28 Blinks Later: Tackling Practical Challenges of Eye Movement Biometrics

Simon Eberz
simon.eberz@cs.ox.ac.uk
University of Oxford

Giulio Lovisotto
giulio.lovisotto@cs.ox.ac.uk
University of Oxford

Kasper B. Rasmussen
kasper.rasmussen@cs.ox.ac.uk
University of Oxford

Vincent Lenders
vincent.lenders@armasuisse.ch
armasuisse

Ivan Martinovic
ivan.martinovic@cs.ox.ac.uk
University of Oxford

ABSTRACT

In this work we address three overlooked practical challenges of continuous authentication systems based on eye movement biometrics: (i) changes in lighting conditions, (ii) task dependent features and the (iii) need for an accurate calibration phase. We collect eye movement data from 22 participants. To measure the effect of the three challenges, we collect data while varying the experimental conditions: users perform four different tasks, lighting conditions change over the course of the session and we collect data related to both accurate (user-specific) and inaccurate (generic) calibrations.

To address changing lighting conditions, we identify the two main sources of light, i.e., screen brightness and ambient light, and we propose a pupil diameter correction mechanism based on these. We find that such mechanism can accurately adjust for the pupil shrinking or expanding in relation to the varying amount of light reaching the eye. To account for inaccurate calibrations, we augment the previously known feature set with new features based on binocular tracking, where the left and the right eye are tracked separately. We show that these features can be extremely distinctive even when using a generic calibration. We further apply a cross-task mapping function based on population data which systematically accounts for the dependency of features to tasks (e.g., reading a text and browsing a website lead to different eye movement dynamics).

Using these enhancements, even while relaxing assumptions about the experimental conditions, we show that our system achieves significantly lower error rates compared to previous work. For intra-task authentication, without user-specific calibration and in variable screen brightness and ambient lighting, we achieve an equal error rate of 3.93% with only two minutes of training data. For the same setup but with constant screen brightness (e.g., as for a reading task) we can achieve equal error rates as low as of 1.88%.

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3354233

CCS CONCEPTS

• **Security and privacy → Biometrics**;

KEYWORDS

authentication, biometrics, eye movements

ACM Reference Format:
Simon Eberz, Giulio Lovisotto, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2019. 28 Blinks Later: Tackling Practical Challenges of Eye Movement Biometrics. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3319535.3354233

1 INTRODUCTION

Authentication based on various biometric modalities has become increasingly popular in recent years. This surge has been mostly driven by the integration of biometric sensors in smartphones, with fingerprint scanning and face recognition being nowadays available in most devices. With the increasing use of deep models, both fingerprint and face recognition can now offer low error rates and convenient recognition times.

While such physiological biometrics offer accurate and fast recognition, they are easily *observable*: both fingerprints and faces can be easily obtained and forged by adversaries. While there have been extensive efforts to detect spoofed samples (e.g., fake fingers), this often unfolds into an arms race with attackers improving their artifact to circumvent liveness detection. In comparison, behavioral biometrics rely on distinctive behaviour rather than physical characteristics. As such, behavioral traits are inherently less observable compared to physiological ones. Different behaviour-based modalities have been proposed by the academic community, including keystroke dynamics [19], touchscreen input dynamics [16], gait [5, 24], mouse movements [38], electrocardiography (ECG) [14] and mobile device pickups (MDP) [15]. These systems may combine both behavioural and physiological components, e.g., touch dynamics make use of touch pressure, which partially depends on the size of the user's finger.

Eye movements have recently gained interest as a behavioural biometric with a strong physiological component [3, 11, 12, 17, 20, 21, 31, 32, 35]. With advances in technology, video-based eye trackers are increasingly cheap and integrated in consumer devices. Eye movement biometrics (not to be confused with iris recognition) combine the steadiness of gaze, rapid short-term movements, the shape and duration of visual fixations and the (changes in) size

of the person's pupil. While eye movements has been shown to achieve relatively low error rates, there are three major practical challenges of using eye movements in realistic settings: (i) the need for a *precise calibration*, (ii) the eye movements' *task-dependency* and (iii) the pupil *light-sensitivity*.

A precise calibration is required in order to obtain good eye tracking accuracy and subsequently accurate eye movements features. While calibration itself may be relatively quick (<10 seconds), it is highly sensitive to changes in the user's interaction (e.g., posture, distance between eyes and screen, head movements), leading to poor stability over time. Additionally, previous work has shown that eye movements are highly task-dependant, and that authentication across different tasks (i.e., enrolling the user on one task and authenticating on another) leads to significant increases in error rates. Finally, given the importance of pupil-based features, changes in pupil diameter caused by changing lighting environment (e.g., ambient light, screen brightness) compromise the stability and accuracy of the recognition.

In this paper, we propose new methods to address the three challenges of using eye movements biometrics in realistic settings and combine them to develop an authentication system. We collect data from 22 participants recruited from the general public, across two different sessions. For each partipant, we collect eye movement data across different tasks, different lighting conditions (including both screen brightness and ambient light) and different calibrations accuracies. We augment the eye movements feature set used in previous work to include binocular features based on the difference of tracking between left and right eye. We show that the full set of features can successfully discriminate users without requiring user-specific calibrations. We propose a pupil diameter correction mechanism that accounts for the screen brightness and level of ambient light in order to refine the pupil diameter measurement coming from the sensor. We further show that using a population based cross-task mapping function, we can automatically adjust for the task-dependent changes in feature distributions, improving the accuracy of authentication across tasks.

The **contributions** of this paper can be summarized as follows:

- We develop a new eye movements recognition pipeline which accounts for imprecise calibrations, changes in lighting environment and cross-task authentication.
- We test the system on 22 participants recruited from the general public, across two separate sessions. We test a set of different lighting conditions (i.e., screen brightness and ambient light), tasks and calibration quality.
- We make our dataset and the code used for the experiments available online[1].

The remainder of the paper is organised as follows: Section 2 outlines background and related work on eye movement biometrics. Section 3 and 4 describe our experimental design and methods. In Section 5 we present our results. We discuss the security of our approach in Section 6 and conclude the paper in Section 7.

2 RELATED WORK

In this section, we provide an overview of the medical foundation of eye movements, eye tracking technology and eye movement authentication systems.

2.1 Eye movement background

The human eye moves within six degrees of freedom with six muscles responsible for the movement of the eyeball. The main types of eye movements can be categorized into *saccades* and *fixations*, while the neural signs controlling these movements can be categorized as voluntary, involuntary and reflexive. Saccades are rapid stepwise movements of both eyes in the same direction that typically last 10-100 ms, depending on the distance covered [8] and are used to shift the gaze to another location. In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100-400 ms and are used to stabilize the retina over a stationary object of interest. Yet, eyes are never perfectly still and exhibit involuntary movements even during visual fixations. The main reason for such movements is to counteract retinal fatigue and to prevent visual fading. One type of such movements are microsaccades, characterized by high velocity and acceleration often away from the fixation centre [29].

Besides the eye movements, the pupil diameter is also an distinctive feature which can be included in the analysis of eye behaviour. The range for this feature in an individual is largely determined by eye physiology, gender and ethnicity and is relatively constant during adulthood [28]. Nevertheless, multiple causes that affect the pupil diameter have been found, including memory and cognitive workload [25], lighting conditions [36] and drug consumption [23]. The pupil size also shrinks as a person ages, an effect which is particularly pronounced in low lighting conditions [37].

2.2 Eye tracking technology

Eye tracking is the process of tracking the position and movements of a person's eye. When these movements are calibrated with regard to an external screen (i.e., the system determines gaze points), the process is called gaze tracking. There are two main approaches to eye tracking: electrooculography (EOG) and video-based tracking. EOG is a tracking technique that measures electrical potential between two adhesive electrodes which are placed around the eyes. This approach is popular in the medical field, as it enables accurate recordings of eye movements even while the eyes are closed (e.g., during blinks or while the subject is sleeping). However, this is a rather invasive technique which is unlikely to be acceptable to users outside medical and research trials.

