A Calibration Framework for Photosensor-based Eye-Tracking System

Dmytro Katrychuk
Department of Computer Science,
Texas State University
San Marcos, Texas
d_k139@txstate.edu

Henry K. Griffith
Department of Computer Science,
Texas State University
San Marcos, Texas
h_g169@txstate.edu

Oleg V. Komogortsev
Department of Computer Science,
Texas State University
San Marcos, Texas
ok@txstate.edu

ABSTRACT

The majority of eye-tracking systems require user-specific calibration to achieve suitable accuracy. Traditional calibration is performed by presenting targets at fixed locations that form a certain coverage of the device screen. If simple regression methods are used to learn a gaze map from the recorded data, the risk of overfitting is minimal. This is not the case if a gaze map is formed using neural networks, as is often employed in photosensor oculography (PSOG), which raises the question of careful design of calibration procedure. This paper evaluates different calibration data parsing approaches and the collection time-performance trade-off effect of grid density to build a calibration framework for PSOG with the use of video-based simulation framework.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *Ubiquitous and mobile computing design and evaluation methods*; • **Computing methodologies → Neural networks**.

KEYWORDS

eye-tracking, calibration procedure, photo-sensor oculography, PSOG, machine learning, ML

ACM Reference Format:
Dmytro Katrychuk, Henry K. Griffith, and Oleg V. Komogortsev. 2020. A Calibration Framework for Photosensor-based Eye-Tracking System. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Short Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3379156.3391370

1 INTRODUCTION

Eye-tracking (ET) technology is a cutting edge tool for touchless human-computer interaction, and has tremendous potential for biometrics and health-assessment applications. Underlying gaze models usually contain a set of subject-dependent parameters that should be derived from pre-collected data or from data gathered during a calibration procedure, performed for every new use of the system. Despite recent advances in using the former approach to build a calibration-free ET device, the later one is currently utilized in order to achieve the highest level of performance of modern ET systems [Kim et al. 2019].

Conventional calibration [Nyström et al. 2013] requires the establishment of the spatial distribution of fixation targets, along with a protocol for detecting the fixation interval for each target location to ensure the validity of ground-truth (GT) data. In the case of video-based oculography (VOG), the number of targets usually varies from 9 to 25 [Kasprowski et al. 2014]. Fixations are typically assumed to occur in a fixed time period offset within the interval of presentation for each target. Prior research has largely utilized empirical values for these offsets, rather than relying on statistics derived from the collected data [Blignaut and Wium 2014], [Akkil et al. 2014], which risks either discarding valid data or including erroneous fixation samples.

For common VOG systems, gaze maps are built from extracted image features using regression models with limited degrees of freedom, which is enough to achieve good performance within certain conditions of restricted head movements, sensor shifts, and lighting changes. Nevertheless, neural networks proved to be an essential tool for enabling eye-tracking in more complex environments, exhibiting state-of-the-art performance in robust non-restrained eye gaze estimation in the wild [Kim et al. 2019]. Next example is a challenging domain of portable virtual reality, that puts substantial restrictions on overall power consumption and hardware complexity of an underlying eye-tracking sensor. Those limitations are unlikely to be met by widely adopted VOG and require a novel approach, such as photosensor oculography (PSOG) [Katrychuk et al. 2019], which is the focus of this paper.

The previous work [Griffith et al. 2019] studied the complexity of the PSOG-based eye gaze estimation which reasons the necessity of using neural networks as the inference mechanism. The greater representation ability of these networks afforded by their large number of parameters introduces a much higher chance of overfitting to data samples which are either erroneously included as fixation intervals, or are the result of poor subject compliance to a presented target.

The aforementioned challenges pose the need for careful design of components that comprise the whole calibration framework. Therefore, this study focuses on the assessment of the learned gaze map performance with respect to different calibration data parsing approaches and grid densities for PSOG-based ET sensors. We keep in mind the hardware limitation on power consumption of the system and mainly focus on software basis of eye gaze estimation, therefore the sensor hardware output is simulated using a VOG-based pipeline.

