Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency

Yusuke Sugano
Perceptual User Interfaces Group
Max Planck Institute for Informatics
sugano@mpi-inf.mpg.de

Andreas Bulling
Perceptual User Interfaces Group
Max Planck Institute for Informatics
bulling@mpi-inf.mpg.de

ABSTRACT

Head-mounted eye tracking has significant potential for gaze-based applications such as life logging, mental health monitoring, or the quantified self. A neglected challenge for the long-term recordings required by these applications is that drift in the initial person-specific eye tracker calibration, for example caused by physical activity, can severely impact gaze estimation accuracy and thus system performance and user experience. We first analyse calibration drift on a new dataset of natural gaze data recorded using synchronised video-based and Electrooculography-based eye trackers of 20 users performing everyday activities in a mobile setting. Based on this analysis we present a method to automatically self-calibrate head-mounted eye trackers based on a computational model of bottom-up visual saliency. Through evaluations on the dataset we show that our method 1) is effective in reducing calibration drift in calibrated eye trackers and 2) given sufficient data, can achieve gaze estimation accuracy competitive with that of a calibrated eye tracker, without any manual calibration.

Author Keywords

Mobile Eye Tracking, User Calibration, Calibration Drift, Electrooculography, Visual Saliency

ACM Classification Keywords

H.5.m Information Interfaces and Presentation (e.g. HCI): Miscellaneous; I.4.9 Image Processing and Computer Vision: Applications

INTRODUCTION

Gaze is a compelling modality for human-computer interaction [28] and explicit gaze interaction techniques on, e.g., hand-held and ambient displays [35, 39, 40] have been studied extensively over many years. The recent advent of lightweight head-mounted eye trackers is starting to cause a paradigm shift towards gaze interaction in daily-life settings and applications in which human gaze behaviour is analysed continuously over hours or even days [10, 34]. Gaze behaviour analysis has significant potential for a range of HCI applications, such as human activity and context recognition [8, 9, 38], life logging [20, 19], mental health monitoring [41], or for quantifying everyday reading behaviour [26].

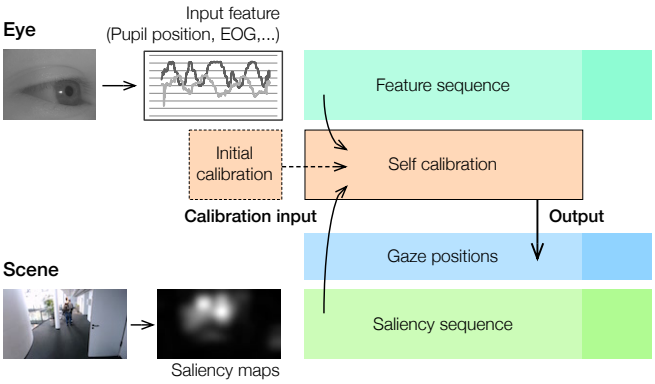


Figure 1: Our method automatically calibrates a head-head mounted eye tracker using gaze input features and visual saliency maps calculated on the egocentric scene video.

One major limitation of state-of-the-art head-mounted eye trackers is that they have to be calibrated to each user prior to first use. This initial calibration typically requires the user to look at a set of predefined targets to establish a mapping from eyeball rotations to gaze positions in the user's visual scene. A second limitation is that this calibration is currently assumed to be static, i.e. not to change (drift) over time. Particularly in mobile daily-life settings, however, shifts of the eye tracker on the head, e.g. caused by physical activities or the user touching or even taking off the eye tracker, are very likely to occur and can cause significant calibration drift, thereby considerably reducing gaze estimation accuracy. To address the problem of calibration drift, the eye tracker could be recalibrated frequently. However, frequent recalibration would be time-consuming and disruptive and therefore impractical for real-world use.

In this work we propose a novel method to self-calibrate head-mounted eye trackers, i.e. to establish an accurate mapping of pupil positions to gaze positions in the visual scene without the need for any explicit (re-)calibration. In contrast to existing calibration routines that use a single initial calibration, our method is designed to run continuously and update the gaze mapping transparently in the background. While such a continuous calibration approach is taken in other wear-

UIST 2015, November 8–11, 2015, Charlotte, NC, USA.

ACM ISBN 978-1-4503-3779-3/15/11 ...$15.00.
http://dx.doi.org/10.1145/2807442.2807445

able systems such as eye alignment for optical see-through displays [22, 31], it has not yet been fully investigated in the mobile eye tracking literature. This approach has two distinct advantages over state-of-the-art methods that only use an initial calibration: Our approach does not require the user to perform any time-consuming and cumbersome explicit calibration actions, and it embraces the inevitable calibration drift and compensates for it permanently.

The core of our method is a visual saliency model that simulates the perceptual mechanisms of human bottom-up visual attention and computes 2D probability maps of where in the visual scene the user is more likely to fixate on [6]. Although even state-of-the-art saliency models cannot perfectly locate exact gaze positions, the statistical correlation between gaze positions and saliency values can be enhanced by aggregating multiple maps [36]. The proposed robust mapping method uses aggregated maps to either compensate for drift in the initial eye tracker calibration or to establish a gaze mapping from scratch over time (see Figure 1).

To better understand calibration drift and the real-world applicability of the saliency-based calibration approach, we present a new gaze dataset of 20 users performing everyday activities in a naturalistic scenario. Our dataset contains synchronised egocentric videos as well as gaze data from a head-mounted video-based and an Electrooculography (EOG)-based eye tracker. We quantify calibration drift on the dataset and compare the performance of the proposed self-calibration method with a standard calibration routine. We show that our method is effective in reducing drift in calibrated eye trackers and given sufficient data, can achieve competitive gaze estimation accuracy without any initial eye tracker calibration.

The specific contributions of this work are threefold. First, we characterise calibration drift during daily-life recordings on a new dataset. Second, based on this analysis, we propose a novel method for eye tracker self-calibration using visual saliency models. To address challenges of saliency prediction in mobile settings, we further introduce a novel robust mapping method that approximates the mapping task as a rotation alignment. Finally, we evaluate our method for both video and EOG-based self-calibration in two scenarios with and without initial user-specific calibration.

