
Implicit User Calibration for Model-based Gaze-tracking System using Face Detection around Optical Axis of Eye

Mamoru Hiroe
Kobe University
Kobe, Japan
173w107w@stu.kobe-u.ac.jp

Shogo Mitsunaga
Kobe University
Kobe, Japan
188w107w@stu.kobe-u.ac.jp

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

ABSTRACT

In recent studies of gaze tracking system using 3D model-based methods, the optical axis of the eye is estimated without user calibration. The remaining problem for achieving implicit user calibration is to estimate the difference between the optical axis and visual axis of the eye (angle κ). In this paper, we propose an implicit user calibration method using face detection around the optical axis of the eye. We assume that the peak of the average of face region images indicates the visual axis of the eye in the eye coordinate system. The angle κ is estimated as the difference between the optical axis of

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5971-9/19/05.

<https://doi.org/10.1145/3290607.3312942>

KEYWORDS

eye tracking, calibration, face detection

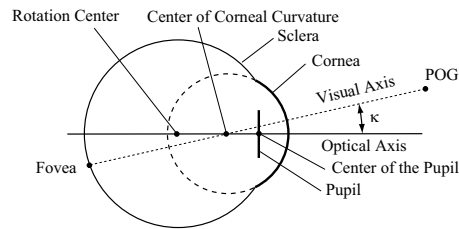


Figure 1: 3D eye model

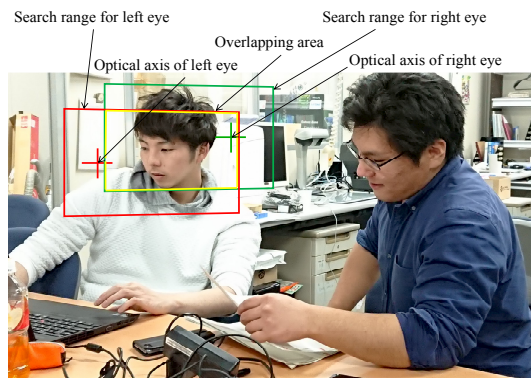


Figure 2: Search range of angle κ for left and right eyes, and overlapping area

the eye and the peak of the average of face region images. We developed a prototype system with two cameras and two IR-LEDs. The experimental results showed that the proposed method can estimate the angle κ more accurately than the method that uses Itti's saliency map instead of face detection.

INTRODUCTION

Gaze-tracking technology serves as a user interface. However, conventional gaze-tracking systems require user calibration which requires the user to gaze at specific points on the screen before the system can be used. Calibration is one of the biggest obstacles to seamless interaction with gaze-based systems. In recent studies using 3D eye model as shown in Figure 1, the optical axis (pupillary axis) of the eye was estimated without user calibration [2]. There is a difference between the optical axis and visual axis (line of sight) of the eye. This difference is called the angle κ . We refer to a method estimating the angle κ without active user participation as implicit calibration. The remaining problem for achieving implicit user calibration is to estimate the angle κ automatically.

In order to estimate the angle κ , a method that exploits the binocular constraint (the visual axes of both eyes intersect each other on a display) was proposed [7]; however, it is sensitive to noise. Another approach is to use the information displayed on the screen. Model and Eizenman [6] proposed a calibration method for infants based on the assumption that when a small attractive stimulus is presented, there is a higher probability that the infant will look at the stimulus. However, this is not an implicit calibration in the strictest sense, because they controlled the position of the stimulus. Hiroe et al. [3] proposed an implicit calibration method using a saliency map around the optical axis of the eye; we call this method a saliency method. Their assumption was that the user is more likely gazing at the salient region near the optical axis of the eye. In their method, a single-point calibration was done using the averaged Itti's saliency map [4] around the optical axis of the eye. However, the salient points were not always good predictions of gaze; there are many objects that attract human attention but have low saliency.

Human faces are known to be one of the most prominent object that attract human attention. Therefore, considering only when the optical axis is near a face, we assume the user gazes at the face.

In this paper, we propose an implicit user calibration method for gaze-tracking systems using face detection.

