Looks Like Eve: Exposing Insider Threats Using Eye Movement Biometrics

SIMON EBERZ and KASPER B. RASMUSSEN, University of Oxford
VINCENT LENDERS, Armasuisse
IVAN MARTINOVIC, University of Oxford

We introduce a novel biometric based on distinctive eye movement patterns. The biometric consists of 20 features that allow us to reliably distinguish users based on differences in these patterns. We leverage this distinguishing power along with the ability to gauge the users' task familiarity, that is, level of knowledge, to address insider threats. In a controlled experiment, we test how both time and task familiarity influence eye movements and feature stability, and how different subsets of features affect the classifier performance. These feature subsets can be used to tailor the eye movement biometric to different authentication methods and threat models. Our results show that eye movement biometrics support reliable and stable continuous authentication of users. We investigate different approaches in which an attacker could attempt to use inside knowledge to mimic the legitimate user. Our results show that while this advance knowledge is measurable, it does not increase the likelihood of successful impersonation. In order to determine the time stability of our features, we repeat the experiment twice within 2 weeks. The results indicate that we can reliably authenticate users over the entire period. We show that lower sampling rates provided by low-cost hardware pose a challenge, but that reliable authentication is possible even at the rate of 50Hz commonly available with consumer-level devices. In a second set of experiments, we evaluate how our authentication system performs across a variety of real-world tasks, including reading, writing, and web browsing. We discuss the advantages and limitations of our approach in detail and give practical insights on the use of this biometric in a real-world environment.

CCS Concepts: • **Security and privacy → Security services; Authentication; Biometrics**

Additional Key Words and Phrases: Biometrics, continuous authentication, metrics

ACM Reference Format:
Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2016. Looks like eve: Exposing insider threats using eye movement biometrics. ACM Trans. Priv. Secur. 19, 1, Article 1 (June 2016), 31 pages.
DOI: http://dx.doi.org/10.1145/2904018

1. INTRODUCTION

In this article, we evaluate the effectiveness of using eye movement biometrics as a novel defense against the "lunchtime attack" by an insider threat.[1] An insider threat in this context refers to a person with physical access to a workstation that he or

[1]This article is an extension of a previous conference paper [Eberz et al. 2015].
This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M50659X/1].
Authors' addresses: S. Eberz, K. B. Rasmussen, and I. Martinovic, University of Oxford, Department of Computer Science, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom; emails: {simon.eberz, kasper.rasmussen, ivan.martinovic}@cs.ox.ac.uk; V. Lenders, armasuisse Science and Technology, Feuerwerkerstrasse 39, 3600 Thun, Switzerland; email: vincent.lenders@armasuisse.ch.

2471-2566/2016/06-ART1 $15.00
DOI: http://dx.doi.org/10.1145/2904018

she is not supposed to use (e.g., using a coworker's workstation while he or she is at lunch). As such, our system serves as a second line of defense after the workstation has already been compromised (i.e., the attacker has physical access and the workstation is either unlocked or he or she is in possession of all necessary passwords and access tokens). Our approach considers both users that are simply careless and users that are actively collaborating with the attacker by giving up information. The second case makes this attack notoriously difficult to defend against. We propose a set of features that can be extracted from human eye movements and analyze their distinctiveness and robustness using a systematic experimental design.

The human eyes offer a rich feature space based on voluntary, involuntary, and reflexive eye movements. Traditionally, the analysis of eye movements has been used in the medical domain to facilitate diagnosis of different ocular and neuronal disorders. Eye-tracking devices have become much cheaper within the last few years, and even low-cost open-source hardware and software is available [San Agustin et al. 2010]. Recent advances in video-based eye-tracking technology make eye tracking applicable to a conventional workplace as it does not require any physical contact with the users (more detail on eye tracking is given in Section 3).

Our experimental design captures the unique characteristics of each user's eye movements as measured by the eye tracker. We also consider ways in which the attacker could use his or her position to gain inside information about the user and the system through observation or social engineering. We define metrics to measure this advance knowledge through eye movement data and determine whether it affects the authentication decision. We consider three scenarios in particular: (1) no prior knowledge, that is, no information advantage; (2) knowledge gained through a description—for example, the adversary is provided with a textual description by a colluding legitimate user; and (3) knowledge gain through observation, for example, by looking over the shoulder of a legitimate user performing a task (shoulder surfing).

We perform these experiments with 30 subjects recruited from the general public and repeat them after 2 weeks to test the time stability of the proposed features. While our experimental results show that an adversary does benefit from an increased level of knowledge when executing a task, the analysis of the proposed features also shows that he or she cannot utilize that knowledge to circumvent the eye movement biometric. In order to evaluate the performance of a continuous authentication system in a real-world environment, we conduct a second set of experiments based on a number of real-world tasks, including reading, writing, and browsing.

Our main contributions are a set of 20 features and measurements that confirm that these features are suitable to perform transparent continuous user authentication. We use different metrics to measure the quality of these features and quantify the effects of increasing time distance on both feature groups and individual features. In order to evaluate whether the eye movement biometric can be used in conjunction with cheap consumer-level hardware, we also determine the impact that the reduced sampling rate of these devices has on both feature quality and the performance of a continuous authentication system. Lastly, we propose two new metrics that give insights into the expected real-world performance of a system based on our biometric.

The rest of the article is organized as follows: Section 2 gives an overview over related work. Section 3 gives background information about the human visual system and eye-tracking techniques. Section 4 outlines our threat model. Our experimental design is described in Section 5. Our features and their quality measures are given in Section 6, while our classifiers and metrics are outlined in Section 7. Section 8 lists our results and discusses the effects of task selection. We discuss possible limitations of our approach in Section 9 and conclude the article in Section 10.

2. RELATED WORK

The body of related work can be divided into papers proposing biometrics and biometric systems, and attacks on either of these.

2.1. Biometric Systems

The idea of using physiological or behavioral biometrics in the context of system security is not new and has been an active research area for many years. Jain et al. [2006] provide a comprehensive overview of hard biometrics (e.g., fingerprints, iris patterns, DNA) in a security context. The use of hard biometrics allows the distinction between users with high accuracy and usually over the entire lifetime of a person. A person's biometric features cannot usually be changed, which makes it harder to mimic another person's features without having to circumvent liveness detection mechanisms. On the other hand, the feature becomes useless once another person is able to copy it. Attacks on fingerprint sensors, including the iPhone's TouchID feature, using mock fingers created of various materials have recently been shown to be feasible under practical conditions [Barral and Tria 2009; Rieger 2013]. This is particularly dangerous as copies of fingerprints can be easily collected in an office environment, for example, by lifting them off a coffee mug. Another downside of hard biometrics lies in poor collectability and high intrusiveness.

Facial recognition may seem like a convenient method to provide continuous authentication but is not feasible in a high-security context due to imperfect liveness detection. Attacks on facial recognition software are possible using simple photographs [Duc and Minh 2009] or more complex 3D video models [Boehm et al. 2013].

Behavioral biometrics are typically less susceptible to these kinds of replication attacks, but their performance with regard to false accept rates (FARs) and false reject rates (FRRs) often makes them unsuitable for standalone authentication. This is a result of the low time stability of human behavior as well as noise effects created by external distractions. One of the oldest behavioral biometrics was proposed in 1980 and exploits distinctive keystroke patterns [Gaines et al. 1980]. Since then, extensive research based on this biometric has been conducted using different classifiers with static and dynamic texts in multiple environments. The error rates are low for static texts but increase rapidly for free-form texts as many unpredictable pauses are introduced in the typing process. Additionally, templates are usually tied to keyboard layouts and even physical devices. Two recent comprehensive surveys of keystroke dynamics can be found in Shanmugapriya and Padmavathi [2009] and Banerjee and Woodard [2012].

Mouse movements have been extensively studied as a potential behavioral biometric that can be combined particularly well with keystroke patterns, as both traits are usually collected at different times. A survey on the extensive body of work can be found in Revett et al. [2008]. The best accuracy has been reported with an FAR of 0.36% and an FRR of 0% [Nakkabi et al. 2010]. As the data was collected on the test subjects' own PCs, it is questionable whether the classifier did not distinguish input devices instead of subjects [Jorgensen and Yu 2011].

Given the increasing share of smartphones and tablets, the keyboard and mouse are no longer used ubiquitously. A recent study reported an equal error rate of 2% to 3% when identifying subjects across sessions based on their stroke patterns on a smartphone touchscreen [Frank et al. 2012]. A similar approach that also tests the resistance to imitation attacks is described in Zheng et al. [2012]. However, the authors only account for observation, not for a compromised user template.

There has been some work on the way the human body modifies electrical currents. Rasmussen et al. [2014] measure the body's response to an electric square pulse signal and report 100% accuracy over a static dataset and 88% over a dataset that contains

samples taken over several weeks. Similar work that uses bioimpedance as a biometric reports a recognition rate of 90% but requires augmentation with hand geometry [Cornelius et al. 2012]. Furthermore, the scope of the study was limited to a family-size study with up to five subjects.

Eye movements have previously been studied as an input channel that is resistant to shoulder-surfing attacks. These systems still rely on a conventional PIN, a password, or a passphrase. Kumar et al. [2007] developed a system using a Tobii 1750 gazetracker and report a password entry time of 9 to 12 seconds with error rates between 3% and 15%. Similar work used eye gestures instead of passwords and reduced the fraction of successful shoulder-surfing attacks to 55% with an average input time of 5.3 seconds [De Luca et al. 2009].

One of the earliest works using eye movements as a biometric was published in 2005 [Bednarik et al. 2005]. The authors use gaze velocity and the distance between pupils as features and achieve identification rates of up to 92%. However, the error rates rapidly increase without relying on the pupil distance, a feature more commonly associated with facial recognition than with eye tracking. Our work is perhaps most closely related to Kinnunen et al. [2010]. The authors use a Tobii X120 gazetracker with a sampling rate of 120Hz to capture a subject's eye movements while he or she is watching a movie and use short-term eye gaze direction to construct feature vectors that are modeled using Gaussian mixtures. Depending on the amount of training data, an equal error rate (EER) of 28.7% to 47.1% is reported. The authors do not state whether the type of video affects the templates (e.g., whether training and testing with different videos is possible). A different approach by Cantoni et al. [2015] attempts to distinguish individuals by the way they look at different images. Unlike our work, this approach requires display of controlled stimuli; as such, it is more suited for one-time authentication (e.g., to replace a password to log in) rather than continuous authentication. Using density and duration of fixations as their main features, they report an EER of 27.06%. Similarly, Liang et al. [2012] measure the eye's tracking behavior when a moving stimulus is displayed. They use the acceleration of eye movements while the subjects are pursuing a moving shape as input to both Support Vector Machines and a Back-Propagation neural network. In an experiment with five subjects, they achieve an identification accuracy of 82%.

2.2. Attacks

There are two broad groups of attacks on biometric systems: zero-effort attacks and imitation attacks. The zero-effort threat model is typically used when measuring a biometric's error rates; it reflects the possibility of the biometric template of two users being close enough that the system might mistake one for the other. It does not consider any active or conscious effort on the attacker's part (hence the name). Conversely, imitation attacks involve the attacker somehow modifying how he or she is perceived by the system in order to appear as the legitimate user. Imitation attacks on biometric systems can be further subdivided into three categories: replay attacks, manual imitation attacks, and automatic imitation attacks.

Replay attacks involve the playback of previously recorded data, for example, through the hijacked USB connection of a keyboard (or an eyetracker). As such, they are attacks on systems rather than the biometric.

Manual imitation attacks involve a human modifying his or her own behavior and consequently changing his or her biometric template to appear similar to the victim. This type of attack has been shown to be highly successful against keystroke dynamics [Tey et al. 2013], while touchscreen inputs appear to be harder to imitate [Zheng et al. 2012]. While the attack itself is performed by a human, the attacker can be assisted by a system during the preparation phase, for example, by providing feedback regarding the closeness to the victim's template.

Automatic imitation attacks seek to circumvent liveness detection. As such, the input to the system measuring the biometric is not generated by a human, but by a machine. Serwadda and Phoha [2013] built a robot based on the LEGO mindstorm platform to automatically generate swipes on a touchscreen that resemble a large number of legitimate users. As the success of behavioral biometrics is partially driven by the availability of sensors in consumer devices, this sort of attack is problematic as good liveness detection is often difficult to achieve without relying on additional sensors.