RELATED WORK

Head-Mounted Eye Tracking

The two most common approaches for head-mounted eye tracking are video-based and electrooculography (EOG)-based (see Figure 2 for a sample video-based and EOG-based system). Video-based approaches provide more accurate gaze estimates and are therefore the most popular method to date [17]. Most recently the trend is toward to low-cost devices, including open-source hardware [25]. Video-based eye trackers typically track pupil positions using an eye camera and map them to gaze positions in the scene camera. This mapping has to be established using a user-specific calibration routine and is currently assumed to be a static relationship. As for stationary remote eye trackers, it is possible to eliminate the effect of changes in eye and head position via 3D gaze estimation. However, such techniques require additional hardware, such as multiple light sources [16], and they often assume a fixed physical relationship between the eye camera and the stimuli (i.e. the scene camera in mobile settings). Hence, they pose strict constraints on the hardware design and are more difficult to implement. In contrast, our self-calibration approach does not rely on additional equipment and can therefore be easily applied to head-mounted eye trackers.

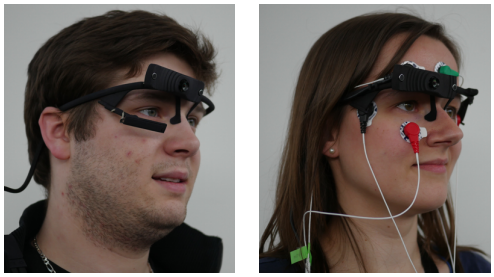


Figure 2: Left: video-based Pupil Pro eye tracker, right: EOG-based tracker based on the TMSI Mobi6.

Electrooculography measures changes in the electric potential field caused by eyeball rotations using electrodes placed around the eye. The resulting signal can be used to track relative eye movements. The key advantage over video is that EOG only requires lightweight signal processing. It can therefore be implemented as a self-contained wearable device, perfectly suited for long-term recordings in daily life [29, 7]. Recently, EOG was implemented in the first commercial glasses-type device [21].

Calibration-Free Gaze Estimation

Since the requirement for eye tracker calibration can be a major technical hurdle for gaze interaction, calibration-free tracking techniques have been a key topic of research interest. For instance, if the 3D pose of the eyeball with respect to the target plane can be directly measured, it is possible to estimate gaze directions [33, 30]. However, as discussed above, such techniques require specialised hardware and 3D information is not always available in mobile settings. As a more general alternative approach for calibration-free gaze estimation, several works used bottom-up visual saliency maps (or actual human fixation patterns) to calibrate eye trackers [36, 13, 1]. However, these methods are designed for stationary remote systems, and they cannot be directly applied to the mobile setting where the performance of saliency models significantly decreases. To the best of our knowledge, this is the first work to evaluate the idea of saliency-driven eye tracker calibration (1) in a daily-life mobile setting and (2) with a generalised formulation that can be applied to both video-based and EOG-based methods. To this end, we propose a novel robust mapping approach based on an error analysis on our real-world dataset.

Characterisation of Gaze Estimation Error

Figure 3: Sample images recorded using the egocentric camera of the video-based eye tracker showing the large variability in environments and activities. Faces were obfuscated for privacy reasons.

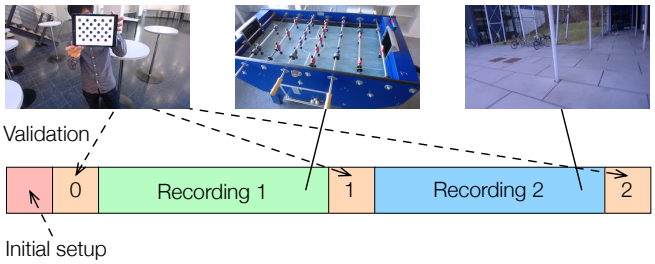


Figure 4: Data collection protocol. Each recording consisted of two recording and three validation sessions. Two recording sessions took place at different locations, and the whole data collection session took about 1 hour.

Although state-of-the-art gaze estimation techniques are claimed to achieve less than one degree error, in practical settings there are several different error sources that affect gaze recording quality. A couple of recent works analysed gaze estimation error arising from different sources, such as disparity and physiological differences [18, 14, 3]. Most similar to this work, John et al. [23] discussed an error model of mobile eye trackers and presented a post-processing method for error correction. However, their experimental setting was still limited to a laboratory setting. Calibration drift in daily-life recording therefore remains unexplored in eye tracking and gaze interaction research. This work presents the first principled study of gaze estimation error using a novel dataset collected during mobile eye tracking recording sessions.

ANALYSIS OF GAZE ESTIMATION ERROR

We first conducted a data collection to study how calibration drift affects eye tracking accuracy in real-world settings. We collected data from both a video-based and EOG-based eye tracker simultaneously (see Figure 2) and compared the initial calibration results with ground-truth gaze positions.

Data Collection

A total of 20 participants (10 female) participated in the data collection. As illustrated in Figure 4, each participant conducted three validation sessions and two recordings at two different locations. We further divided 20 participants into two groups of 10 participants which did not share recording locations. Hence, the whole dataset contains four different locations, including one outdoor location. For the recording sessions, the only instruction provided to the participants was the location where the recording took place, and that they could behave freely. Accordingly, the data contains various daily-life activities, such as walking, talking with other persons, using mobile phones, or reading posters, leaflets, and public displays (see Figure 3). Each recording session took at least 15 minutes; the whole data collection took about 1 hour per participant.

Error measurements

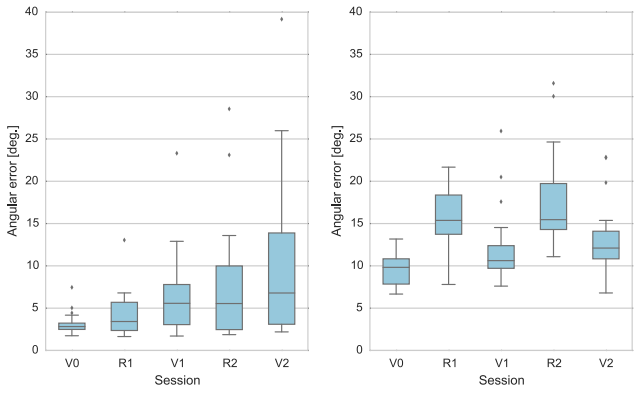
In the validation sessions we showed 3D fixation targets to evaluate gaze estimation errors. Participants were instructed to look at the differently colored circle in the camera calibration pattern (Figure 4), and the camera calibration routine recovered both the 3D pose of the calibration pattern and the camera projection matrix.