On the other hand, video-based eye tracking involves the recording of the user's eyes through cameras with high frame rates. While tracking is possible with conventional RGB cameras, accuracy is usually enhanced by using an infrared camera and an additional infrared light source. The users retinas reflect the light source allowing for a more accurate tracking of the eye's positions. As infrared light is invisible to the human eye, the tracking itself is completely non-invasive and not noticeable to the user. In order to determine the user's gaze point on a screen, the system has to be calibrated. The calibration process requires the user to look at a sequence of

[1] https://simonizor.github.io/28blinkslater

Study	Mode	Stimulus	Feature types	Cross-task	Calibration optional	Pupil correction	EER [%]
[32]	Login	Movie trailer	Fixation density map	✗	✗	✗	14
[17]	Login	Human faces	Distribution of area of interest	✗	✗	✗	36.1
[31]	Login	Human faces	Graph matching	✗	✗	✗	30
[3]	Login	Human faces	Scan paths	✗	✗	✗	25
[35]	Login	Moving dot	Fixation and saccade shape	✗	✗	✗	6.3
[12]	Continuous	Various tasks	Fixation and pupil features	✓	✗	✓	0.04 - 4.9
[11]	Continuous	Moving dot	Fixation and pupil features	✗	✗	✗	3.98
[20]	Continuous	Reading	Scan paths	✗	✗	✗	23
[21]	Continuous	Reading	Fixation and saccade shape	✗	✗	✗	16.5

Table 1: Summary of biometric eye movement authentication. Cross-task indicates whether the study measures cross-task authentication accuracy. Pupil correction indicates whether the study corrects for the effect of light on the pupil diameter.

points shown on the screen and is sensitive to posture, including the distance to the screen. Video-based eye tracking is increasingly available in consumer devices, laptops in particular, but even integrated in virtual or agumented reality headsets.

2.3 Eye movement authentication

The body of work on eye movement biometrics can be divided into three parts: Eye movements as an input channel, continuous authentication and login-time authentication.

Eye movements as an input channel. In the past, eye movements have been used as a mechanism to input conventional credentials (such as PINs [26, 27], passwords [26] and patterns). The main benefit lies in increased resistance to shouldersurfing (performed either by a human or through CCTV).

Bulling et al. propose an image-based gaze authentication system [2]. During enrolment, the user is shown a specific image and chooses a gaze path within the image as their secret. During authentication, the user is then shown the same image and has to replicate their enrolment-time gaze trace. In order to increase the entropy of these traces, the authors use saliency masks. The mask covers parts of the image that are most likely to attract the user's attention (such as faces) to prompt them to choose more random gaze paths. In the second part of the study, the authors showed participants close-up videos of another person's gaze, while entering an image-based password, and asked them to guess the "password". Users were successful in guessing a PIN-based password in 19 out of 81 cases, which dropped to 1/82 and 8/72 for image based passwords with and without saliency masks, respectively.

Login time biometric authentication. While the techniques in the previous section are used at login time, they merely use gaze as an input channel without making use of the biometric component of eye movements. Their benefit lies in their resistance to replay attacks (e.g., shouldersurfing) but they still require memorizing a secret (a PIN, password or image gaze sequence). If this secret is revealed, these techniques do not provide any further protection.

There are several proposed technique to achieve login time biometric recognition based on eye movement patterns, a summary is given in Table 1. Login time authentication systems have the advantage of being able to use controlled stimuli (rather than having to work with the user's normal system interactions). Therefore, they can measure the user's visual response to a controlled and fixed stimulus, without the user's eye movement patterns being influenced by changing stimuli. In addition, as the system knows the screen content at any moment it can use "high-level" features that make use of the user's gaze positions, rather than the more "low-level" saccadic or fixational movements. These high-level features include scan paths (i.e., the shape and position of the user's time-varying gaze points) or distribution of areas of interest and density maps (i.e., which part of an image the user focuses on the most). Techniques based on these high-level features exhibit relatively high error rates with the EER ranging from 6.3% to 30%. In addition, it is not yet known how time-stable these patterns are as users become more familiar with the (static) stimuli used for authentication.

The best error rates (an EER of 6.3%) have been achieved by Sluganovic et al., who propose a login time system using low-level eye movement features. They test their approach using a desktop-based eyetracking system with an SMI RED500 eye tracker and 30 participants [35]. During login and enrolment, users are asked to look at a red dot on the screen. Once the user's gaze focuses on the dot, it moves to a new, random position on the screen. The authentication process is then two-fold: To prevent replay attacks, the system confirms whether the newly recorded gaze positions match the (randomized) positions of the dots. If the data were simply replayed from a prior login, the positions would be unlikely to match (the reported success rate for the replay attack is 0.06%). The actual biometric verification is based on raw eye movement data without using the state of the stimulus.

All the systems [3, 17, 31, 32, 35] heavily rely on accurate calibrations. In fact, these studies make use of the relationship between the user's gaze and the visual stimulus position (e.g., red dot on a black screen). Imprecise calibrations, which could occur with slight posture changes or head movements, will affect the system performance, leading to recognition errors [17, 31] or to compromised replay detection [35].

Continuous authentication. The idea of continuous authentication is to establish the user's identity not just once at login time but also continuously while the person is using the system. As such, it is able to detect a change in user identity even after the initial login. Continuous authentication is only possible when the

authentication system does not rely on creating specific stimuli (as the process of authentication would otherwise interfere with the user's work). In this case, the system needs to account for the fact that a person's visual response may change as the stimulus changes, therefore it is necessary to choose biometric features that are as independent of the stimulus as possible.

There are a number of papers that propose eye movement-based continuous authentication systems [11, 12, 20, 21] (see Table 1), but most of them focus on scenarios where the user is working on a single specific task. In this case, the system can use the specific characteristics of the users' responses within such a task to develop distinctive task-dependent features (e.g., scan paths in [20]). However, as users carry out several different tasks while using a computer, a continuous system should be able to authenticate users across tasks, either by training the classifier for each task or by making cross-task predictions. Eberz et al. [11, 12] investigated authentication with several real-world tasks (i.e.,reading, typing, web browsing and watching different videos). Similarly to [35], the biometric features are low-level, i.e., do not relate to the state of a stimulus. The authors use three feature types: (a) spatial features reflect the size and shape of fixations, (b) temporal features measure the speed of (micro-)saccades and (c) pupil features measure changes in the size of the pupil. Out of this feature set, the pupil diameter contributed the biggest amount of information, followed by temporal features and spatial features.

In [11, 12], same-task recognition rates vary depending on the task: 0.04% (browsing) to 4.9% (typing). However, recognition performance drops significantly when authenticating on a task the system was not trained on (e.g., enrolment data is for browsing, test data is for typing). In particular for typing, the study shows that typing is quite problematic, leading to EERs close to 50% (i.e., random guessing) when using eye movements data collected during typing to authenticate the user during other tasks. For some tasks combination, an improvement is achieved by correcting the pupil diameter for the brightness of the screen content, see Section 2.1. While (non-controlled) changes in screen brightness are accounted for, the authors do not consider changes in ambient light. In addition, they only use a small user sample (10 users) for the task dependence and brightness adjustment experiment which questions the robustness of their results.

3 EXPERIMENTAL DESIGN

Here we describe our experimental goals and the setup and outline of our data collection procedure.

3.1 Design Goals

The main challenges raised by previous work are three-fold: (a) requirement of a precise calibration, (b) effects of changing ambient light and screen brightness and (c) task dependence of features. The objective to eliminate these effects is reflected in our design goals:

- **Calibration-free operation:** The system should not rely on an accurate calibration. As all commercial eyetrackers require calibration data, this design goal can be satisfied by either using random calibration data or by using a generic calibration not tailored to the current user.

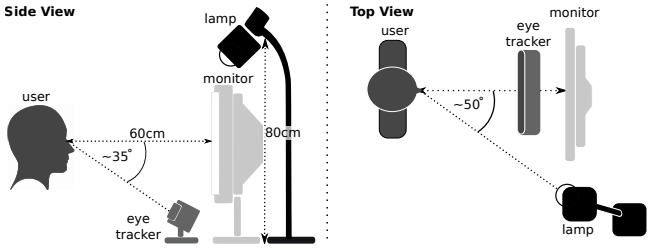


Figure 1: Experimental setup. Users are sat on a chair around 60cm away from the monitor. The eye tracker sits at the base of the monitor. The lamp is positioned around 80cm above the desk and to the right of the user. The reported angles vary slightly depending on the participant's height and posture.