Despite the growing popularity of behavioral biometrics (and attacks on them), there has been, to the extent of our knowledge, no published work describing either automatic or manual imitation attacks on eye movement biometrics.

3. VISUAL SYSTEM BACKGROUND

This section provides a brief introduction to the specifics of the human visual system (HVS) required to understand the rationale behind this work, and discusses the eye-tracking and gaze-tracking technologies to justify its applicability to the security domain. For a systematic overview of the HVS and eye-tracking related research, see, for example, Duchowski [2007].

The HVS has been part of neurophysiological research for many decades. The current understanding of the human brain includes considerable knowledge about the connections between the retina and the brain regions that are responsible for generating eye movements. The experimental design defined in this work is inspired by neuroscientific insights related to fixational eye movements. These movements are a particular type of eye behavior involved in processing static scenes that are typical for working with a desktop machine and processing static stimuli shown on a display, such as navigating through the file system or reading documents. Another important requirement of this work is the technology that allows one to capture the eye movements of a person working on everyday tasks. The technology should not pose significant usability disadvantages and should be applicable in a conventional working environment.

3.1. Characteristics of Eye Movements

The human eyes move within 6 degrees of freedom with six muscles responsible for the movement of the eyeball. The main types of eye movements used in perceiving a stationary object or scene or reading a document can be categorized into *saccades* and *fixations*. The neural signals controlling these eye 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 to 100ms, depending on the distance covered [Duchowski 2007]. They are used to move the fovea[2] to another location. Once the saccadic movement has been signalized by the related neurons, the movement must be completed; that is, neither the saccade's position nor its velocity can be consciously altered, even if the target has changed its position [Cassin et al. 1984].

In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100 to 400ms. They are used to stabilize the retina over a stationary object of interest. Yet, the eyes are never perfectly still; they make involuntary movements even during visual fixations. The main reason for such movements is to counteract retinal fatigue and to prevent visual fading; that is, if a person attempts to artificially fixate the eyes on an image by strongly focusing on a single fixation point, the image would start to fade away and the scene would become blank. One type of such movements are microsaccades, characterized by high velocity and acceleration often away from the fixation center [Martinez-Conde et al. 2006]. Related to microsaccades are movements called Saccadic Intrusions (SIs), which consist of involuntary

[2]The fovea is a part of the retina that allows the central, high-resolution vision.

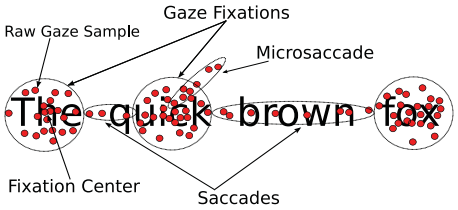


Fig. 1. A simplified example of gaze tracking: the raw gaze samples are collected by the eye-tracking device and subsequently clustered into fixations, saccades, and microsaccades.

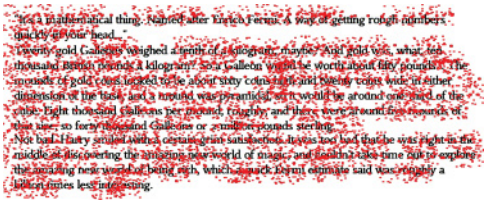


Fig. 2. Real-world gaze data from a subject reading a document.

movements away from the previous eye position, followed by a return to that position after a short duration [Abadi and Gowen 2004]. SIs are characterized through a high velocity and significantly higher amplitude compared to microsaccades. This terminology is visualized in Figure 1 and an example of real-world gaze-tracking data from a person reading a block of text is shown in Figure 2.

Conventionally, the studies of fixational eye movements have been concerned with medical diagnoses, such as Alzheimer's disease [Jones et al. 1983] and schizophrenia [Clementz et al. 1990]. Yet, with an advance in eye-tracking technologies, analyzing eye movements has proven to be valuable in many other areas, such as marketing (e.g., for analyzing visual attention as a measure of effective advertising) [Rayner et al. 2001; Wedel and Pieters 2000], human-computer interface design [Jacob 1995], pilot training [Ottati et al. 1999], or detecting fatigue and drowsiness in drivers [Tock and Craw 1996; Ito et al. 2002; Devi and Bajaj 2008].

Besides the eye movements, the pupil diameter is also an interesting feature that can be included in the analysis of eye behavior. The *range* for this feature in a single subject is largely determined by eye physiology, gender, and ethnicity and usually remains constant during adulthood [MacLachlan and Howland 2002]. Nevertheless, multiple causes that affect the pupil diameter have been found, including memory and cognitive workload [Kahneman and Beatty 1966], lighting conditions [Taptagaporn and Saito 1990], and drug consumption [Jasinski et al. 1978].

3.2. Eye- and Gaze-Tracking Techniques

Eye tracking is the process of capturing a person's eye movements and measuring their positions. If the eye positions are calibrated with respect to an external display, then the process is called gaze tracking. There are many types of eye-tracking techniques, with the main tradeoff between temporal/spacial accuracy versus intrusiveness and usability. Traditional eye-tracking techniques require either a head-mounted device or electrodes attached to the subject's face. One such example is electrooculography (EOG), which is a technique for recording eye movements by measuring the electric potential at the electrodes placed around the eyes. While this technique can be used to capture the eye movements even during sleep (e.g., to monitor REM sleep), its main disadvantage is its high intrusiveness since the electrodes must be attached to a person's face.

Recently there has been significant progress in eye-tracking technology driven by its importance in many commercial scenarios, such as advertising and usability studies. The gaming and entertainment industries also show a trend toward consumer-level eye-tracking devices not only as an additional control channel but also to enhance computer-human interaction. The most widely used eye-tracking technology today is video based. Video-based eye tracking uses a video camera that focuses on the pupils and records their movements and size. To improve the tracking accuracy, these devices

usually use a source of controlled infrared or near-infrared light to create distinctive reflection patterns (see Figure 3). Importantly, the current video-based eye tracking is noninvasive and remote, operating without any contact with the subject. The required hardware is only a standard webcam capable of recording infrared light. For example, the ITU Gaze Tracker [ITU Gaze Group 2015] is an open-source project that offers eye-tracking software that can be used by many low-cost webcams. Some smartphone manufacturers such as Samsung have also recently started to include basic eye-tracking capabilities in their phones.

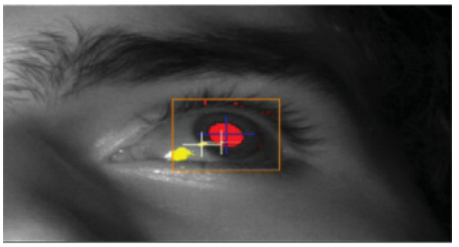


Fig. 3. Video-based gaze tracking: the tracking of eye movements is software based and does not require any physical contact with a subject. The gaze position is calculated using the distance between pupil position and the corneal reflections (shown as two white crosses).

Given the increasing availability and simplicity of eye tracking, it is likely that the trend of using eye tracking outside of the medical and research domain will continue. The current noninvasive eye-tracking technology already enables easy access to a rich and distinctive feature space of fixational eye movements. Their distinctive capabilities and involuntary nature make them a potentially valuable biometric.

4. THREAT MODEL

The adversary model considered in this article focuses on insider threats. A well-known example of an insider threat is the so-called "lunchtime attack" where an adversary temporarily gains access to a coworker's workstation while the coworker is away for lunch. Other examples include cleaning staff getting access to workstations after hours and the trivial case where one employee simply allows another employee to use his or her workstation or access credentials. In all these scenarios, an adversary might gain access to a fully operational system, already logged in to a privileged account, and with access to everything that the legitimate user of the workstation would normally have access to. Any subsequent attack mounted from such a compromised workstation can be very hard to trace back to the real attacker. A 2011 study showed that 33% of electronic crimes in companies are committed by insiders [CERT 2011]. Sixty percent of these attacks use a compromised account; in the remaining cases the attacker uses his or her own account [Keeney et al. 2005]. Account compromise is particularly difficult to detect as the account used to carry out the attack typically was not associated with suspicious activity before. Furthermore, it is more difficult to trace back the attack (and investigation may even put false blame on the victim). Most organizations allow their employees remote access (e.g., via SSH or a VPN connection); nevertheless, 43% of attacks are performed locally using *physical access* to the workstation [Keeney et al. 2005].

In our model, the adversary is aware of the gaze-tracking system and will do his or her best to imitate the behavior of the legitimate user. This can be done by familiarizing oneself with the system before sitting down at the terminal, thus trying to appear to the gaze-tracking system as an experienced user. From the attacker's perspective, there are two incentives to obtain this kind of information: If he or she manages to observe how the user accesses sensitive data or performs some sort of transaction, he or she will most likely be able to carry out the attack much faster, helping to avoid detection. Besides this, performing a task in a similar way may result in ocular characteristics being closer to the legitimate user. The adversary will win if he or she can circumvent the gaze-tracking system, that is, exhibit ocular characteristics that are similar enough to the legitimate user.

We consider two models of knowledge transfer to help the adversary familiarize him- or herself with a system: (1) the adversary has gained knowledge about the system by reading (or being told) how the system works and (2) the adversary has seen (e.g., by shoulder surfing) how a legitimate user operates the system.

We assume the adversary cannot disable the gaze-tracking system, nor can he or she interfere with its operation in any way, as doing so would quickly lead to the detection of his or her attack. We don't consider insider threats that involve the attacker using his or her own workstation. These attacks can always be traced back to the actual attacker and are better dealt with through behavioral monitoring [Kandias et al. 2010]. The aim here is to show that gaze tracking is a viable way of identifying users, as well as gauging a user's level of knowledge and familiarity with a particular task.

5. EXPERIMENTAL DESIGN AND DATA COLLECTION

In this section, we give an overview of our design goals and show how our experimental design meets those goals. We describe our test subject population and discuss how features change over time, as well as the best way to capture these changes. Finally, we describe a downsampling procedure designed to model data captured with low-cost eye-tracking devices.

5.1. Design Goals

The experiments described in this section are designed to test the feasibility of building an authentication system based on the distinctiveness of human eye movement patterns. Such a system should continuously monitor the user's eye movement behavior in the background without requiring any modifications in the user's behavior or even his or her knowledge or consent. In addition, the experiments should allow us to test whether eye movements reveal information about a user's task familiarity, that is, whether eye movement behavior changes significantly between familiar and unfamiliar users. This distinction could be used to detect outside attackers as they can be assumed to be significantly less familiar with the system they attempt to access than legitimate users (or inside attackers).

In order to design experiments that show whether or not gaze tracking is suitable as an authentication mechanism, we have to determine which tasks the test subjects should perform while they are being monitored. One option is to give them an entirely free environment in which the subjects can choose what to do for the duration of the experiment. This is probably the experiment that best captures actual user behavior, but since it is likely that each subject will choose a different set of tasks, it is very hard to guarantee that the distinguishing power of the resulting classifier is really capturing differences in users, rather than differences in behavior or tasks. While we choose features such that their computation does not depend on specifics of the task, it is difficult to rule out that some differences in their *distributions* are due to the user-chosen task. As even the variations within a single task (such as different types of websites for a web browsing task) might already cause features to change, the number of subtasks to test would be enormous. If each user chose a different task, which possibly results in specific feature characteristics, this would lead to an overestimation of classification accuracy, as the classifier performs task distinction instead of user distinction. Conversely, a fixed task for all users means that any differences between the datasets are due to differences between users.

Another approach is to fix a set of general tasks and let all the users perform those the way they prefer. This will limit the influence of user-chosen tasks, but the visual stimuli presented to the subjects will still be different. For example, if the subjects are

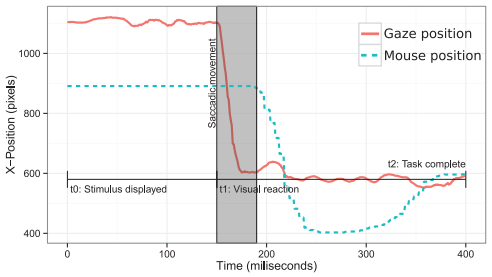


Fig. 4. At time t_0 a new dot appears followed by a period of inactivity (reaction time) in which neither the gaze nor the cursor moves significantly. After about 150ms, at t_1, a visual reaction in the form of a large saccade occurs (the gray area) and the gaze and cursor converge to the position of the stimulus.

