) against the screen, and resize the displayed reference image until it was

[1] https://github.com/hysts/pytorch_mpiigaze
[2] https://github.com/NVlabs/few_shot_gaze
[3] https://github.com/xucong-zhang/ETH-XGaze

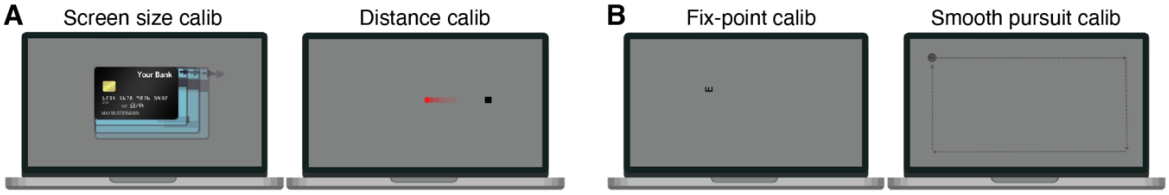

Figure 1: Types of calibration used in the experiment. A. Device calibration procedures used to estimate the screen size (left) and distance (right) of participants. B. Procedures for calibrating participants' gaze location. B, left: Fix-point calibration. The target "E" (black) occurred randomly in one of four orientations (up, down, left, right) at one of 16 possible locations. B, right: Pursuit calibration. The target moved in a rectangular trajectory, indicated by the dotted path (path not visible to participants).

the same size as the card. Since the physical dimensions of the card and pixel resolution of the image are known, the procedure allows to calculate the pixel density per mm for a display, and provides a pixel-to-mm conversion factor.

For distance calibration (Fig. 1A, right), we implemented the blind spot distance estimation task [Li et al., 2020], which leverages the fact that the human eye blind spot is located at a relatively consistent angle (α). The distance of a participant can be estimated using simple trigonometric calculations given α and the corresponding distance projected on screen. The task consisted of a fixed black square and a moving red circle that swept from right to left in the horizontal direction. Participants were instructed to sit at a distance of 50 cm from the screen, fixate on the black square with their right eye closed, and respond as soon as they perceived that the red circle disappeared. Participant responses were averaged over 5 repetitions with $\alpha = 13.5°$. If participants were estimated to be seated +/- 3.5 cm away from the required 50 cm distance, they were instructed to change their distance accordingly and the procedure was repeated.

2.5.2 Gaze Calibration. We customized and adapted two calibration methods (fix-point and smooth-pursuit) in our online experiment (Figure 1B). Fix-point calibration collects paired-data (recorded webcam frame and gaze target position) by presenting a sequence of evenly spread fixation points on the screen, whereas smooth-pursuit calibration presents a gradually moving target to be tracked by the participant's gaze. To ensure improved data quality, implicit attention detection mechanisms were integrated in both methods. Fix-point calibration provides a strict check for participants' fixation on the displayed target by validating their responses, while smooth-pursuit calibration filters data samples by calculating the correlation between gaze predictions and the moving target trajectory. We used the collected data to fit a second-order polynomial function which was chosen based on the calibration comparisons done by Harezlak et al. [2014].

In the fix-point calibration task (Fig. 1B, left), a stationary target (letter "E") occurred randomly across an evenly spread 4x4 grid on the screen. The target appeared in one of four possible orientations (up, down, left, and right) and participants were instructed to press the associated arrow key corresponding to the displayed orientation. The size and contrast of display required participants to make a saccade to the target and fixate it for correct identification. Target position was updated only when a correct response was registered. The last 10 frames recorded before correct keystrokes for each fixation target were used to fit the calibration model. The task ended after each of the 16 positions was fixated once.

In the smooth-pursuit calibration (Fig. 1B, right), participants followed a moving target with their gaze. A rectangular trajectory for target movement was chosen, as it was shown to be highly efficient in previous studies [Bace et al., 2020; Pfeuffer et al., 2013; Hassoumi et al., 2019]. Predicted gaze coordinates were compared over a moving window with the coordinates of the target by calculating Pearson's product-moment correlation and thresholding over a limit to check if eye movements followed the moving target. The size of the moving window (15 frames) and threshold limit (0.2) were chosen based on the analysis by Vidal et al. [2013].

2.5.3 Fixation Task. Participants completed a battery of tasks. Details about all tasks can be found in our pre-registration [Saxena et al., 2021]. In the current paper, we focus solely on results from the fixation task to assess the accuracy of different calibration strategies because this task required the most spatial precision of all tasks in the battery.