NEW USER CALIBRATION METHOD USING FACE DETECTION ON DISPLAY IMAGE

We propose a method that estimates the angle κ (horizontal: α , vertical: β) using face detection around the optical axis of the eye. The optical axis of the eye can be estimated using the model-based method [2]. Figure 2 shows a scene of discussing together that are displayed on a computer display. The cross indicates the intersection of the optical axis of the each eye and the display plane

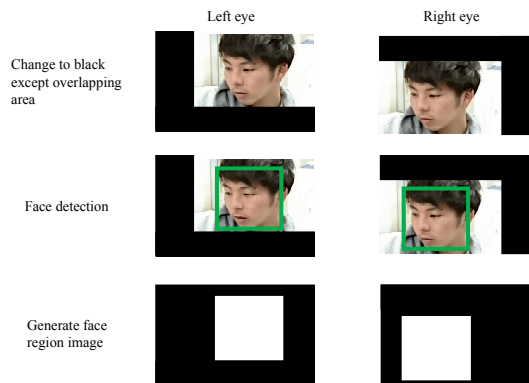


Figure 3: Processing cropped images

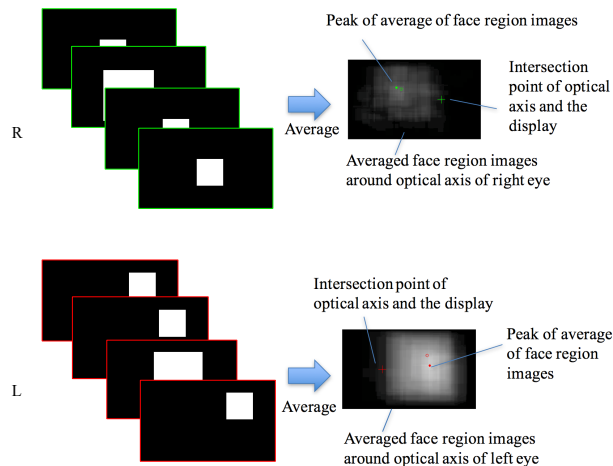


Figure 4: Average of face region images around the optical axis of the eye

(green : right eye, red : left eye). The red and green quadrilaterals show the search range of gazing point around the optical axis of the left and right eye, respectively. We decided that the search ranges around the optical axes of both eyes are from -3 to 7 degrees for the right eye (-7 to 3 degrees for the left eye) horizontally, and -3 to 3 degrees vertically as same as Hiroe’s work [3] .

After estimating the optical axis of the eye, the search range is cropped and transformed into the each eye coordinates. In Hiroe’s method, they just cropped the image around the optical axis of the eye from the display image in just rectangles. Since Hiroe’s method did not take into account the perspective transformation, those cropped images were not the images the user actually looked at. Therefore, we perform homography transformation in order to convert the cropped images to the images of the user’s viewpoint. The calculation takes into account Listing’s law. The search ranges move according to movement of the optical axes of both eyes.

Same as in Hiroe’s method [3] we used both eyes restriction (both eye gaze at the same point). Figure 3 shows image processing for left and right eyes that were cropped and homography-transformed. When a human gazes at something on a display, both eyes are directed toward the same object. Therefore, the search ranges of κ for both eyes can be reduced to the overlapping search range of both eyes. We change the value of the image where the search range for both eyes do not overlap with each other to 0 (black), as shown in the top row of Fig.3. Next, we detect faces in the images as shown in the middle row of Fig. 3. Afterwards, we convert the images so that the face area is white and the rest is black, as shown in the bottom row of Fig. 3. We call these black and white images as the “face region image.”

Then, we calculate the average images of the face region images as shown in Fig.4. We determine the visual axis of the eye based on the point of maximum value of the averaged face region images. If at least one of the face region images at a certain time has no white area, then the face region images at that time are not used to calculate the averaged face region images. This is because the optical axes of both eyes may not be detected correctly, or face detection may fail. With this selection, the probability that the wrong data is used can be decreased.

The difference between the peak of average of face region images and the point that the optical axis directs is considered the angle κ for each eye.

EVALUATION

System

A prototype system was implemented, as shown in Figure 5. This system consists of two monochrome GigE digital cameras (HXG20NIR, Baumer GmbH), three displays, and a Windows-based PC (Windows 7). One display is for the participant (a 19” LCD), and the others are for the experimenter. Each

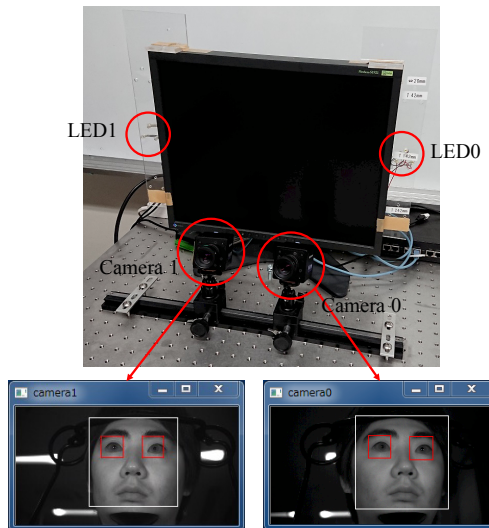


Figure 5: Developed system

camera is equipped with a $2/3''$ CMOS image sensor with a resolution of 2048×1088 pixels. A 16-mm lens and a visible light cut filter were attached to each camera. These cameras were positioned under the display. IR-LEDs were attached to the left and right sides of the display, and the positions were measured. The camera parameters were determined beforehand. The software was developed using OpenCV in C++ language. The diameter of the pupil in the captured image is approximately 30 pixels.