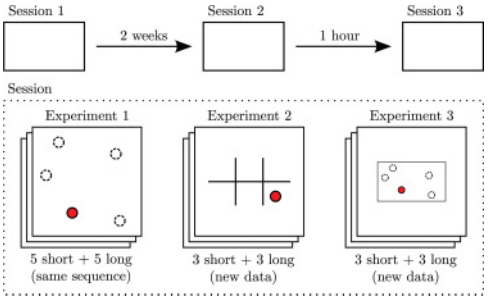


Fig. 5. Experiment structure. Each session is divided into three experiments, each of which is repeated a number of times. The entire session is repeated after 2 weeks, and again an hour after the second repetition.

asked to browse the web but not restricted in what pages to visit or specifically what to read, different subjects would have very different experiences. Even if the task is as simple as watching a movie, different subjects will focus on different things and the resulting classification might be biased by genre preference and other factors.

In order to overcome these sources of error, we define a specific set of tasks that all users must complete. Our goal is to determine whether the users' eye movements are distinguishable, even if they are completing the same task the same way with the same knowledge. If this is indeed the case, this means that there are *inherent* differences between users that cannot be attributed to different ways of completing a single task. Nevertheless, we choose our features such that their *computation* does not make any assumptions about the task.

5.2. Experiment Structure

We first introduce terminology to make it easier to refer to different parts of our interaction with test subjects; see Figure 5 for a visualization. We refer to one sitting of a test subject as a *session*. Two weeks after the first session, the test subject comes back for a second session. This is done to make sure our results are consistent over time. To verify that our results are consistent not only over longer periods but also across two subsequent sessions on the same day, our test subjects do a third session about an hour after completing session 2. All three sessions are identical, and each consists of three different *experiments*.

Each experiment has a similar structure. The test subject is initially presented an empty screen with a gray background. Once the experiment begins, a red dot with a white center appears at a random location on the background. The user is then asked to click on the dot as fast as possible. Once the dot is clicked, the next one appears after a short delay, during which the screen is reset to the gray background. All instructions are displayed on-screen before the experiment begins, and the experiments differ in the nature of the instructions given to the subject. Additionally, each experiment comes in a short and a long version. This allows us to capture potential effects of training and memory for both simple and more complex tasks.

Experiment 1 (no prior knowledge) provides no instructions to the test subjects beyond asking them to click the dots as fast as possible. The short version has five dots and the long version has seven dots. The idea behind Experiment 1 is to model a scenario in which an adversary sits down at a workstation without prior knowledge of

the task he or she is facing. We assume that the subject's performance is affected by increasing task familiarity and that there are memory-based learning effects when he or she completes the *same* sequence of dots multiple times. These effects reflect those observed in real environments when users become accustomed to their typical working environment. During the experiment, the test subject learns the position of the dots over time but in addition gains a general familiarity with the nature of the experiments. This experiment can also be transferred to an attacker that performs tasks he or she is accustomed to on a victim's workstation to cover his or her own tracks.

At each repetition the test subject is informed that the sequence will remain the same for the next iterations. We would expect the learning effects in short sequences to be bigger compared to long sequences. In order to test this, each user performs five repetitions of the short sequence and five repetitions of the long sequence. The random seed used to generate the position list was kept identical for all subjects in order to eliminate distortion effects caused by the dot positions. The five-dot and seven-dot sequences are chosen independently; as such, the long sequence is not merely an extension of the short sequence. This design ensures that the user does not benefit from sequence-specific knowledge gained during the previous sequence.

Experiment 2 (knowledge through description) provides the test subject with textual information about the dot positions before the dot sequence is shown. The screen is divided into six areas, numbered 1 through 6, and the positions of the dots in the sequence is given in terms of a sequence of numbers that correspond to an area. This experiment models a scenario where a trusting (or even actively collaborating) user provides the adversary with knowledge about his or her workstation, as outlined in our threat model. Such information transfer is rarely perfect, so we model the transferred knowledge by giving the test subject the rough location of the dots, that is, one of the six areas, before they appear on the screen. The test subject has no time limit when looking at and trying to remember the dot positions. We repeat the experiment 3 times with different five-dot sequences and 3 times with different seven-dot sequences, to capture both simple and more complex tasks. Unlike the previous experiment, we consider knowledge transfer rather than natural learning; consequently, each sequence is only used once. We make this choice as repetitions of identical sequences would combine the effects of knowledge transfer (i.e., giving external information to the user) and natural learning (i.e., the user learning from completing a sequence more than once). This combination would then make it hard to isolate the individual contributions of each source of information.

Experiment 3 (knowledge through observation) provides the test subject with a visual representation of the exact dot positions before the dot sequence is shown. This models the case where the adversary is able to observe the legitimate user while he or she performs tasks, also known as "shoulder surfing." While a legitimate user's gaze position is not visible through observation in an office environment, things like the cursor position are still likely to reveal some information. This experiment represents the maximum amount of information an adversary is able to obtain before attempting the task him- or herself.

5.3. Feature Stability Over Time

For eye tracking to be a useful defense against insider threats, the features measured from our test subjects must be relatively stable over time; otherwise, false rejects would occur frequently as the template becomes outdated. While this can be countered by sporadically retraining the classifier, this constitutes a serious challenge, as the user identity has to be established reliably during this time. We present a full list of features in Section 6 (Table I). In this section, we present the main reasons that time stability is a challenging problem:

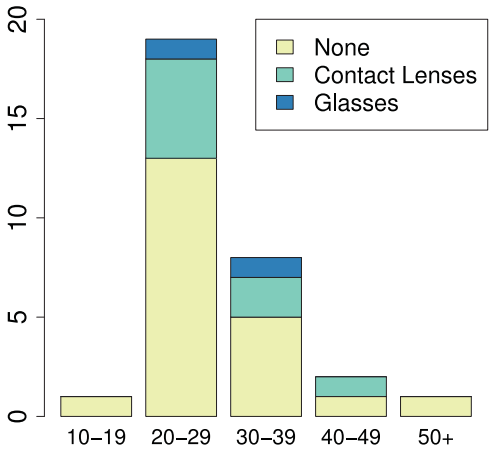


Fig. 6. Participant age distribution in decades. Out of 30 participants, two are wearing glasses and nine are wearing contact lenses.

Fig. 7. Our experimental setup consists of an SMI RED500 gazetracker that determines the user's gaze position on a 24-inch screen with a 1920x1200 resolution.

Changes in the environment. Features like the pupil diameter may change depending on lighting conditions. While the screen brightness is kept constant across all subjects and all sessions, the level of daylight may change. It is important that the classifier accounts for these changes.

Changes in the user's physical and mental state. Neuroscientific research shows that a person's eye movement behavior can change depending on states like drowsiness, exhaustion, stress, or euphoria (see Section 3 for details).

Technical artifacts. A recent study shows that the duration and number of fixations and saccades can depend on the gazetracker precision and the fraction of missing samples [Holmqvist et al. 2012]. As these values rely on the calibration of the gazetracker, they may change slightly across different sessions.

The changes described previously can manifest themselves both within the same session and across multiple days or weeks. Technical artifacts may be particularly prevalent when using data collected in different sessions due to the fact that a separate calibration has to be performed before each session. Despite these difficulties, we show in Section 7 that we are able to collect a classifier training dataset that is rich enough to reduce the influence of these error sources. By including training data from several sessions, we are able to capture, and adjust for, both long-term and short-term feature decay.

5.4. Participant Recruitment

Our data is collected from 30 participants, 20 male and 10 female. Participants were recruited through public advertisements, mailing lists, and social media. Aside from a minimum age of 18, there were no further exclusion criteria. The age distribution, as well as whether the subjects are wearing glasses or contact lenses, is given in Figure 6. The experiments are conducted with the approval of the ethics committee of the University of Oxford, reference SSD/CUREC1/13-064.

5.5. Task Selection

The set of tasks described in Section 5.2 is suitable to both detect the effect of learning on eye movement behavior and to measure the amount of identifying information that

can be obtained through eye movements. However, it does not reflect tasks that would typically be performed in an office environment. As such, it is hard to draw reliable conclusions with regard to the performance of a continuous authentication system in a real-world environment. In order to amend this, we define a second set of experiments that more closely resembles such an environment. In line with our initial reasoning, we keep the tasks identical for all participants in order to avoid classification of tasks rather than individuals. Our task set consists of reading, writing, two videos, and web browsing.

Reading: As part of this task, the participants are presented an excerpt from Daniel Defoe's *Robinson Crusoe*. The black text is displayed on a white background in a single column centered horizontally on the screen, as is common with many types of e-book software.

Writing: During this task, the users are asked to copy part of the text they read before. To this end, they are presented with the original text on the top half of the screen and an empty text box below the text. One might argue that restricting users in what they are typing might potentially influence features; however, it is difficult to truly capture daily typing behavior in a lab experiment. Even when displaying a writing prompt (such as asking participants to recapitulate their day), many participants would likely be at a loss as to what to write, thus greatly limiting the fraction of time spent typing. While it would be possible to ask participants to perform tasks they were intending to perform regardless of the experiment, the unfamiliar environment would likely affect behavior and constitute a confounding factor. Besides the test of task dependence, we hope to gain another insight from this task: due to the nature of optical eye tracking, samples might be lost when the user is looking at the keyboard (which is likely to be frequent for inexperienced typists). Based on this assumption, we will also quantify the fixation rate, as it directly impacts the speed with which authentication decisions can be made (see Section 7 for details).

Browsing: An obvious choice for this task would be to give users a fixed time limit and not to restrict the websites they visit. Naturally, this might lead to users choosing wildly different sites, such as streaming sites (e.g., YouTube), news sites, or online games. Compared to the number of possible groups of websites, the number of users is relatively small, likely leading to profiling of website types rather than users. To address the tradeoff between a real-world environment and the need to fix the task, we used a Wikipedia browsing game for this task. As part of this game the user is initially shown a random Wikipedia article and asked to exclusively use links within that page to reach the article "University of Oxford." Once this goal is accomplished, the user is asked to use Wikipedia's "Random Article" function to start over until the task's time limit is reached.

Videos: With the increasing popularity of online streaming sites such as YouTube and Netflix, we considered it important to include watching a video as one of the tasks. Because it is infeasible to include all varieties and genres of videos in a single lab study, we selected two different videos as representatives. The first video shown to participants is "Big Buck Bunny," a popular short computer-animated comedy film produced by the Blender Foundation.[3] The rationale behind this choice is that the film is released under an open-source licence (and can be shown as part of the experiment without further legal restrictions) and is likely to keep participants engaged in the experiment. The film itself features both slow fading and rapid cuts, together with frequently changing color schemes. Our second choice is an educational video titled "The Problems with First Past the Post Voting Explained," which is also freely available

[3]https://peach.blender.org/., last visited 01/25/2016

on YouTube.[4] Unlike Big Buck Bunny, there is very limited movement, quick scene changes are mostly absent, and the color scheme is bright and lacks frequent changes.

The data for this experiment is collected from 10 participants, six male and four female. The recruitment process was identical to that used for the previous experiment (see Section 5.4). The experiments are conducted with the approval of the ethics committee of the University of Oxford, reference SSH_C1A_15_139.

5.6. Experimental Setup

Figure 7 shows our experimental setup. We use an SMI RED500 eye-tracking device with a sampling rate of 500Hz to collect the raw gaze data. The stimuli are displayed on a 24-inch Dell U2412M monitor with a resolution of 1920x1200 pixels. The viewing distance between the subjects and the screen is approximately 50cm. In order to reduce distractions and to minimize the influence of the experimenter on the subjects, all instructions were displayed on-screen during the session. Although the gazetracker compensates for minor head movements during the data collection, we asked the participants to move as little as possible.

Before the session, the gazetracker has to be calibrated for each test subject. This stage consists of a calibration phase and a verification phase in which the error between actual and estimated viewing angle in degrees is determined. In order to ensure as high a data quality as possible, we reject calibrations with a viewing angle error of more than 1°, either horizontally or vertically. If the error is too high, the calibration has to be repeated. At the end of the session, we repeat the verification phase in order to test whether the initial calibration is still valid. A large verification error at this stage indicates low-quality data, most likely due to excessive movements during the experiments. During testing, we observed an average error of 0.49° in the X-direction and 0.52° in the Y-direction immediately after calibration. These errors increased to 0.74° and 0.72°, respectively, over the course of the experiment. Given that the error rates are lower than our threshold even at the end of the experiment, we are confident in the quality of our data.

