Preventing Lunchtime Attacks: Fighting Insider Threats With Eye Movement Biometrics

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

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

Vincent Lenders
Armasuisse, Switzerland
vincent.lenders@armasuisse.ch

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

***Abstract*—We introduce a novel biometric based on distinctive eye movement patterns. The biometric consists of 21 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, i.e., 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 identification and 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 two weeks. The results indicate that we can reliably authenticate users over the entire period. We show that the classification decision depends on all features and mimicking a few of them will not be sufficient to trick the classifier. 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.**

I. Introduction

In this paper, we evaluate the effectiveness of using eye movement biometrics as a novel defence against the "lunchtime attack" by an insider threat. An insider threat in this context refers to a person with physical access to a workstation that he is not supposed to use (e.g., using a coworker's workstation while he 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 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 years and even low-cost open-source hardware and software is available [1]. Recent advances in video-based eye tracking technology makes 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 II).

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 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: (i) no prior knowledge, i.e., no information advantage; (ii) knowledge gained through a description, e.g., the adversary is provided with a textual description by a colluding legitimate user; and (iii) knowledge gain through observation, e.g., 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 two 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 cannot utilize that knowledge to circumvent the eye movement biometric.

Our main contributions are a set of 21 features and measurements that confirm that these features are suitable to perform user authentication. We carefully consider various error sources and validate our design by looking at the learning behavior of our test subjects. We further show that it is possible to gauge the level of familiarity with a specific task through the eye tracker biometric. This property is very useful when dealing with an insider threat. Finally we also present a basic authentication system based on this biometric as well as a discussion of the robustness of our results over time.

The rest of the paper is organized as follows: Section II gives an overview over the relevant background on the human visual

NDSS '15, 8-11 February 2015, San Diego, CA, USA

http://dx.doi.org/10.14722/ndss.2015.23203

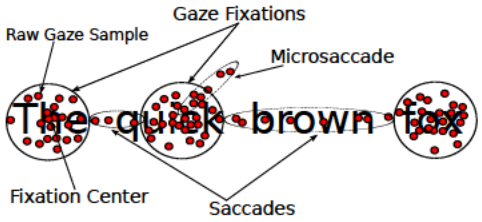


Fig. 1: A simplified example of gaze tracking: the raw gaze samples are collected by the eye tracking device and subsequently clustered into fixations, saccades, and microsaccades. The clustering depends on the spatial and temporal distribution of the gaze data (for the sake of clarity, the temporal distribution is not shown in the figure).

system. Section III describes the threat model, the definition and a discussion of features is given in Section IV. Section V describes our experimental design. Section VI presents our results and related work is summarized in Section VII. We conclude the paper in Section VIII.

II. 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, e.g., [2].

The Human Visual System 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 which 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 which 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 which allows 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.

A. Characteristics of Eye Movements

In general, the human eyes move within six 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 10-100ms, depending on the distance covered [2]. They are used to move the fovea[1] to another location. Once the saccadic movement has been signalized by the related neurons, the movement must be completed, i.e., neither the saccade's position nor its velocity can be consciously altered, even if the target has changed its positon [3].

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

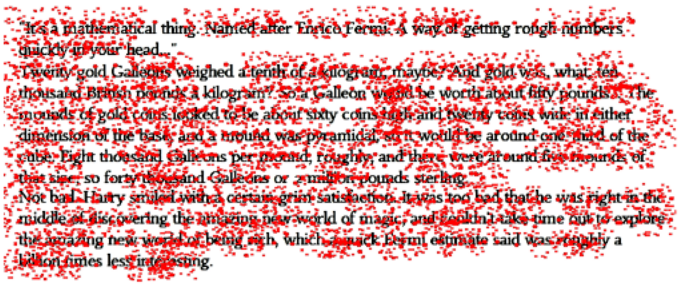


Fig. 2: Real-world gaze data from a subject reading a document.

In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100-400 ms. 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, i.e., if a person attempts to artificially fixate 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 centre [4]. Related to microsaccades are movements called Saccadic Intrusions (SI), which consist of involuntary movements away from the previous eye position, followed by a return to that position after a short duration [5]. 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 diagnosis, such as Alzheimer's [6] and schizophrenia [7]. 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) [8], [9], human-computer-interface design [10], pilot training [11], or detecting fatigue and drowsiness in drivers [12], [13], [14].

Besides the eye movements, the pupil diameter is also an interesting feature which 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 [15]. Nevertheless, multiple causes that affect the pupil diameter have been found, including memory and cognitive workload [16], lighting conditions [17] and drug consumption [18].