For the validation sessions, gaze estimation error is directly calculated from the recovered 3D positions of the gaze targets. Recovery of 3D gaze vectors from estimated 2D gaze positions was done using the camera calibration information at validation sessions, and the angle from the 3D gaze targets was evaluated as the error metric. For the recording sessions, 2D ground-truth gaze positions were obtained by linearly interpolating recalibration results from two validation sessions before and after each recording session (e.g. V1 and V2 are used for R2). Then, 2D pixel errors were converted to 3D angular errors based on the camera calibration information.

Apparatus

Figure 2 shows the recording setup. For the video-based setting, we used a Pupil Pro eye tracker [25]. The headset features a 720p world camera and an infrared eye camera with an adjustable camera arm. Egocentric videos were recorded using the world camera and synchronised via hardware timestamps. The Pupil recording software automatically detects pupil positions in the eye videos. The calibration is done by showing fixation targets to the participant and the default calibration routine establishes a polynomial mapping between pupil and gaze positions. We used physical markers at different distances from the participant.

For EOG data collection we used a TMSi Mobi6 that transferred the data via Bluetooth to a laptop running the Context Recognition Network (CRN) toolbox for data logging [2]. At each self-calibration, baseline drift was corrected by subtract-



(a) Video-based setting (b) EOG-based setting

Figure 5: Calibration drift statistics for (a) video-based and (b) EOG-based settings.

(a) Eye video frame at session V0 (b) Eye video frame at session V2

Figure 6: Example eye video frames of the participant with the largest calibration drift.

ing the running average of the input EOG signal. Blinks were removed in the same manner as described in Bulling et al. [8].

Drift Characteristics

The box plots in Figure 5 provide an overview of the calibration drift for all 20 participants for the video-based (5a) and EOG-based (5b) eye tracker for the three validation sessions (V0, V1, V2) and two recording sessions (R1, R2).

As can be seen from Figure 5a, the gaze estimation error clearly increased over the course of the recording. While at V0 the mean error was 3.2° (SD=1.3°), in V2 the maximum estimation error reached up to 40 degrees (M=10.1°, SD=9.2°). Figure 7 shows a more detailed analysis for the video-based tracker. We divided participants into two groups according to the error at V2. A participant is categorised into the *calibration drifted* group if the error was outside the 95% confidence interval of the initial error distribution at V0, while the others become the *calibration maintained* group. 12 out of 20 participants showed calibration drift which in most cases increased gradually at a more or less constant rate. This error is likely caused by shifts of the eye tracking headset and thus pupil positions (see Figure 6 for an example). More significant drift occurred in the top two cases and was likely caused by external factors, such as users touching the headset.

Figure 5b shows the corresponding results for the EOG-based eye tracker. In this case, the first validation session (V0) is used to calibrate the baseline estimator. We take a linear

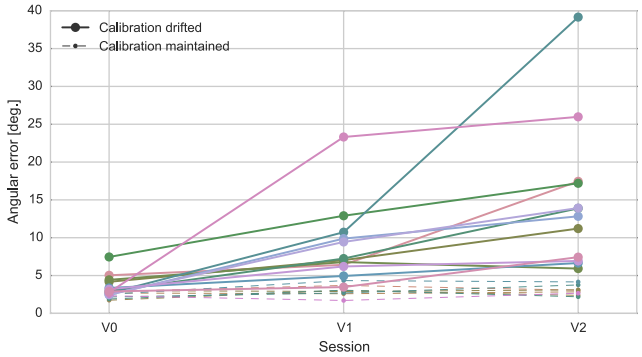
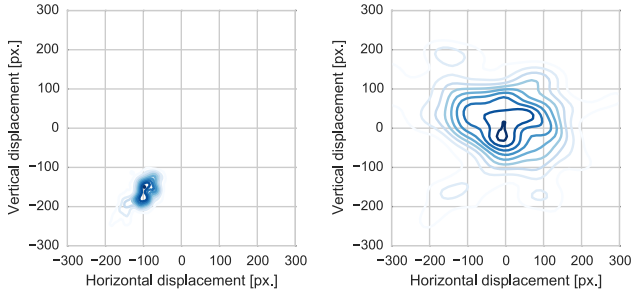


Figure 7: Time series variations of estimation errors of the video-based setting.



(a) Video-based setting (b) EOG-based setting

Figure 8: Examples of spatial distributions of the calibration drift. Each graph shows a kernel density estimation of the offset between ground-truth and estimated gaze positions.

regression approach, and the plot V0 represents the residual error of the mapping. The difficulty in EOG-based estimation is the baseline signal drift, and that is present from the beginning of the recording sessions. It can be also seen that the error becomes larger in recording sessions. This is mainly because validation sessions have uniform gaze distribution and it makes the baseline level estimation task easier than in natural recording cases. Also, EOG signals during daily-life recordings can contain irregular drift patterns and baseline drift removal becomes a nontrivial task.

Figure 8 further illustrates the difference between the video and EOG-based setting. Each graph shows a kernel density estimation of the offset between ground-truth and estimated gaze positions in the image space, and both correspond to the same participant's same validation session. While the calibration drift happens as a constant shift in the video-based setting, the EOG-based setting shows a broader distribution.

Insights into the Self-Calibration Approach

From this analysis we can draw two important insights into how automatic self-calibration can be used in the mobile eye tracking scenario. First, the analysis demonstrated that calibration drift is indeed a critical problem even for short-term recordings of one hour. Expected estimation accuracy after a long recording session can be significantly lower even with an

initial calibration. In the worst case, video-based estimation can become less accurate than EOG-based estimation. If a similar performance can be expected from a fully bottom-up calibration which provides a significant advantage in usability, it can be considered a promising design choice.

