Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage

Philipp Müller
Max Planck Institute for Informatics
Saarland Informatics Campus
pmueller@mpi-inf.mpg.de

Daniel Buschek
LMU Munich
daniel.buschek@ifi.lmu.de

Michael Xuelin Huang
Max Planck Institute for Informatics
Saarland Informatics Campus
mhuang@mpi-inf.mpg.de

Andreas Bulling
University of Stuttgart
Institute for Visualisation and Interactive Systems
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

Automatic saliency-based recalibration is promising for addressing calibration drift in mobile eye trackers but existing bottom-up saliency methods neglect user's goal-directed visual attention in natural behaviour. By inspecting real-life recordings of egocentric eye tracker cameras, we reveal that users are likely to look at their phones once these appear in view. We propose two novel automatic recalibration methods that exploit mobile phone usage: The first builds saliency maps using the phone location in the egocentric view to identify likely gaze locations. The second uses the occurrence of touch events to recalibrate the eye tracker, thereby enabling privacy-preserving recalibration. Through in-depth evaluations on a recent mobile eye tracking dataset (N=17, 65 hours) we show that our approaches outperform a state-of-the-art saliency approach for automatic recalibration. As such, our approach improves mobile eye tracking and gaze-based interaction, particularly for long-term use.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Ubiquitous and mobile computing**;

KEYWORDS

Mobile eye tracking; Eye tracker recalibration

ACM Reference Format:
Philipp Müller, Daniel Buschek, Michael Xuelin Huang, and Andreas Bulling. 2019. Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '19), June 25–28, 2019, Denver , CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3314111.3319918

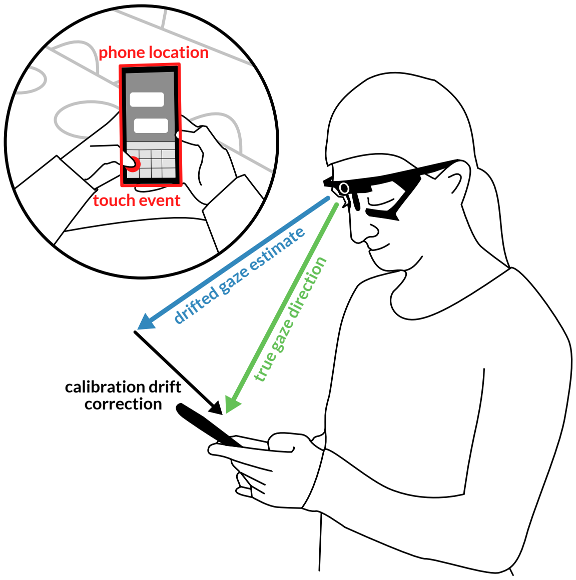

Figure 1: Mobile eye trackers suffer from calibration drift and inaccurate gaze estimates (blue arrow), for example caused by headset slippage. Our two novel automatic recalibration methods correct for calibration drift (black arrow) by either using the phone's location or users' touch events (red) to infer their true gaze direction (green arrow).

1 INTRODUCTION

The ubiquity of smartphones and the increasing availability of mobile eye trackers enables novel interaction concepts and applications [Bulling and Gellersen 2010; Duchowski 2002; Pfeuffer et al. 2015; Pfeuffer and Gellersen 2016], and has lead to a growing body of research on gaze-based interaction with mobile devices [Khamis et al. 2018]. Unfortunately, more widespread adoption of mobile gaze-based interaction in everyday life remains challenging due to *calibration drift*. This describes the accumulating deterioration in eye tracking accuracy after an initial manual calibration was performed. For example, the eye tracking headset may slightly slip and shift on the user's head due to body movement throughout the day, rendering gaze interactions inaccurate.

ETRA '19, June 25–28, 2019, Denver , CO, USA

ACM ISBN 978-1-4503-6709-7/19/06…$15.00
https://doi.org/10.1145/3314111.3319918

Common calibration procedures are tedious and impractical to perform several times a day in everyday life. To address this problem, recent work has proposed an *automatic* recalibration approach that uses saliency maps computed from a mobile eye tracker's scene video [Sugano and Bulling 2015]. A saliency map is a 2D probability map of where in the visual scene the user is likely to fixate on [Borji and Itti 2013]. The recalibration approach exploits the (assumed) correlation between saliency maps and gaze to continuously recalibrate the eye tracker in the background.

The performance of this approach inherently depends on the quality of the computed per-frame saliency maps. The authors in [Sugano and Bulling 2015] studied a free-viewing and, hence, artificial scenario in which users were walking around a building without any concrete task in mind. Natural daily-life settings, however, are dominated by task-driven attentive behaviour (e.g. grabbing something, pressing a button in an elevator, using computers or phones). In such situations, the user's task is more likely to determine attention than the visual saliency of the object [Hayhoe and Ballard 2005]. It therefore remains unclear how the original saliency-based approach performs in real-world contexts and whether it can be improved to better exploit task-driven behaviour.

This paper aims to address both questions and proposes novel improvements for a pervasive everyday mobile interaction use case. By inspecting real-life recordings of a recent mobile eye tracking dataset [Steil et al. 2018b], we observed that users are likely *to attend to their phones* once these appear in the view of the egocentric camera. However, using phone presence for automatic recalibration is challenging because interaction "on the go" leads to frequent attention switches between the phone and the environment [Oulasvirta et al. 2005; Steil et al. 2018b]. Thus, it is unlikely that users will *always* look at the phone, even if it is present in their field of view [Steil et al. 2018b]. We address this challenge by making use of phone detections in a robust state-of-the-art saliency-based recalibration approach [Sugano and Bulling 2015].

Moreover, we propose a second approach ("blind recalibration"), where we use the occurrence of *touch events* on the user's phone as an indicator for gazing at the assumed phone location. This approach does not require a scene camera, and may thus prove useful for privacy-sensitive applications or contexts in which recording egocentric video is not desirable (cf. [Steil et al. 2018a]).