Method

We conducted an experiment to compare the single-point calibration (base line) [8], our proposed, and the saliency methods [3]. The participants were six adults (Participants A–F). Only participant B wore soft contact lenses. The stimuli were three videos included in the YouTube-8M Dataset [1]. Video 1 is a TV news show that includes studio scenes and interviews. The length is 198 s. Video 2 is the coverage of a professional basketball team including scenes from games. The length is 150 s. Video 3 is the trailer of a fantasy movie. The length is 235 s. Although the proposed method can work when the head moves, the participants' heads were supported by a chin rest during the experiment to reduce errors caused by image processing. The participants' eyes were approximately 600 mm from the display. All three videos were presented to each participant, who was asked to look at the display freely for each video.

Results

Figure 6 shows the estimation of the angle κ (α_L , β_L , α_R , and β_R) in degrees for 6 participants. The graphs are plotted in the eye coordinate system and the origin indicates the optical axis of the eye. The horizontal axis indicates α while the vertical axis indicates β in degrees. The cross indicates the α and β estimated by a single point calibration. In the calibration process, the participants intentionally gazed at a single point to calibrate the angle κ , i.e., α and β . The red and blue square indicate the estimated values when participants looked at video 1 using our proposed and the saliency methods, respectively. The red and blue triangle indicate the estimated values when participants looked at video 2. The red and blue x-mark indicate the estimated values when participants looked at video 3.

Figure 7 shows the comparison of the error angles between the single-point calibration, our proposed, and the saliency methods for the left and right eyes when the participants looked at different stimuli. The data are the average error angle when the participants gazed at a grid of nine points (± 8.9 degrees horizontally, ± 7.1 degrees vertically) on the display. The red and blue-tinged color graphs indicate our proposed and the saliency methods, respectively. In each color series, 1, 2, and 3 correspond to the error angles in cases where each participant watched videos 1, 2, and 3. The average error angles with the single-point calibration, our proposed, and the saliency methods were 1.25, 1.71, and 2.65 degrees, respectively.

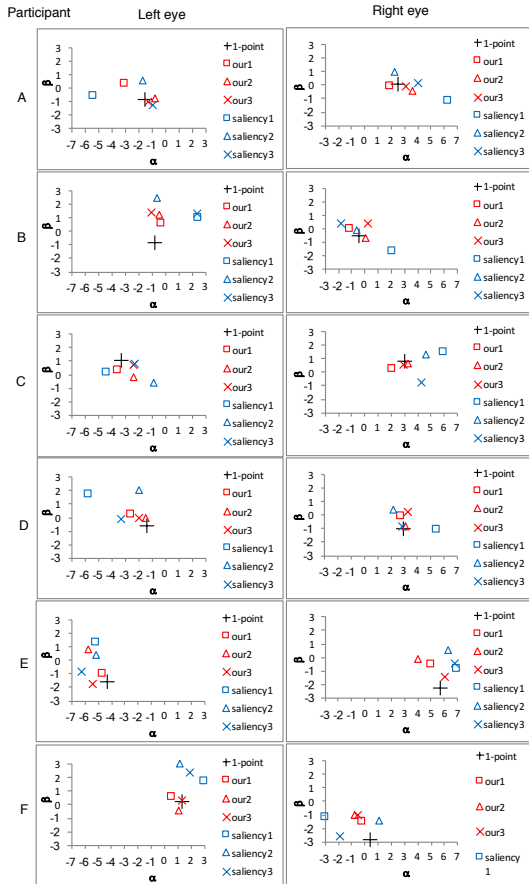


Figure 6: Estimation of α_L , β_L , α_R , and β_R in the eye coordinate system in degrees. The black crosses, red symbols, and blue symbols indicate the angle κ (α and β) estimated by the single point calibration, our proposed, and the saliency methods, respectively.

Table 1 and Table 2 show the converged time of the saliency and our proposed methods, respectively. We judged that it converged when all α s and β s for each case reached a value within 0.5 degree from the final value. The number of face detection is shown in Table 3. The number indicates the number when the system could detect faces in both images of search range around the optical axis of the eye.