5.7. Modifying Sampling Rate

As outlined previously, all the data in our study was collected at a sampling rate of 500Hz. Capturing data at the highest available sampling rate provides the benefit of exploring exactly which distinctive features are contained in human eye movements, even though this sampling rate might not be available in equipment used in many productive environments. While it would be possible to repeat the experiments with different hardware, this would make comparisons more difficult due to external factors (e.g., lighting, different individuals) that are virtually impossible to control for. Even when keeping all factors identical by simultaneously collecting data with multiple devices, the number of datasets that is collected is inherently limited by space constraints when placing the devices. In order to provide insights into both the maximum distinctiveness of the biometric and the performance that can be expected with consumer-level hardware, we perform downsampling on our dataset to simulate the sampling rate of these devices. We employ downsampling factors of 1, 2, 5, and 10. A downsampling factor of n means that every n^{th} sample is *kept*. Consequently, based on the initial sampling rate of 500Hz, we generate individual datasets with 500, 250, 100, and 50Hz.

[4]https://www.youtube.com/watch?v=s7tWHJfhiyo., last visited 01/25/2016.

6. FEATURE SELECTION

In this section, we describe different types of features, explain the reasoning behind each choice, and link them to the foundations in neuroscientific work described in Section 3. We will describe three measures of feature quality and analyze both the feature distinctiveness and their time stability using these measures. Additionally, we analyze the effects of reduced sampling rate and explain these effects using insights gained from neuroscientific work (see Section 3).

6.1. Feature Selection Criteria

An important consideration when choosing features is what data is required to compute them and whether there are any constraints regarding the environment in which they are collected. In order to make the authentication system usable in a standard working environment, the calculation of the features must only use raw eye-tracking data without relying on controlling, or even being aware of, running applications or screen content. This assumption distinguishes our approach from related work, which measures the user's reactions to controlled stimuli, and is therefore unsuitable for transparent continuous authentication [Cantoni et al. 2015; Liang et al. 2012].

It is important to know to what degree features are influenced by the task the user performs while the features are collected. As eye movements are always a reaction to a stimulus, perfect task independence can never be guaranteed, but some features are more susceptible to such influences than others. Largely task-independent features allow conducting the training phase with a task different from the one performed during the system's actual operation. This is particularly desirable in an office environment, as a wide variety of tasks are performed on a daily basis. A higher degree of task independence will significantly reduce the error rates exhibited by the system.

We choose our features such that their computation does not depend on any specific experimental design. As such, we don't use features that depend on the dot-clicking game (e.g., the position or even the presence of the red dots) or any of the real-world tasks. The main advantage of this approach is that the experimental design (i.e., the tasks performed by the subjects) is interchangeable and the authentication can be transparent and occur without the user's cooperation or even knowledge. While the features can be computed regardless of the task, their *distributions* might still be affected. We evaluate the effect of task selection and changing feature distributions in Section 8.3

6.2. Grouping of Samples

The gazetracker reports raw samples containing x/y coordinates and the current pupil diameter. As a single raw sample does not contain any distinguishing information, it is necessary to combine multiple raw samples and use the relationships between these samples (i.e., movements instead of static positions) as features. Given the nature of the data, we consider fixations to be the most natural level of abstraction. The gazetracker groups samples collected over at least 50ms that lie within a 30-pixel radius into a fixation (see Figure 1). In the context of this section, the term "sample" will refer to one fixation (i.e., a set of raw samples). In our data, we observe one fixation on average every 250ms, yielding a sampling rate of 4Hz. It is important to note that this rate may change depending on the experimental design (e.g., reading will lead to longer fixations and a lower sampling rate) and across different users.

6.3. Feature Types

A complete list of our features is given in Table I. We consider three different types of features: pupil features, temporal features, and spatial features.

Table I. List of Pupil, Temporal, and Spatial Features That Are Computed for Each Fixation

Feature	RMI	K-S Statistic	Bhattacharyya Distance
Pupil features			
Pupil Diameter - Max	19.84%	0.61±0.28	0.78±0.88
Pupil Diameter - Mean	20.27%	0.62±0.28	0.84±0.97
Pupil Diameter - Min	20.26%	0.61±0.29	0.82±0.97
Pupil Diameter - Range	1.19%	0.12±0.07	0.02±0.02
Pupil Diameter - Stdev	0.98%	0.11±0.06	0.02±0.01
Temporal features			
Acceleration - Max	2.49%	0.18±0.12	0.05±0.06
Acceleration - Mean	0.35%	0.07±0.03	0.01±0.00
Duration of Saccade	1.09%	0.12±0.05	0.02±0.02
Duration of Fixation	0.9%	0.10±0.06	0.01±0.02
Pairwise Speed - Max	4.95%	0.25±0.16	0.10±0.12
Pairwise Speed - Mean	5.36%	0.26±0.17	0.11±0.14
Pairwise Speed - Stdev	1.77%	0.14±0.09	0.03±0.04
Spatial features			
Distance from Center - Max	1.2%	0.12±0.06	0.02±0.02
Distance from Center - Mean	2.52%	0.20±0.12	0.04±0.05
Distance from Center - Min	0.72%	0.11±0.06	0.01±0.01
Distance from Center - Stdev	1.21%	0.13±0.07	0.02±0.02
Distance from Previous Fixation	0.66%	0.10±0.05	0.01±0.01
Max Pairwise Distance	1.23%	0.13±0.07	0.02±0.02
Max Pairwise Distance X Only	1.06%	0.13±0.07	0.02±0.02
Max Pairwise Distance Y Only	0.84%	0.11±0.05	0.02±0.01

For each feature, we report the relative mutual information (RMI) shared with the user ID. Additionally, we compute the average and standard deviation (as indicated by the ± sign) of the Kolmogorov-Smirnov statistic of the two-sample KS test, and the Bhattacharyya distance for all pairs of users. For all metrics, higher values indicate higher feature quality.

Pupil features can be split into static and dynamic features. As outlined in Section 3, the *range* of the pupil diameter is largely constant for each person. We capture this static range using the maximal, minimal, and mean pupil diameter that is observed during one fixation. The dynamic component is reflected by the short-term changes of the pupil diameter. These changes can be caused by cognitive load or different stimulation through lighting. While these external stimuli are equal for all participants, their *reactions* to them may not be. We model these changes through the standard deviation and the difference between the minimal and maximal pupil diameter observed during a fixation.

Temporal features include the duration of saccades and fixations as well as speed and acceleration. Both the peak and the average velocity of movements within a fixation have been shown to differ greatly between people in related neuroscientific work (see Section 3). These differences are mainly caused through different prevalences of saccadic intrusions and microsaccades, both of which are characterized by high velocity and acceleration. Different studies report similar ranges for these values, even though their experimental designs differ significantly. This suggests that these features show a high degree of task independence, which makes them particularly desirable for classification. We compute the velocity between each pair of consecutive samples and only use the magnitude of acceleration (i.e., we do not use the direction). The reasoning behind this is that the direction of acceleration depends on the location of the target stimulus and is therefore task dependent [Hafed and Clark 2002].

Spatial features are a method to measure the steadiness of a person's gaze. A fixation is a group of samples within a fixed-size circle, which consists of the samples

and a center point (see Figure 1 for an illustration). While the total area that can be covered by a fixation is limited by this definition, the spatial distribution of samples within this area can still be different. If a person's gaze is steady, the samples will be clustered closely around the fixation center, with few samples outside of this group. If a person has trouble focusing his or her gaze, the samples will be spread more evenly. We compute both the distance between each raw sample and the center point and the distance between each pair of raw samples. As some movements may be more pronounced in the vertical or horizontal direction, we also make this distinction. The distance between two fixations (as measured by the Euclidean distance between their center points) allows us to measure how many points between two areas of interest (i.e., target stimuli) are actively focused and processed by the subject.

6.4. Determining Feature Quality

Having a measure of feature quality is important for two reasons: (1) to be able to select the best features when the entire set is too high-dimensional and (2) to gain better insights into *why* the biometric works. Additionally, it allows one to measure how external factors (such as different hardware or collection time spans) affect each feature. Even initially highly distinctive features might be unusable if one or more of these factors severely degrade its performance. In order to ensure the robustness of the ranking of the features in our set, we employ three different measures: the relative mutual information (RMI), the Kolmogorov-Smirnov statistic of a two-sample KS-test, and the Bhattacharyya distance. Initially, an amount of uncertainty is associated with the user ID (its entropy). This amount depends on the number of classes (i.e., users) and the distribution of the samples between users. Each feature reveals a certain amount of information about the user ID, and this amount can be measured through the mutual information (MI). In order to measure the mutual information relative to the entire amount of uncertainty, we use the relative mutual information (RMI), which measures the percentage of entropy that is removed from the user ID when a feature is known [Frank et al. 2012]. The RMI is defined as

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

where H(A) is the entropy of A and H(A|B) denotes the entropy of A conditioned on B. The range of this feature is between 0 (indicating that the feature contains no information about the user) and 1 (meaning that all users can be uniquely identified through this feature). In order to calculate the entropy of a feature, it has to be discrete. As most features are continuous, we perform discretization using an Equal Width Discretization (EWD) algorithm with 20 bins [Dougherty et al. 1995]. This algorithm typically produces good results without requiring supervision. In order to limit the drastic effect that outliers can have when using this approach, we use the $1st$ and $99th$ percentile instead of the minimal and maximum values to compute the bin boundaries. A high RMI indicates that the feature is distinctive on its own, but it is important to consider the correlation between features as well when choosing a feature set. Additionally, several features that are not particularly distinctive on their own may be more useful when combined.

The relative mutual information relies on discretization of feature values; the number of bins and the algorithm used to filter outliers might change not only the absolute values of the measure but also the relative ranking of features. In order to gain additional insights, we calculate the Kolmogorov-Smirnov statistic of a two-sample KS test. We consider the two 1-dimensional probability distributions for two users with regard to a single feature. The KS-test then tests whether the two samples are drawn from the same distribution (null hypothesis). A feature would only be distinctive if the null hypothesis can be rejected for a high number of user pairs. As the information

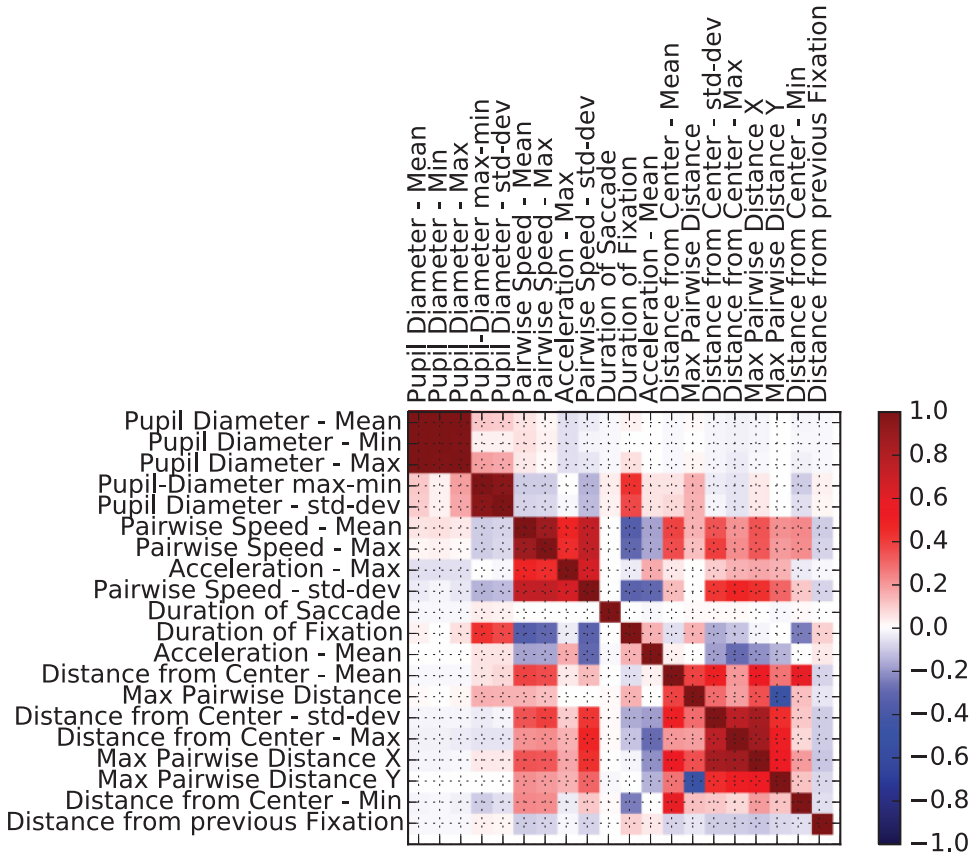


Fig. 8. Feature correlation measured by the Pearson correlation coefficient. A value of 0 indicates no correlation; values of 1 and −1 signify positive and negative correlation, respectively.

whether the null hypothesis is rejected at a certain confidence does not provide any information about the *magnitude* of the differences between the samples, we use the Kolmogorov-Smirnov statistic as a metric, computed as

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of two different subjects. The measure is computed for all pairs of subjects. Table I provides the averages and standard deviations. Defined as the difference between two empirical distribution functions (that lie between 0 and 1 at each point), the KS-statistic also takes a value in that interval. A value of 1 indicates that the average difference between two users regarding this feature is maximal, thus suggesting a distinctive feature. None of the features used in the biometric follow a normal distribution ($p < 0.001$); as such, the fact that the test does not assume any specific distribution of the data is critical.

