Implicit User Calibration for Gaze-tracking Systems Using Saliency Maps Filtered by Eye Movements

Hiroe, Mamoru

Yamamoto, Michiya

Nagamatsu, Takashi

(Citation)
ETRA '23: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications:1-5

(Issue Date)
2023-05-30

(Resource Type)
conference paper

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/0100490245

Implicit User Calibration for Gaze-tracking Systems Using Saliency Maps Filtered by Eye Movements

Mamoru Hiroe
Kwansei Gakuin University
Sanda, Japan
hiroe@ieee.org

Michiya Yamamoto
Kwansei Gakuin University
Sanda, Japan
michiya.yamamoto@kwansei.ac.jp

Takashi Nagamatsu
Kobe University
Kobe, Japan
nagamatu@kobe-u.ac.jp

ABSTRACT

In recent studies on gaze tracking systems using 3D model-based methods, the optical axis of the eye was estimated without user calibration. The remaining challenge in achieving implicit user calibration is estimating the difference between the optical and visual axes of the eye (angle κ). In this study, we propose two methods that improve the implicit user calibration method using saliency maps, focusing on eye movement to reduce calculation costs while maintaining accuracy.

CCS CONCEPTS

• **Human-centered computing → User interface management systems**.

KEYWORDS

Eye tracking, calibration, saliency map, model-besed eye tracking

ACM Reference Format:
Mamoru Hiroe, Michiya Yamamoto, and Takashi Nagamatsu. 2023. Implicit User Calibration for Gaze-tracking Systems Using Saliency Maps Filtered by Eye Movements. In *2023 Symposium on Eye Tracking Research and Applications (ETRA '23), May 30–June 02, 2023, Tubingen, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3588015.3589204

1 INTRODUCTION

Gaze tracking technology is not only used as a user interface for research purposes (e.g. entertainment, usability of mobile devices).. However, conventional gaze tracking systems necessitate user calibration, which requires the user to gaze at specific points on the display before the system can be utilized. Calibration is one of the most challenging obstacles to achieving seamless interactions with gaze-based systems. Although some gaze tracking systems are head-mounted, this paper deals with a desktop video-based gaze tracking system that is non-invasive and can be easily replaced by the user.

Methods that use binocular geometric constraints to perform implicit calibration have been proposed [Model and Eizenman 2010; Nagamatsu et al. 2009]. Nagamatsu et al. proposed a method in which the intersection between the optical axis of the eye and the display was calculated for both eyes and calibration was performed by assuming that the eyes were gazing at their midpoint [Nagamatsu et al. 2009]. Although this method was able to stably estimate the gaze point, it could not reliably measure the gaze point with high accuracy because of individual differences in the accuracy of gaze estimation. Model et al. proposed a calibration method to determine the displacement between the optical and visual axes of the two eyes, using the constraint that the measured visual axes of the two eyes intersect on the display [Model and Eizenman 2010]. However, this method is sensitive to the accuracy of the optical axis estimation and uses a pyramid-shaped display for stable estimation. Therefore, it is currently difficult to perform an accurate and stable calibration using an approach based on binocular geometric constraints.