B. 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

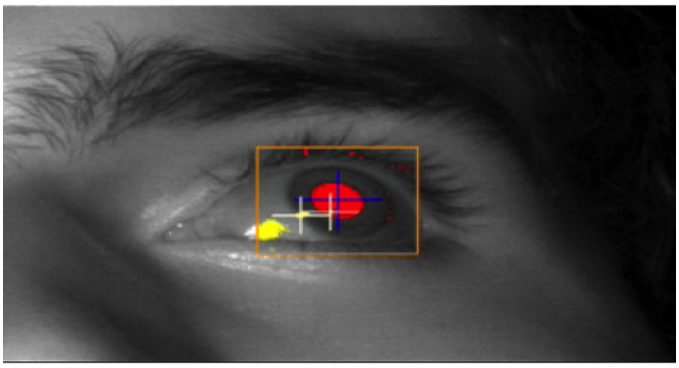


Fig. 3: Video-based gaze tracking: the tracking of eye movements is software-based and does not require any physical contact with a subject. The gaze position is calculated using the distance between pupil position and the corneal reflections (shown as two white crosses).

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 trade-off between temporal/spacial accuracy vs. 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 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 the 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 towards 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 which 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 reflexion patterns (see Figure 3). Importantly, the current video-based eye tracking is non-invasive 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 [1] is an open source project which 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 to their phones.

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 non-invasive eye tracking technology already enables an easy access to a rich and distinctive feature space of fixational eye movements. Their distinctive capabilities and involuntary nature makes them a potentially valuable biometric.

III. Threat Model

The adversary model considered in this paper 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 co-worker's workstation while the co-worker is away for lunch. Other examples include cleaning staff getting access to workstations after hours, or the trivial case where one employee simply allows another employee to use his workstation or access credentials. In all these scenarios, an adversary might gain access to a fully operational system, already logged into 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 has shown that 33% of electronic crimes in companies are committed by insiders [19]. 60% of these attacks use a compromised account, in the remaining cases the attacker uses their own account [20]. 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 organisations 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 [20].

In our model the adversary is aware of the gaze tracking system and will do his best to imitate the behavior of the legitimate user. This can be done by familiarizing himself 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 manages to observe how the user accesses sensitive data or performs some sort of transaction he will most likely be able to carry out his attack much faster, helping him 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 can circumvent the gaze tracking system, i.e., exhibit ocular characteristics that are similar enough to the legitimate user.

We consider two models of knowledge transfer to help the adversary familiarize himself 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 shouldersurfing) how a legitimate user operates the system.

We assume the adversary cannot disable the gaze tracking system, nor can he interfere with its operation in any way, as doing so would quickly lead to the detection of his attack. We don't consider insider threats which involve the attacker using his own workstation. These attacks can always be traced back to the actual attacker and are better dealt with through behavioural monitoring[21]. The aim here is to show that gaze tracking is a viable way of identifying users, as well as gauge a user's level of knowledge and familiarity with a particular task.

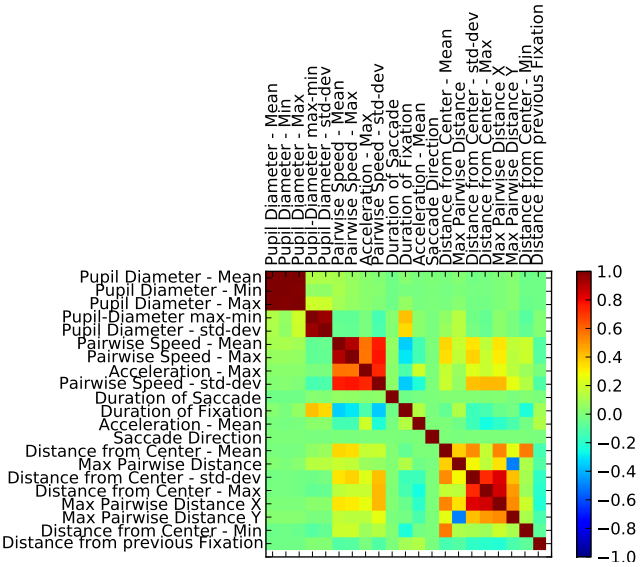


Fig. 4: Feature correlation measured by the pearson correlation coefficient. A value of 0 indicates no correlation, values of 1 and -1 signify positive and negative correlation, respectively.

IV. Feature Definition

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 II. We will rank these features according to the information they reveal about the user ID and discuss the implications of using different sets of features for classification.

A. Design 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 eyetracking 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 [22], [23].

It is important to know to which 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 to 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 they are as task-independent as possible and do not require any controlled stimuli. The main advantage of this approach is that the experimental design (i.e., the tasks performed by the subjects) is interchangeable, and the results are transferable to a wide set of general tasks.

B. 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.

C. 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.

Pupil features can be split into static and dynamic features. As outlined in Section II 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 both 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 II). These differences are mainly caused through different prevalence 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 [24].