- **Task independence:** The system should reduce the effect of task dependent features on error rates during cross-task authentication (i.e., when training and testing on different tasks).
- **Light-invariant features:** The system's error rates should not change significantly if the levels of ambient light or the brightness of the screen content changes during training, testing or between training and testing.

3.2 Experiment Setup

Figure 1 shows a representation of our experimental setup. We use an SMI RED500 eye tracker capturing samples at 500Hz. Unlike previous work, we use binocular eye tracking, which reports separate gaze positions and pupil diameters for both eyes. The user is facing a 22" screen with a 1920x1080 resolution positioned about 60cm away. The screen is set to the same brightness for all users (although the screen *content* brightness varies throughout the experiment as discussed below). In order to vary the ambient light, we use a desk lamp with a Philips Hue light bulb placed to the right of the screen. The bulb's brightness can be programmatically controlled. The lamp is angled towards the keyboard to avoid blinding the user on higher brightness settings. The room itself is illuminated with a ceiling light dimmed to a moderate level in order to achieve sufficient variation in brightness by using the desk lamp. Table 2 shows the complete experiment data collection for a single participant. In the following we explain how we designed the experiment to collect data which allows us to test our goals: (i) effect of calibration, (ii) effect of task selection and (iii) effect of light sensitivity.

Calibration. The RED500 eye tracker requires to be calibrated in order to collect accurate gaze. The calibration maps different rotations of each eyeball to the respective gaze positions on the screen. As such, calibration depends on a set of factors including the size and resolution of the screen, position of the eyetracker, the distance of the user to the screen, the distance between the eyes and the user's viewing angle.

In order to test the effect of calibration on the collected data, we conduct two separate sessions for each study participant. In the first session, we calibrate the eye tracker with a user-specific calibration. To create such calibration, we use a 9-point calibration procedure (T0 in Table 2), followed by a 4-point validation procedure that

	TaskId	Performed task	User-specific calibration	Screen brightness	Ambient light	Duration (s)
Calibrated session	T0	User calibration	—	—	—	60
	T1	Calibration validation (pre)	✓	—	—	10
	T2	Slideshow	✓	increasing	constant	180
	T3	Reading	✓	constant	increasing	300
	T4	Browsing	✓	constant	increasing	300
	T5	Slideshow	✓	random	increasing	300
	T6	Calibration validation (post)	✓	—	—	10
Uncalibrated session	—	Load random calibration	—	—	—	—
	T7	Calibration validation (pre)	✗	—	—	10
	T8	Slideshow	✗	increasing	constant	180
	T9	Reading	✗	constant	increasing	300
	T10	Browsing	✗	constant	increasing	300
	T11	Slideshow	✗	random	increasing	300
	T12	Calibration validation (post)	✗	—	—	10

Table 2: Outline of the complete experiment for each participant. Each participant undergoes two sessions: in the first one we compute a user-specific calibration (T0), in the second one we load a random calibration profile. Within one session, each participant completes four tasks: slideshow (T2, T8), reading (T3, T9), browsing (T4, T10) and slideshow again (T5, T11). During each task we vary the screen brightness, the ambient light, or both. At the beginning and at the end of each session, we measure the calibration error with a validation procedure (T1, T6, T7, T12). The table reports the tasks in chronological order (from the top) and the two sessions for one participant are collected at least two hours apart from each other.

measures the accuracy of the calibration (T1). In order to obtain precise data, we repeat the calibration (T0 and T1) until the mean calibration error across both eyes and the X- and Y-direction is less than 1 degree in the validation phase. We then save both the calibration accuracy and the calibration coefficients of the accepted calibration. We perform another 4-point validation at the end of the session (T6) to test whether the tracking accuracy changed over the course of the session (e.g., due to changes in posture or excessive head movements).

In the second session, rather than computing a user-specific calibration, we load a different participant's calibration profile instead. While the position of the screen and eye tracker are fixed for each session, the remaining factors affecting the calibration are uncontrolled (e.g., participants height, posture, head angle). Similarly to the first session, we measure the accuracy resulting from the different calibration profile with the same 4-point validation procedure, both at the beginning (T7) and at the end of the session (T12). We always use the previous participant's calibration profile, i.e., the calibration profile of the participant who was measured last. The reason we use the previous user's calibration, rather than a single generic calibration for all users, is to limit the effect of how the calibration is chosen. As the eye tracker can not retroactively apply different calibrations to raw video data, we are limited to one calibration setting per session. Throughout the rest of the paper, we refer to these two sessions, the one with a user-specific calibration and the one using a different user calibration, as the *calibrated* and the *uncalibrated* sessions, respectively.

Task selection. Similar to previous work [11, 12], we choose three main tasks inspired by day-to-day activities: reading (T3 and T9 in Table 2), web browsing (T4, T10) and an image slideshow (T2, T5, T8, T11). Each task lasts approximately 3-5 minutes, after which the

Figure 2: Four images used in the slideshow task. Images are sorted by increasing brightness. The brightnesses computed with the root mean squared method are 47.4, 123.8, 142.2, 192.2 (left to right, top to bottom).

experiment continues automatically. The reading task consists of reading an excerpt of Alice in Wonderland [4]. The text is shown in a centred column on the screen on a grey background. We instructed users to flip pages using the keyboard once they reached the end of a page. Typically, users flipped page around three times during the task. For the web browsing task, users browse Wikipedia: they are shown a random Wikipedia article and asked to use (chains of) links within the article to reach a target (Wikipedia) article. This type of activity involves both skimming and reading and is therefore similar to typical browsing patterns. During the slideshow task, users watch a sets of images in a slideshow, where each image is shown for two seconds before being substituted with the next one. While conceptually very close to the videos used in previous

work (e.g., [12, 32]), using images instead of videos allows to better control the variation of brightness levels. We repeat the slideshow task twice within one session to collect additional information about the effect of varying light (see next paragraph). We choose a set of nature-themed images while filtering images that may elicit extreme user responses, such as spiders. Figure 2 shows four of the images used in the slideshow, sorted in terms of increasing brightness (left to right, top to bottom).

Light variability. As mentioned before, two factors mainly affect the amount of light perceived by the pupil: the screen brightness and the ambient light. Therefore, within one session, we vary one or both factors within each task. In particular, for the first repetition of the slideshow (T2, T8) we increase the brightness of the images shown on the screen (these are sorted beforehand and shown in increasing brightness order) while keeping the ambient light constant. During the reading and browsing tasks (T3, T4, T9, T10), the screen brightness is constant (the largely text-based nature of both tasks results in negligible brightness differences), but we increase the amount ambient light. For the second repetition of the slideshow (T5, T11), we instead choose a random order for the images, while again increasing the amount of ambient light. We determine an image's brightness by calculating the average root mean squared pixel brightness of its greyscale representation. We always vary the light (both ambient and screen) with increments rather than decrements. We choose this ascending order as the pupil's adaptation to increasing light is near-instantaneous, whereas adaption to darkness occurs over time. For tasks with increasing ambient light, we linearly increase the brightness of the desk lamp from the minimum value to the maximum one. Varying the amount of light which the participant is subjected to allows us to re-create realistic uncontrolled lighting conditions.

3.3 Data Collection Process

We recruited 22 participants (11 male, 11 female) from the general public, the only selection criteria were a minimum age of 18 and normal or corrected-to-normal vision. The age distribution and the presence of glasses and contact lenses are shown in Figure 3. We collect whether the user is wearing glasses or contact lenses as we found that, these often lead to less precise calibrations (due to the glasses lenses reflecting or altering the reflection of the infrared light used by the eye tracker). We advertised the study through social media and participants were compensated for their time. The data collection was approved by Oxford's Interdivisional Research Ethics Committee, reference R50977/RE002.

4 METHODS

In this section, we present the methods used to authenticate users based on their eye movement patterns. The source code for each of the steps and the data needed to precisely reproduce our results are available online.