ETRA '20 Short Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7134-6/20/06…$15.00
https://doi.org/10.1145/3379156.3391370

2 DATA COLLECTION

The data was collected using the custom VOG setup introduced in [Abdulin and Komogortsev 2017] that allows access to the full video stream of the whole recording, which is essential for our study. The device has the operational range of (−16.7°; 16.7°) vertically and (−20.51°; 20.51°) horizontally in degrees of the visual angle and was set to $120Hz$ sampling frequency.

The recorded data set consists of 40 subjects that performed an oblique saccades stimulus task. The task consists of 174 fixation targets that are depicted in Fig. 1. The targets $[C1..C29]$ that are drawn as red crosses are shown on the screen for a fixed time of 1.75 sec. For the rest of the targets, the time was drawn from the uniform distribution with the range of [1..1.25] sec. To reduce fatigue effects, groups labeled as $[C1..C29]$, $[V1..V16]$, which were used in learning gaze maps, were presented first and second, respectively. The targets are presented in the randomized (but the same for all subjects) order within each of the three groups.

3 METHODOLOGY

3.1 Pre-processing

PSOG sensor outputs were simulated using cropped images of the near eye region from the VOG system. As neural networks are employed for PSOG gaze mapping to improve shift robustness [Rigas et al. 2018], a workflow simulating translation shifts was enacted on each image. Head movements were compensated to ensure precise control over the simulated sensor shift range. Without having an explicit marker to track, we based the compensation algorithm on the observation that the relationship between eye gaze and pupil center is near linear for the fixed head position and slight transitional head movement change only the intercept of the relation. Therefore, the steps performed are the following:

(1) Use the linear regression of eye gaze $(pos_x; pos_y)$ to pupil center position $(pc_x; pc_y)$ reported by the ET system for the whole recording:

$$(pc_x; pc_y) = (A_x * pos_x + b_x; A_y * pos_y + b_y)$$

(2) For every sample i perform the head movement correction of $(-A_x * pos_x^i; -A_y * pos_y^i)$ pixels.

The rest of the shifts and PSOG output simulation pipeline remains the same as described in [Griffith et al. 2019].

3.2 Calibration data parsing

In previous PSOG studies [Katrychuk et al. 2019], [Griffith et al. 2019], partitioning data into train/validation/test sets was performed randomly due to limitations of the common data set. This scenario does not reflect the target use case of the device, where gaze maps are trained using data collected only during calibration procedure. Therefore, that split neither reflected the distribution of gaze samples across the screen nor the fact that calibration target position should be used as the ground-truth labels in the train set instead of eye gaze reported by VOG. In the current study the proper care was taken during the stimulus design to overcome these issues.

The following algorithms were used for identifying valid fixation intervals on a per-target basis.

(V) VOG eye gaze was classified using I-VT [Salvucci and Goldberg 2000] with adaptive threshold and modified merging stage. The $85th$ percentile of radial velocity distribution computed from the whole recording was used as a decision boundary to pre-classify every sample as a fixation or a saccade. As the purpose of the algorithm is the detection of data intervals corresponding to fixations on a calibration target and not the eye movement classification by itself, fixations were merged on a per-target basis using the clustering DBSCAN algorithm [Ester et al. 1996] with $\varepsilon = 0.85$.

(P) Uncalibrated PSOG signal was classified into fixation and non-fixation samples using the sliding temporal window of raw sensor output that is fed to convolution neural network (CNN) [Hoppe and Bulling 2016]. The model consists of three 1D convolution layers with 24 feature maps and kernel size of 5, followed by five fully connected layers with 20 neurons in each. Dropout with 0.1 probability was applied to every layer except the output one. The input sliding window captures the 11 samples before and after the current one (23 samples total). Output of (V) with $\varepsilon = 0.4$ for more fine-grained fixation intervals was used as a ground-truth during the training phase.

The first research question that we are going to address to build the PSOG calibration framework is how to detect data regions that correspond to fixation on a calibration target based on raw PSOG output? The different ways for such detection are entitled as calibration data parsing approaches and are presented next:

(1) 'VOG_GT' - for every calibration target estimates the corresponding fixation interval start Fix_{beg} and end Fix_{end} based on the output of the algorithm (V). Used as a benchmark unrealistic scenario when the gaze map is already available prior to calibration.