Second, especially in the video-based setting where calibration drift can be modelled as a spatial shift from the initial calibration, it is also worth considering an alternative approach assuming the initial calibration data. If the initial mapping function is also given as an additional input, the task for the saliency-based calibration process is to find the spatial drift. Although this approach still requires the initial calibration action, it is expected to provide better performance than the fully bottom-up approach.

SELF-CALIBRATING EYE TRACKER

As illustrated in Figure 1, self-calibration can be done by continuously repeating processing of the input data. Our method requires two synchronised input sources: images from the scene camera and signals from the eye sensor, i.e. in our case either pupil positions in the eye camera images or EOG signals. From these synchronised data, the self-calibration process gives a mapping function between the eye sensor signal and gaze position. The mapping can be used in two ways: either to refine the gaze estimation on the input data or as a new gaze estimation function for the next frames.

In the following, we propose two approaches to self-calibration with and without initial calibration, i.e. given a certain amount of an input calibration sequence, the system either 1) calibrates the estimation function from scratch or 2) adjusts the initial estimation function.

Fully Automatic Self-Calibration

Figure 9 provides an overview of the proposed method for fully automatic self-calibration without initial calibration. The scene images are first converted to visual saliency maps, and the mapping function between the input feature and gaze position is obtained through self-calibration. The self-calibration process consists of two sequential steps. In the aggregation step, saliency maps are clustered according to the similarity between associated input features in order to improve the low fixation prediction reliability of raw saliency maps. The clustering is done in a similar but simplified manner as in Sugano et al. [36] using the mini-batch k-means algorithm [32]. This results in a smaller set of cluster mean input features and corresponding mean saliency maps, which is used in the following mapping step to find the relationship between the input feature space and gaze coordinate space.

Let $\boldsymbol{x}$ and $\boldsymbol{s}$ be input pairs (mean feature and saliency map respectively) after the aggregation step. In prior work, saliency-based self calibration was formulated as a direct optimisation between gaze mapping output and saliency values. In the simplest form, the task is to find the gaze estimation function f that globally maximises saliency values at the output gaze positions. The underlying assumption is that the output gaze positions lie inside the image (saliency) area, and each $\boldsymbol{x}$ can be always mapped to somewhere inside $\boldsymbol{s}$. This assumption holds for stationary eye trackers where images and videos are displayed on a monitor. However, in mobile cases users might naturally look outside the scene camera as well, which can lead to input features that cannot be mapped to any point in the saliency map.

Robust Rotation Mapping

In order to handle the unreliability of egocentric saliency maps, we propose a robust rotation regression approach. The key idea is to introduce the RANSAC [15] approach and additional approximations and constraints to the mapping task.

We first assume that the input signal can be embedded in a 2D subspace, whose axes can be seen as rotations of horizontal and vertical axes of the scene camera. This approximation is inspired by the fact that human gaze patterns follow a normal distribution in the natural viewing condition. In the video-based setting, the 2D subspace of the polynomial feature can be constructed via principal components analysis given a certain number of observations. In the EOG-based setting, signals are usually measured using two different electrode pairs and they already correspond to gaze directions.

Since the range of gaze directions is constrained by physical limitations, we can further assume standard deviations of the gaze distribution are preliminarily obtained as a user-independent, hardware-specific setting. This assumption can be understood as a generalised version of Chen and Ji's approach using a Gaussian distribution around the image centre [13]. In contrast to their stationary setting, gaze patterns do not share the centre position in our mobile setting, and we can only rely on the scale of gaze distribution.

Under these assumptions, the mapping function only has to consider shift and rotation between input and output spaces, which can be solved by the Kabsch algorithm [24]. To find the mapping function, we take the RANSAC approach as follows. Each mean saliency map is represented by the position of its maximum value $\boldsymbol{m}$, and the mapping function is estimated by finding an optimal random subset of the sample pairs $(\boldsymbol{x}_i, \boldsymbol{m}_i)$. Random samples are selected according to the associated saliency value, i.e. $\boldsymbol{m}$ obtained from higher saliency peaks are used more frequently during the RANSAC evaluation. At each random selection, shift and rotation from the input space $\boldsymbol{x}$ to output space $\boldsymbol{m}$ are computed by the Kabsch algorithm. The mapping error is evaluated only with inlier-subset samples, and the best mapping function with minimum error is selected after a fixed number of trials.

Self-Calibration with Initial Calibration

If we can also assume an initial calibration data, it brings further constraints to this mapping task. As illustrated in Figure 10, the input feature can be mapped to 2D gaze positions $\boldsymbol{g}$ using the calibration data, instead of the 2D subspace $\boldsymbol{x}$. Then, we can employ a simpler mapping model between $\boldsymbol{g}$ and $\boldsymbol{m}$, according to the calibration drift model suited to the input modality. For example, calibration drift of the video-based tracker can be modelled as a spatial shift (Figure 8a).

Visual Saliency Models

As shown in Figure 11, we use four saliency map models in our method. We select two fast bottom-up saliency methods

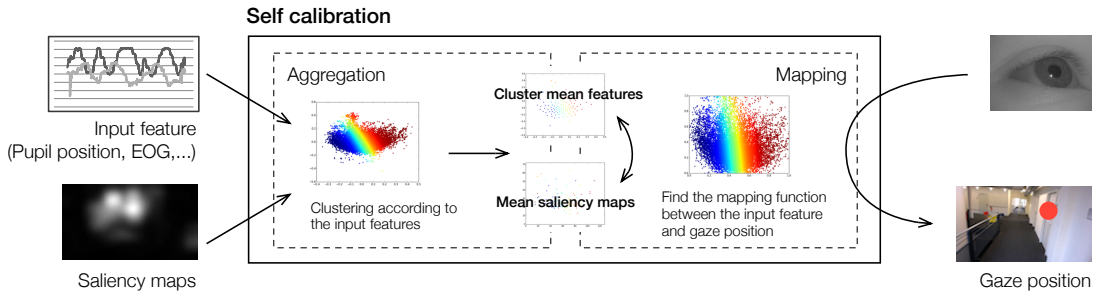


Figure 9: Method overview. Our self calibration process consists of two sequential steps, aggregation and mapping.