4.1 Preprocessing

The SMI RED500 used for this study reports two different types of samples: raw gaze samples and fixation events. Raw samples are measured at a rate of 500Hz while fixations are computed automatically as they occur. Raw samples consist of a timestamp, X/Y

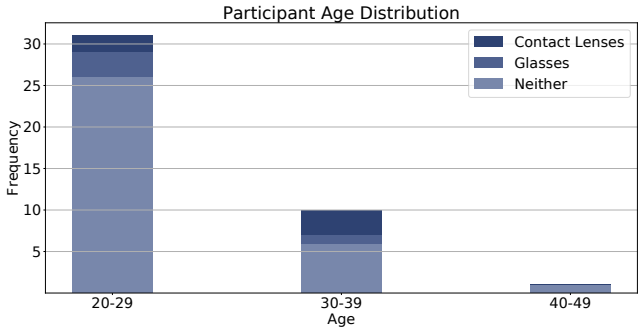


Figure 3: Age and eye sight correction (glasses, lenses or neither) distribution among the experiment participants. Note that each participant's session is counted separately as some participants wore glasses only for one of them.

coordinates and the pupil diameter. Since we use the tracker in binocular mode, the coordinates and pupil diameters are reported separately for each eye. For some samples, the eyetracker is unable to determine the pupil diameter for one or both eyes (which leads to them being reported as zero). This indicates an incorrectly tracked sample (e.g., following or during a blink). We therefore discard these samples. Fixations are calculated by the eyetracker using a proprietary algorithm. Conceptually, raw samples are clustered into fixations if they occur within a short window of low-velocity movements. Each fixation is associated with a centre point as well as start and end timestamps which enable us to find the associated raw samples. We only use an event if it contains at least 10 samples. This filters both unnaturally short fixations (10 samples correspond to 20 milliseconds) and those with an excessive number of missing or corrupt samples. Since our features are based on fixations, we discard all raw samples not belonging to a fixation (i.e., saccades, blinks and various noise).

4.2 Feature Extraction

Following the preprocessing, we compute a set of features for each fixation, so that each fixation leads to a feature vector used by the system classifier. In our data, we observe roughly five fixations per second on average, which leads to five biometric samples per second. In the following we describe the features used in our system.

Spatial-based features. These features relate to the spatial distribution of samples within a fixation. To capture the size of a fixation, we calculate each sample's distance to the fixation centre and use this measure's 10th percentile, 90th percentile, mean and standard deviation as features. As a measure of a fixation's shape, we compute the maximal pairwise distance between any two (potentially not consecutive) samples in the fixation. We use both Euclidean distance as well as individual distance in X and Y direction.

Temporal-based features. These features represent the eye movement speed during fixations. We measure pairwise speed and acceleration between consecutive samples and use the 10th percentile, 90th percentile, mean and standard deviation as features. We also compute the duration of a fixation.

Pupil-based features. Pupil features of the min, max and mean of the pupil diameter of the respective eye during the fixation. As pupil diameter measurements are far less noisy than coordinates, we use the min and max values, rather than percentiles. An individual's pupil diameter is not constant over their lifespan, in fact, as a person ages, their pupil diameter shrinks [6]. However, the timescale of these changes is too long to significantly impact the authentication system. A far bigger concern is its susceptibility to light. Both the light of the screen (which is changed by the brightness of the image shown) and ambient light change an individual's pupil size. The issue of screen light has been previously identified as a problem for eye movement authentication [12] and changing ambient light has been used as an attack vector [18]. We address these challenges in Section 4.3.

Binocular-based pupil features. We augment the set of features leveraging the binocular tracking offered by the eye tracker. In particular, we focus on pupil based features for each individual eye as medical work has shown that, even given stable lighting conditions, the pupil diameter of the left and right eye is not always identical [30]. Besides possible differences in actual pupil size, the tracking itself may cause differences between the left and right eye (e.g., depending on the user's posture, or head inclination). We use the min, max and mean tracking difference between the left and right eye as well as the difference in pupil size as binocular features.

Measuring feature quality. In order to compute the distinctiveness of each feature, we use the Relative Mutual Information (RMI). The RMI is defined as follows:

$$\text{RMI}(uid, F) = \frac{H(uid) - H(uid|F)}{H(uid)}$$

where H(A) is the entropy of A and H(A|B) denotes the entropy of A conditioned on B. The *uid* (i.e., the set of user identities) is discrete, but the feature space for most features is continuous. Therefore, binning would be required to discretize the features (as is done in, e.g., [11, 12]). However, the reported RMI would depend on the binning strategy and number of bins (with more bins leading to a higher calculated RMI). To avoid this problem, we use the non-parametric approach proposed by Ross et al. to estimate the mutual information between the discrete user ID and continuous features [33].

4.3 Pupil diameter correction

We use linear regression to model the pupil's response to changing levels of light, both screen brightness and ambient light. Figure 4 shows an example of the effect of increasing ambient light on one participants pupil diameter. As expected, across the entire dataset, we find on average a negative correlation between pupil diameter and amount of light: r-values of -0.5818 ± 0.05 and -0.2140 ± 0.07 for screen brightness and ambient light, respectively. In order to infer an approximation of one user's sensitivity to screen brightness, we use the data recorded during first slideshow task (T2 and T8, separately for the two sessions) to fit a regression model. As described in Section 2.3, during these tasks the ambient light is constant, allowing us to isolate the effect of varying screen brightness. We pair each pupil diameter measurement with the image on the screen brightness at the time it was recorded. Using all of these pairs, we use linear regression to determine the slope of the fitted line (see Figure 4 for an illustration). For each newly recorded sample, we determine the corrected pupil diameter $diam_{cor}$ as follows:

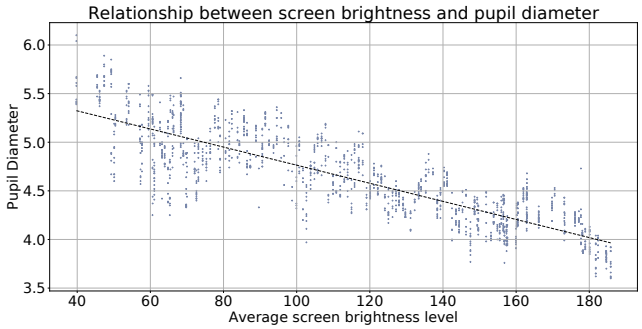


Figure 4: Relationship between (increasing) screen brightness and measured pupil diameter during the first slideshow task (T2) under constant ambient light ($r = -0.8417$).

$$diam_{cor} = diam_{raw} - s_{scr} * br_{scr}, \quad (1)$$

where $diam_{raw}$ is the original measurement, s_{scr} is the slope obtained through linear regression and br_{scr} is the screen brightness at the time the sample was measured. We use the same method to establish the user's sensitivity to ambient light, by using the dim setting of the lamp as input. In this case, for reading (T3, T9) and browsing (T4, T10) tasks we use data from the same task to fit the regression, while for the second slideshow (T5, T11) we use the data from the reading task. This results in a separate slope, s_{amb}.

With these two adjustment factors, we can correct for both image brightness and ambient light changes as follows:

$$diam_{cor} = diam_{raw} - s_{amb} * br_{amb} - s_{scr} * br_{scr}. \quad (2)$$

In particular we use the complete correction for the second slideshow tasks (T5, T11), where both image brightness and ambient light change over the course of the task. We attempted to fit a single light-sensitivity model considering the totality of users, but we found in our data that the slope coefficients s_{amb}, s_{scr} vary greatly across users, suggesting that individual models will perform significantly better.

4.4 Cross-task feature prediction

Medical work shows that eye movement patterns vary according to the task performed by the user. The fixation duration is of particular interest and well-studied for a variety of tasks [34]. While task-specific changes in pupil diameter are partially corrected through the light-based adjustment proposed in the previous section, the task itself can also have an influence.

This method follows the assumption that task-specific changes in feature distributions exhibit a certain consistency between users. For example, fixation times are expected to be longer when reading a text compared to watching a video. Naturally, the *magnitude* of these differences can be user-specific and any prediction without user-specific knowledge will be an approximation.

Eberz et al. presented a system to automatically predict changes in Electrocardiography (ECG) features caused by different measurement devices [9, 10]. We adapt their extended approach to predict

changes in features caused by changing tasks (rather than different devices). The goal is to train the user model on a *source* task and authenticating on a different *target* task without the need for re-enrolment. In order to achieve this, we use population data to train a mapping function that transforms a feature vector measured in the *source* task to account for the task-specific changes expected during the *target* task.