(2) 'Blind_Empirical' - the data within time range of

$$[Tar_{beg} + 1000ms; Tar_{end} - 250ms]$$

is blindly considered as the fixation interval on the calibration target that starts at Tar_{beg} and ends at Tar_{end}. The constants are picked as a conservative range from the literature [Blignaut and Wium 2014].

(3) 'Blind_Temporal' - the same as (2) but with the time range of

$$[Tar_{beg} + Lat_{beg}; Tar_{end} + Lat_{end}]$$

, where Lat_{beg}, Lat_{end} are constants based on statistics of the data from (1). For all calibration fixations of subject from the **train+validation** sets, the Lat_{beg} is equal to the $95th$ percentile of the distribution of $Fix_{beg} - Tar_{beg}$, and the Lat_{end} is the $95th$ percentile of $Tar_{end} - Fix_{end}$ (i.e., 95% of calibration fixations started before the Lat_{beg} ms after the calibration target appeared and ended after the Lat_{end} ms after the target disappeared).

(4) 'All_Fix_Uncalib_PSOG' - **all** "fixation" samples based on the output of the algorithm (P) go into the train set, which can be seen as the way to alleviate the non-ideal performance of the classification algorithm - the vast majority of fixations should correspond to their calibration target and even if the

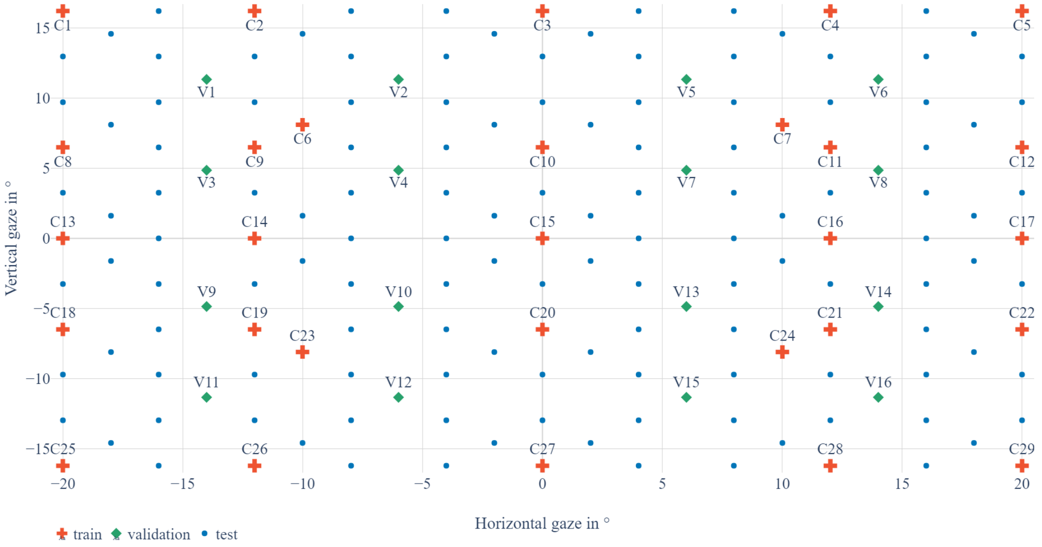

Figure 1: All fixation targets presented to the subject during the task. Different set of targets correspond to the within-subject data split into train (red crosses)/validation (green diamonds)/test (blue circles) for the gaze map learning purposes.

true fixation is split into many segments, all of them will be included.

(5) 'Longest_Fix_Uncalib_PSOG' - the (4) is applied first and then **only** the longest fixation on the calibration target is kept in the train set.

(6) 'Longest_Fix_Temporal' - the output of (5) bounded by the time range used in (3). So, only the part of the longest classified calibration fixation within the timestamp range when it is the most likely to happen is kept in the train set.