In summary, our contribution is two-fold: First, we present *two novel approaches for automatic mobile eye tracker recalibration* that use a) smartphone screen locations and b) occurrence of touch events to counter calibration drift in everyday use of mobile eye trackers. Second, we report *in-depth evaluations of these approaches* on a recent dataset collected in-the-wild (N=17, 65 hours), which show that our approaches consistently outperform the previously proposed state-of-the-art saliency-based approach.

2 RELATED WORK

Our work is related to previous work on 1) phone use in everyday life and 2) automatic eye tracker calibration.

2.1 Phone Use in Everyday Life

We use phone interactions to recalibrate mobile eye trackers, since they come with several beneficial properties: For example, related work has shown that many phone users are *highly responsive*, attending to mobile messages 12 hours a day (84 hours a week) [Dingler and Pielot 2015]. Typing in general happens throughout the whole day, both on weekdays and weekends [Buschek et al. 2018]. Thus, messaging and typing already cover a large timeframe in which recalibration via phone use is possible.

People also develop *usage habits*, such as frequently checking for new messages and content updates [Oulasvirta et al. 2012]. Many interactions also result in repeated notifications later on (e.g. chat, email, social networks, music), bringing users back to their phones. For instance, Shirazi et al. [Sahami Shirazi et al. 2014] found that 50 % of interactions with incoming notifications happen within 30 seconds. This "checking behaviour" supports our approach, since self-calibration benefits from phone use spread out across time and many situations, to get up-to-date and diverse samples. Moreover, many people even interact with their phone if they have no specific task in mind, that is, if they seek stimulation in situations of *boredom* [Pielot et al. 2015]. Nevertheless, mobile phone use leads to frequent *switches of attention* between phone and environment [Oulasvirta et al. 2005; Steil et al. 2018b]. Thus, exploiting phone use for recalibration needs to deal with uncertain user attention, even if the phone is in sight of the scene camera.

In summary, related work on mobile phone use in everyday life reveals unique challenges and opportunities for self-calibrating mobile eye trackers via phone use and thus motivates our research questions in this paper.

2.2 Automatic Eye Tracker Calibration

Despite continuing advances in eye tracking technology, e.g. by improved pupil detection algorithms [Dierkes et al. 2018; Swirski and Dodgson 2013], wider adoption of the technology is still prevented by the need for repeated manual calibration of eye trackers. Therefore, automatically calibrating (i.e. without initial calibration) and recalibrating (i.e. with initial calibration) eye trackers has been of interest to the HCI community. Initial work on automatic calibration focused on stationary settings. While [Yamazoe et al. 2008] used an eyeball model, other works used mouse clicks, and more diverse associations between interaction patterns and users' visual attention [Huang et al. 2016; Sugano et al. 2008; Zhang et al. 2018]. Subsequently, more general self-calibration approaches exploited bottom-up saliency maps or gaze patterns obtained from other users [Alnajar et al. 2013; Chen and Ji 2015; Sugano et al. 2010]. A different way to self-calibrate mobile eye trackers uses corneal images [Lander et al. 2017; Takemura et al. 2014a]. Such approaches require specialised hardware, as they rely on RGB eye cameras to extract the corneal image. Consequently they also struggle more with suboptimal lighting conditions [Takemura et al. 2014a]. Moreover, they cannot be used for privacy-preserving recalibration, as scene properties can be decoded from RGB eye images [Backes et al. 2008; Lander et al. 2017; Takemura et al. 2014b]. In contrast, no such approach is known for active illumination infrared eye cameras.

The closest work to ours is from Sugano et al. who were first to analyse the severe calibration drift in mobile eye trackers and proposed saliency-based recalibration to retain the quality of an initial manual calibration over a longer period of time [Sugano and Bulling 2015]. The employed saliency maps consisted of bottom-up

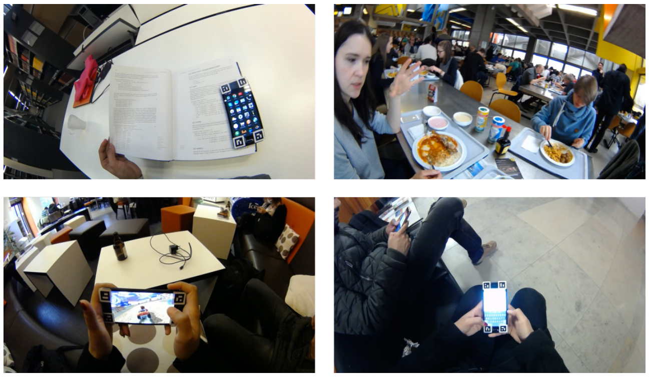

Figure 2: Example images from the scene camera in different situations from the dataset of Steil et al. [Steil et al. 2018b].

components along with face- and person detectors. However, they evaluation focussed on a free-viewing setting and mobile device usage was neither incorporated in the approach nor occured during the study.

3 DATASET

To investigate automatic recalibration in natural environments, we used a recent 20-participant mobile eye tracking dataset originally recorded to study visual attention forecasting in natural situations [Steil et al. 2018b]. We chose this dataset because it contains relatively long recordings of interactive behaviour with mobile phones during everyday situations, including studying in a library, working in an office, eating in a canteen, or drinking a coffee in a café. The dataset was subsequently ground-truth annotated for users’ current environment [Steil et al. 2018a], which we used to evaluate our methods in different daily-life situations.

3.1 Apparatus

For recording, participants were equipped with a state-of-the-art PUPIL mobile eye tracker [Kassner et al. 2014] featuring an infrared eye (640 × 480 pixels) and a fisheye scene camera (1280 × 720 pixels). Participants interacted with a mobile phone that was augmented with visual markers on its corners to obtain groundtruth phone location in the egocentric scene camera view. Logging software was used to monitor users’ phone interactions, including all touch events. A messaging application was used as a means of communication between experimenter and participant.