The core of this method is to find an optimal mapping, i.e., a set of transformation functions $F = \{f_j\}_{j \in J}$ with $f_j : \mathbb{R} \rightarrow \mathbb{R}$, such that, for each feature j and subject i, they minimise the statistical distance between the transformed source distribution $f_j(D^S_{i,j})$ and the corresponding target distribution $D^T_{i,j}$. In other words, f_j transforms values of feature j from task S in order to be as close as possible, statistically speaking, to the values of the same feature from task T. As in [10], we restrict the search to linear functions, of the form:

$$f_j(x) = a_j x + b_j \tag{3}$$

Naturally, a mapping function can not be specific to the user, as it would require samples from both tasks to train it and if these were available one could train on the target task directly. Instead, we find the ideal mapping function for the set of remaining users in a leave-one-out fashion. In practice, this means that a mapping function for any task combination can be derived based on population data.

4.5 Authentication pipeline

Following feature extraction (see Section 4.2), we use the following methods in our experimental evaluation.

Training data selection. Given all the data collected for a certain user, the training (enrolment) data for the authentication system can be chosen either randomly or sequentially. Previous work has shown that random selection (e.g., repeated random sampling or stratified cross-validation) leads to greatly overstated performance as the temporal distance between training and testing samples is kept artificially low [1, 13]. In order to avoid this, within each task, we select the first part of the data for training and all following samples for testing. We analyze the effect of varying training data amounts in the following section. For completeness and easier comparison with previous work, we also report the results obtained with random training data selection. For cross-task authentication, we select the complete source task for training and the complete target task for testing.

Pupil diameter correction. We apply the method described in Section 4.3 to both the training and testing data. In the following section, we report results with and without this correction.

Cross-task mapping function. For cross-task authentication (i.e., different training and testing tasks), we apply the corresponding mapping function to the target task to resemble the feature distribution of the source. In practice, when we are looking for a mapping from a source task to a target task for a specific user, we take the remaining users data from the target and source task and solve an optimization problem to find the coefficients in Equation 3, see [10].

Classification. We choose to use a support vector machine (SVM) for the recognition. Since our goal is authentication rather than identification, a one-class model is the natural choice. Some previous works instead train a binary model using data from all users in the dataset (e.g., [11, 16]). This is disadvantageous as data from other users is required for training and, depending on which users are included in the negative class, the classifier performance may not represent accurately the actual error rates of the system. We use a radial basis function kernel and set the SVM hyper-parameters ν to 0.5 and γ to $\frac{1}{|J|}$, J being the feature set. At test time, rather than using the (binary) output of the classifier for the decision, we use each sample's distance to the learnt decision boundary.

Normalization. In order to account for the varying feature ranges of different features, we independently normalize all feature values in input to the classifier:

$$z_i = \frac{x_i - \mu}{\sigma}.$$

This way, each feature is replaced by the number of standard deviations it lies away from the distribution mean. The values for μ and σ are computed on the training data, the transformation is applied to both training and testing data.

Sample aggregation. Similarly to previous work, we aggregate multiple samples into a single window to make an authentication decision. For eye movements, this does not particularly slow down the authentication time as the eye tracker produces several samples over short windows (on average we obtain around five samples per second). We choose to use a fixed-size sliding window, i.e., each window contains exactly n samples. Within each window, we feed each sample to the classifier, collect the distance of each sample from the decision boundary and select the median distance for the decision. Selecting the median rather than the mean allows us to better account for outliers. As a result, we would expect an attacker to go undetected for roughly $\frac{n}{2}$ samples. As n should be chosen based on the system security requirements, we further investigate the choice of n in the following section.

Setting the decision threshold. Using the median boundary distance obtained through sample aggregation, we then select a threshold for acceptance. If the median of the aggregation window is above the threshold, this window of samples is accepted. Varying this threshold controls the tradeoff between the system's FAR and FRR. Note that this threshold is selected on a per-user basis, as users with more erratic behaviour (i.e., a larger mismatch between training and testing data) will require a more lenient threshold to achieve acceptable performance.

5 ANALYSIS AND RESULTS

In this section, we first present an analysis of the features and then show the authentication performances of the system. It should be noted that throughout the results, we always present and treat calibrated and uncalibrated sessions separately. Additionally, we refer to "reading" as tasks T3, T9 (see Table 2), to "browsing" as tasks T4, T10 and to "slideshow" as tasks T5, T11, and consider these six tasks as the ones we use to evaluate the authentication. The first slideshow (tasks T2, T8) is only used to fit the screen brightness model of Equation 1.

Table 3: RMI values for each individual feature. The values are computed by using the data from all the tasks and users, and are shown separately for the calibrated and uncalibrated session.

	RMI [%]	
Feature	Cal	Ucal
Pupil Diameter Difference (min)	21.90	27.03
Pupil Diameter Difference (mean)	21.54	26.27
Pupil Diameter Difference (max)	21.07	26.15
Pupil Diameter (min)	14.19	15.82
Pupil Diameter (mean)	13.88	15.61
Pupil Diameter (max)	13.87	15.84
Left-Right difference (max)	8.81	34.58
Left-Right difference (mean)	8.62	34.92
Left-Right difference (min)	6.85	33.94
Pupil Diameter Difference (std-dev)	6.29	8.97
Pupil Diameter (std-dev)	3.29	6.47
Duration of Fixation	3.28	4.46
Pairwise Speed (mean)	2.09	2.52
Pairwise Acceleration (10 Perc)	1.96	2.22
Pairwise Speed (90 Perc)	1.94	2.43
Pairwise Acceleration (90 Perc)	1.69	2.35
Pairwise Speed (10 Perc)	0.87	1.38
Pairwise Speed (std-dev)	0.87	0.89
Distance to Centre (mean)	0.55	0.85
Distance to Centre (90 Perc)	0.42	0.77
Maximal Pairwise Distance (Y-direction	0.27	0.61
Maximal Pairwise Distance (X-direction)	0.25	0.30
Distance to Centre (std-dev)	0.16	0.24
Maximal Pairwise Distance	0.12	0.33
Distance to Centre (10 Perc)	0.04	0.37

5.1 Feature Analysis

The RMI (see Section 4.5 for details of its computation) of each feature is given in Table 3.

Pupil-based features. Features coming from the pupil diameter measurements contribute the highest amount of information in the calibrated dataset. This is consistent with previous work [11, 12, 18]. Similar to previous work, the static ranges (e.g., min, max and mean) are significantly more distinctive than the changes within a fixation (as measured by the standard deviation).

Temporal-based features.. These features exhibit minimal changes in distinctiveness when using uncalibrated data. This confirms our initial hypothesis that our features depend on precision (i.e., the gaze tracker reports similar coordinates for similar gaze values) rather than accuracy (i.e., the gaze tracker reports the correct gaze position) and that linear shifts in gaze positions will not affect them.

Binocular-based pupil features.. These features have not been explored in previous work. We found that the difference in size between the left and right pupil diameter is even more distinctive than the raw measurements themselves. This can be explained through two factors: inherent size differences between the left and right pupil and different light exposure. In our experimental setup, the desk lamp is placed on one side of the screen (to the right), which leads to each pupil being exposed to different amounts of light. The difference in size between both pupils would therefore be a function of their baseline size, their light sensitivity (which has been shown to be different between individuals) and the user's posture (e.g., when a user is not sitting centred in front of the screen). The results show that the features using the pupil diameter contribute the highest amount of information.

A noteworthy observation is that most binocular features perform significantly better in the uncalibrated setting (RMI of 34.58% vs 8.81%). This suggests that the feature is at least partially dependent on the quality of the calibration. During the experiment we observed that inaccurate calibrations often led to one eye being tracked more accurately than the other and the distance between left and right eye gaze positions being large, but relatively consistent. In the calibrated session, we require a minimum calibration accuracy before starting the tasks (see Section 2.3). This limits the range of calibration errors between users. Despite this apparent relationship, we argue that this feature is not merely a technical "fluke", but still reflects user-specific properties. As outlined in Section 3, a design goal is that the system can be set up with a "generic" calibration in order to avoid having to recalibrate it for each user. Due to a multitude of changes in the user's height, posture, distance between eyes and distance to the screen, this generic calibration will lead to a unique calibration error for each user and result in high distinctiveness for the relevant features. Based on our data, it is evident that these factors remain stable enough during our 20-minute session. We leave a further exploration of the long-term stability of these binocular features for future work.

	calibrated		*uncalibrated*	
Task	raw	corrected	raw	corrected
Reading	4.79	1.88	5.54	2.18
Browsing	7.24	3.92	5.01	2.82
Slideshow	7.57	4.97	6.85	3.93

Table 4: EER [%] for intra-task authentication, considering both calibrated and uncalibrated session, and for both raw pupil measurements and pupil-corrected measurements. Values are computed using 100 aggregated samples and a 50% training data percentage.