Implicit calibration methods using on-screen information have been proposed [Chen and Ji 2011; Hiroe et al. 2018; Sugano et al. 2013; Wang et al. 2016]. Chen et al. proposed a method for applying saliency map-based calibration to model-based eye tracking [Chen and Ji 2011]. They formulated a model based on Bayesian estimation that expresses the correlation between ocular parameters, including the optical and visual axes of the eye, and developed a method for probabilistic estimation of individual parameters. Extending the work of Chen et al., Wang et al. proposed a method to create a fixation map trained with a regression based deep convolutional neural network and calibrated it by focusing on the distribution of gaze points [Wang et al. 2016]. To compute the distribution of gaze points, each image in the experiment was examined for 4 s. The accuracy of gaze estimation was high, but the user was required to look at the still image for a certain amount of time and complex calculations were required. Hiroe et al. proposed implicit user calibration for gaze-tracking systems using an averaged saliency map around the optical axis of the eye [Hiroe et al. 2018]. The assumption underlying this method is that individuals are more likely to gaze at salient regions near the optical axis of the eyes. In this method, single-point calibration was performed using the peak of the average of Itti's saliency maps [Itti et al. 1998] around the optical axis of the eye. Hiroe et al. improved the above method by utilizing a machine-learning-based saliency map [Hiroe et al. 2021], achieving an accuracy of approximately 1.5°. While the above methods use an infrared camera, Sugano et al. proposed an appearance-based method to estimate the gaze point by matching learning from pairs of eye images obtained from a visible-light camera and a screen saliency map [Sugano et al. 2013]. Although this method has the advantage of being able to measure gaze with a commonly used RGB camera, it is not as accurate as the model-based method.

Hiroe's method has advantages in that it can be applied to a situation in which the user is watching a video, and the method

ETRA '23, May 30–June 02, 2023, Tubingen, Germany

ACM ISBN 979-8-4007-0150-4/23/05…$15.00
https://doi.org/10.1145/3588015.3589204

is performed with a simple calculation of averaging the saliency maps. However, because this method calculates saliency maps for all frames, the computational cost is high. In this paper, we propose two methods to reduce the computational cost in Hiroe's method.

2 IMPLICIT USER CALIBRATION FOR GAZE-TRACKING SYSTEMS USING SALIENCY MAPS AROUND OPTICAL AXIS OF THE EYE

The implicit calibration method [Hiroe et al. 2018] is based on the one-point calibration method [Nagamatsu et al. 2008] that is used for estimating the offset between the visual and optical axes of the eye. This offset is called angle κ. The optical axis is defined as the axis connecting the center of the corneal curvature and pupil center of the eye, while the visual axis is defined as the axis connecting the fixation point (explicit calibration point) and fovea of the eye (Fig.1). Angle κ is estimated by calibration as a unique individual parameter.

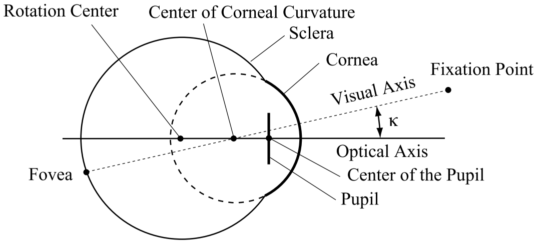

Figure 1: 3D eye model

The implicit calibration method estimates the position of the visual axis of the eye using saliency maps. The saliency maps generated for the entire display images are cropped around the optical axis of the eye; the optical axis can be estimated without user calibration [Guestrin and Eizenman 2006; Nagamatsu et al. 2010; Shih and Liu 2004]. The cropping range was 7°, 3°, and ±3° on the nose side, ear side, and vertical direction, respectively. The cropped saliency maps (Fig. 2 (a), (b)) were transformed to the coordinate system based on the optical axis of the eye by homography transformation; the range became a rectangular area (Fig. 2 (a'), (b')). Next, all the saliency maps around the optical axis of the eye were averaged. The peak of the averaged saliency map provides the gaze point in the coordinate system based on the optical axis of the eye, as shown in Fig. 2 (c). The visual axis was estimated by the line connecting the corneal center and gaze point. After the visual axis was estimated, a conventional one-point calibration was performed to obtain angle κ.

A breakthrough in Hiroe's method [Hiroe et al. 2021] is the use of state-of-the-art machine-learning-based saliency maps (UNISAL [Droste et al. 2020], MSI-Net [Kroner et al. 2020], DeepGazeⅡ [Kümmerer et al. 2017]), which achieves high accuracy. The best accuracy was 1.54° when using UNISAL. Saliency maps were computed over all frames and were aggregated to obtain a position that was more likely to be the visual axis, thus enhancing the robustness and accuracy.