3.2 Procedure

Each participant took part in three consecutive recording blocks, each lasting on average for 77 minutes. Before each recording block, a calibration sequence was recorded in which the participants were instructed to gaze at visual markers that were manually presented by the experimenter at at least nine different locations in order to cover the field of view of the scene camera. For 17 participants, additional calibration sequences were recorded at the end of every recording block for.We restrict our analysis to those 17 participants because the additional calibration sequences allow us to quantify the error of automatic recalibration approaches. The top of Figure 3 gives an overview over the structure of the dataset. During the recordings, participants were free to roam the university campus under the conditions that they did not stay at a single place for more than 30 minutes and to visit the canteen, the library and the coffee shop at least once during the recording. Apart from this, participants were not given any instructions or otherwise constrained in their behaviour. In particular, they were also allowed to put off the eye tracker during short breaks between recording blocks. Figure 2 shows sample images obtained using the egocentric camera.

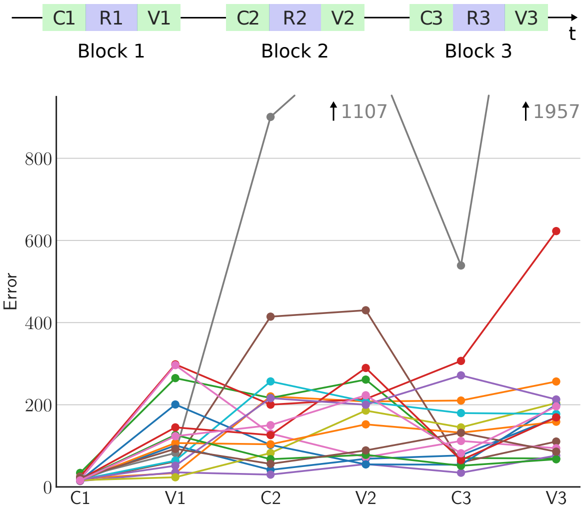

Figure 3: Top: Illustration of the dataset structure consisting of three recording blocks each comprised of calibration sequence (CX), recording (RX) and calibration sequence used for validation (VX). Bottom: Gaze estimation error in pixels measured on different calibration sequences when using the first calibration sequence (C1) to calibrate the eye-tracker. Lines are added to connect corresponding measurements.

3.3 Analysis

We used the calibration sequences recorded after each recording block to evaluate the calibration drift compared to the initial calibration recorded at the beginning. Gaze estimation is performed using a seven dimensional polynomial pupil feature based on pupil detections provided by the PUPIL software [Kassner et al. 2014]. We then use ridge regression to learn the mapping of pupil features to marker locations. This is in line with the approach taken in [Sugano and Bulling 2015], except that we perform 2D gaze estimation, as the calibration sequences on the dataset only provide 2D information. To not weight errors differently at different eccentricities of the field of view as a result of using a fisheye camera, we undistort gaze estimates and calibration markers before error measurements.

Figure 3 shows the gaze estimation error of the calibration obtained from the initial calibration session measured on all available calibration and validation sessions in the dataset. Each participant is represented by a line connecting the corresponding measurements. Calibration drift is present for a participant if the error at later points in time is larger than the error of the initial calibration. In line with [Sugano and Bulling 2015], some participants only showed a minor calibration drift while others showed a large increase in

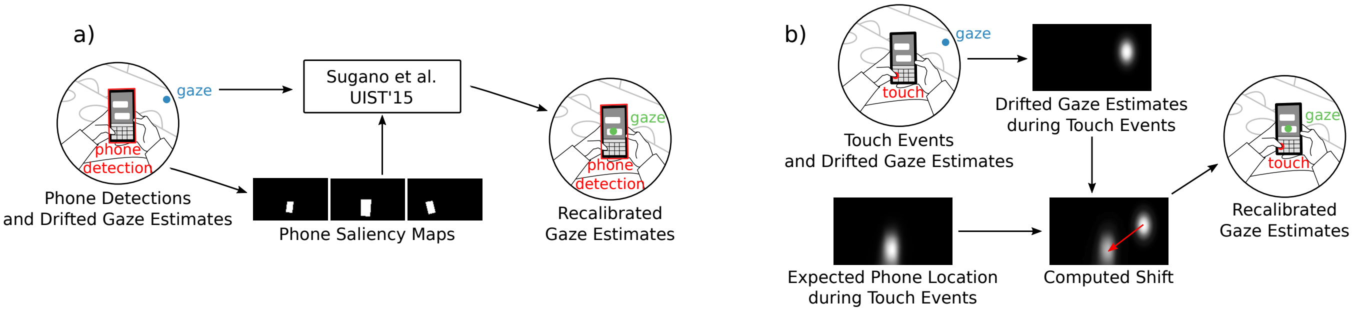
a)
gaze
phone detection
Phone Detections and Drifted Gaze Estimates
Sugano et al. UIST'15
Phone Saliency Maps
gaze
phone detection
Recalibrated Gaze Estimates
b)
gaze
touch
Touch Events and Drifted Gaze Estimates
Drifted Gaze Estimates during Touch Events
Expected Phone Location during Touch Events
Computed Shift
gaze
touch
Recalibrated Gaze Estimates

Figure 4: Overview over the two proposed methods. a) From phone detections we build corresponding phone saliency maps that serve as input to the method of [Sugano and Bulling 2015] together with an initial eye tracker calibration. b) We combine a model of where we expect the phone to be located during touch events with an aggregation of the drifted gaze estimates during touch events. From this we compute the shift we have to apply to the drifted gaze estimates in order to obtain correct gaze estimates.

gaze estimation error over the recording blocks. One participant exhibited a particularly large increase in error, reaching a gaze estimation error of 1957 pixels during the last validation sequence. Closer inspection of this participant's eye video showed that the eye camera is severely shifted in the calibration sequence after the second and third recording blocks. We opted to keep this outlier in our analysis because such severe shifts in the relation between eye and camera are precisely the challenge in mobile eye tracking that we are trying to solve. Removing the participant from the dataset does not change the general pattern of results. The large increase in error over time that exists for many participants illustrates the need for automatic recalibration methods.

Further analysis revealed that the probability of gazing at the phone is 0.59 when the phone is detected. The presence of touch events increased this probability to 0.7. This strong relationship between phone interaction and gaze is the basis of our proposed automatic recalibration methods.