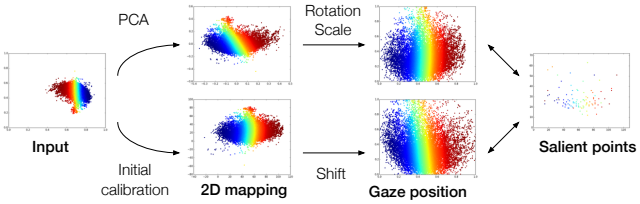


Figure 10: Detail of the mapping step. In both cases with and without initial calibration, we approximate the mapping task as 2D space alignment to the output space. The input feature space is first mapped to a 2D space using either PCA or initial calibration, and aligned with the gaze position space using data sampled from saliency maps.

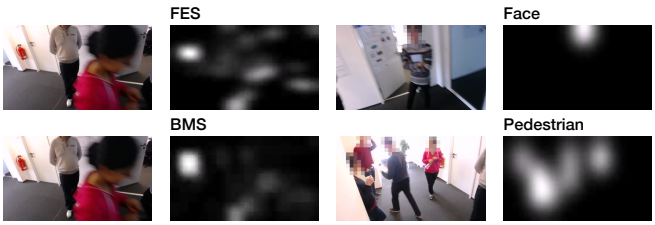


Figure 11: Examples of saliency models used in our framework. We combine two bottom-up methods (FES and BMS) and two object (face and pedestrian) detection methods.

which are highly ranked in the MIT saliency benchmark [11] since computational cost is the most important factor, and two object detectors to further incorporate high-level saliency. These four maps are linearly combined to compute the input saliency maps to our framework.

Fast and Efficient Saliency

Fast and Efficient Saliency (FES) [37] is a model with a simplified scheme to compute centre-surround color differences, and achieves near real-time performance while keeping a moderate fixation prediction accuracy. Similarly to most state-of-the-art methods, FES uses an average fixation map as a prior distribution to take into account the strong fixation bias towards the centres of images. As we discussed earlier, however, an egocentric camera's field of view does not match the user's perspective and such a spatial bias becomes purely user-dependent. Hence, we replace the centre bias with a uniform distribution and use bias-free saliency maps. In our own C++ implementation, the model runs at about 100 frames per second.

Boolean Map Saliency

BMS (Boolean Map Saliency) [42] is one of the top-ranked methods in the MIT saliency benchmark [11]. The model is designed based on a Gestalt principle of figure-ground segmentation and finds regions with closed contours in the feature space as salient regions. Using the author-provided C++ implementation, the model also runs at 100 frames per second. As in the case of FES, we use bias-free maps from the BMS model.

Face Detection

Human faces are known to be one of the most prominent objects that attract human attention, and face detection results can be used as a high-level saliency map [12]. In our system, we employed one of the state-of-the-art face detection methods by Li et al. [27]. Using the implementation of the ccv library with default parameters[1], the model can run at 33 frames per second.

Pedestrian Detection

Similarly to the face detection, human or pedestrian detection can be another prominent saliency indicator, especially when their faces cannot be observed clearly. We also used one of the state-of-the-art pedestrian detectors [4, 5] implemented in the ccv library, and the model runs at 17 frames per second.

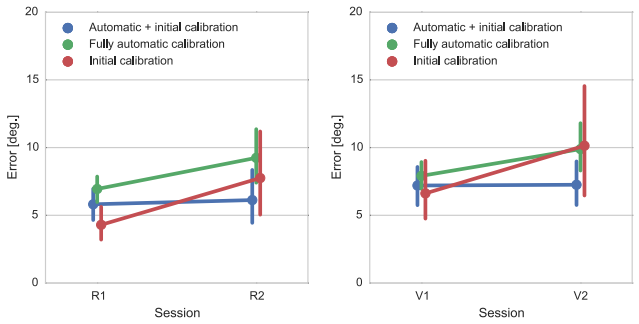
EVALUATION

We performed evaluations as to the effectiveness of the proposed approach to (1) achieve a practically useful gaze estimation accuracy in the fully automatic self-calibration setting and (2) maintain gaze estimation accuracy in the self-calibration setting with initial calibration.

Performance with Video-Based Eye Tracker

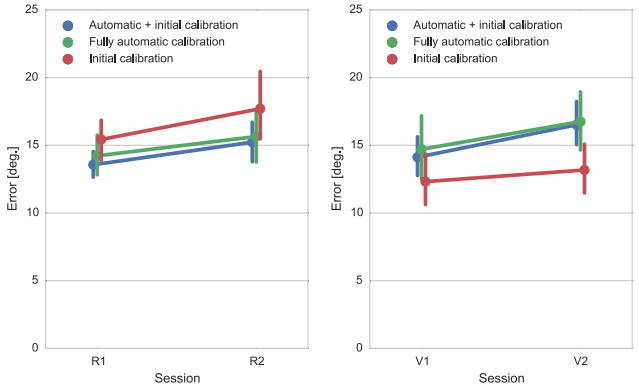
Figure 12 summarises the corresponding estimation error for the video-based eye tracker. In addition to two variants of our self-calibration approach with and without initial calibration, the figure shows a baseline result only with the initial calibration. For the recording sessions, the self-calibration methods took each of the whole recording session as an input calibration sequence. For validation sessions, the previous recording session were used as input. Our method used pupil

[1] http://libccv.org/



(a) Recording sessions (b) Validation sessions

Figure 12: Gaze estimation performance in the video-based setting. Error bars correspond to 95% confidence intervals.



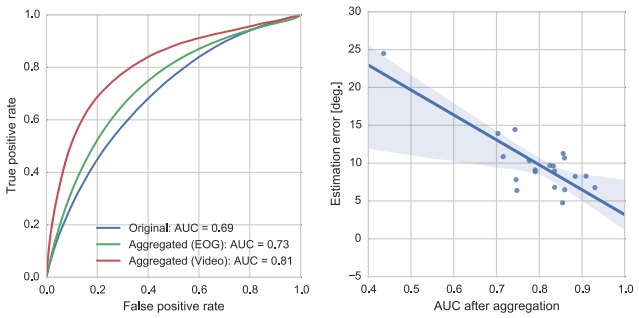
(a) Recording sessions (b) Validation sessions

Figure 13: Gaze estimation performance in the EOG-based setting.

detection results from the Pupil software and the same seven-dimensional polynomial pupil position feature as the baseline method.