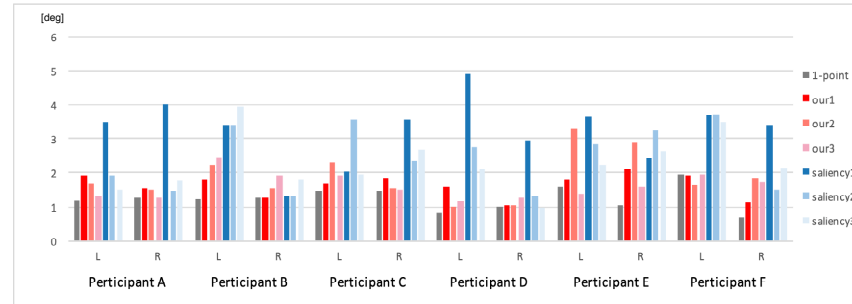


Figure 7: Comparison between the error angles of single-point calibration, our proposed, and the saliency methods with different stimuli. The black, red, and blue bars indicate the errors with the single point calibration, our proposed, and the saliency methods, respectively.

Discussion

By averaging simple rectangles that indicate the face position in eye coordinate system, we could estimate the angle κ .

Since the effectiveness of our proposed method depends on the stimulus that the participants looked at, we conducted the experiment using three kinds of stimuli (videos). As shown in Fig. 6, the estimated α and β with our proposed method (red symbols) are similar to the α and β determined by single-point calibration (black cross), but the estimated α and β with the saliency method (blue symbols) are relatively scattered. In addition to this, from Fig. 7, the error angles of our proposed method is smaller than the saliency method in most cases. Therefore, we think if face is detected, our proposed method is more accurate than the saliency method.

As for the converged time, our proposed method took longer time than the saliency method comparing Table 1 and Table 2. Our proposed method only uses frames that include faces, but the saliency method uses all frame images. Therefore, the saliency method converged fast. However, our proposed method is more accurate than the saliency method as shown in Fig. 6 and Fig. 7.

Table 1: Converged time using the saliency method [sec.].

participant	video1 (news)	video2 (sports)	video3 (movie)
A	26	52	36
B	34	36	20
C	25	41	22
D	36	12	33
E	26	31	41
F	20	28	36

Table 2: Converged time using our proposed method [sec.].

participant	video1 (news)	video2 (sports)	video3 (movie)
A	172	76	88
B	66	40	211
C	154	67	91
D	188	113	102
E	188	85	147
F	161	115	229

Table 3: The number of face detection.

participant	video1 (news)	video2 (sports)	video3 (movie)
A	127	52	47
B	123	50	52
C	78	58	70
D	281	125	120
E	14	17	9
F	30	31	27

The number of face detection is affected by the size of the angle κ . If the angle κ is large as the participant E, the overlapping search range becomes small and the number of face detection decreases.

The saliency method by Hiroe et al. uses Itti's saliency map, we will compare to the saliency method that uses the state-of-the-art saliency map in future work.

There is a limitation of the proposed method. In cases where the face image does not appear on the display, the proposed method is not effective. In order to solve this problem, we will introduce other probable fixation targets [5] such as texts, icons, logos, etc. Furthermore, we can use a saliency map, when there are no probable fixation targets.

CONCLUSION

We proposed an implicit user calibration method for gaze-tracking systems using face detection around the optical axis of the eye. We assumed that the peak of the average of face region images indicates the visual axis of the eye in the eye coordinate system. The experimental results showed that the proposed method can estimate the angle κ more accurately than the method that uses Itti's saliency map.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 16H02860.

REFERENCES

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv:1609.08675 [cs.CV]* (2016), <http://arxiv.org/abs/1609.08675>.
- [2] 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.
- [3] 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*. ACM, Article 56, 5 pages.
- [4] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 11 (1998), 1254–1259.
- [5] Pawel Kasprowski and Katarzyna Harezlak. 2018. Comparison of mapping algorithms for implicit calibration using probable fixation targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 1–8.
- [6] Dmitri Model and Moshe Eizenman. 2010. An Automated Hirschberg Test for Infants. *IEEE Transactions on Biomedical Engineering* 58, 1 (2010), 103–109.
- [7] 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.
- [8] Takashi Nagamatsu, Junzo Kamahara, and Naoki Tanaka. 2008. 3D Gaze Tracking with Easy Calibration Using stereo Cameras for Robot and Human Communication. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN) 2008*. 59–64.