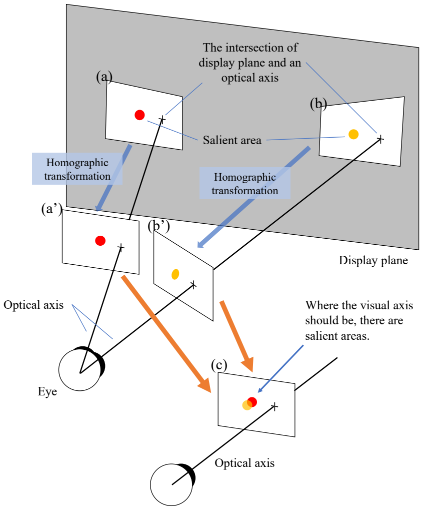
The intersection of display plane and an optical axis
(a)
(b)
Salient area
Homographic transformation
Homographic transformation
(a')
(b')
Display plane
Optical axis
Where the visual axis should be, there are salient areas.
Eye
(c)
Optical axis

Figure 2: Average of Saliency Map around the Optical Axis of the Eye.

3 FILTERING OF DATA BY ESTIMATING EYE MOVEMENT FROM THE OPTICAL AXIS

In the method of Hiroe et al., the estimation of the optical axis and the estimation of the gaze point with the transformation and averaging of the saliency maps do not require as much computational cost; however, the cost of generating the saliency map makes the real-time calibration difficult. If the computational effort could be reduced, it would be feasible to perform implicit calibration online in the background.

In this study, we propose two methods for filtering gaze data in order to reduce the number of saliency maps used in the estimation of the gaze point. This is achieved by identifying the type of eye movement based on the movement of the optical axis and selecting the frames to be used in the estimation. Eye movements can be categorized into several types; however, we focus on fixation, where the user is definitely looking at an object.

3.1 Filtering method with focusing on the speed of eye movement (Velocity Filter)

This method filters saliency maps when the user's eyes produce saccades. During saccades, the user might not be looking at a particular object but might be searching for objects of interest. Excluding data from this time reduces the number of saliency maps to be generated. We defined a saccade as a movement of the optical axis by $30^\circ/s$ or more. An image diagram of this method is shown in Fig. 3. The top graph shows the relationship between the velocity of movement of the optical axis and time. The lower bar represents the frames of the saliency maps used for calibration. The time when the optical axis moves more than $30^\circ/s$, represented by the blue dotted lines, are considered as saccades and the saliency maps for these frames (shown as grey parts of the lower bar) are excluded from the calibration calculation.

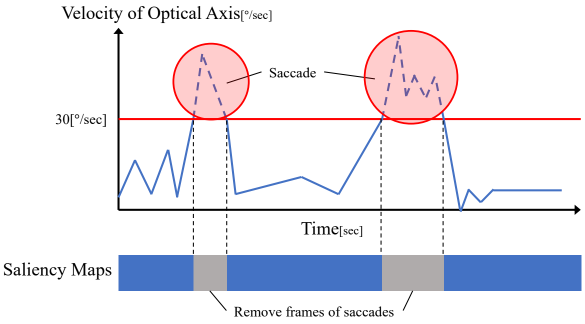

Figure 3: Velocity filter

3.2 Filtering method with focusing on the position and duration (Fixation Filter)

Explicit calibration calculates the κ angle from the difference between the presented gaze point and optical axis. Therefore, we believe that implicit calibration could be performed efficiently by collecting only the difference between the position of the optical axis and the gaze objects as gaze points during fixation. Thus, we proposed using for calibration only the saliency map of the first frame for each fixation, assuming that the user was fixating if the gaze point was within 1.0° for 0.3s or longer. An image diagram of this method is shown in Fig. 4. The top graph shows the relationship between the rotation angle of the optical axis and time. The lower bar represents the frame of the saliency map used for calibration. The lines represented by the solid blue line are estimated as fixations, and only the saliency map of the first frame of each fixation (shown as red parts of the lower bar) is used in the calibration calculation.