Additionally, for each pair of users (p,q) we compute the Bhattacharyya distance of a feature as

$$D_B(p,q) = -ln\left(\sum_{x \in X} \sqrt{p(x)q(x)}\right).$$

The Bhattacharyya distance measures the similarity of two continuous probability distributions, in this case the probability distributions of the same feature for two

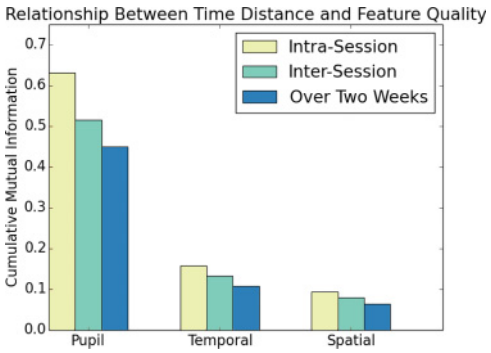


Fig. 9. Influence of data collection time span.

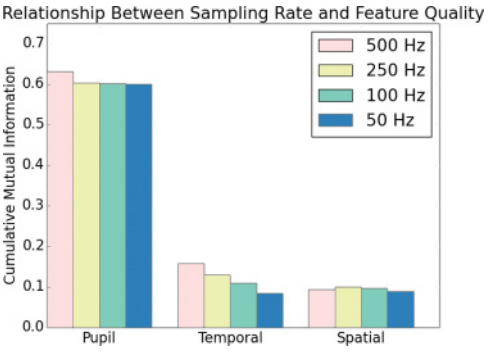


Fig. 10. Influence of sampling rate.

users. Higher values indicate bigger differences between the distributions, resulting in higher distinctiveness of this feature. This metric has been shown to correlate well with classification accuracy for a number of classifiers and datasets [Choi and Lee 2003].

6.5. Discussion

Table I shows how each feature performs with regard to the three metrics discussed in the previous section. Each metric has different value ranges; as such, the values are not directly comparable. However, the relative rankings of features in the set are comparable between the three metrics, which suggests that the feature discretization performed before computing the RMI does not distort the overall ranking and all three metrics are indicative of the features' quality.

The static pupil diameter features (i.e., min, mean, and max) share the most information with the user ID. The dynamic pupil diameter features (i.e., the standard deviation and the min-max difference) are less distinctive, which suggests that the pupil diameter is more a result of different genders, ethnicities, and eye shapes than a behavioral feature.

While the behavioral features, both temporal and spatial ones, show a lower distinctiveness than the pupil diameter, they still contribute significant amounts of information. The fact that both peak speed and acceleration exhibit comparatively high scores with regard to all metrics shows that we accurately model the distinctive capabilities of saccadic intrusions and microsaccades.

When selecting which feature candidates should form the final feature set, there are several aspects that have to be considered: each of the features should be hard to imitate in a given threat model. As we focus on insider threats, this rules out features that can be easily observed and copied. Given the insights from Section 3, we suspect that it may be possible for a sophisticated attacker to modify his or her own pupil diameter to a certain degree. Specifically, it has been demonstrated that the pupil reacts almost instantaneously to external light stimulation, while the reversion to the baseline occurs slowly [Herbst et al. 2011]. Consequently, an attacker could decrease his or her own pupil diameter by constantly shining a bright light in his or her own eyes, even though the reverse might be harder to achieve. While it might be possible to recognize such a stimulation (e.g., by monitoring ambient light intensity), we still consider this a valid threat. In order to address this issue, we also investigate the performance of a feature set that does not make use of the pupil diameter features. When putting the system into operation, it can then be decided which feature set should be used, depending on the threat model and the capabilities of potential attackers. We will discuss the impact of not using the pupil diameter as a feature (and thereby raising the bar for an attacker

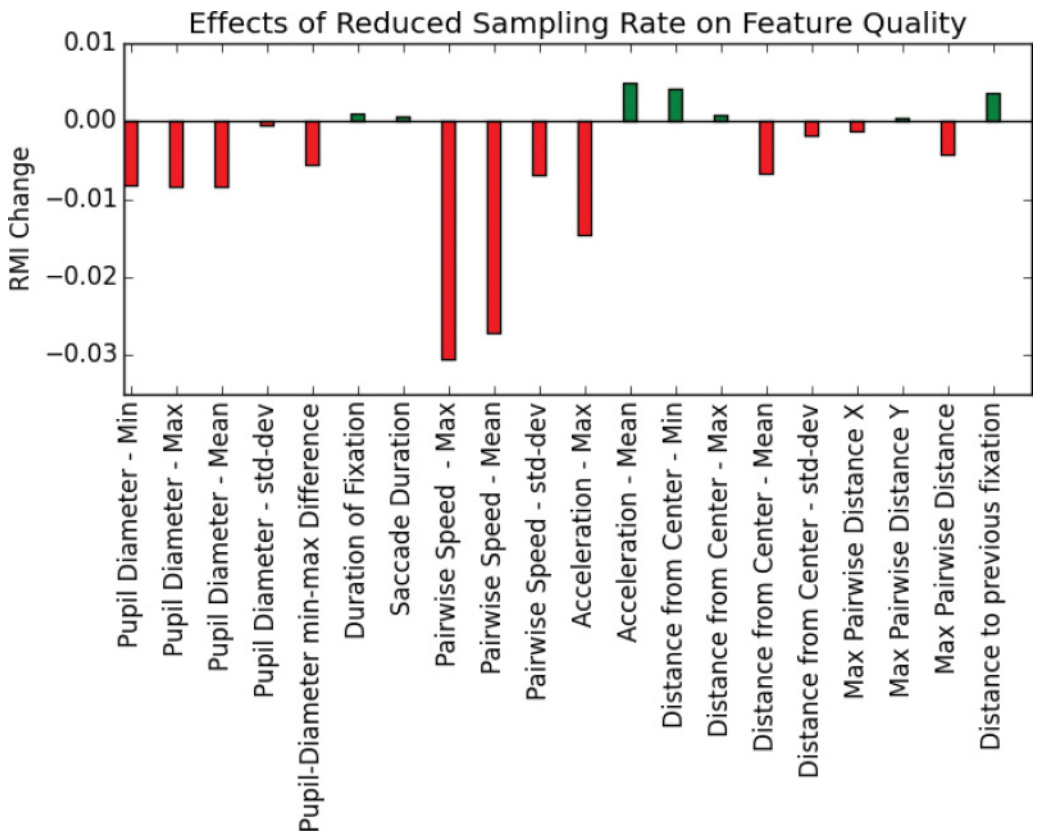


Fig. 11. Changes to the RMI of individual features when reducing the sampling rate to 50Hz. The temporal features reflecting the properties of microsaccades (speed and acceleration) are most strongly affected.

trying to perform an imitation attack) in Section 8. Figure 8 shows that the correlation between features belonging to the same group (i.e., pupil diameter, temporal or spatial) is relatively high, while the intergroup correlation is considerably lower. This suggests that all three groups contribute to the distinctiveness of the biometric and no group can be replaced entirely by another. This also makes sophisticated imitation attacks more difficult, as a number of very distinct features have to be emulated simultaneously.

6.6. Feature Degradation

As outlined in the previous sections, we consider two main factors that could negatively impact feature quality: (1) increased data collection time span and (2) reduced sampling rate.

Figure 9 shows the effects of increased data collection time spans on the cumulative RMI of the three feature groups. Not surprisingly, the intrasession dataset (which only includes the first session) results in the highest total information. After that, the features' information content declines with increasing time distance between the sessions, with the effect being more pronounced for the features involving the pupil diameter. As outlined in Section 3, the pupil diameter is susceptible to external factors such as lighting, so this degradation is to be expected.

In line with our expectations, reducing the sampling rate also reduces the information content of features (see Figure 10). Pupil diameter features suffer slightly from

halving the sampling rate, and any further reduction has no measurable effect. Conversely, temporal features decrease almost linearly with every further increase of the downsampling factor. This effect can be explained through the short-lived nature of the main physiological processes causing the distinctiveness of this feature group. As described in Section 3, microsaccades and saccadic intrusions are distinctive but only last a few milliseconds. As such, it is not surprising that reducing the sampling rate below a certain threshold prevents their distinctive capabilities from being harvested. Figure 11 breaks down the degradation of features when reducing the sampling rate to 50Hz. Maximal and average speed and peak acceleration, the features most strongly associated with microsaccades, suffer from the biggest degradation of all features in the set. The degradation is statistically significant for all features in this group ($p < 0.05$). Conversely, the spatial features, with the exception of the mean distance to center, show no statistically significant changes ($p > 0.05$) even for the lowest sampling rate setting.

7. CHOICE OF CLASSIFIERS AND METRICS

In this section, we will describe both open-set and closed-set classifier candidates and discuss their individual advantages when used with the eye movement biometric, as well as the impact that feature selection, sampling rate, and time distance have on the classifier performance. In addition to the error rates, we will present metrics that make it possible to gauge the real-world performance of the system. Finally, we will give insights on how different parameters of our system can be chosen to reflect different security requirements.

7.1. Closed-Set Classifiers

Training of closed-set classifiers requires samples for each potential user of the system. The output of the classifier for a new sample is then the predicted class, chosen out of the set of classes it was initially trained with. This type of classifier is useful in an insider threat scenario, as an employer using a biometric system likely has collected templates for all employees. The advantage of such a system is that it not only detects an unauthorized user (by virtue of the claimed identity, as established through a user name, not matching the user recognized by the classifier), but also can reveal the attacker's identity. The major disadvantage is that once an external attacker (i.e., somebody not enrolled in the system) attempts to access a workstation, the classifier will recognize him or her as the user with the closest template, which might lead to incorrectly granting access or framing an innocent user for the failed attack.

We consider two closed-set classifiers, the k-nearest-neighbors (knn) algorithm and Support Vector Machines (SVMs). In order to determine the optimal parameters for these classifiers, we perform a grid search through a defined subset of the parameter space. For the knn classifier, we tested values of k between 1 and 20 and weighting samples uniformly or by Euclidean distance. For the SVM, we tested a linear, a polynomial, and a radial kernel (rbf) function. For all three kernels, we varied the soft margin constant C in powers of 10 between 1 and 10,000. The polynomial kernel was used with degrees between 2 and 5, and for the radial kernel function, we tested values of γ between 0.00001 and 10. The best results were achieved with $k = 5$ and weights based on Euclidean distance for knn and an rbf-kernel with $C = 10{,}000$ and $\gamma = 0.001$ for the SVM.