4 METHOD

We propose two different methods for automatic recalibration by exploiting phone usage: Our first approach uses *phone detections* from the scene camera. These phone detections are transferred into saliency maps, and the method of [Sugano and Bulling 2015] for automatic recalibration using visual saliency is applied. Our second approach recalibrates *"blindly"* without using the scene image. Here, we compensate for calibration drift by computing the shift between the (potentially drifted) gaze estimates when *touch events happen* and the expected location of the mobile phone. An overview over both methods is given in Figure 4. We next describe our two approaches in more detail. In all cases, initial gaze estimates are obtained as described in the previous section.

4.1 Approach 1: Phone Saliency Maps

To use phone detections in the scene camera, we follow the approach to automatic recalibration starting from an initial calibration as proposed by [Sugano and Bulling 2015] (see also [Sugano et al. 2010]). We give an overview of this method and then describe our adaptation to integrate phone detections.

4.1.1 Visual saliency based recalibration. The approach of [Sugano and Bulling 2015] relies on the association of saliency maps extracted from the scene video with pupil positions and polynomial pupil features extracted from the eye camera. It consists of two steps, namely aggregation and robust mapping. In the aggregation step, the polynomial pupil features are clustered using the mini-batch k-means algorithm [Sculley 2010]. The clustering on pupil features also defines a clustering of the corresponding saliency maps, from which a mean saliency map is computed for every cluster. The goal of the robust mapping step is to find a mapping from the clusters of pupil features to locations in the scene video by making use of the mean saliency maps. To this end, 2D gaze predictions are obtained from the polynomial pupil features by applying the initial calibration. Subsequently, RANSAC [Fischler and Bolles 1981] is employed to find a shift from this 2D space of initial predictions to the output space consisting of the positions of maximum values in the mean saliency maps. Applying this shift to the initial predictions removes the calibration drift. For further details we kindly refer the reader to [Sugano and Bulling 2015].

4.1.2 Phone saliency maps. Our approach based on phone detections relies on a saliency map in which we set the area of the detected phone in the scene video to the maximum value and everything else to zero. The area of the phone is defined as the convex polygon that has the detections of the phone corner markers as its vertices (see red polygon in Figure 1). In frames without phone detections, the corresponding saliency map is all zero. We call these saliency maps *phone saliency maps.* Figure 5 (left) shows an average phone saliency map for illustration. Phone saliency maps along with pupil detections and the initial calibration are then used as input to the approach of [Sugano and Bulling 2015].

For phone detection, the dataset [Steil et al. 2018b] uses four visual markers attached to the phone corners, and the marker detection implemented in PUPIL [Kassner et al. 2014]. This simulates a robust phone detection method. However, methods for detecting screens without markers exist as well (see e.g. [Korayem et al. 2016]) and could be integrated for practical deployments without markers.

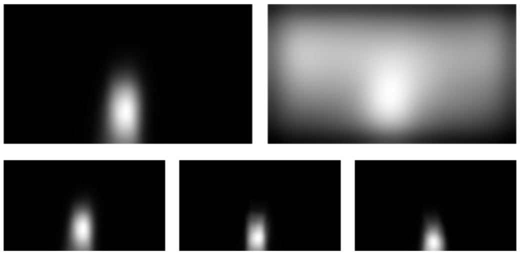

Figure 5: Different saliency maps averaged over all participants. Top left: Phone saliency maps for moments when touch events happen. Top right: Saliency map constructed according to [Sugano and Bulling 2015]. Bottom, left to right: Phone saliency maps for sitting, standing and walking.

4.2 Approach 2: Blind Recalibration

Our second proposed approach uses information about touch events taking place on the mobile phone in order to automatically recalibrate the eye tracker. This approach is motivated by privacy concerns about scene recordings using body-worn cameras [Koelle et al. 2018; Steil et al. 2018a], since it does not need such a scene camera for recalibration. It is based on two assumptions.

4.2.1 Assumption 1: Touch events indicate attention. We assume that when touch events take place the user is likely to look at the phone. This assumption is confirmed by our analysis on the dataset by Steil et al. [Steil et al. 2018b] showing that the probability of gazing at the detected phone is 0.7 when a touch event takes place.

4.2.2 Assumption 2: Common phone location in scene view. We assume that phones are most of the time positioned at a similar area in the scene view while interacting with them via touching. The top row of Figure 5 supports this assumption by showing the localised average phone saliency map when touch events take place in comparison to the average saliency map following the approach of [Sugano and Bulling 2015]. While the maximum at the bottom middle in both cases, for the phone saliency map it is closer to the bottom. Furthermore, the bottom row of Figure 5 shows that the average phone saliency map during touch events only slightly changes when people are sitting, standing or walking.

4.2.3 Recalibration and evaluation. Exploiting the two assumptions, we correct calibration drift in the following way:

1. Estimating usual phone location: We estimate the usual location of the mobile phone during touch events by averaging phone saliency maps for moments in time that are within a one second window centred on a touch event. Taking the argmax of this saliency map, we obtain the most likely location of the phone in the scene view when touch events take place (cf. Figure 5 top left). For our evaluation, we compute this in a leave-one-out cross-validation fashion: When testing on the data of a participant, we estimate that participant's mean saliency map on all other participants.

2. Estimating expected phone location: For a given test recording (i.e. in practice: during use), we retrieve all the (possibly drifted) initial gaze estimates that are within a one second window centred at a touch event. By taking their median, we obtain the expected location of the phone in the space of initial gaze estimates.

3. Estimating shift for recalibration: We can now estimate the shift (between 1. and 2.) by subtracting the expected phone location in the space of initial gaze estimates (2.) from the usual location of the phone in the scene view during touch events (1.). We recalibrate the initial gaze estimates by applying this shift.

5 EVALUATION

We evaluated both methods for 1) short and long-term calibration, 2) performance in different environments, and 3) influence of forced phone use (i.e. chat blocks in the dataset). We give an overview of our evaluation as follows:

In all evaluations, we measure gaze estimation error on the validation sequences that were recorded at the end of each recording block. For recalibration, we always use the data recorded right before the validation sequence that is used for measuring the error.

When evaluating the influence of forced phone use or environments, we restrict the data that is used for recalibration to certain phone usage conditions or environments, respectively.