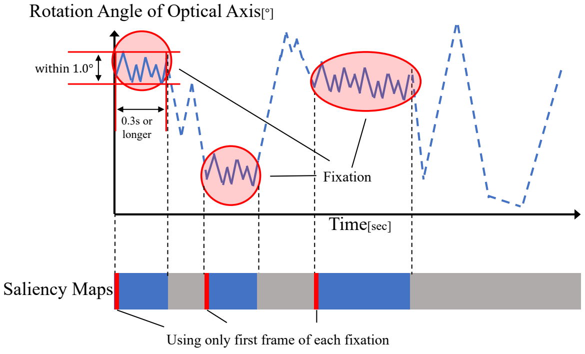

Figure 4: Fixation filter

4 EVALUATION

4.1 System

A prototype was developed, as shown in Fig. 5, consisting of two monochrome GigE digital cameras (HXG20NIR, Baumer GmbH), three monitors, and a Windows-based PC (Windows 7). One monitor was for the participant (19" LCD) and the remaining monitors were for the experimenter. Each camera was equipped with a 2/3" CMOS image sensor with a resolution of 2048 × 1088 pixels. A 16 mm lens and a visible light cut-off filter were attached to each camera, which were positioned under the monitor. IR LEDs were mounted on the left and right sides of the monitor. The software was developed using OpenCV in C++. The pupil diameter in the captured image was approximately 30 pixels. The camera parameters were determined beforehand.

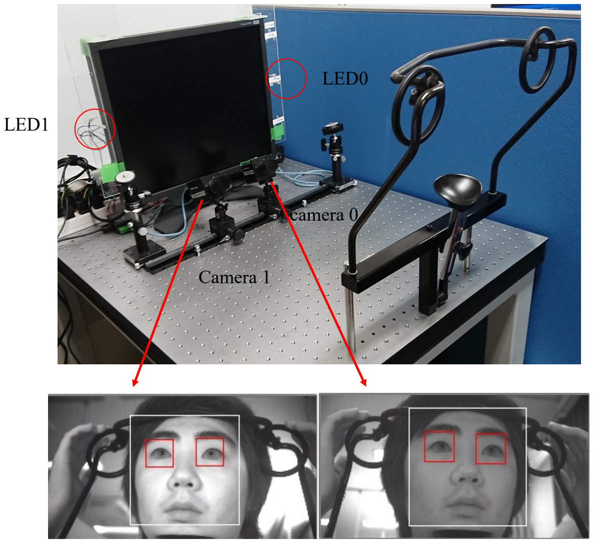

Figure 5: Eye tracking system used in experiment

4.2 Method

We conducted an experiment to compare and evaluate the two filtering methods proposed in Section 3 with the method using all frames of the saliency map and the single-point calibration method [Nagamatsu et al. 2008] as the baseline.

The participants were seven adult men (participants A-G), with only Participant A wearing soft contact lenses(six of whom were in their 20s and one in his 40s). The stimuli comprised three videos from the YouTube-8M Dataset (https://research.google.com/youtube8m/) in the order Video 1 to 3. Video 1 was a TV news program with studio scenes and interviews lasting 198 s. Video 2 was a report about a professional basketball team, including scenes from games, with a duration of 150 s. Video 3 was a trailer of a fantasy film with a duration of 235 s. Although the proposed method can work when the head is moving, the participants' heads were stabilized with a chin rest during the experiment to minimize image-processing errors. The participants' eyes were positioned approximately 600 mm from the monitor. All three videos were presented to each participant, who was instructed to view each video freely.