5.2 Classification Results

The system overall performance depends on several factors, including the combination of training and testing tasks, aggregation window size, whether calibration was used or not, proportion of training data, pupil diameter correction and cross-task mapping adjustment. For brevity, we report in Table 4 the EER results of a reasonable combinations of these factors, where we perform intra-task authentication, use a window size of 100 aggregated samples and 50% of data for training. 100 samples are collected, on average, after 20 seconds, which means an attacker would be detected after roughly 10 seconds (i.e., once half the sliding window is filled

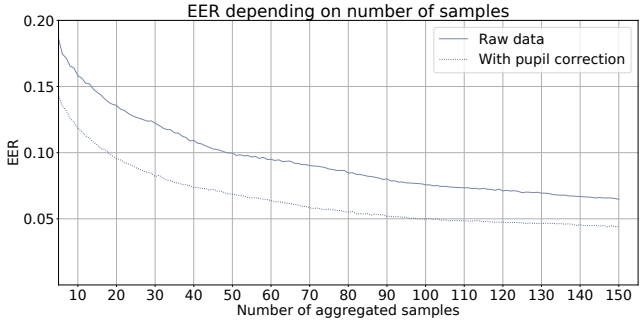


Figure 5: EER depending on number of aggregated samples for the calibrated slideshow (T5) task.

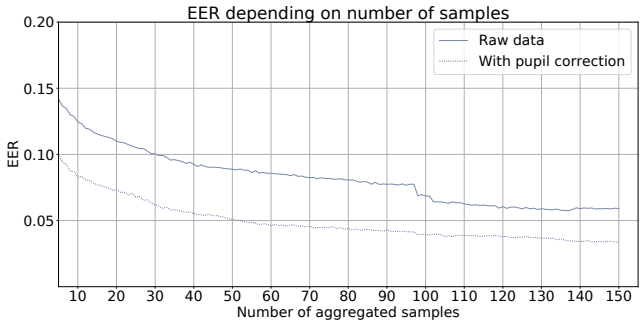


Figure 6: EER depending on number of aggregated samples for the uncalibrated slideshow (T11) task.

with the attacker's samples) due to the median-score aggregation strategy. Table 4 shows the system EER for both calibrated and uncalibrated session and for both light-corrected pupil diameter and raw pupil diameter. The table shows that error rates are lowest for the reading task across all configurations. The slideshow, which includes randomly changing ambient and screen light, shows the highest error rates. All tasks benefit significantly from the pupil diameter correction.

Sample Aggregation. The influence of the number of aggregated samples is shown in Figures 5 and 6 for the calibrated and uncalibrated case, respectively. In both cases, the EER is reduced with increasing window size. The effect becomes less pronounced over time. This is intuitive, as the EER of most users reaches zero after a moderate number of samples (i.e., a further increase in the number of aggregated samples won't improve it further).

Amount of training data. The effect of using varying fractions of the entire dataset for training is shown in Figure 7. Across all three tasks, we can observe that the average EER decreases as more training data is used. While the EER barely changes beyond 10% training data for the browsing and reading tasks, diminishing return only set in after about 40% for the slideshow task. This shows that it is beneficial for the classifier to observe different illumination patterns despite the pupil diameter correction. As discussed in Section 4.5, our system uses sequential training data in order to closely reflect how it would be run in the real world. The fact that

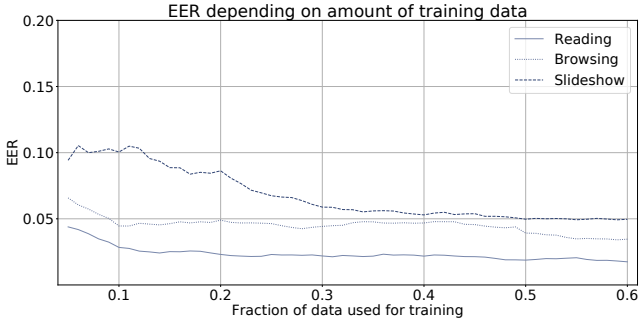


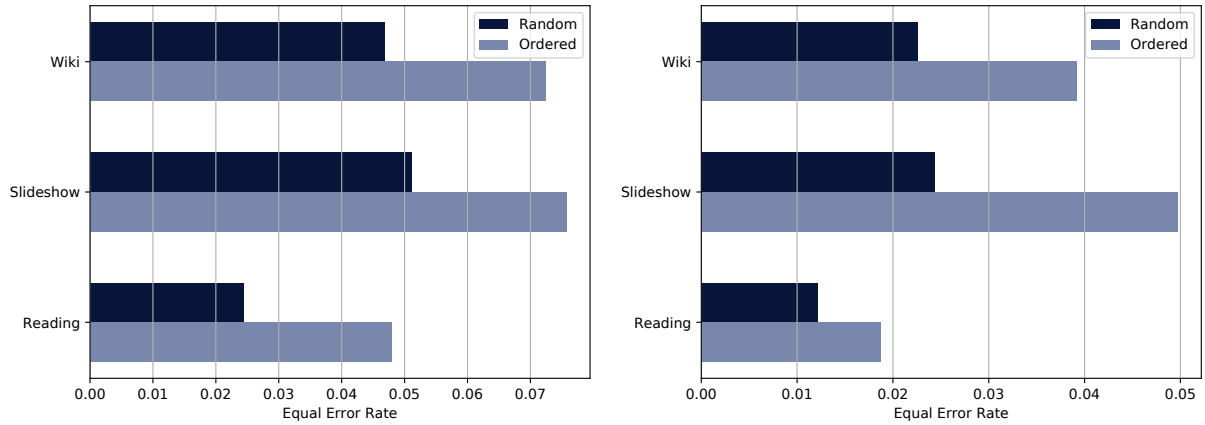
Figure 7: EER depending on the amount of training data for the slideshow task (T5).

low error rates can be achieved even with comparatively small amounts of training data shows that the user's behaviour does not vary significantly across the duration of each task.

Error rate distribution. While a system's average EER gives a rough idea of its expected security against zero-effort impersonations, it is insufficient in the context of continuous authentication without knowing the distribution of errors between users. The highly skewed nature of error rates of biometric systems and the resulting security implications has been previously shown by Doddington et al. in 1998 [7]. Figure 9 shows that the average EER is highly skewed by few users while most users show an EER close to 0%. Previous work has suggested the use of the Gini Coefficient (GC) to capture this property, with a high GC close to 1 indicating skewed error rates [13]. Figure 11 shows a graphical representation of our system's Gini Coefficient for both the FAR and FRR. The FAR in particular is highly skewed with a GC of .94. Despite this skew, the highest FAR achieved across all victim-attacker pairs is 72%. Due to the continuous nature of the authentication system, even this attacker would be detected after a short time span. Unlike previous work [11, 12], we did not observe any systematic false negatives (i.e., perpetually undetected attackers).

Impact of training data selection. As discussed in Section 4.5, we use sequential training data in order to make our analysis as realistic as possible. Nevertheless, in order to allow comparison with other works that use random training data selection we show both selection methodologies in Figure 8. It is evident that randomly sampling the training data improves the overall system performance. The effect is particularly pronounced when not applying the pupil diameter correction. This is a result of the system not observing the entire range of lighting changes when sequential data is used. The effect is particularly strong for the reading and wiki tasks without pupil diameter correction as the lighting changes sequentially rather than randomly (see Table 2).

Cross-task authentication and mapping function. The results of using one task for training and another for testing (i.e., cross-task authentication) can be seen in Figure 10. In the raw data case (no pupil diameter correction and no mapping function, see Figure 10a) the error rates are, not surprisingly, the highest. Intuitively, using the pupil diameter correction only marginally affects the error rates between the reading and browsing task as the brightness differences



(a) Raw data **(b) With pupil diameter correction**

Figure 8: Effect of training data selection, random vs sequential.