The long- and short-term calibration settings differ with respect to which initial calibration sequence is used: In the long-term case we only use the first calibration sequence for every participant to extract an initial calibration, allowing us to investigate the effect on gaze estimation accuracy over an extended period of time. In the short-term case we always use the calibration sequence at the beginning of the recording block on which we evaluate our methods. This lets us investigate whether our methods are already useful after wearing the eye tracker for a shorter amount of time.

All evaluations use all 17 participants, with the exception of the evaluation for different environments where we make use of additional annotations which are present only for a subset of participants. The next sections report on these evaluations in detail.

5.1 Long-term Recalibration

We first evaluated the recalibration over an extended period of time, i.e. over the whole recording. To compare our proposed methods to the state of the art, we measured their performance after every recording block using the corresponding validation sequence. Our methods as well as the comparison methods used the calibration obtained from the calibration sequence before the first recording as a starting point. Saliency maps and touch events were always extracted from the recording block on which the error was measured. To robustly compare the different methods, we averaged the error over all recording blocks for each subject. See Figure 6 for a visualisation of the evaluation scheme and the resulting performances.

As can be seen in the Figure, our method using *phone detections* achieves an error of 117 pixels, which is significantly lower than both the initial calibration at 212 pixels (t=-2.13, p=0.049, df=16, two-tailed) and the state-of-the-art method by [Sugano and Bulling 2015] at 145 pixels (t=-2.33, p=0.033, df=16, two-tailed). Our *blind recalibration* method achieves an error of 121 pixels, reaching statistical significance compared to the method by Sugano et al. (t=-2.45, p=0.026, df=16, two-tailed), but not quite compared to the initial calibration (t=-1.86, p=0.082, two-tailed).

We also evaluated a saliency map incorporating touch events, which was generated by doubling the magnitude of activations on the detected phone at moments in time lying within a one second

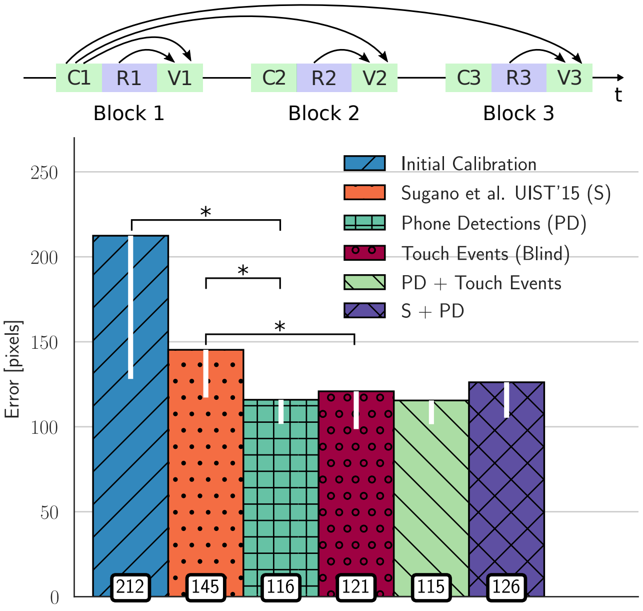

Figure 6: Top: Visualisation of the long-term recalibration setting. Arrows from Calibration segment C1 to validation segments VX indicate usage of the manual calibration from C1 and evaluation on VX. Arrows from the recording segments RX to validation segments VX indicate extraction of saliency maps and touch events from RX when evaluating on VX. Bottom: Our methods outperform baselines in this setting. Stars indicate statistically significant differences, white lines the lower parts of 95% confidence intervals (upper parts are symmetric).

window around each touch event. This approach reaches an error of 115 pixels, which is only slightly better than the plain phone saliency map. As incorporating touch events makes additional assumptions with respect to the recording setup (a phone needs to be equipped with recording software and synchronised with the eye tracker), we do not consider this approach further. Finally, we added the phone detection based saliency map to the saliency map constructed according to Sugano et al., reaching an error of 126 pixels. This combination thus did not reduce error further than our phone detection approach alone.

To quantify how stable our methods are under growing distance in time to the initial calibration, we also analysed their performances for each recording block separately (see Figure 7). While the error of the initial calibration increased strongly as time progressed, the error of all automatic recalibration methods was relatively stable. Our approach based on phone detections consistently achieved the lowest error, followed by our method on touch events, and by the state of the art by [Sugano and Bulling 2015].

5.2 Performance in Different Environments

To investigate the robustness of our method with respect to different environments, we evaluated it using only data from a specific environment for recalibration. To this end, we made use of the additional annotations that were collected for 14 of the participants [Steil et al. 2018a]. We evaluated our recalibration methods for three environments that each participant was asked to visit at least once during the study, namely the canteen, a coffee shop and the library. These environments are interesting for evaluation, as they correspond to different tasks participants perform alongside phone interactions. Furthermore, they differ significantly with respect to the amount of other people that are present. For the canteen, on average we count 694 frames with face detections per minute, whereas it is 289 for the coffee shop and only 94 for the library. For each participant we selected one recording block in which the participant visited a specific environment for evaluation. If a participant visited the same environment in more than one recording block, we chose the recording block containing the longest visit. Additionally, to ensure that the environment "canteen" was behaviourally distinct from the other environments, we excluded four participants in this conditions who did not have a meal during their visit to the canteen.

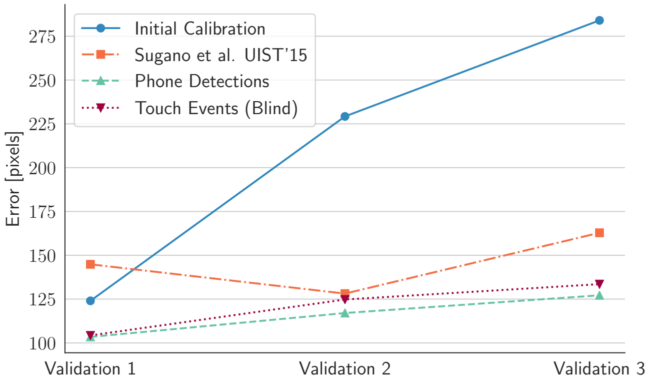

Figure 7: Our methods showing better performance than the baselines on every validation sequence after each of the three recording blocks, when using the manual calibration at the beginning of the first recording block as a starting point. Lines are added to connect individual measurements.