After completing the training phase, the classifier continuously assigns labels (i.e., user IDs) to fresh samples. If the system is used for authentication (which is the focus of this work) rather than identification, this decision can be transformed to either an accept (i.e., the predicted user matches the claimed user) or a reject (i.e., the predicted user is different from the claimed user). The decision's robustness can be increased

by combining multiple samples before making a final decision. Combining multiple samples will increase the accuracy of the decision but also introduces a delay before an imposter can be detected. As eye tracking provides a stream of new samples at a constant and high rate, we choose to combine several samples for each authentication decision. Our authentication system is parameterized through the size n of a sliding window and the threshold t, which defines how many of the samples in this window must support the current user. As such, a fresh sample is accepted only if at least t out of the last n labels output by the classifier match the claimed user ID. Therefore, a higher value of n increases the system's robustness but delays the detection of an attacker, while the value of t controls the tradeoff between the false accept rate and false reject rate.

7.2. Open-Set Classifiers

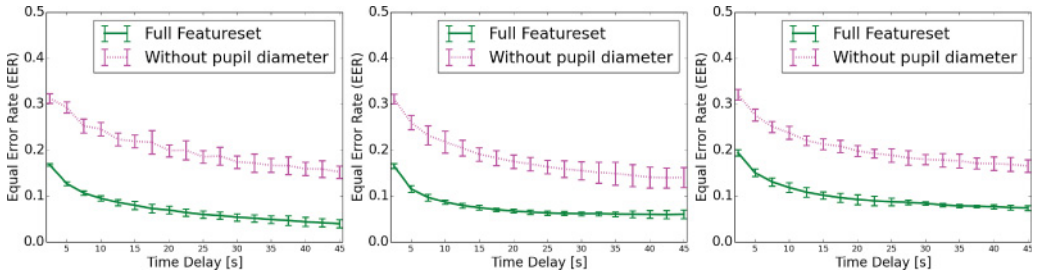
When used for authentication, one-class classifiers are only trained with data belonging to a single legitimate user. For each new sample, the classifier determines whether it is sufficiently close to the data observed during training. As such, the output of the classifier is only a yes/no decision, rather than a user ID (the output of a closed-set classifier); therefore, it is unable to reveal the identity of an attacker. However, identification of attackers is impossible even for a closed-set classifier unless reference data is available for all potential attackers. While this may well be the case for an insider threat scenario, it is unrealistic for a more general setting. Additionally, there are usually other means of identifying an attacker who is physically present, such as video surveillance. The major advantage of one-class classifiers is that training does not require any knowledge about either the set of possible attackers or their individual biometric templates. As such, it is well suited to detect both inside and outside attackers.

In this section, we analyze the performance of the one-class Support Vector Machine. This classifier is parameterized by the kernel coefficient γ and the regularization parameter ν. The value of ν is an upper bound on the fraction of *training* errors. Higher values of ν will increase false rejects while reducing false accepts. While it would be possible to tailor the value of ν to achieve a certain split between the two error rates, this is generally undesirable as changing the parameter requires retraining the classifier. Therefore, we choose both values to minimize the total number of errors within the development dataset. Using a grid search on a development set different from the data used for training and testing, we identify $\gamma = 0.001$ and $\nu = 0.18$ as optimal parameters. The actual authentication decision is made using the same sliding-window technique we use for the closed-set classifier. A sample is accepted only if at least t out of the last n classifier decisions are accepts.

7.3. Metrics

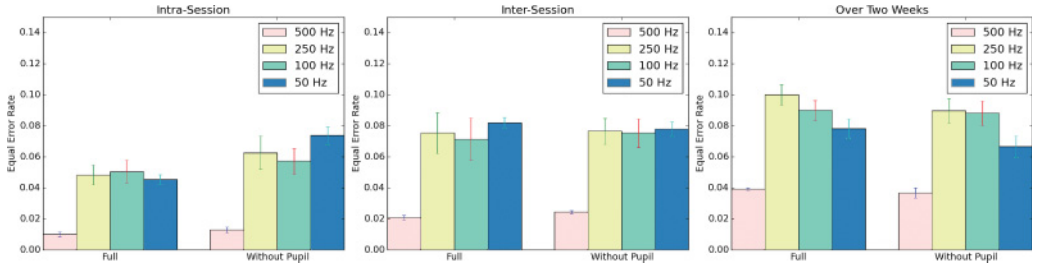
Most papers proposing new biometric systems provide both the FAR and FRR of their chosen biometric. Most of the time some sensitivity parameters can be adjusted to trade a lower FAR for a higher FRR, and vice versa. In an attempt to make different systems more comparable, the EER is often given as well. The EER is the error rate of the system when its parameters are adjusted such that both the FAR and FRR are equal.

In order to ensure comparability of our biometric with previous work, we also provide the EER (see Section 8). However, this metric suffers from a number of practical problems that make it difficult to draw valid conclusions from it: in a continuous authentication scenario (which is the environment in which behavioral biometrics often compare most favorably with hard biometrics), it is crucial to know how the errors are distributed between both legitimate users and user-attacker pairs. An FRR of 5% could



(a) Intra-Session dataset (b) Inter-Session dataset (c) 2-weeks dataset

Fig. 12. Average equal error rates obtained through fivefold stratified cross-validation on three different datasets using the closed-set SVM classifier. The error bars indicate 95% confidence intervals.



(a) Intra-Session dataset (b) Inter-Session dataset (c) 2-weeks dataset

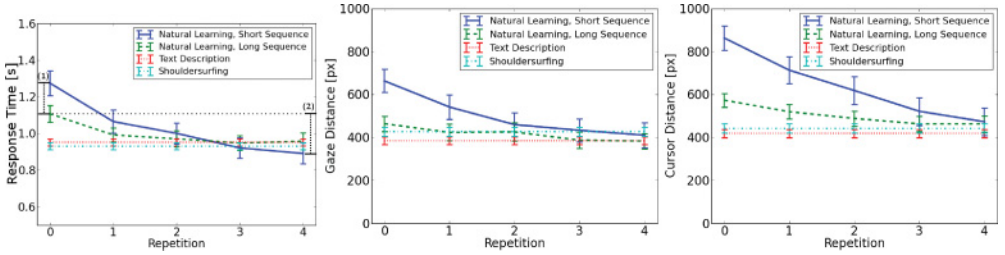
Fig. 13. Average equal error rates obtained with the one-class SVM classifier for different sampling rates. The error bars indicate 95% confidence intervals. Reducing the sampling rate to 250Hz has a large effect on the error rates; any subsequent reduction does not produce a statistically significant increase in the EER.

signify that all users are rejected exactly once every 20 samples, or that 5% of the users are rejected consistently (or, most likely, something in between). Obviously, these cases pose very distinct challenges. The latter case could be addressed by authenticating the users that are consistently rejected using a different mechanism (such as another biometric), while the first case renders the entire biometric largely useless. The same can be said for the FAR, questioning the usefulness of the EER as a measure of security. Additionally, it is impossible to derive a biometric's typical attack detection speed without knowing its sampling rate.

To address these issues, we provide two more metrics: the systematic false-negative rate (sys-fn) and the median time until an attacker is detected (med-ttd). The systematic false-negative rate is the fraction of attackers that are never detected (within the scope of our data). This is usually due to their biometric template being close to that of a legitimate user. As with the conventional measures (FAR and FRR), these values depend on the system's sensitivity settings. In order to provide results that are easy to compare, we report them at a setting that never rejects legitimate users. As such, any additional security provided by the system comes at no cost in terms of user inconvenience or needing a mechanism to handle false alarms.

8. RESULTS

In this section, we will discuss the usefulness of eye movement patterns to both derive task familiarity and perform continuous authentication. The latter will be evaluated using data from both the task familiarity experiments (Section 5.2) and the task dependence experiments (Section 5.5).



(a) Response Time (b) Distance from Gaze to Stimulus (c) Distance from Cursor to Stimulus

Fig. 14. Changes of three different performance metrics caused by natural learning, text descriptions, and shoulder surfing: (1) shows the effects of increased familiarity with game mechanics, and (2) shows the memory effects of learning a sequence. The error bars indicate the 95% confidence intervals.

8.1. Task Familiarity

As stated in Section 5, the goal of the experiment is to distinguish familiar and unfamiliar users by analyzing eye movements. In order to make this distinction, we first have to confirm that users are actually improving (gaining familiarity) over the course of the experiment, as all users are initially unfamiliar with the dot-clicking game. In order to provide ground truth for this assessment, we define two metrics to measure task performance. This allows us to test whether any improvements in the gaze-based metrics correlate with improvements in actual user performance. (1) *Response Time* is the time it takes a test subject to complete one dot of the sequence in the experiment. It is important to note that this metric refers to a *single dot*, rather than a *sequence of dots*. This definition allows us to compare both short and long sequences using the same metric. As the users are asked to complete the sequences as quickly as possible, this measure is the most natural measure of performance. (2) *Cursor Distance* is the distance between the cursor location and the position of the stimulus, right before it is displayed. Reaction time and mouse movement time are the most significant components of the response time (Figure 4 illustrates this for a single user). As such, using information about the (predicted) dot location to position the cursor closer to the dot will significantly improve the task completion time. A (significant) decrease in the cursor distance can only be due to a better prediction of the dot location by the user, showing that the user has gained and used information either through natural learning or knowledge transfer.

In addition to these ground-truth measures, we introduce the gaze distance, defined as the distance between a user's gaze position and the position of the stimulus (the dot) just before it is displayed. If this measure follows the same trend set by the two ground-truth metrics, it shows that it is possible to learn about users' task familiarity by observing their eye movement behavior.

Figure 14 shows the results of our experiment. As we do not perform repetitions with identical sequences for Experiments 2 and 3 (text descriptions and shoulder surfing), the figure shows the average over all sequences.

There are two ways in which the users improve over the course of the experiments: (1) by learning the game mechanics and (2) by predicting the location of the dot before it appears. The second component then allows the participant to use the blank period between dots to reposition the cursor. With an increasing number of repetitions, the users improve with regard to all three metrics. This improvement shows in the response time decreasing from 1.33 seconds to 0.9 seconds for the short sequence, with similar effects on gaze and cursor distance. Generally, this effect could be due to either (1) or (2), as users become more familiar with both the game and the sequence. However, once they proceed from the short sequence to the (different) long sequence, they only benefit

Table II. Performance of the One-Class SVM Classifier

			Full			Without Pupil Diameter		
Dataset	**Subjects**	**Sampling Rate**	**EER**	**Sys-fn**	**Med-ttd**	**EER**	**Sys-fn**	**Med-ttd**
Intrasession	30	500Hz	1.00%	15.44%	30.50s	1.29%	51.80%	∞
Intrasession	30	250Hz	4.83%	24.88%	37.50s	6.26%	60.04%	∞
Intrasession	30	100Hz	5.05%	25.35%	37.50s	5.71%	63.63%	∞
Intrasession	30	50Hz	4.52%	13.63%	22.25s	7.36%	50.28%	∞
Intersession	20	500Hz	2.06%	8.24%	27.50s	2.44%	42.88%	370.00s
Intersession	20	250Hz	7.53%	16.14%	35.50s	7.65%	42.81%	378.25s
Intersession	20	100Hz	7.14%	17.08%	37.50s	7.53%	44.50%	384.25s
Intersession	20	50Hz	8.19%	10.74%	27.00s	7.77%	35.37%	263.50s
2 weeks	20	500Hz	3.92%	2.69%	27.50s	3.67%	31.70%	323.75s
2 weeks	20	250Hz	10.00%	25.14%	84.25s	8.96%	68.38%	∞
2 weeks	20	100Hz	8.99%	25.52%	105.50s	8.81%	66.38%	∞
2 weeks	20	50Hz	7.82%	9.90%	31.50s	6.64%	54.57%	∞

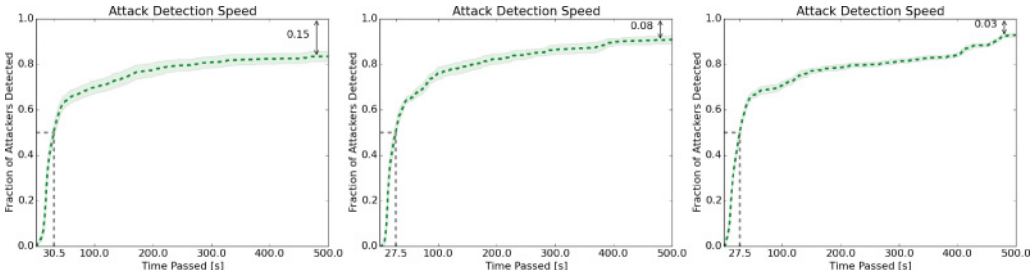
Sys-fn is the fraction of attackers that are not detected within the scope of the data; med-ttd is the median time until an attacker is detected. Note that all metrics are computed at a parameter setting that never rejects a legitimate user (i.e., not at the parameter setting that achieves the equal error rate). For a visualisation of these metrics see Figure 15.

from an improved grasp of the game mechanics (as the short and long sequences are independent from each other). This improvement, shown as (1) in Figure 14, is present for both of the ground-truth metrics (response time and cursor distance), as well as the gaze-based metric. On average, users improve from 1.33 seconds for the first iteration of the short sequence to 1.11 seconds for the first iteration of the long sequence. In both cases, the sequence was unknown, suggesting that this improvement is only due to an improved grasp of the game mechanics. The improvement resulting from learning the sequence (i.e., becoming better at predicting the position of the next dot) is marked as (2) in Figure 14, showing the difference in performance between a familiar and an unfamiliar sequence. While users take an average of 0.90 second for the last repetition of the short sequence, this increases to 1.11 seconds as the effect of having learned the sequence disappears.