The fixation task was an online adaptation of the small-grid task by Ehinger et al. [2019]. In the task, a fixation target appeared randomly at one of 13 points chosen from a 7 x 7 grid, equally spaced across -6.2° to 6.2° visual degrees vertically and -11.1° to 11.1° horizontally. Participants had to fixate the target and keep fixating until it moved to a new location. They indicated the initiation of the fixation by a mouse click. Following the mouse click, the target remained stationary for 2500 ms before moving to the next location. A trial consisted of the target appearing once in all 13 locations, in random order, with each block starting and ending in the central position. Participants completed a total of 10 trials of this task and one additional practice trial that was discarded from further analyses.

2.6 Procedure

Data collection started with giving informed consent. Then, device calibration (screen and distance calibration tasks) was performed (Fig. 1A). The experimental part consisted of two blocks, created by equally splitting the number of trials for each task in the battery of eye-tracking tasks (for a full description see Saxena et al. [2021]). The sequence of experimental tasks was balanced between participants and kept constant for the two blocks. Gaze calibration procedures (Fig. 1B) were performed before (beginning) and after the first task block (middle), and again after the second task block (end). Each calibration block consisted of randomized single trials

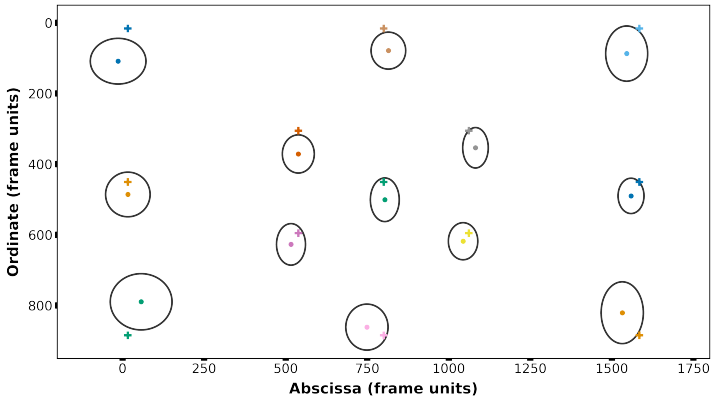

Figure 2: Fixation task predictions from the FAZE model, using Beg+Mid+End calibration samples and the E+SP calibration routine. Crosses (+) represent the location of displayed fixation targets on screen. Dots and ellipse axes represent the mean and standard deviation in x and y directions, respectively, of predicted fixation points, over all 41 participants.

of both fix-point and smooth pursuit calibration. In the current paper, we only report results with respect to calibration and the fixation task.

2.7 Data Treatment

We analyzed the effect of gaze-tracking model and calibration strategy on the fixation task accuracy. All video frames were resized to a 640x480 resolution and processed sequentially by each model. Intrinsic camera parameters were approximated based on image resolution. Camera placement was assumed to be in the top-center of the screen – a requirement confirmed by participants prior to beginning the online experiment. These parameters were used to geometrically map the model outputs in normalized space to 2D points on the screen coordinate system, which were then corrected following a personal gaze calibration procedure for each participant. The calibration models were trained using data from calibration trials and evaluated on their prediction accuracy for the fixation task. Accuracy for the fixation task refers to the offset between the displayed target position and the estimated fixation position from gaze predictions. We estimated the fixation point as the median of calibrated gaze predictions over a fixed duration of 2500 ms and calculated offsets as the Euclidean distances between presented targets and estimated fixation points. The distances were calculated in presentation frame units and converted to visual angles using a conversion factor (see 2.3). The final scores were aggregated by calculating 20% winsorized average values, first over all 13 points, then over the 10 fixation trials.

3 RESULTS

In Figure 2 we provide an example of the results from one of the gaze prediction models. It can be seen that fixation accuracy decreases from center to periphery, likely because calculating fixation location is more error-prone when the eyeball turns relative to a central camera – an effect well-documented in infra-red camera based eye-tracking studies.

3.1 Effect of Calibration Sample Size

We compared the difference in fixation accuracy when calibration data of different sizes were used, drawn from the three-time points of recording, referred to as beginning (Beg), beginning+middle (Beg+Mid), beginning+middle+end (Beg+Mid+End). These samples were used to fit the calibration models, and the calibration models were then used to get final gaze predictions for the fixation task.