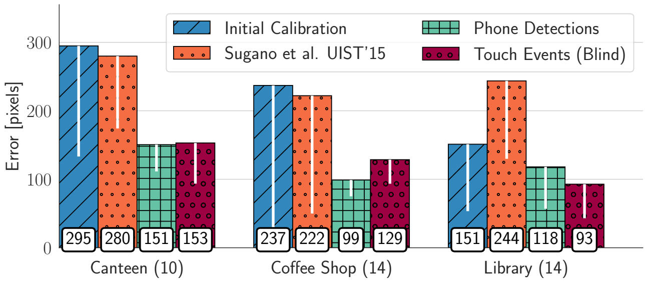

Figure 8: Analysis of robustness showing consistent results of our methods in different environments. The number of subjects from which the results for a specific environment are obtained is given in brackets. White lines show the lower parts of 95% confidence intervals (upper parts symmetric).

The results of this evaluation are shown in Figure 8. Errors achieved in different environments cannot be compared directly, as different recording blocks are chosen for different environments.

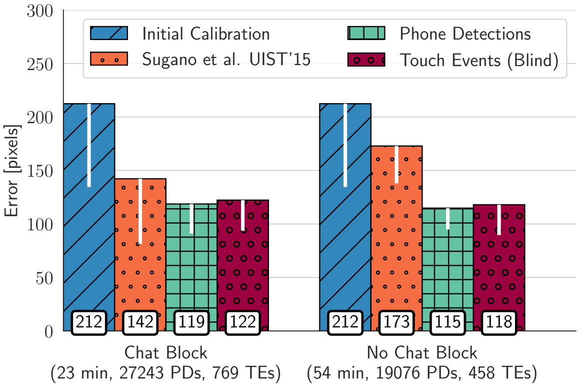

Figure 9: Similar patterns of results when exclusively using data from chat blocks versus non-chat block time periods. Numbers in brackets indicate average length, average number of phone detections and average number of touch events in chat blocks / outside chat blocks during a recording block. White lines show the lower parts of 95% confidence intervals (upper parts are symmetric).

The general pattern, however, shows that our two proposed methods perform consistently better than both the initial calibration and the state-of-the-art method [Sugano and Bulling 2015].

5.3 Influence of Chat Blocks on Performance

In each recording block of the dataset, several chat blocks took place [Steil et al. 2018b], in which the experimenter chatted with the participant. This implies that the participant is forced to use the phone, thereby generating phone detections and touch events. We thus investigated the influence of chat blocks on the performance of our methods: We analysed if the pattern of results stayed the same regardless of whether we restrict our saliency map generation and touch event usage to 1) the chat blocks contained in a recording, or 2) the other parts of the recording (i.e. no chat blocks).

The split of the data resulted in the following numbers of detections: On average, chat blocks took up 23 out of 77 minutes of a recording block. During the chat block portion of a recording block, there were on average 27,243 frames with phone detections and 769 touch events, while the non-chat block portion contained on average 19,076 phone detections and 458 touch events.

Figure 9 shows that the pattern of results is indeed the same in both conditions: Both our methods achieved a lower error than the initial calibration and the state of the art by [Sugano and Bulling 2015]. It is important to note that direct performance comparisons between the "chat block" and "no chat block" conditions must not be drawn, since the amount of data in each of the conditions is different. The most likely explanation of the slightly worse performance of our proposed methods for chat blocks compared to the "no chat block" condition is this: Although the number of phone detections and touch events is higher during chat blocks, time spent outside of chat blocks is much higher, potentially leading to more diverse samples of phone detections and touch events.

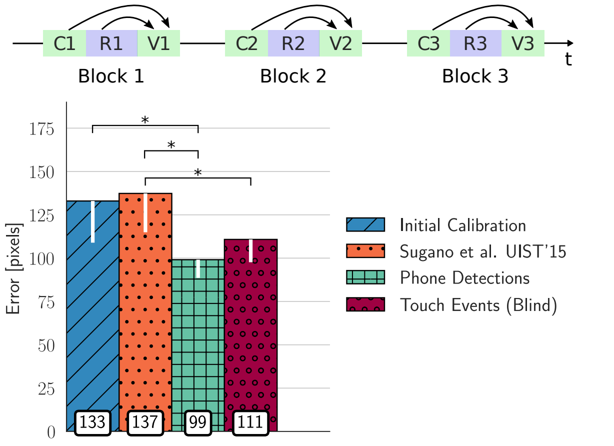

Figure 10: Top: Illustration of the short-term recalibration scenario (see Figure 6 for an explanation). Bottom: Our methods outperform baselines in this scenario. Stars indicate statistically significant differences, white lines lower parts of 95% confidence intervals (upper parts symmetric).

5.4 Short-term Recalibration

Finally, we evaluated whether our method was useful in rather short eye-tracking recordings by treating each recording block in the same way: We extracted an initial calibration from the calibration sequence right before the recording block started. Saliency maps and touch events for the evaluated recalibration approaches were extracted from the recording block, and the performance of methods was evaluated on the calibration sequence at the end of the recording block. Errors were averaged over all recording blocks for a given participant, yielding a more robust estimate of the performance compared to the analysis presented in Figure 7.

The results are shown in Figure 10. As can be seen from the figure, our method based on *phone detections* achieved the lowest error with 99 pixels, significantly outperforming the initial calibration at 133 pixels (t=-2.78, p=0.013, df=16, two-tailed) and the state of the art at 137 pixels error (t=-2.66, p=0.017, df=16, two-tailed). Our method based on *touch events* performed slightly worse with an error of 111 pixels. It still reached statistical significance compared to the state of the art (t=-2.22, p=0.041, df=16, two-tailed), yet not compared to the initial calibration (t=-1.35, p=0.195, df=16, two-tailed). Interestingly, the state-of-the-art saliency based method was not able to improve above the initial calibration in this evaluation.