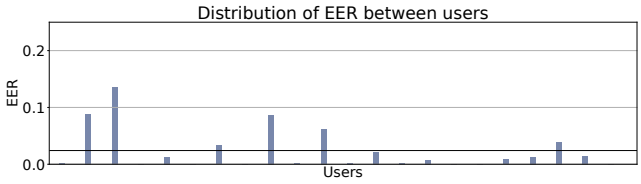


Figure 9: Distribution of EER between users for the calibrated reading task (T3). The horizontal line shows the average EER.

	pre		*post*		*change*	
Task	r	p-value	r	p-value	r	p-value
Reading	0.31	0.16	-0.14	0.55	-0.38	0.08
Browsing	0.31	0.16	-0.17	0.44	-0.41	0.06
Slideshow	0.21	0.34	-0.13	0.55	-0.31	0.16

Table 5: Correlation between calibration error and EER for the calibrated sessions. Values are computed using the calibration accuracy measured at the beginning of the session (T1) and the calibration accuracy measured at the end of the session (T6).

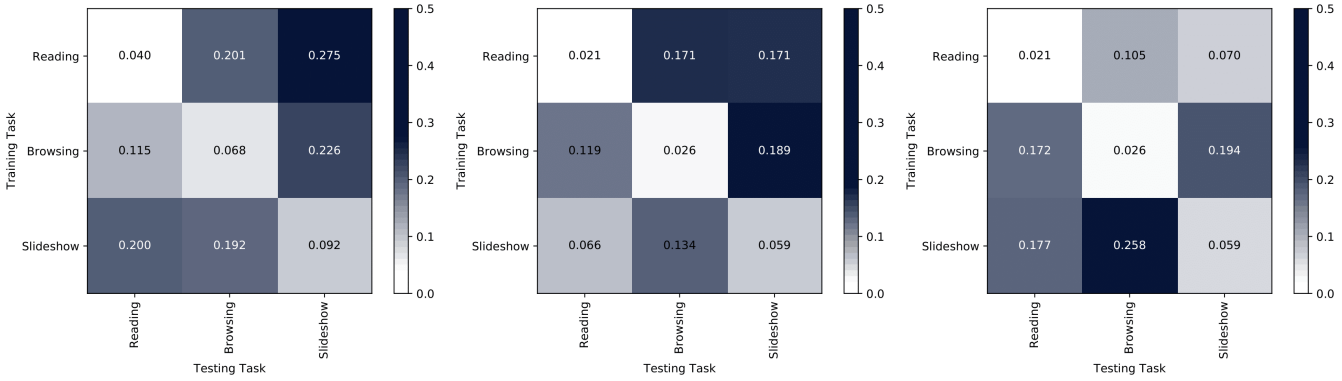
are low. For cross-task authentication between the slideshow task and the others, the system benefits greatly from the pupil diameter correction as the slideshow images are, on average, much darker than text on white background. Applying the mapping function to the reading task greatly reduces error rates for both target tasks (by 39% and 59%, respectively). Interestingly, the EER increases when applying the function to the slideshow task as a source. In practice, it would be sensible to use the mapping function only for relatively predictable source-target combinations (e.g., reading to browsing). This test can be performed on population statistics without input from a particular user. Since the system can be trained on an arbitrary task, choosing a training task that allows easy predictions of other tasks' features is particularly valuable.

Influence of calibration. Table 4 showed that similar or even lower error rates are possible when using a generic (i.e., highly inaccurate) calibration. However, this decrease is partially driven by binocular features which grow in distinctiveness if users show highly diverse calibration errors. In order to measure the effect of calibration error in the calibrated experiment (where users will generally have similar, high-quality calibrations), we compute the correlation between the EER and calibration errors. The results of this computation are shown in Table 5. The table shows the correlation of the EER with the pre-experiment accuracy (i.e., measured directly after calibration), post-experiment accuracy (i.e., after the final task) and their absolute difference. While we observed a moderate positive correlation between pre-experiment calibration error and EER, this was not statistically significant ($p > 0.05$) for any task. This result supports our hypothesis that the system's effectiveness is not significantly affected by the quality of eye tracker calibration.

6 DISCUSSION AND SECURITY ANALYSIS

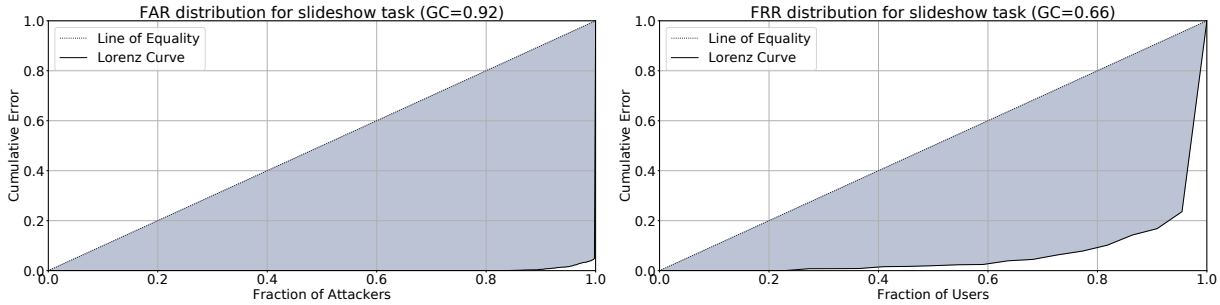
In this section, we will discuss four possible attacks on this system and possible countermeasures.

Manual imitation. Imitation attacks involve the imposter modifying their own eye movement behaviour to appear more similar to the victim. This first requires the attacker to obtain information about the victim's eye movement patterns. This can be achieved through observation if the victim is using an attacker-controlled or otherwise compromised device with a (covert) eye tracker. However, the involuntary nature of eye movements make them hard to consciously control. Microsaccades have been shown to be extremely hard to consciously suppress and controlling them to such a degree to deliberately alter biometric features seems virtually impossible. The pupil diameter is probably the most likely target, as some conscious actions (such as memory cognitive load) cause



(a) Raw data (b) Pupil diameter correction (c) Pupil diameter correction and mapping function

Figure 10: EERs for cross-task authentication.



(a) False Accept Rate (b) False Reject Rate

Figure 11: The Gini Coefficient (the fraction of the area under the line of equality that is shaded) shows the skewness of error rate distributions.

dilations and contractions of the pupil. Nevertheless, this is still difficult to achieve, especially if the attacker is focusing on the attack at the same time. Assuming the legitimate user's calibration configuration is unknown to the attacker, it will be difficult to reproduce the binocular tracking-based features in the uncalibrated setting. While it might be possible to infer some calibration information based on the user's height, seat position and posture, we believe this is unlikely to be sufficient.

Light stimulation. Attacks which use light stimulation have been presented in [18]. The idea is to change the ambient light (in this case, through a dimmable desk lamp) to cause changes in the attacker's pupil diameter. While this has been shown to be effective in [18], the system the authors attacked did not use ambient light correction. In order to defeat this attack, it would be possible to use an ambient light sensor, rather than the light source's dim settings, as input to the pupil diameter correction. Any attacker-induced changes in ambient light would then lead to an increased correction of the pupil diameter and therefore be unable to affect the corrected measurement that is used for authentication. Therefore, a much more targeted light source (such as a laser pointer) would be needed. While this would avoid the ambient light detection, it might still be possible to detect by analysing the illumination differences between the eyes and the rest of the face. This could be performed automatically by the eye tracker's camera.

Artificial eyes. An eye tracker precision and accuracy can be measured without the noisy influence of human eyes [22] using artificial eyes. If two such eyes were attached to a high-precision motor, it would arguably be feasible to reproduce even short-lived movements (such as microsaccades). Dynamically changing the pupil diameter of such an eye could be achieved with a controllable shutter around the pupil. Similar to the other attacks, this still requires the attacker to obtain a (close to) perfect copy of the legitimate user's eye movement behaviour. In addition, liveness detection methods can be used to distinguish an artificial eye from a real one.

7 CONCLUSION

In this paper, we have proposed a continuous authentication system based on eye movement biometrics. This work addressed three practical concerns overlooked by previous work: the need for a precise calibration, the effect of light sensitivity and the task dependence of biometric features. We proposed new eye tracking features based on binocular tracking, showing that their distinctiveness remains

even in presence of generic (i.e., not user-specific) calibrations. We showed a pupil diameter correction mechanism based on linear regression can account for the differences in pupil diameter caused by varying screen brightness and ambient light. Lastly, we addressed task dependence through a cross-task mapping function trained on population data.