We first report a two-factor ANOVA with the factors model type (MPIIGaze, FAZE, ETHXGaze) and sample size (Beg+Mid+End; Beg+Mid; Beg). The main effect of model was not significant, ($F(2,80) = 2.493$, $p = 0.121$, $\eta 2 = 0.057$); however, the factor sample size was ($F(2,80) = 24.423$, $p < 0.001$, $\eta 2 = 0.008$). The factors did not interact ($F(4,160) = 1.833$, $p = 0.190$, $\eta 2 = 0.001$). Even though Figure 3A indicates that FAZE resulted in the best accuracy (smallest deviation), followed by MPIIGaze, followed by ETHXGaze, this tendency was not significant. To explore the main effect of sample size, we calculated post-hoc t-tests. Increasing the sample size by adding the Mid calibration block to the Beg, improved accuracy, (Beg and Beg+Mid: $t(40) = 4.894$, $p < 0.001$), as did adding the End calibration: Beg+Mid and Beg+Mid+End: $t(40) = 2.503$, $p = 0.049$, Beg and Beg+Mid+End: $t(40) = 5.401$, $p < 0.001$. See Figure 3A.

3.2 Effect of Calibration Time

Based on the sample size analysis (section 3.1), we selected the Beg+Mid+End sample for further analyses, as it provided minimum calibration error (maximum accuracy). The calibration strategy of selecting all calibration blocks (Beg+Mid+End) together was compared with a block-specific (Blockwise) strategy in which single calibration trials were temporally assigned to the following experimental block. That is, the beginning calibration sample was used to predict fixation performance in the first task block, and the middle calibration for the second. A two-factor ANOVA with the factors model type (MPIIGaze, FAZE, ETHXGaze) and calibration strategy (Beg+Mid+End and Blockwise) showed no significant effect of model type ($F(2,80) = 2.489$, $p = 0.121$, $\eta 2 = 0.057$) and no interaction