6 DISCUSSION

6.1 Recalibration Performance

Our approaches to automatic recalibration outperform the state of the art significantly and consistently across different evaluation scenarios. They improve eye tracking accuracy both in short- and in long-term recordings and in different situations like eating in a canteen, sitting in a library or visiting a coffee shop. Our approach based on saliency maps built from phone detections performs slightly better than our blind calibration approach based on touch events, which still significantly outperforms the state of the art.

6.2 Initial Manual Calibration

Our approach requires initial manual calibration (see evaluations). We also tested phone saliency maps without initial calibration but observed very inaccurate results. This is explained by the relatively narrow area near the bottom centre of the scene view in which phones occur most of the time (cf. Figure 5), thereby not providing diverse enough samples to estimate the full calibration parameters. Nevertheless, we have shown that these samples are still suitable to estimate and correct calibration drift. Moreover, we specifically exploited this "peak phone area" in our blind recalibration approach.

6.3 Dataset and Study Setting

We used the mobile eye tracking dataset provided by Steil et al. [Steil et al. 2018b], which contains a rich diversity of everyday situations. One particular aspect of the study setting and dataset are the "chat blocks" in which the experimenter triggered text messaging with the participant. It is worth reflecting on whether this yields an unrealistically high degree of phone use. Considering the findings on phone use, mobile messaging, and typing in the literature (e.g. [Buschek et al. 2018; Dingler and Pielot 2015; Sahami Shirazi et al. 2014]), we argue that the covered extent of chatting is not unrealistic. Moreover, we evaluated our approaches also on the parts of the dataset that involved no such study-triggered phone use and found comparable results (see Figure 9).

6.4 Applications

By significantly decreasing calibration drift, our recalibration approaches facilitate everyday use of interaction techniques that require precise gaze estimation: Examples include multi-modal mobile interaction that combines touch and gaze input, for example to redirect direct touch to a cursor at the gaze position on a table [Pfeuffer and Gellersen 2016]. Another proposed concept combines pen and gaze input in a similar way [Pfeuffer et al. 2015]. With our novel recalibration methods we take an important step towards enabling such interaction techniques in daily life.

6.5 Privacy

A privacy concern of mobile eye tracking is the scene camera, which might record sensitive information, in particular if it is also used to record lifelogging videos [Steil et al. 2018a], and does not indicate its recording status to bystanders [Koelle et al. 2018]. Korayem et al. [Korayem et al. 2016] used a CNN-based computer vision approach to detect displays (phone, PC, etc.) in egocentric lifelogging videos, which users perceive as sensitive content [Hoyle et al. 2015, 2014]. Such scenes or image regions could then be blurred or redacted. This could easily be integrated with phone-based recalibration: The combined system would detect the phone display, recalibrate the eyetracker, and redact the display area in the lifelogging video. Moreover, Steil et al. [Steil et al. 2018a] used Deep Learning and both scene video and eye movement data to inform when to start/stop recording to avoid capturing sensitive content. This leads to interruptions in the scene video. Our touch-based approach could recalibrate the eye tracker during such interruptions.

In summary, if the eye tracker's scene camera recordings are stored (e.g. for lifelogging), phone detection in the scene can be exploited *both* for recalibration and privacy redaction. In contrast, if the scene recordings are not needed or momentarily interrupted, then our touch-based approach avoids the need for input from a scene camera altogether and thus helps to preserve privacy.

6.6 Outlook: Generalising our Approaches

While we utilised phone interactions, both our recalibration approaches could be extended to other devices, such as mobile devices like tablets, smart watches and laptops [Zhang et al. 2018], or stationary devices like public displays. For the extension of our approach based on phone saliency maps, related work on automatic detection of screens can be helpful [Korayem et al. 2016].

A main conceptual generalisation of our blind recalibration approach would no longer assume visual attention and interaction at the *same* location, but rather include cases with *separate* locations: For example, we might assume visual attention on a desktop monitor while keystrokes appear at the keyboard, or attention on a smart TV while using the remote control. For these cases, we need to be able to make robust assumptions about locations of these objects in the scene camera view. This might hold for some devices and contexts but not for others. For example, a laptop might commonly be located at the bottom to centre area of the camera view, while a smart TV might appear in different areas depending on where the user is sitting. These considerations present ample opportunities for future work, which could systematically collect such use cases beyond phone and touch, and investigate how to generate and exploit saliency maps for them.

Finally, our approach might be generalised beyond interaction with computing devices, such as reaching for a coffee mug, operating an elevator or a vending machine, and so on. Related, future work might exploit gaze behaviour in social situations, such as looking at faces and speakers (cf. [Müller et al. 2018; Siegfried et al. 2017]), or following pointing hands or handing over objects (e.g. money at a counter).

7 CONCLUSION

In this work we presented two novel methods to recalibrate mobile eye trackers by exploiting mobile phone usage. Our first method is based on saliency maps built from phone detections obtained from the scene camera. Our second method "blindly" recalibrates the eye tracker using touch events registered on the mobile phone. We evaluated both methods against the state of the art on a recent dataset of in-the-wild mobile eye tracking recordings. Both our methods reduced calibration drift and significantly outperformed the state-of-the-art method. While our blind recalibration approach performs slightly worse than our phone detection-based one, it offers advantages in privacy-sensitive situations, as it does not rely on images obtained from a scene camera. As such, we believe our work represents an important step towards enabling gaze-based interaction techniques in daily life.

ACKNOWLEDGMENTS

This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, Germany, as well as by a JST CREST research grant under Grant No.: JPMJCR14E1, Japan.