The following between-subject split is used for the assessment and parametric fine-tuning of approaches listed above:

- *train*: subjects with id from 1 to 20
- *validation*: subjects with id from 21 to 25
- *test*: subjects with id from 26 to 40

3.3 Calibration grid density

The amount of calibration points naturally reflects the trade-off between the quality of the learned gaze map and time to collect calibration data for every new use of the ET system. To assess this trade-off, we pick the best approach from the Calibration data parsing study and repeat the gaze map learning with gradual decrease of the amount of calibration targets used to comprise the within-subject train set. The second research question of building the PSOG calibration framework is how the performance of the learned gaze map is influenced by calibration grid density? In order to address it, the following grid configurations are tested (using labels from the Fig. 1):

(1) 29 points: [*C*1..*C*29]
(2) 25 points: the (2) **without** *C*6, *C*7, *C*23, *C*24
(3) 21 points: the (3) **without** *C*9, *C*11, *C*19, *C*21
(4) 15 points: the (4) **without** *C*8, *C*10, *C*12, *C*18, *C*20, *C*22
(5) 5 points: *C*1, *C*5, *C*15, *C*25, *C*29
(6) 9 points: the (5) **with** *C*3, *C*13, *C*17, *C*27
(7) 13 points: the (6) **with** *C*6, *C*7, *C*23, *C*24

3.4 Eye gaze map learning

The low-power architecture from [Katrychuk et al. 2019],targeted for deployment on embedded device, was used herein. The CNN model has the following architecture: two convolution layers with 4 feature maps of size 3*x*3 in each and 'same' padding followed by four fully-connected layers with 20 neurons in each.

The following within-subject split was utilized in the subject-specific gaze map learning:

- *train*: gaze samples that correspond to up to 29 targets labeled as [*C*1..*C*29] in Fig. 1
- *validation*: gaze samples that correspond to 4 targets picked from [*V*1..*V*16] in Fig. 1
- *test*: the rest of subject's gaze samples

To prevent overfitting to the train set which represents only small fraction of the screen, the early stopping technique based on the validation loss was utilized. The patience parameter was set to 50, only relative decrease of the loss for more than 1% is considered as an improvement. As the subsampling from the train set will not provide the ability to assess the generalization of the learned map across the screen, the validation set was constructed from the different grid of points that are labeled as [*V*1..*V*16] in the Fig. 1. That procedure ensures the best possible subject compliance to presented targets in the absence of subject-controlled validation step during the data collection:

(1) Put targets into 4 groups based on their quadrant location: [$V1..V4$], [$V5..V8$], [$V9..V12$], [$V13..V16$].
(2) For every group, pick the target with the minimum misalignment between validation target position and eye gaze during the corresponding fixation on it.

4 RESULTS

The results of the calibration data parsing study are presented in Tab. ??. The main observations from them are:

- The 'VOG_GT' approach showed the best performance as expected, so it is picked as the 'baseline' and all other results in Tab. ?? are normalized with respect to it.
- It was obtained from the data that $Lat_{beg} = 735ms$ and $Lat_{end} = -155ms$ for 'Blind_Temporal' approach. Even though the 'Blind_Empirical', which is based on previous studies, performs just marginally worse than 'Blind_Temporal', we suggest that future VOG studies also estimate offset values using statistics from the collected data to improve reliability.
- The 'All_Fix_Uncalib_PSOG' approach exhibited the worst performance of all techniques. We hypothesize that this is related to the sensitivity of the learned gaze map to outliers. This performance degradation suggests that some fixations occur prior to settling on the calibration target, resulting in the inclusion of erroneous ground-truth data.
- The remaining approaches produced similar spatial errors. The 'Longest_Fix_Temporal' that combines the blind temporal-based removal of the data together with results of the classification of the uncalibrated PSOG signal is marginally the best with a mean spatial error of 2.15°.

The results of the calibration grid density study are summarized in Tab. ?? with '29 points' configuration being the best performing, as expected. As shown, spatial error increases substantially when the number of targets is reduced from 21 to 15 (spatial error degradation of 15% and 42%, respectively). Given this observation, we suggest that 21 targets constitutes the optimal trade-off between calibration time and performance of the learned gaze map.

5 DISCUSSION

The major assumption behind the described calibration procedure is the expectation that user will follow the target to the best of their skill, which is not always the case and has the significant influence of inevitable micro eye movements. Overall, inability to obtain the well-defined ground-truth for conventional eye-tracking methods led to the research in promising retinal image-based eye-tracking [Bowers et al. 2019] which has not matured yet to be widely adopted. In conventional systems, the ground-truth labels are derived from the target position, which may result in outliers in the train set. The way to detect those outliers during the calibration procedure itself can be used to automatically obtain points for the extra recalibration step or to lower the confidence in corresponding samples during the learning phase. The effect of subject compliance on the performance of the learned map is an open research question that should be addressed in the future work.