The gaze-based metric follows similar patterns, showing an improvement of 132px for (1) and 30px improvement for (2), although only the former is statistically significant ($p < 0.01$). For all metrics, there were no statistically significant differences ($p > 0.05$) in user performance, regardless of whether information was obtained through natural learning, shoulder surfing, or text descriptions. This suggests that gauging task familiarity through gaze patterns is particularly effective against outside attackers without access to these sources of information.

8.2. Continuous Authentication

The performance of the open-set classifier for all combinations of dataset, feature set, and sampling rate for the task familiarity experiment (see Section 5.2) is given in Table II. Most notably, the equal error rate decreases greatly when compared to the closed-set classifier (Figure 12). The relative relationships between the error rates for different datasets (i.e., intrasession, intersession, over 2 weeks) is preserved, and increasing time distance also increases the equal error rate. This effect is present for all sampling rates, although it is most pronounced for the 500Hz setting as the intrasession error rates are much lower. Interestingly, the same cannot be said when switching from the full feature set to a set without the pupil diameter. While the error rates increase for two out of the three datasets, they even *decrease* slightly when using data gathered over 2 weeks. This suggests that the pupil diameter is not only a potential security



(a) Intra-Session dataset (b) Inter-Session dataset (c) 2-weeks dataset

Fig. 15. Attack detection speed of the open-set classifier with parameter settings that never reject legitimate users. The dotted line represents the average, and the shaded area the 95% confidence interval over the five folds of the cross-validation. These results were obtained using the full feature set and sampling rate; see Table II for the remaining results.

concern but also an actual source of error when the biometric is used over longer time spans. This is most likely due to the fact that the effects of external stimuli (such as lighting) overshadow the initial distinctiveness provided by physical characteristics. Consequently, this feature group should not be used for long-term operation without regular classifier retraining (e.g., at the start of a session).

As would be expected, reducing the sampling rate reduces classification accuracy and consequently increases the equal error rate. The sharpest increase can be observed when reducing the initial sampling rate of 500Hz to 250Hz. Any subsequent changes don't result in a statistically significant increase in error rate. Surprisingly, the reduction from 100Hz to 50Hz *lowers* the equal error rate for some datasets.

Regarding the newly introduced metrics "median time to detection" and "systematic false negative rate" (see Section 7.3 for details), it is interesting to note that they don't correlate well with the equal error rate. While sys-fn increases when reducing the sampling rate, the effect is not nearly as strong as would be expected given the enormous increase in EER. The effect is even more pronounced when not using the pupil diameter for classification. While the error rates increase only marginally (or even decrease in the 2-weeks dataset), the fraction of undetected attackers increases immensely. In some datasets, this fraction increases to over 50%, resulting in a median time to detection of ∞. Note that these metrics are computed for a parameter setting that never results in legitimate users being rejected within the scope of our data. As such, the detection of attackers comes at no additional costs with regard to user inconvenience or handling of false alarms.

8.3. Task Dependence Experiment

8.3.1. Pupil Diameter Correction. Medical research has shown that a person's pupil diameter is greatly affected by light stimulation (see Section 3 for details). Naturally, the tasks described in Section 5.5 will result in different screen brightness (with the videos being generally darker than text on white background). Consequently, it is likely that classification accuracy would suffer when the image or screen brightness differs between enrollment and operation. However, a software-based continuous authentication system can monitor the brightness of the currently displayed image and correct the raw pupil diameter reported by the eyetracker accordingly. In order to perform this correction, two pieces of information are needed: (1) the brightness of an image and (2) the way a person's pupil diameter depends on this brightness.

During each of the main tasks, we continuously record the image brightness along with the eye-tracking data. This is necessary, as the screen brightness during a task

might depend on user actions (e.g., the set of webpages visited during the browsing task). To compute the brightness of the image, we compute the average brightness of all RGB pixels. As different components of an RGB color contribute differently to the overall brightness (green being perceived as brighter than blue, for instance), we use the following formula, as proposed by the W3C[5]:

$$Br = (R \times 0.299) + (G \times 0.587) + (B \times 0.114).$$

Given a maximum value of 255 for each of the color components, the brightness is a value between 0 and 255, with 0 being black and 255 being white. When examining the correlation between screen brightness and pupil diameter for the combination of all tasks, an average Pearson correlation coefficient of -0.68 was observed. This correlation is statistically significant ($p < 0.001$) and suggests that a correction for this relationship would be beneficial.

In order to perform the actual correction, it is crucial not to use information derived from the tasks used for classification. Before starting the main tasks described in Section 5.5, we display a black screen, followed by a white screen for 20 seconds. We choose this order as the pupil's adaptation to bright light is almost instantaneous, whereas adaptation to darkness is gradual [Herbst et al. 2011]. Based on these tasks, we observe that, on average, the pupil diameter decreases by 0.005 when brightness is increased by one. The difference in this slope between users is minimal; as such, we use a single value for all users. While classification accuracy might improve marginally by using a user-tailored approach, a single value greatly simplifies the enrollment phase. The corrected pupil diameter is obtained according to the following formula:

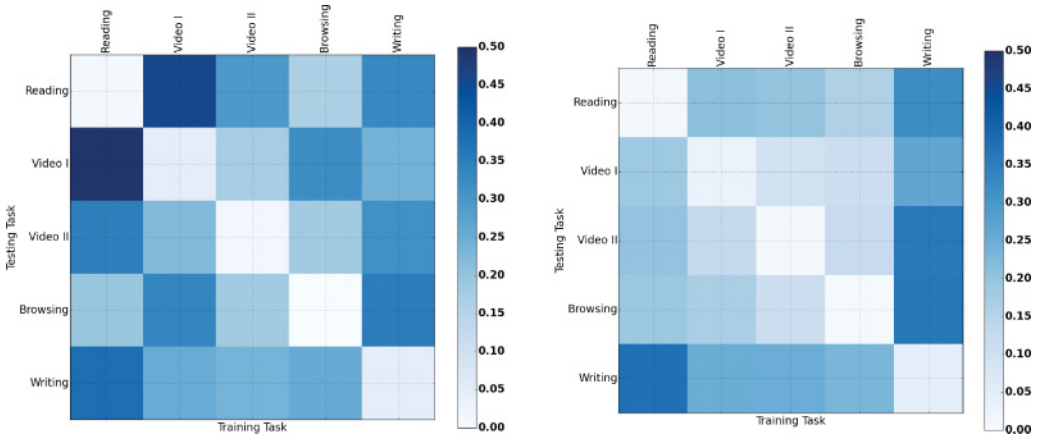
$$d_{new} = d_{raw} + (Br \times 0.005).$$

When using the pupil diameter correction, this formula is applied to all samples, both in the training and testing sets. After applying the correction, we observe no statistically significant correlation between corrected pupil diameter and screen brightness ($p > 0.05$).

8.3.2. Evaluation. Figure 16 shows the results of our analysis. It is apparent that the EER is low when using data from the same task for training and testing, with values between 0.04% for browsing and 4.9% for the first video. These error rates are comparable to those exhibited with the task familiarity experiment described in Section 5.2. Within a single task, the pupil diameter correction only provides a meaningful change for the first video, decreasing the EER from 4.9% to 3.3%. This result is intuitive, as the other tasks exhibit significantly smaller variations in brightness.

Classifier performance changes significantly when using one task for enrollment and another for operation. The most apparent increase in EER results in the combination of the first video and the reading task, regardless of which of these two tasks is used for training. In this scenario, the EER approaches 50%, thus not providing a benefit over random guessing. This is likely the result of the pupil diameter being greatly affected by the bright background of the reading task compared to the relatively dark video. Consequently, the classification accuracy is actually worse than not using the pupil diameter at all, as the pupil diameter of Alice reading a text might match that of Bob watching a video. This EER is drastically lowered when applying the pupil diameter correction, reducing the error rate from 49% to 18.7%. A similar reduction can be observed for the other task pairs, with the one exception being the writing task. When using writing for either training or testing, performance is poor, regardless of whether or not the pupil diameter correction is used. The most likely cause of this lies in the

[5]https://www.w3.org/TR/2000/WD-AERT-20000426#color-contrast, last visited 01/25/2016.



(a) Without pupil diameter correction (b) With pupil diameter correction

Fig. 16. Equal error rates for different combinations of training and testing tasks. The right figure uses the pupil diameter correction described in Section 8.3.1.

imperfect nature of eye tracking when the user is not looking directly at the screen. During the experiment, we noticed that many users were either inexperienced typists or not familiar with the keyboard layout, resulting in them frequently looking away from the screen. This apparent behavior is backed by a drop in the sampling rate (when only counting valid samples) from an average of 470Hz to 360Hz for the writing task. However, once the device loses track of a person's eye, there is a brief recovery period that results in low-quality tracking. These two factors lead to a fixation rate of 0.9Hz, as opposed to the average of 4Hz over the other tasks. Both the sampling rate and the EER suggest that the eye movement biometric is not suitable for use during tasks centered largely on writing or other tasks that frequently draw the user's gaze away from the screen. While the gaps in the data created by these distractions will ultimately remain a problem, we believe that improvements in gaze-tracking technology (especially given the rise of eye tracking in the entertainment sector) will likely reduce the length of these gaps and reduce the loss of tracking accuracy around them, thereby mitigating this additional source of error.

The results show that the error rates are comparable to, or even lower than, those observed for the task familiarity experiment. While the error rates increase significantly when performing enrollment on a different task, this issue can be mitigated through pupil diameter correction. Additionally, the results suggest that grouping similar tasks (e.g., text-based tasks) for the purpose of enrollment might yield sufficiently low error rates, thereby balancing acceptable error rates with simplified enrollment. The generation of templates for different task groups could be coupled with automated task detection to further improve accuracy without giving up the transparent authentication.

9. LIMITATIONS AND DISCUSSION

There are a number of possible limitations to consider when evaluating this work. All data has been collected in a lab study; while the features can be computed in any environment and error rates are low for all tested real-world tasks, the participants were still restricted in their actions. While this may seem like an obvious limitation, we argue that it is a necessary first step to draw meaningful conclusions about the biometric. Regardless of the number of subjects, there is always the danger of each

subject choosing an individual (variation of a) task, which could lead to classification accuracy being overestimated as the classifier distinguishes tasks instead of users. The same reasoning applies for the need to conduct the study in a controlled environment (i.e., in a lab study). Under the insider threat model, the attacker would always use the workstation in the *same environment* as the victim, as the biometric is meant to secure local access. Consequently, the environment (and resulting factors such as lighting) have to remain the same for all users and this level of control can only be reliably established in a lab study. While a field study would certainly give interesting additional insights, for example, by instructing users to attempt to impersonate their coworkers, we consider a lab study a necessary first step into investigating the suitability of eye movements as a biometric.