Our results show significantly lower error rates than previous work while allowing the system to be used in less controlled environments. We achieve an intra-task EER of 3.93% while requiring only two minutes of uncalibrated training data even with random and frequent changes of lighting conditions. We show that our proposed cross-task mapping can reduce the EER of cross-task authentication by up to 59% when enrolling on a reading task and authenticating on an arbitrary task.

ACKNOWLEDGEMENTS

This work was supported by a grant from Mastercard.

REFERENCES

[1] Kevin Allix, Tegawendé F Bissyandé, Jacques Klein, and Yves Le Traon. 2015. Are your training datasets yet relevant?. In *International Symposium on Engineering Secure Software and Systems*. Springer, 51–67.

[2] Andreas Bulling, Florian Alt, and Albrecht Schmidt. 2012. Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3011–3020.

[3] Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. 2015. GANT: Gaze analysis technique for human identification. *Pattern Recognition* 48, 4 (2015), 1023–1034. https://doi.org/10.1016/j.patcog.2014.02.017

[4] Lewis Carroll. 1930. *Alice in Wonderland.*

[5] Guglielmo Cola, Marco Avvenuti, Fabio Musso, and Alessio Vecchio. 2016. Gait-based authentication using a wrist-worn device. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 208–217.

[6] Véronique Daneault, Gilles Vandewalle, Marc Hébert, Petteri Teikari, Ludovic S Mure, Julien Doyon, Claude Gronfier, Howard M Cooper, Marie Dumont, and Julie Carrier. 2012. Does pupil constriction under blue and green monochromatic light exposure change with age? *Journal of biological rhythms* 27, 3 (2012), 257–264.

[7] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds. 1998. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *National Institut of Standards and Technology Gaithersburg* (1998), 1–4.

[8] Andrew T. Duchowski. 2017. *Eye tracking methodology: Theory and practice: Third edition.* Springer International Publishing, Cham. 1–366 pages. https://doi.org/10.1007/978-3-319-57883-5 arXiv:arXiv:1011.1669v3

[9] Simon Eberz, Giulio Lovisotto, Andrea Patane, Marta Kwiatkowska, Vincent Lenders, and Ivan Martinovic. 2018. When your fitness tracker betrays you: Quantifying the predictability of biometric features across contexts. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 889–905.

[10] Simon Eberz, Nicola Paoletti, Marc Roeschlin, Andrea Patani, Marta Kwiatkowska, and Ivan Martinovic. 2017. Broken Hearted: How To Attack ECG Biometrics. In *Proceedings 2017 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2017.23408

[11] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2015. Preventing Lunchtime Attacks: Fighting Insider Threats With Eye Movement Biometrics. In *Proceedings 2015 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2015.23203

[12] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2016. Looks Like Eve: Exposing Insider Threats Using Eye Movement Biometrics. *ACM Transactions on Privacy and Security* 19, 1 (2016). https://doi.org/10.1145/2904018

[13] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2017. Evaluating Behavioral Biometrics for Continuous Authentication. In *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security - ASIA CCS '17*. ACM Press, New York, New York, USA, 386–399. https://doi.org/10.1145/3052973.3053032

[14] S. Zahra Fatemian, Foteini Agrafioti, and Dimitrios Hatzinakos. 2010. HeartID: Cardiac biometric recognition. In *IEEE 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2010*. https://doi.org/10.1109/BTAS.2010.5634493

[15] Tao Feng, Xi Zhao, and Weidong Shi. 2013. Investigating mobile device picking-up motion as a novel biometric modality. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–6.

[16] Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. 2013. Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *IEEE Transactions on Information Forensics and Security* 8, 1 (Jan. 2013), 136–148. https://doi.org/10.1109/TIFS.2012.2225048

[17] Chiara Galdi, Michele Nappi, Daniel Riccio, Virginio Cantoni, and Marco Porta. 2013. A new gaze analysis based soft-biometric. In *Mexican Conference on Pattern Recognition*. Springer, 136–144.

[18] Isaac Griswold-Steiner, Zakery Fyke, Mushfique Ahmed, and Abdul Serwadda. 2018. Morph-a-Dope: Using Pupil Manipulation to Spoof Eye Movement Biometrics.

[19] Daniele Gunetti and Claudia Picardi. 2005. Keystroke analysis of free text. *ACM Transactions on Information and System Security* 8, 3 (2005), 312–347. https://doi.org/10.1145/1085126.1085129

[20] Corey Holland and Oleg V Komogortsev. 2011. Biometric identification via eye movement scanpaths in reading. In *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 1–8.

[21] Corey D Holland and Oleg V Komogortsev. 2013. Complex eye movement pattern biometrics: Analyzing fixations and saccades. In *Biometrics (ICB), 2013 International Conference on*. IEEE, 1–8.

[22] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. 2012. Eye tracker data quality: what it is and how to measure it. In *Proceedings of the symposium on eye tracking research and applications*. ACM, 45–52.

[23] Donald R. Jasinski, Jeffrey S. Pevnick, and John D. Griffith. 1978. Human Pharmacology and Abuse Potential of the Analgesic Buprenorphine: A Potential Agent for Treating Narcotic Addiction. *Archives of General Psychiatry* 35, 4 (1978), 501–516. https://doi.org/10.1001/archpsyc.1978.01770280111012

[24] Andrew H Johnston and Gary M Weiss. 2015. Smartwatch-based biometric gait recognition. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 1–6.

[25] Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science* 154, 3756 (1966), 1583—-1585.

[26] Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. 2007. Reducing shoulder-surfing by using gaze-based password entry. In *Proceedings of the 3rd symposium on Usable privacy and security - SOUPS '07*. 13. https://doi.org/10.1145/1280680.1280683

[27] Dachuan Liu, Bo Dong, Xing Gao, and Haining Wang. 2015. Exploiting eye tracking for smartphone authentication. In *International Conference on Applied Cryptography and Network Security*. Springer, 457–477.

[28] Colleen MacLachlan and Howard C. Howland. 2002. Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years. *Ophthalmic and Physiological Optics* 22, 3 (May 2002), 175–182. https://doi.org/10.1046/j.1475-1313.2002.00023.x

[29] Susana Martinez-Conde, Stephen L. Macknik, Xoana G. Troncoso, and Thomas A. Dyar. 2006. Microsaccades counteract visual fading during fixation. *Neuron* 49, 2 (2006), 297–305. https://doi.org/10.1016/j.neuron.2005.11.033

[30] Mihai Pop, Yves Payette, and Emma Santoriello. 2002. Comparison of the pupil card and pupillometer in measuring pupil size. *Journal of Cataract & Refractive Surgery* 28, 2 (2002), 283–288.

[31] Ioannis Rigas, George Economou, and Spiros Fotopoulos. 2012. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recognition Letters* 33, 6 (2012), 786–792.

[32] Ioannis Rigas and Oleg V Komogortsev. 2014. Biometric recognition via probabilistic spatial projection of eye movement trajectories in dynamic visual environments. *IEEE Transactions on Information Forensics and Security* 9, 10 (2014), 1743–1754.

[33] Brian C. Ross. 2014. Mutual information between discrete and continuous data sets. *PLoS ONE* 9, 2 (Feb. 2014), e87357. https://doi.org/10.1371/journal.pone.0087357

[34] Hildur EH Schilling, Keith Rayner, and James I Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition* 26, 6 (1998), 1270–1281.

[35] Ivo Sluganovic, Marc Roeschlin, Kasper B. Rasmussen, and Ivan Martinovic. 2016. Using Reflexive Eye Movements for Fast Challenge-Response Authentication. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*. ACM Press, New York, New York, USA, 1056–1067. https://doi.org/10.1145/2976749.2978311

[36] Sasitorn Taptagaporn and Susumu Saito. 1990. How display polarity and lighting conditions affect the pupil size of VDT operators. *Ergonomics* 33, 2 (Feb. 1990), 201–208. https://doi.org/10.1080/00140139008927110

[37] B Winn, D Whitaker, D B Elliott, and N J Phillips. 1994. Factors affecting light-adapted pupil size in normal human subjects. *Investigative ophthalmology & visual science* 35, 3 (1994), 1132–1137.

[38] Nan Zheng, Aaron Paloski, and Haining Wang. 2011. An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security - CCS '11*. 139. https://doi.org/10.1145/2046707.2046725