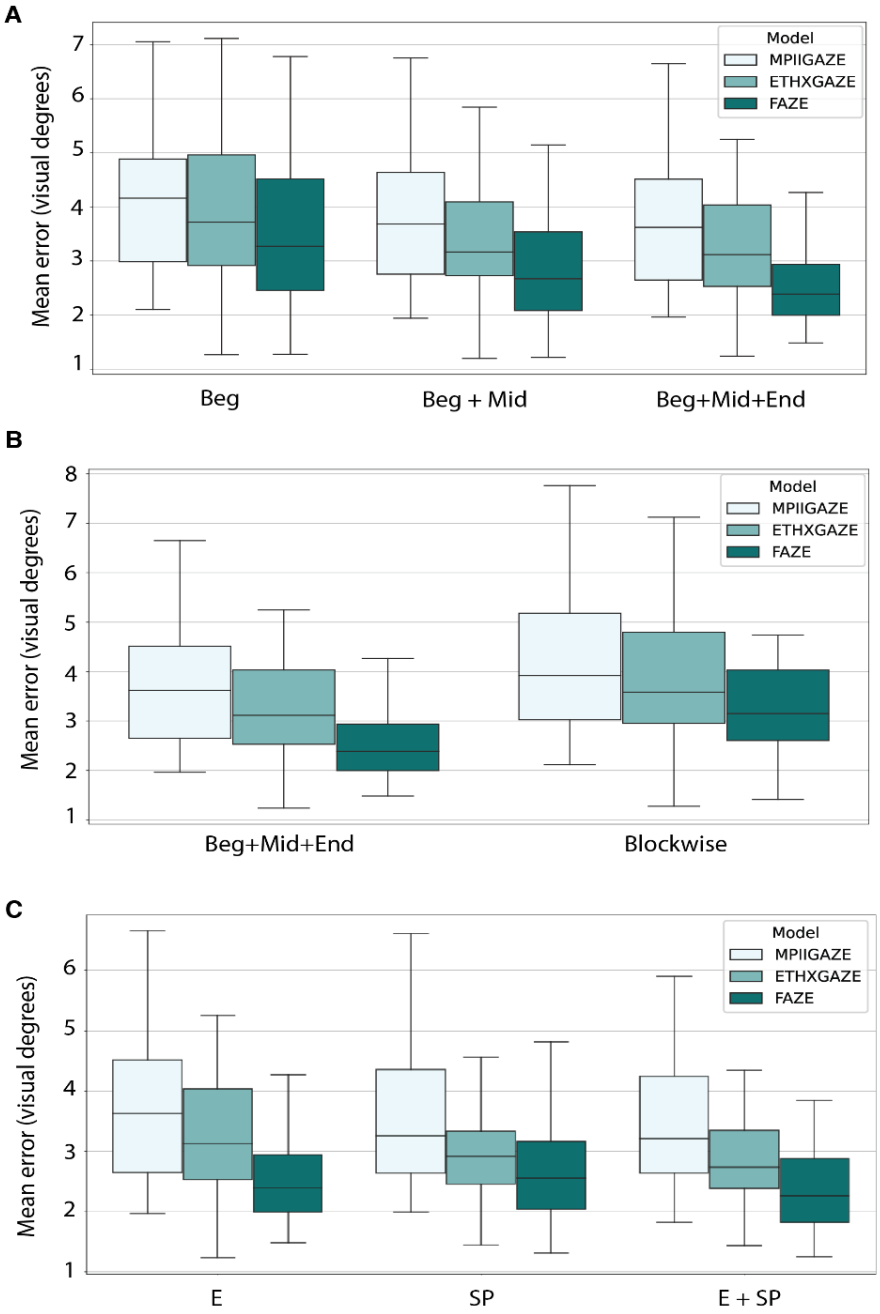

Figure 3: Mean error (in visual degrees; less error = better accuracy) for the three deep learning models (MPIIGaze, FAZE, ETHXGaze; colored bars), using different calibration strategies. Error bars represent the variability between participants. A. Fixation accuracies for the three different sample sizes of calibration data (beginning, beginning + middle, and beginning + middle + end). B. Comparison of the best strategy from A (Beg+Mid+End) to a blockwise (i.e., temporally contingent) calibration strategy. C. Comparison of calibration task type (fixation task (E) vs. smooth pursuit task (SP) vs. E+SP).

effects ($F(2,80) = 0.393$, $p = 0.601$, $\eta 2 < 0.001$). The main effect of calibration strategy was significant ($F(1,40) = 57.544$, $p < 0.001$, $\eta 2 = 0.010$). Fixation accuracy was higher when using Beg+Mid+End as compared to using the Blockwise calibration strategy, $t(40) = 7.586$, $p < 0.001$. See Figure 3B

3.3 Effect of Calibration Task

We proceeded with the Beg+Mid+End calibration sample and checked for the effect of calibration task on fixation accuracy (fix-point calibration on the letter E: E, and smooth pursuit: SP). Fixation trials were evaluated using calibration data from the two calibration tasks separately (E, SP) and pooled (E+SP). We analyzed the data again by the two-factor ANOVA, with model (MPIIGaze, FAZE, and ETHXGaze) as one factor and calibration type (E, SP, E+SP) as the other. The main effect of model was not significant, ($F(2,80) = 2.657$, $p = 0.110$, $\eta 2 = 0.060$), but that of calibration strategy was, ($F(2,80) = 9.038$, $p = 0.001$, $\eta 2 = 0.002$). The factors did not interact, ($F(4,160) = 1.883$, $p = 0.184$, $\eta 2 = 0.001$). Post-hoc t-tests comparing different calibration strategies showed a significant difference between E and E+SP, ($t(40) = 4.417$, $p < 0.001$), as well as between SP and E+SP ($t(40) = 3.583$, $p = 0.003$), meaning the combined E+SP strategy was better than either strategy alone. There was no significant difference between E and SP calibration ($t(40) = 1.371$, $p = 0.534$). Results are plotted in Figure 3C.

4 DISCUSSION

Our results are highly informative for the practical usage of deep learning for web-based eye-tracking and demonstrate a clear improvement over existing methods for online eye-tracking with a mean error of around 2.6 visual degrees, in comparison to existing tools such as WebGazer's 4.17[0]. Given the computational limitations of real-time model inference in web browsers, our approach of splitting data collection from model processing allows the use of more complex and accurate eye-tracking models. Comparing to high-speed video-based infra-red eye trackers, such a deviation is arguably big (Eyelink1000: 0.57° and Pupil Core: 0.82° winsorized mean error in the same fixation task [Ehinger et al., 2019]). However, for taking the lab into the wild, using webcam eye-tracking in an unsupervised online experiment, we were surprised by the high accuracy of $\sim 2.6°$. Such deviation means that a variety of experimental paradigms can be conducted online. Knowing the critical deviation, future experiments can be set up accordingly by taking it into account when selecting and placing stimuli and defining regions of interest.