It was shown in [Katrychuk et al. 2019] that transfer-learning from the pre-collected data can be used to reduce the training time but it did not help in improving the spatial error. We speculate that the reason for that is the unrealistically simplified data split utilized in that study that provides much higher coverage of the screen by the train set which does not reflect the real calibration procedure. Therefore, re-evaluation of transfer-learning should be done in the context of our study with a proper calibration framework, with the expectation of lowering the required calibration data collection time for the same performance level.

Table 1: Study results as spatial errors in degrees of visual angle. All relative changes are computed with respect to the best performing option.

(a) Calibration data parsing approaches study

Approach	Spatial error	Δ in error
VOG_GT	1.99°	-
Blind_Empirical	2.19°	+9.9%
Blind_Temporal	2.16°	+8.3%
All_Fix_Uncalib_PSOG	2.43°	+21.9%
Longest_Fix_Uncalib_PSOG	2.19°	+10.1%
Longest_Fix_Temporal	2.15°	+7.9%

(b) Calibration grid density study

№ of targets	Time	Δ in time	Spatial error	Δ in error
29	51 *s*	-	2.15°	-
25	44 *s*	−14%	2.22°	+3%
21	37 *s*	−28%	2.46°	+15%
15	26 *s*	−48%	3.04°	+42%
13	23 *s*	−55%	3.04°	+42%
9	16 *s*	−69%	3.79°	+77%
5	9 *s*	−83%	5.68°	+165%

6 CONCLUSION

In the paper we studied two important components of the calibration framework that need to be considered when the eye gaze map is built using highly parametric machine learning approaches. Those components were evaluated in the case of PSOG-enabled ET system and the resulting calibration framework uses:

- The longest fixation classified from uncalibrated PSOG signal bounded by the time range where it is most likely to be on the calibration target
- $21\,(train) + 4\,(validation) = 25$ points calibration grid

The fixation time bounds are statistically supported by the collected data. The achieved mean spatial error of that configuration is 2.46°. The up-to-date codebase and the data set to replicate the results is available on GitHub. [1]

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant Numbers CNS-1250718 and CNS-1714623. It was inspired by the Google Global Faculty Research Award bestowed on Dr. Komogortsev in 2019.

[1] https://github.com/pseudowolfvn/psog_nn/tree/etra2020

REFERENCES

Evgeniy R Abdulin and Oleg V Komogortsev. 2017. Study of Additional Eye-Related Features for Future Eye-Tracking Techniques. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1457–1463.

Deepak Akkil, Poika Isokoski, Jari Kangas, Jussi Rantala, and Roope Raisamo. 2014. TraQuMe: a tool for measuring the gaze tracking quality. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 327–330.

Pieter Blignaut and Daniël Wium. 2014. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior research methods* 46, 1 (2014), 67–80.

Norick R Bowers, Agostino Gibaldi, Emma Alexander, Martin S Banks, and Austin Roorda. 2019. High-resolution eye tracking using scanning laser ophthalmoscopy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ACM, 58.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

Henry Griffith, Dmytro Katrychuk, and Oleg Komogortsev. 2019. Assessment of Shift-Invariant CNN Gaze Mappings for PS-OG Eye Movement Sensors. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

Sabrina Hoppe and Andreas Bulling. 2016. End-to-end eye movement detection using convolutional neural networks. *arXiv preprint arXiv:1609.02452* (2016).

Pawel Kasprowski, Katarzyna Harężlak, and Mateusz Stasch. 2014. Guidelines for the eye tracker calibration using points of regard. In *Information Technologies in Biomedicine, Volume 4*. Springer, 225–236.

Dmytro Katrychuk, Henry Griffith, and Oleg Komogortsev. 2019. Power-efficient and shift-robust eye-tracking sensor for portable VR headsets. (2019).

Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. NVGaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 550.

Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost Van De Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior research methods* 45, 1 (2013), 272–288.

Ioannis Rigas, Hayes Raffle, and Oleg V Komogortsev. 2018. Photosensor oculography: survey and parametric analysis of designs using model-based simulation. *IEEE Transactions on Human-Machine Systems* 99 (2018), 1–12.

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