Compared to the initial calibration result that shows a significant calibration drift, our proposed self-calibration methods show nearly consistent performance. In particular, in the last validation session (V2), the fully automatic calibration approach without initial calibration achieved nearly the same performance (M=9.9°, SD=4.1°) as the baseline method (M=10.1°, SD=9.2°). This error can be further decreased by incorporating the initial calibration data, in which case the mean estimation error goes down to around 6.0° during the recording sessions (R1: M=5.8°, SD=2.7°, R2: M=6.1°, SD=4.2°). Obviously, the proposed self-calibration method can only improve calibration drift for those cases where calibration drift actually happened. For the corresponding subset of 12 participants who showed calibration drift in session V2 (see Figure 7), our joint calibration method (M=8.0°, SD=3.5°) performed significantly better than the initial calibration result (M=14.9°, SD=9.2°); $t(12) = 2.50$, $p = .03$ by paired t-test.

Performance With EOG-Based Eye Tracker



(a) (b)

Figure 14: Relationship between saliency performances and self-calibration results. (a) ROCs of saliency maps before and after the aggregation step. (b) Self calibration accuracy with respect to AUC values.

Figure 13 summarises the corresponding estimation error comparison for the EOG-based eye tracker. While the improvement made by our method is smaller than for the video-based tracker, initial calibration is also helpful in this case. As discussed before, the estimation error increases during recording sessions. Through a paired t-test to compare estimation errors in session R2, our joint self-calibration method (M=15.2°, SD=3.2°) showed a significant performance improvement from the initial calibration method (M=17.7°, SD=5.7°); $t(20) = 2.73$, $p = .01$.

Performance of Saliency Models

Intuitively, gaze estimation accuracy depends on how well saliency models can predict fixation positions. Figure 14a shows receiver operating characteristics (ROCs) of saliency maps before and after the aggregation step. The vertical axis corresponds to the true positive rate, i.e. the rate of fixated pixels in saliency maps, while the horizontal axis corresponds to the false positive rate, i.e. the rate of non-fixated pixels. Hence, the area under the ROC curves (AUC) becomes larger if the maps have sharp peaks around ground-truth fixation positions. The blue curve indicates the performance of the original saliency maps, and the green and red curves correspond to the maps after the aggregation step using EOG-based and video-based input signals, respectively. It can be seen that the aggregation step increases on the original AUC value (0.69) in both cases (0.81 and 0.73 in video- and EOG-based settings, respectively). However, these are still significantly lower than the AUC values reported in prior work [36] (0.82 and 0.93 before and after the aggregation) in a stationary setting, and illustrate the core difficulty of the mobile setting.

These saliency performance metrics are directly related to the final performance of the self-calibrated gaze estimator. Figure 14b shows the relationship between AUC values after the aggregation step at the R2 session (horizontal axis) and final estimation accuracy of the video-based fully-automatic method at the V2 session (vertical axis). The overlaid line is a linear model relating two values with 95% confidence intervals. The two variables are strongly correlated, $r(20) =$

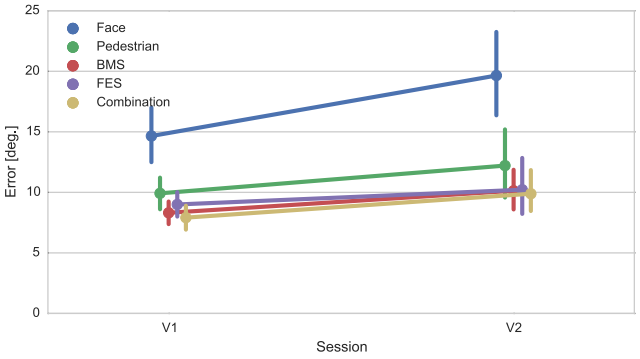
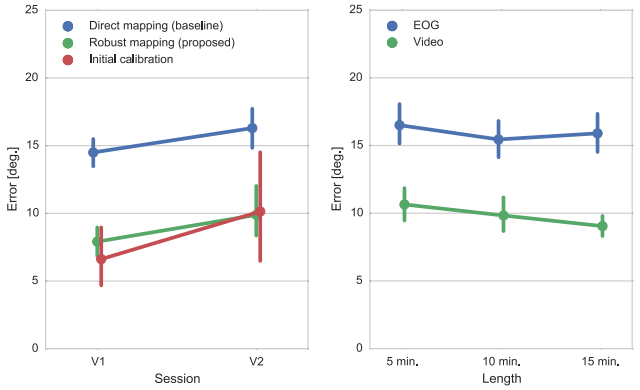


Figure 15: Self-calibration performance using different saliency models: two bottom-up models (BMS, FES), two object detection models (Face, Pedestrian) and their combination.



(a) (b)

Figure 16: Detailed analysis of calibration performance. (a) Comparison between different mapping approaches. (b) Calibration accuracy with respect to the amount of input data.

-0.83, $p < 0.01$, clearly indicating the relationship between saliency and calibration performance.

While in this work we used a combination of four different saliency models, each saliency model contributes differently to the self-calibration performance. Figure 15 compares the estimation accuracy of the video-based fully-automatic method using different saliency models. While the face detection model is known to be very effective in stationary settings, in our setting it showed the worst performance (M=17.2°, SD=7.2° across two validation sessions). The pedestrian detection model showed a moderate performance (M=11.1°, SD=5.0°), and bottom-up models in general achieved better performance (FES: M=9.6°, SD=4.0°, BMS: M=9.2°, SD=3.2°) than object detection models. Although it is not statistically significant, the combination model achieved the best performance among them (M=8.9°, SD=3.4°).