REFERENCES

Fares Alnajar, Theo Gevers, Roberto Valenti, and Sennay Ghebreab. 2013. Calibration-Free Gaze Estimation Using Human Gaze Patterns. In *Proc. of the IEEE International Conference on Computer Vision.* 137–144. https://doi.org/10.1109/ICCV.2013.24

Michael Backes, Markus Dürmuth, and Dominique Unruh. 2008. Compromising Reflections -or- How to Read LCD Monitors around the Corner. In *Proc. of the IEEE Symposium on Security and Privacy.* 158–169. https://doi.org/10.1109/SP.2008.25

Ali Borji and Laurent Itti. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Andreas Bulling and Hans Gellersen. 2010. Toward Mobile Eye-Based Human-Computer Interaction. *IEEE Pervasive Computing* 9, 4 (2010), 8–12. https://doi.org/10.1109/MPRV.2010.86

Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* Article 255, 14 pages. https://doi.org/10.1145/3173574.3173829

Jixu Chen and Qiang Ji. 2015. A Probabilistic Approach to Online Eye Gaze Tracking Without Explicit Personal Calibration. *IEEE Transactions on Image Processing* 24, 3 (2015), 1076–1086. https://doi.org/10.1109/TIP.2014.2383326

Kai Dierkes, Moritz Kassner, and Andreas Bulling. 2018. A novel approach to single camera, glint-free 3D eye model fitting including corneal refraction. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications.* Article 9, 9 pages. https://doi.org/10.1145/3204493.3204525

Tilman Dingler and Martin Pielot. 2015. I'll Be There for You: Quantifying Attentiveness Towards Mobile Messaging. In *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services.* 1–5. https://doi.org/10.1145/2785830.2785840

Andrew T Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 455–470. https://doi.org/10.3758/BF03195475

Martin A. Fischler and Robert C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (1981), 381–395. https://doi.org/10.1145/358669.358692

Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9, 4 (2005), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Roberto Hoyle, Robert Templeman, Denise Anthony, David Crandall, and Apu Kapadia. 2015. Sensitive Lifelogs: A Privacy Analysis of Photos from Wearable Cameras. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 1645–1648. https://doi.org/10.1145/2702123.2702183

Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy Behaviors of Lifeloggers Using Wearable Cameras. In *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing.* 571–582. https://doi.org/10.1145/2632048.2632079

Michael Xuelin Huang, Tiffany CK Kwok, Grace Ngai, Stephen CF Chan, and Hong Va Leong. 2016. Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 5169–5179. https://doi.org/10.1145/2858036.2858404

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication.* ACM, 1151–1160. https://doi.org/10.1145/2638728.2641695

Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices: Survey and Lessons Learned. In *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services.* Article 38, 17 pages. https://doi.org/10.1145/3229434.3229452

Marion Koelle, Katrin Wolf, and Susanne Boll. 2018. Beyond LED Status Lights - Design Requirements of Privacy Notices for Body-worn Cameras. In *Proc. of the International Conference on Tangible, Embedded, and Embodied Interaction.* 177–187. https://doi.org/10.1145/3173225.3173234

Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. 2016. Enhancing Lifelogging Privacy by Detecting Screens. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 4309–4314. https://doi.org/10.1145/2858036.2858417

Christian Lander, Markus Löchtefeld, and Antonio Krüger. 2017. hEYEbrid: A Hybrid Approach for Mobile Calibration-free Gaze Estimation. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4, Article 149 (2017), 29 pages. https://doi.org/10.1145/3161166

Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications.* 31:1–31:10. https://doi.org/10.1145/3204493.3204549

Antti Oulasvirta, Tye Rattenbury, Lingyi Ma, and Eeva Raita. 2012. Habits Make Smartphone Use More Pervasive. *Personal and Ubiquitous Computing* 16, 1 (2012), 105–114. https://doi.org/10.1007/s00779-011-0412-2

Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 919–928. https://doi.org/10.1145/1054972.1055101

Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. In *Proc. of the ACM Symposium on User Interface Software and Technology.* 373–383. https://doi.org/10.1145/2807442.2807460

Ken Pfeuffer and Hans Gellersen. 2016. Gaze and Touch Interaction on Tablets. In *Proc. of the ACM Symposium on User Interface Software and Technology.* 301–311. https://doi.org/10.1145/2984511.2984514

Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing.* 825–836. https://doi.org/10.1145/2750858.2804252

Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale Assessment of Mobile Notifications. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 3055–3064. https://doi.org/10.1145/2556288.2557189

David Sculley. 2010. Web-scale k-means clustering. In *Proc. of the International Conference on World Wide Web.* 1177–1178. https://doi.org/10.1145/1772690.1772862

Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. 2017. Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation. In *Proc. of ACM International Conference on Multimodal Interaction.* ACM. https://doi.org/10.1145/3136755.3136793

Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2018a. *PrivacEye: Privacy-Preserving First-Person Vision Using Image Features and Eye Movement Analysis.* Technical Report.

Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018b. Forecasting User Attention During Everyday Mobile Interactions Using Device-integrated and Wearable Sensors. In *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services.* Article 1, 13 pages. https://doi.org/10.1145/3229434.3229439

Yusuke Sugano and Andreas Bulling. 2015. Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency. In *Proc. of the ACM Symposium on User Interface Software and Technology.* 363–372. https://doi.org/10.1145/2807442.2807445

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2010. Calibration-free gaze sensing using saliency maps. In *Proc. of the IEEE Conference on Computer Wision and Pattern Recognition.* 2667–2674. https://doi.org/10.1109/CVPR.2010.5539984

Yusuke Sugano, Yasuyuki Matsushita, Yoichi Sato, and Hideki Koike. 2008. An Incremental Learning Method for Unconstrained Gaze Estimation. In *Proc. of the European Conference on Computer Vision.* 656–667. https://doi.org/10.1007/978-3-540-88690-7_49

Lech Swirski and Neil Dodgson. 2013. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting. In *Proc. of the European Conference on Eye Movements.*

Kentaro Takemura, Shunki Kimura, and Sara Suda. 2014a. Estimating point-of-regard using corneal surface image. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications.* 251–254. https://doi.org/10.1145/2578153.2578197

Kentaro Takemura, Tomohisa Yamakawa, Jun Takamatsu, and Tsukasa Ogasawara. 2014b. Estimation of a focused object using a corneal surface image for eye-based interaction. *Journal of Eye Movement Research* 7, 3 (2014), 1–9. https://doi.org/10.16910/jemr.7.3.4

Hirotake Yamazoe, Akira Utsumi, Tomoko Yonezawa, and Shinji Abe. 2008. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications.* ACM, 245–250. https://doi.org/10.1145/1344471.1344527

Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. 2018. Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems.* 624:1–624:12. https://doi.org/10.1145/3173574.3174198