For researchers interested in applying these methods, we wish to highlight the importance of the distance and screen calibration strategies we have proposed here. For online eye-tracking studies, which provide less control of the surrounding environment, proper estimates of user distance and screen size are essential to have reliable conversions between gaze predictions and stimulus presentation coordinates. The device calibration tasks (see section 2.5.1) robustly calculated these measures and are highly recommended for similar unsupervised setups. The blind spot task for estimating user distance (2.5.2) tackles the complex problem of monocular depth estimation from a single webcam, providing a simple and reliable alternative to computer vision-based approaches.

With respect to gaze calibration, we show that pooling calibration data collected at multiple time points, and using multiple calibration methods, proved to be the best gaze prediction strategy, as compared to re-calibrating at each time point–the standard practice in eye-tracking studies. It is not the temporal relation between calibration and data collection that makes the additional blocks important, as shown by the missing benefit of Blockwise calibration, but rather the increased amount of data. Similarly, the increased amount of data and sampled screen locations that come with pooling both fixation and smooth pursuit calibration strategies resulted in better gaze prediction. It is interesting to note that a single trial of the fix-point calibration task took on average 14 seconds and collected 160 data points, while a single trial of the smooth-pursuit calibration took around 25 seconds to complete and collected 433 data points on average. Therefore, the total time spent for the two calibration tasks by each participant was less than 2 minutes (the total study time was around 35 minutes without any breaks), which is significantly less than previous online eye-tracking procedures where calibration took almost 50% of the study time [Semmelmann and Weigelt, 2018]. This demonstrates the efficiency of deep learning networks in real-world noisy environments; they are less susceptible to over-time gaze prediction drifts commonly experienced in eye-tracking studies and overcome the design restrictions of short trials and frequent calibrations adopted in previous studies to deal with such issues [Xu et al., 2015].

From the results of the fixation task, it could be descriptively seen that the FAZE model reports the best accuracy for all conditions followed by ETHXGaze and then by MPIIGaze, though these effects were not statistically significant (comparing MPIIGaze and FAZE, mean fixation accuracy dropped from 3.63° deviation to 2.58° (Beg+Mid+End)). However, note that our target sample size for between-model comparisons is higher (n = 64) than the sample size used in this report (n = 41), and data collection is still ongoing (see addendum in Saxena et al. [2021]). Also, the effect sizes ($\eta 2$) were of rather medium size for the main effect of the factor model in all three ANOVAs, indicating that it might be wise to wait for final data collection before drawing firm conclusions on the between-model comparisons. Also note that the fine-tuning procedure for FAZE was deactivated to keep the inference steps consistent (see section 2.4). The final accuracy of predictions after personalized fine-tuning is, therefore, expected to further improve the gaze predictions (forthcoming paper in preparation). Additionally, the prediction accuracy may also be dependent on the type of eye movement being analyzed. Therefore, our planned thorough investigation of these model predictions under different task settings (smooth pursuit, free view, fixation etc.) will help to elucidate the applicability of these models further.

The performance of FAZE demonstrates the effectiveness of highly parameterized networks for the task of gaze prediction, since FAZE applies a more sophisticated DenseNet based encoder-decoder model as compared to the ResNet and AlexNet-based architectures applied in ETHXGaze and MPIIGaze. However, the increase in accuracy comes at a higher computational expense. The FAZE model takes considerably more time and memory, which should be considered before choosing models for similar studies. For instance, inferencing a single frame (640*480 resolution) on our setup enabled with NVIDIA Titan RTX GPU took nearly, 50 ms for MPIIGaze, 200 ms for ETHXGaze, and 350 ms for FAZE. Nonetheless, researchers who are interested in fixation accuracies - using tasks analogous to our fixation task - would therefore benefit from a similar offline processing setup as compared to a real-time setup with light-weight regression models. Overall, the current results motivate further development of deep learning-based methods for gaze-tracking and

highlight a useful application of these models for researchers who wish to conduct online studies.

REFERENCES

Andrew T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments, & Computers 34, 455–470. https://doi.org/10.3758/BF03195475.

Orval Hobart Mowrer, Theodore Cedric Ruch and Neal E. Miller. 1935. The corneo-retinal potential difference as the basis of the galvanometric method of recording eye movements. American Journal of Physiology. American Journal of Physiology-Legacy Content 114, 423–428. https://doi.org/10.1152/ajplegacy.1935.114.2.423.