We performed the experiment with 30 users for the task familiarity experiment and 10 users for the analysis of task dependence. Naturally, a higher number of participants would give greater confidence in the robustness of our results. Nevertheless, our participants cover a wide variety of age groups, both genders, and a number of participants with glasses or contact lenses (see Figure 6). Together with the narrow confidence intervals shown in Figure 12, this suggests that similar results could be expected in a larger study. Our recruitment process is based on social media and mailing lists, which, along with the natural selection of people willing to participate in experiments in general, might introduce a bias that influences our results. Additionally, our sample might be subject to additional unknown sampling bias (as some subsets of the entire population may be particularly hard or particularly easy to distinguish), although this cannot necessarily be avoided even with higher sample sizes if the source of the bias is unknown. So far there has been, to the best of our knowledge, no research exploring how the distribution of our features changes across different subsets of the population, with the exception of the pupil diameter. Without establishing these effects first, it is hard to draw a sample from the entire population that is representative with regard to the biometric.

The threat model in this work assumes a zero-effort attack, with participants not actively trying to modify their own eye movements with regard to our feature set. If an attacker is able to record a victim's eyes, he or she might attempt to match his or her victim's eye movement patterns when attempting to access the victim's workstation. While it is, by the nature of the problem, impossible to show that such an attack is infeasible, we consider it unlikely to succeed in practice. Medical research shows that subjects have not been able to permanently suppress microsaccades (the type of eye movement likely responsible for the distinctiveness of temporal features) and that temporary suppression leads to a higher rate of microsaccades shortly after [Engbert and Kliegl 2004]. Modifying the exact duration and magnitude of acceleration would arguably be even more difficult. To the best of our knowledge, the only feature that has been shown to be susceptible to influence through stimulation is the pupil diameter. Besides manual imitation, another attack vector would be the creation of an artificial eye that moves according to the attacker's specification. However, due to the millisecond scale of fixations, the control would have to be extremely precise and the attack could be countered by implementing liveness detection (i.e., distinguishing between human and artificial eyes).

There might be factors influencing a person's eye movements we have not accounted for (such as fatigue or the effects of medication or alcohol). While we have likely captured many different confounding factors by recording data across three sessions, more research is needed to measure potential long-term changes in eye movement patterns.

The transparent nature of the proposed authentication system allows the establishment or confirmation of a user's identity without his or her active cooperation or

even knowledge. This property, which is shared by other behavioral biometrics (such as mouse movement behavior), poses a privacy concern. However, the work presented here performs authentication, and as such it is necessary for the user to make an initial identity claim (e.g., by entering a user name). This necessity mitigates the privacy concern, as the system does not allow identifying an anonymous user. The use of eye movement technology in general still raises concerns due to their diagnostic capabilities (see Section 3), although these are not the focus of this article.

10. CONCLUSION

In this work, we have contributed a set of 20 discriminative features based on a person's eye movement patterns. The features are based on fundamentals set by related neuroscientific and medical work. We have shown that the features provide identifying information on individuals both for artificial and real-world tasks. We designed a controlled experiment that accounts for different ways an inside attacker can obtain information from a naïve or colluding user, to aid in impersonation attacks. Using gaze-tracking data from our experiments, we quantify the advantage an adversary has in impersonating a user and test if the adversary has obtained knowledge about the task the user normally performs. The data collected during our experiments comes from 30 members of the general public. In order to test the time stability of our features, we performed two repetitions of the experiments, 2 weeks apart. By downsampling our data to sampling rates typically provided by low-cost hardware, we show that using the eye movement biometric is feasible even with cheap consumer-level devices. By employing different metrics, we evaluate our feature set and measure the effects that increased time distance and reduced sampling rate have on the quality of both feature groups and individual features. Using an open-set classifier with data from a single session, we achieve an equal error rate of 1.00%. When used with a parameter setting that never rejects legitimate users, the system takes a median of 33.5 seconds to detect an attacker. Under the same zero-reject settings, the biometric detects 84.56% of all attackers within the scope of our data. To demonstrate the suitability of the biometric for continuous authentication during everyday tasks, we collected a second dataset consisting of reading, writing, browsing, and video tasks. The results show that reliable authentication is possible for any of these task groups, with error rates comparable to the artificial task set. The combination of these results with the increasing availability of cheap eye trackers suggests that eye movements are an excellent candidate biometric to provide additional security at little extra cost.

REFERENCES

Richard Abadi and Emma Gowen. 2004. Characteristics of saccadic intrusions. *Vision Research* 44, 23 (2004), 2675–2690.

Salil P. Banerjee and Damon L. Woodard. 2012. Biometric authentication and identification using keystroke dynamics: A survey. *Journal of Pattern Recognition Research* 7, 1 (2012), 116–139.

Claude Barral and Assia Tria. 2009. Fake fingers in fingerprint recognition: Glycerin supersedes gelatin. In *Formal to Practical Security*. Springer, 57–69.

Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. 2005. Eye-movements as a biometric. In *Image Analysis*. Springer, 780–789.

Arman Boehm, Dongqu Chen, Mario Frank, Ling Huang, Cynthia Kuo, Tihomir Lolic, Ivan Martinovic, and Dawn Song. 2013. SAFE: Secure authentication with face and eyes. In *IEEE PRISMS 2013*.

Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. 2015. GANT: Gaze analysis technique for human identification. *Pattern Recognition* 48, 4 (April 2015), 1027–1038.

Barbara Cassin, Melvin L. Rubin, and Sheila Solomon. 1984. *Dictionary of Eye Terminology*. Wiley Online Library.

CERT. 2011. CyberSecurity Watch Survey. Software Engineering Institute, Carnegie Mellon University. Retrieved from http://resources.sei.cmu.edu/asset_files/Presentation/2011_017_001_54029.pdf.

Euisun Choi and Chulhee Lee. 2003. Feature extraction based on the Bhattacharyya distance. *Pattern Recognition* 36, 8 (2003), 1703–1709.

Brett A. Clementz, John A. Sweeney, Michael Hirt, and Gretchen Haas. 1990. Pursuit gain and saccadic intrusions in first-degree relatives of probands with schizophrenia. *Journal of Abnormal Psychology* 99, 4 (1990), 327.

Cory Cornelius, Jacob Sorber, Ronald Peterson, Joe Skinner, Ryan Halter, and David Kotz. 2012. Who wears me? Bioimpedance as a passive biometric. In *Proceedings of the 3rd USENIX Workshop on Health Security and Privacy*.

Alexander De Luca, Martin Denzel, and Heinrich Hussmann. 2009. Look into my eyes!: Can you guess my password? In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 7.

Mandalapu Sarada Devi and Preeti R. Bajaj. 2008. Driver fatigue detection based on eye tracking. In *Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology, 2008 (ICETET'08)*. IEEE, 649–652.

James Dougherty, Ron Kohavi, and Mehran Sahami. 1995. Supervised and unsupervised discretization of continuous features. In *Proceedings of the International Conference on Machine Learning*. 194–202.

Nguyen Minh Duc and Bui Quang Minh. 2009. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings* (2009).

Andrew Duchowski. 2007. *Eye Tracking Methodology: Theory and Practice*. Vol. 373. Springer.

Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2015. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *Proceedings of the 2015 Network and Distributed System Security (NDSS'15) Symposium*.

Ralf Engbert and Reinhold Kliegl. 2004. Microsaccades keep the eyes' balance during fixation. *Psychological Science* 15, 6 (2004), 431–431.

Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. 2012. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE Transactions on Information Forensics and Security* 8, 1 (2012), 136–148.

R. Stockton Gaines, William Lisowski, S. James Press, and Norman Shapiro. 1980. *Authentication by Keystroke Timing: Some Preliminary Results*. Technical Report. DTIC Document.

ITU Gaze Group. 2015. Eye tracking and gaze interaction. Retrieved from www.gazegroup.org.

Ziad M. Hafed and James J. Clark. 2002. Microsaccades as an overt measure of covert attention shifts. *Vision Research* 42, 22 (2002), 2533–2545.

Kristina Herbst, Birgit Sander, Dan Milea, Henrik Lund-Andersen, and Aki Kawasaki. 2011. Test–retest repeatability of the pupil light response to blue and red light stimuli in normal human eyes using a novel pupillometer. *Frontiers in Neurology* 2, 10 (2011), b30.

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.

Takehiro Ito, Shinji Mita, Kazuhiro Kozuka, Tomoaki Nakano, and Shin Yamamoto. 2002. Driver blink measurement by the motion picture processing and its application to drowsiness detection. In *Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, 2002.* IEEE, 168–173.

Robert J. K. Jacob. 1995. Eye tracking in advanced interface design. *Virtual Environments and Advanced Interface Design* (1995), 258–288.

Anil K. Jain, Arun Ross, and Sharath Pankanti. 2006. Biometrics: A tool for information security. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 125–143.

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.

A. Jones, R. P. Friedland, B. Koss, L. Stark, and B. A. Thompkins-Ober. 1983. Saccadic intrusions in Alzheimer-type dementia. *Journal of Neurology* 229, 3 (1983), 189–194.

Zach Jorgensen and Ting Yu. 2011. On mouse dynamics as a behavioral biometric for authentication. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*. ACM, 476–482.

Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science* 154, 3756 (1966), 1583–1585.

Miltiadis Kandias, Alexios Mylonas, Nikos Virvilis, Marianthi Theoharidou, and Dimitris Gritzalis. 2010. An insider threat prediction model. In *Trust, Privacy and Security in Digital Business*. Springer, 26–37.

Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. 2010. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 187–190.

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*. ACM, 13–19.

Zhen Liang, Fei Tan, and Zheru Chi. 2012. Video-based biometric identification using eye tracking technique. In *Proceedings of the 2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC'12)*. IEEE, 728–733.

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 (2002), 175–182.

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.

Michelle Keeney, Eileen Kowalski, Dawn Cappelli, Andrew Moore, Timothy Shimeall, and Stephanie Rogers. 2005. Insider threat study: Computer system sabotage in critical infrastructure sectors. Retrieved from http://www.cert.org/archive/pdf/insidercross051105.pdf.

Youssef Nakkabi, Issa Traoré, and Ahmed Awad E. Ahmed. 2010. Improving mouse dynamics biometric performance using variance reduction via extractors with separate features. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 40, 6 (2010), 1345–1353.

W. Leigh Ottati, Joseph C. Hickox, and Jeff Richter. 1999. Eye scan patterns of experienced and novice pilots during visual flight rules (VFR) navigation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 43. SAGE Publications, 66–70.

Kasper B. Rasmussen, Marc Roeschlin, Ivan Martinovic, and Gene Tsudik. 2014. Authentication using pulse-response biometrics. In *Proceedings of the 21st Network and Distributed System Security Symposium (NDSS'14)*.

Keith Rayner, Caren M. Rotello, Andrew J. Stewart, Jessica Keir, and Susan A. Duffy. 2001. Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 219.

Kenneth Revett, Hamid Jahankhani, Sérgio Tenreiro de Magalhães, and Henrique M. D. Santos. 2008. A survey of user authentication based on mouse dynamics. In *Global E-Security*. Springer, 210–219.

Frank Rieger. 2013. Chaos Computer Club Breaks Apple TouchID. Retrieved from http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid/.

Javier San Agustin, Henrik Skovsgaard, Emilie Mollenbach, Maria Barret, Martin Tall, Dan Witzner Hansen, and John Paulin Hansen. 2010. Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 77–80.

Abdul Serwadda and Vir V. Phoha. 2013. When kids' toys breach mobile phone security. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. ACM, 599–610.

D. Shanmugapriya and Ganapathi Padmavathi. 2009. A survey of biometric keystroke dynamics: Approaches, security and challenges. *Arxiv Preprint Arxiv:0910.0817* (2009).

Sasitorn Taptagaporn and Susumu Saito. 1990. How display polarity and lighting conditions affect the pupil size of VDT operators. *Ergonomics* 33, 2 (1990), 201–208.

Chee Meng Tey, Payas Gupta, and Debin Gao. 2013. I can be You: Questioning the use of keystroke dynamics as biometrics. In *The 20th Annual Network & Distributed System Security Symposium (NDSS'13)*.

David Tock and Ian Craw. 1996. Tracking and measuring drivers' eyes. *Image and Vision Computing* 14, 8 (1996), 541–547.

Michel Wedel and Rik Pieters. 2000. Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science* 19, 4 (2000), 297–312.

Nan Zheng, Kun Bai, Hai Huang, and Haining Wang. 2012. *You Are How You Touch: User Verification on Smartphones via Tapping Behaviors*. Technical Report. Tech. Rep. WM-CS-2012-06.

Received June 2015; revised March 2016; accepted March 2016