4.3 Result

Table 1 shows the error angles of gaze estimation when the participants look at nine points on a screen calibrated with angle

κ calculated using each method and each saliency map. Except for the single-point calibration, values are the mean of error angle calculated in estimating the visual axis when calibrated with three different stimuli (video 1-3). Looking at the averages across all participants, the gaze estimation accuracy when implicit calibration with saliency maps was performed for frames filtered by the proposed methods were comparable to that when calibration with saliency maps was performed for all frames. Focusing on the individual values, it can be noted that the fixation filter method increased the accuracy of six out of seven participants, while the velocity filter method increased the accuracy of four out of seven participants.

Table 2 lists the number of frames used for calibration reduced by the proposed method. The number of frames after filtering with the proposed method is shown as a percentage, with 100% for all frames. Averaged over all participants, the velocity filter method used approximately 90% of the saliency maps for calibration, i.e. approximately 10% of the maps were reduced. The fixation filter method used approximately 13% of the saliency maps for calibration, i.e. approximately 87% of the maps were reduced. Therefore, significant computational savings can be achieved with the fixation filter method.

5 DISCUSSION

The velocity filter method focused on reducing the amount of data not needed for calibration, while maintaining as much data as possible. On the other hand, The fixation filter method focused on reducing the number of saliency maps as much as possible to reduce computational cost. The results show that there is little difference in accuracy between the two methods. The fixation filter method used significantly fewer saliency maps in its computations. From these results, we conclude that the fixation filter method was more effective.

The proposed methods were expected to produce results close to the accuracy of single-point calibration, but there was no clear improvement in accuracy. The implicit calibration method is based on the single-point calibration method, which is usually applied when the user is gazing at an explicit point near the center of the screen. The proposed method did not consider the position of the object the user was gazing at. Accuracy could be improved by prioritising data when the user is gazing close to the center. Furthermore, accuracy could also be improved by balancing the use of data when the user is looking in different directions when the user's eyes are rotating widely (not looking at the center). There is a problem that the saliency map methods, which increase in accuracy with the proposed method, differ from one participant to another. In the future, we would like to solve this problem by developing a method for generating saliency maps specifically for calibration, rather than saliency maps for estimating gaze distribution.

In this study, evaluation experiments were carried out on a desktop video-based system with videoes on the screen, but the method is applicable to any eye measurement system calibrated using the difference between the optical and visual axes method. In head-mounted systems, the video captured by the world camera corresponds to the videoes.

6 CONCLUSION

The implicit calibration method proposed by Hiroe et al. could be applied even when the user is watching a video, but it was difficult to apply in real-time due to the problem of the high computational cost of generating many saliency maps. In order to improve the computational cost of the implicit calibration method proposed by Hiroe et al, we proposed two methods to filter the data by estimating the type of eye movement from the movement of the optical axis, reducing the number of saliency maps generated while maintaining accuracy. We achieved a reduction of approximately 10% for the velocity filter method and of approximately 87% for the fixation filter method. Both filtering methods were able to estimate angle κ with the same level of accuracy as that when using saliency maps for all frames.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP20H04229 and JP22K21317.

REFERENCES

J. Chen and Q. Ji. 2011. Probabilistic gaze estimation without active personal calibration. In *CVPR 2011*. 609–616.

Richard Droste, Jianbo Jiao, and J. Alison Noble. 2020. Unified Image and Video Saliency Modeling. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*.

Elias Daniel Guestrin and Moshe 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.

Mamoru Hiroe, Michiya Yamamoto, and Takashi Nagamatsu. 2018. Implicit user calibration for gaze-tracking systems using an averaged saliency map around the optical axis of the eye. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Article No.56*. ACM, Article 56, 5 pages.

Mamoru Hiroe, Michiya Yamamoto, and Takashi Nagamatsu. 2021. Implicit User Calibration Method for Gaze-tracking Systems using Saliency Maps around the Optical Axis of Eye. *the Transaction of Human Interface Society* 23, 4 (11 2021), 431–442. https://doi.org/10.11184/his.23.4_431

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions Pattern Anal. Mach. Intell.* 20, 11 (1998), 1254–1259.

Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. 2020. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks* 129 (2020), 261–270.

M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. 2017. Understanding Low- and High-Level Contributions to Fixation Prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 4799–4808.

Dmitri Model and Moshe Eizenman. 2010. User-calibration-free remote gaze estimation system. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 29–36.

Takashi Nagamatsu, Yukina Iwamoto, Junzo Kamahara, Naoki Tanaka, and Michiya Yamamoto. 2010. Gaze Estimation Method based on an Aspherical Model of the Cornea: Surface of Revolution about the Optical Axis of the Eye. In *Proceedings of the 2010 Symposium on Eye Tracking Research & Applications*. 255–258.

Takashi Nagamatsu, Junzo Kamahara, Takumi Iko, and Naoki Tanaka. 2008. One-point Calibration Gaze Tracking Based on Eyeball Kinematics Using Stereo Cameras. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*. 95–98.

Takashi Nagamatsu, Junzo Kamahara, and Naoki Tanaka. 2009. Calibration-free gaze tracking using a binocular 3D eye model. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*. ACM, 3613–3618.

Sheng-Wen Shih and Jin Liu. 2004. A novel approach to 3-D gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 34, 1 (2004), 234–245.

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2013. Appearance-Based Gaze Estimation Using Visual Saliency. *IEEE Transactions Pattern Anal. Mach. Intell.* 35, 2 (2013), 329–341.

Kang Wang, Shen Wang, and Qiang Ji. 2016. Deep Eye Fixation Map Learning for Calibration-free Eye Gaze Tracking. In *Proceedings of the 2016 ACM Symposium on Eye Tracking Research & Applications*. ACM, 47–55.

Table 1: Error angle of gaze estimation calibrated with angle κ calculated by method and saliency map (The value in bold type indicates the case where the accuracy is better than when all frames are used.).

Method	Velocity filter			Fixation filter			All frames			Single-
Saliency map	UNISAL	MSI-Net	DeepGazeⅡ	UNISAL	MSI-Net	DeepGazeⅡ	UNISAL	MSI-Net	DeepGazeⅡ	point
ParticipantA	**2.01**°	**1.99**°	2.04°	**1.89**°	**1.97**°	**1.90**°	2.04°	2.02°	2.01°	1.83°
ParticipantB	1.17°	1.35°	1.24°	1.19°	1.23°	1.23°	1.16°	1.09°	1.05°	0.96°
ParticipantC	**1.44**°	1.61°	**1.49**°	**1.45**°	1.58°	**1.43**°	1.53°	1.51°	1.61°	0.87°
ParticipantD	2.08°	**2.06**°	2.12°	2.19°	2.26°	2.28°	2.07°	2.15°	2.05°	2.03°
ParticipantE	**0.88**°	**0.93**°	1.02°	1.06°	1.24°	1.06°	0.89°	0.96°	0.95°	0.91°
ParticipantF	1.25°	**1.24**°	**1.47**°	1.33°	**1.32**°	**1.34**°	1.25°	1.35°	1.52°	1.11°
ParticipantG	**2.26**°	**1.96**°	2.34°	**2.19**°	2.13°	2.01°	2.27°	2.03°	2.19°	1.26°
Average	**1.58**°	1.59°	1.67°	1.62°	1.68°	**1.61**°	1.60°	1.59°	1.62°	1.28°

Table 2: Percentage of the number of saliency maps used for calibration.

	Velocity filter	Fixation filter
ParticipantA	87.7%	12.5%
ParticipantB	90.3%	12.0%
ParticipantC	90.1%	12.5%
ParticipantD	90.8%	11.3%
ParticipantE	90.8%	13.6%
ParticipantF	91.1%	12.6%
ParticipantG	88.3%	13.8%
Average	89.9%	12.6%

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009