Yihua Cheng, Haofei Wang, Yiwei Bao and Feng Lu. 2021. Appearance-based gaze estimation with deep learning: a review and benchmark. arXiv:2104.12668. Retrieved from http://arxiv.org/abs/2104.12668.

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M. Bhandarkar, Wojciech Matusik and Antonio Torralba. 2016. Eye tracking for everyone. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2176–2184. https://doi.org/10.1109/CVP.

Tobias Fischer, Hyung Jin Chang and Y. Demiris. 2018. RT-GENE: Real-time eye gaze estimation in natural environments. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11214 LNCS, 339–357. https://doi.org/10.1007/978-3-030-01249-6_21.

Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang and James Hays. 2016. WebGazer: Scalable webcam eye tracking using user interactions. International Joint Conference on Artificial Intelligence, 3839–3845.

XLabsGaze - Webcam eye tracking software.2016. Retrieved from https://github.com/xLabsgaze.

Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni and Jianxiong Xiao. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. arXiv:1504.06755. Retrieved from http://arxiv.org/abs/1504.06755.

Kilian Semmelmann and Sarah Weigelt. 2018. Online webcam-based eye tracking in cognitive science: A first look. Behavior Research Methods, 50(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7.

Benjamin T. Carter and Steven G. Luke. 2020. Best practices in eye tracking research. International Journal of Psychophysiology, 155, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010.

Benedikt V. Ehinger, Katharina Groß, Inga Ibs and Peter König. 2019. A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. PeerJ, 7, 1–43. https://doi.org/10.7717/peerj.7086.

Rui Rodrigues, Joao P. Barreto, and Urbano Nunes. 2010. Camera pose estimation using images of planar mirror re#ections. In Proceedings of the 11th European Conference on Computer Vision. 382–395.

G. Bradski. 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 120; 122-125.

Ken Pfeuffer, Mélodie Vidal, Jayson Turner, Andreas Bulling and Hans-Werner Gellersen. 2013. Pursuit calibration: Making gaze calibration less tedious and more flexible. UIST 2013 - Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, 261–270. https://doi.org/10.1145/2501988.2501998.

Almoctar Hassoumi, Vsevolod Peysakhovich and Christophe Hurter. 2019. Improving eye-tracking calibration accuracy using symbolic regression. PloS one, 14(3), e0213675. https://doi.org/10.1371/journal.pone.0213675.

Xucong Zhang, Yusuke Sugano, Mario Fritz and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4511–4520. https://doi.org/10.1109/CVPR.2015.7299081.

Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges and Jan Kautz. 2019. Few-shot adaptive gaze estimation. Proceedings of the IEEE International Conference on Computer Vision, 9367–9376. https://doi.org/10.1109/ICCV.2019.00946.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang and Otmar Hilliges. 2020. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12350 LNCS, 365–381. https://doi.org/10.1007/978-3-030-58558-7_22.

Shreshth Saxena, Lauren Fink and Elke Lange. 2021. An empirical experiment on deep learning models for tracking eye movements via webcam. OSF Pre-registration. https://osf.io/qh8kx.

Holger Finger, Caspar Goeke, Dorena Diekamp, Kai Standvoßand Peter König. 2017. LabVanced: a unified JavaScript framework for online studies. International Conference on Computational Social Science (Cologne).

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In CVPR, 2014.

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2018. Re- visiting data normalization for appearance-based gaze estimation. In ETRA, 2018.

Qisheng Li, Sung Jun Joo, Jason D. Yeatman and Katharina Reinecke. 2020. Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. Scientific Reports, 10, 904 2020. https://doi.org/10.1038/s41598-019-57204-1

Katarzyna Harezlak, Pawel Kasprowski and Mateusz Stasch. 2014. Towards Accurate Eye Tracker Calibration - Methods and Procedures. Procedia Computer Science, 35, 1073–1081. https://doi.org/10.1016/j.procs.2014.08.194.

Mihai Bace, Vincent Becker, Chenyang Wang, and Andreas Bulling. 2020. Combining Gaze Estimation and Optical Flow for Pursuits Interaction. In ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers). Association for Computing Machinery, New York, NY, USA, Article 2, 1–10. DOI:https://doi.org/10.1145/3379155.3391315.

Mélodie Vidal, Andreas Bulling and Hans-Werner Gellersen. 2013. Pursuits: spontaneous interaction with displays based on smooth pursuit eye movement and moving targets. Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, (439–448. https://doi.org/10.1145/2493432.2493477.