Effect of the Proposed Robust Fitting

The low accuracy of saliency maps shown in Figure 14a explains why our proposed robust fitting approach is required. Figure 16a further compares the proposed robust mapping method with a direct mapping method for the video-based setting. The direct mapping refers to the most straightforward approach, i.e., applying a polynomial regression function between salient points and pupil positions. Robust mapping refers to the same fully automatic calibration method shown in Figure 12, and the initial calibration result is also shown as a reference. At session V2, the direct mapping method achieved a mean error of 16.4° (SD=3.6°). The proposed robust mapping method (M=9.9°, SD=4.1°) significantly improved the estimation accuracy; $t(20) = 6.13$, $p < .01$ by paired t-test.

Effect of the Amount of Input Data

The aggregation step generally improves for larger amounts of input data. However, since it assumes a one-to-one correspondence between input feature and output gaze position, it can suffer from sensor drift. Hence, there can be a fundamental trade-off between aggregation performance and input data amount. Figure 16b shows estimation performance with respect to the amount of input to the self-calibration framework. Mean errors of fully-automatic calibration across two validation sessions are shown using different amounts of input data (the last 5, 10, and 15 minutes of the previous recording session) for video-based and EOG-based settings. Although the difference is not statistically significant, the largest input size results in the best performance in the video-based setting (M=9.0°, SD=2.4°). In contrast, in the EOG-based setting where the baseline drift is a critical problem, more data is not always helpful, and the 10-minute setting achieved the best performance (M=15.4°, SD=4.5°).

DISCUSSION

Analyses on our new dataset confirmed what was probably assumed but never shown or quantified before: Even for the relatively short recordings as in our data collection, calibration drift can be severe and considerably reduce gaze estimation accuracy for both eye tracker types. Calibration drift therefore poses an important challenge for gaze recordings in mobile daily-life settings, and can be expected to be even more severe for long-term recordings.

To address this problem, we proposed a robust method to automatically self-calibrate head-mounted eye trackers. Results from our evaluations suggest that our method is effective in reducing drift in a calibrated eye tracker. Given sufficient data, the method can also achieve competitive gaze estimation accuracy compared to a state-of-the-art initial calibration method but without the need for any manual eye tracker calibration.

Our analyses revealed that in mobile settings, the gaze estimation accuracy of head-mounted eye trackers cannot stay true to the best case performance reported by device manufacturers obtained under controlled conditions. The mean estimation error of the best case (joint self-calibration together with initial calibration) was about 6.0° during recording sessions. In the physical space, this error corresponds to, e.g., 11cm at 1m distance from the user and is already good enough to identify objects of interest in the scene. Although traditional gaze

interaction techniques often assume accuracy high enough to interact with objects inside screen spaces, this result suggests that we need to change our way of thinking about technical requirements for pervasive real-world gaze interaction. In light of this disparity, in the future it will be important to take into account both calibration drift and low gaze estimation accuracy and to design applications that do not require high accuracy.

Considering that the method performs nearly the same only with the fastest saliency model, it already has a good potential for practical use. The joint self-calibration approach can significantly reduce users' effort to maintain high gaze estimation accuracy, and together with low-cost devices, it makes it easier to integrate eye tracking capabilities into smart glasses. The fully automatic self-calibration approach is more beneficial to, e.g., elderly and child care applications, where more natural and unobtrusive eye tracking technology is required.

Since our method aggregates saliency maps to improve prediction accuracy, long-term recording can further improve the self-calibration performance as indicated in Figure 16b. However, if the amount of calibration drift is significant, fitting a single mapping function to the whole recording session cannot result in good performance. Instead, it would be better to split the recording into small segments and repeat the self-calibration process. Studying the self-calibration performance during longer recordings would be an important subject for future work.

Limitations and Future Work

The biggest bottleneck of the proposed self-calibration approach is fixation prediction accuracy of the saliency model. Given the correlation between saliency performance and final estimation accuracy, there is great potential that performance can be further improved by enhancing the baseline saliency performance. Egocentric saliency prediction is a relatively new topic in the field of computer vision, and there is a huge potential for future research investigation.

Current object detection-based models perform relatively poorly in egocentric videos, for which there may be two possible explanations. First, even for state-of-the-art computer vision algorithms, we cannot expect perfect accuracy of face and pedestrian detection in egocentric videos. Second, compared to human-edited images and videos used in prior studies, videos captured by egocentric cameras often contain meaningless scenes without any salient objects. In this sense, the performance of object detection-based models is expected to be less significant even with a perfect detection algorithm. It will be also important to conduct more fundamental study on gaze behaviour during daily-life situations.

CONCLUSION

We introduced a method for self-calibrating head-mounted eye trackers and demonstrated the effectiveness of our approach for both video- and EOG-based eye trackers with and without initial calibration. We further presented a 20-participant dataset containing synchronised egocentric videos as well as gaze data. We used the dataset to characterise calibration error during mobile daily-life recordings and quantitatively compare the proposed self-calibration method with an initial calibration method. To the best of our knowledge, this is the first attempt to show the real-world applicability of a saliency-based calibration approach for video-based and EOG-based head-mounted eye trackers. These results are promising and underline the significant potential of the self-calibration approach to enable novel gaze-based applications not yet feasible in unconstrained daily-life settings.

ACKNOWLEDGMENTS

This work was supported, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, the Alexander von Humboldt-Foundation, and a JST CREST research grant.

REFERENCES

1. F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab. 2013. Calibration-Free Gaze Estimation Using Human Gaze Patterns. In *Proc. ICCV*. 137–144.
2. D. Bannach, P. Lukowicz, and O. Amft. 2008. Rapid prototyping of activity recognition applications. *Pervasive Computing, IEEE* 7, 2 (2008), 22–31.
3. M. Barz, A. Bulling, and F. Daiber. 2015. *Computational Modelling and Prediction of Gaze Estimation Error for Head-mounted Eye Trackers*. Technical Report. German Research Center for Artificial Intelligence (DFKI). 10 pages.
4. R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. 2012. Pedestrian detection at 100 frames per second. In *Proc. CVPR*. 2903–2910.
5. R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. 2013. Seeking the Strongest Rigid Detector. In *Proc. CVPR*. 3666–3673.
6. A. Borji and L. Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1 (2013), 185–207.
7. A. Bulling, D. Roggen, and G. Tröster. 2009. Wearable EOG goggles: Seamless sensing and context-awareness in everyday environments. *J. Ambient Intell. Smart Environ.* 1, 2 (2009), 157–171.
8. A. Bulling, J. Ward, H. Gellersen, and G. Troster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 4 (2011), 741–753.
9. A. Bulling, J. A. Ward, and H. Gellersen. 2012. Multimodal Recognition of Reading Activity in ransit Using Body-Worn Sensors. *ACM Trans. Appl. Percept.* 9, 1 (2012), 2:1–2:21.
10. A. Bulling, C. Weichel, and H. Gellersen. 2013. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proc. CHI*. 305–308.

11. Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. 2014. MIT Saliency Benchmark. http://saliency.mit.edu/. (2014).
12. M. Cerf, J. Harel, W. Einhaeuser, and C. Koch. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Proc. NIPS*. 241–248.
13. J. Chen and Q. Ji. 2015. A Probabilistic Approach to Online Eye Gaze Tracking Without Explicit Personal Calibration. *IEEE Trans. Image Process.* 24, 3 (2015), 1076–1086.
14. J. Drewes, G. S. Masson, and A. Montagnini. 2012. Shifts in Reported Gaze Position Due to Changes in Pupil Size: Ground Truth and Compensation. In *Proc. ETRA*. 209–212.
15. M. A. Fischler and R. C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (June 1981), 381–395.
16. E. Guestrin and E. Eizenman. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Trans. Biomed. Eng.* 53, 6 (2006), 1124–1133.
17. D. Hansen and Q. Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 3 (2010), 478–500.
18. A. Hornof and T. Halverson. 2002. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behav. Res. Meth. Instrum. Comput.* 34, 4 (2002), 592–604.
19. Y. Ishiguro, A. Mujibiya, T. Miyaki, and J. Rekimoto. 2010. Aided Eyes: Eye Activity Sensing for Daily Life. In *Proc. AH*. Article 25, 25:1–25:7 pages.
20. Y. Ishiguro and J. Rekimoto. 2012. GazeCloud: A Thumbnail Extraction Method Using Gaze Log Data for Video Life-Log. In *Proc. ISWC*. 72–75.
21. S. Ishimaru, K. Kunze, Y. Uema, K. Kise, M. Inami, and K. Tanaka. 2014. Smarter Eyewear: Using Commercial EOG Glasses for Activity Recognition. In *Proc. Ubicomp*. 239–242.
22. Y. Itoh and G. Klinker. 2014. Interaction-free calibration for optical see-through head-mounted displays based on 3d eye localization. In *Proc. 3DUI*. 75–82.
23. S. John, E. Weitnauer, and H. Koesling. 2012. Entropy-based correction of eye tracking data for static scenes. In *Proc. ETRA*. 297–300.
24. W. Kabsch. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* 32, 5 (09 1976), 922–923.
25. M. Kassner, W. Patera, and A. Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proc. Ubicomp*. 1151–1160.
26. K. Kunze, S. Ishimaru, Y. Utsumi, and K. Kise. 2013. My Reading Life: Towards Utilizing Eyetracking on Unmodified Tablets and Phones. In *Proc. Ubicomp*. 283–286.
27. J. Li and Y. Zhang. 2013. Learning SURF Cascade for Fast and Accurate Object Detection. In *Proc. CVPR*. 3468–3475.
28. P. Majaranta and A. Bulling. 2014. *Eye Tracking and Eye-Based Human-Computer Interaction*. Springer, 39–65.
29. H. Manabe and M. Fukumoto. 2006. Full-time Wearable Headphone-type Gaze Detector. In *CHI EA*. 1073–1078.
30. D. Model and M. Eizenman. 2010. User-calibration-free Remote Gaze Estimation System. In *Proc. ETRA*. 29–36.
31. A. Plopski, Y. Itoh, C. Nitschke, K. Kiyokawa, G. Klinker, and H. Takemura. 2015. Corneal-imaging calibration for optical see-through head-mounted displays. *IEEE Trans. Vis. Comput. Graphics* 21, 4 (2015), 481–490.
32. D. Sculley. 2010. Web-scale K-means Clustering. In *Proc. WWW*. 1177–1178.
33. S.-W. Shih, Y.-T. Wu, and J. Liu. 2000. A calibration-free gaze tracking technique. In *Proc. ICPR*, Vol. 4. 201–204 vol.4.
34. J. Steil and A. Bulling. 2015. Discovery of Everyday Human Activities From Long-term Visual Behaviour Using Topic Models. In *Proc. UbiComp 2015*.
35. S. Stellmach and R. Dachselt. 2012. Look & Touch: Gaze-supported Target Acquisition. In *Proc. CHI*. 2981–2990.
36. Y. Sugano, Y. Matsushita, and Y. Sato. 2013. Appearance-Based Gaze Estimation Using Visual Saliency. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2 (2013), 329–341.
37. H. R. Tavakoli, E. Rahtu, and J. Heikkilä. 2011. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Proc. SCIA*. 666–675.
38. B. Tessendorf, A. Bulling, D. Roggen, T. Stiefmeier, M. Feilner, P. Derleth, and G. Tröster. 2011. Recognition of hearing needs from body and eye movements to improve hearing instruments. In *Proc. PerCom*. 314–331.
39. J. Turner, J. Alexander, A. Bulling, D. Schmidt, and H. Gellersen. 2013. Eye Pull, Eye Push: Moving Objects between Large Screens and Personal Devices with Gaze & Touch. In *Proc. INTERACT*.
40. J. Turner, A. Bulling, J. Alexander, and H. Gellersen. 2014. Cross-Device Gaze-Supported Point-to-Point Content Transfer. In *Proc. ETRA*. 19–26.
41. M. Vidal, J. Turner, A. Bulling, and H. Gellersen. 2012. Wearable Eye Tracking for Mental Health Monitoring. *Computer Communications* 35, 11 (2012), 1306–1311.
42. J. Zhang and S. Sclaroff. 2013. Saliency Detection: A Boolean Map Approach. In *Proc. ICCV*. 153–160.