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,

Feature	RMI	F	R	W
Pupil features				
Pupil Diameter - Max	19.84%	×	×	
Pupil Diameter - Mean	20.27%	×	×	
Pupil Diameter - Min	20.26%	×	×	
Pupil Diameter - Range	1.19%	×	×	
Pupil Diameter - Stdev	0.98%	×		
Temporal features				
Acceleration - Max	2.49%	×	×	×
Acceleration - Mean	0.35%	×		×
Duration of Saccade	1.09%	×	×	×
Duration of Fixation	0.9%	×		×
Pairwise Speed - Max	4.95%	×	×	×
Pairwise Speed - Mean	5.36%	×	×	×
Pairwise Speed - Stdev	1.77%	×		×
Spatial features				
Distance from Center - Max	1.2%	×		×
Distance from Center - Mean	2.52%	×		×
Distance from Center - Min	0.72%	×		×
Distance from Center - Stdev	1.21%	×		×
Distance from previous fixation	0.66%	×	×	×
Max Pairwise Distance	1.23%	×		×
Max Pairwise Distance X only	1.06%	×		×
Max Pairwise Distance Y only	0.84%	×	×	×
Saccade Direction	0.08%	×		×

TABLE I: List of pupil, temporal and spatial features that are computed for each fixation. For each feature we report the relative mutual information (RMI) with the user ID. A value of 0 indicates that the feature carries no information about the user ID, while a value of 1 means that the feature completely defines the user ID. For each feature we report whether it is included in the Full (F), Reduced (R) or without-pupil (W) feature set.

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 focussing their gaze the samples will be spread more evenly. We compute both the distance between each raw sample and the center point as well as 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. The saccade direction, measured in degrees, allows a distinction between stepwise and more diagonal eye movements.

D. Determining Feature Quality

Having a measure of feature quality is important for two reasons: (a) to be able to select the best features when the entire set is too high-dimensional and (b) to gain better insights into *why* the biometric works. 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, 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 [25]. The RMI is defined as

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

where H(A) is the entropy of A and H(A|B) denotes the entropy of A conditioned on B. 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 [26]. 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 1^{st} and 99^{th} 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.

E. Feature Selection

Table I lists the RMI for each of our features. 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 a comparatively high RMI with the user ID 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 II we suspect that it may be possible for a sophisticated attacker to modify his own pupil diameter to a certain degree. 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. Besides the security considerations it is also important to note that a high-dimensional feature set will slow down the classification and cause a higher resource consumption. If the feature redundancy is high or many non-distinctive features are included in the original set feature selection is particularly useful. Figure 4 shows that the correlation between features belonging to the same group (i.e., pupil diameter, temporal or spatial) is relatively high, while the inter-group 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. Therefore an optimal reduced feature set would most likely contain features from all three groups. In order to determine this feature set we used the Minimum Redundancy Maximum Relevance (mRMR) algorithm[27]. This algorithm selects those features that share a high amount of information with the classification variable (i.e., the user ID) while showing low redundancy with other features in the set. In order to achieve a good trade-off between classification speed and accuracy we chose the best ten features as computed by the algorithm. The list of those features can be seen in Table I. In line with our hypothesis features from all groups are part of this set. This also makes sophisticated imitation attacks more difficult, as a number of very distinct features have to be emulated simultaneously. We will discuss the impact of using different feature sets in Section VI-C.

V. Experimental Design

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, discuss how features change over time, as well as the best way to capture these changes.

A. Design Goals

With the experiments described in this section we test the hypothesis that a biometric based on gaze tracking is feasible. The goal is to analyze how well an authentication system can distinguish users within a test group, and to identify what impact, if any, training and knowledge transfer has on the authentication process.

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 designed our features to be as task-independent as possible it is impossible to rule out that *some* differences are due to the user-chosen task. 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 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.

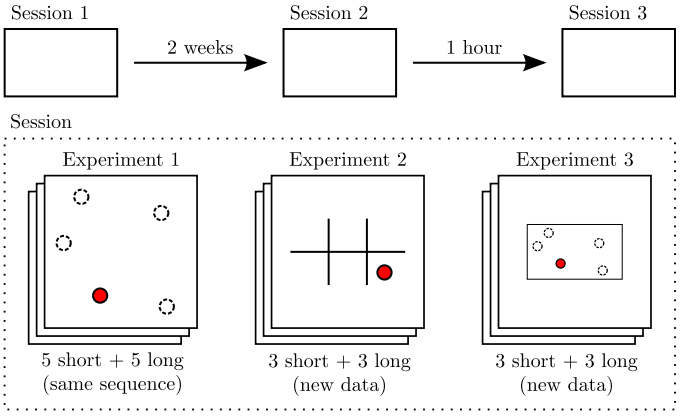


Fig. 5: Experiment structure. Each session is divided into three experiments, each of which is repeated a number of times. The entire session is repeated after two weeks, and again an hour after the second repetition.

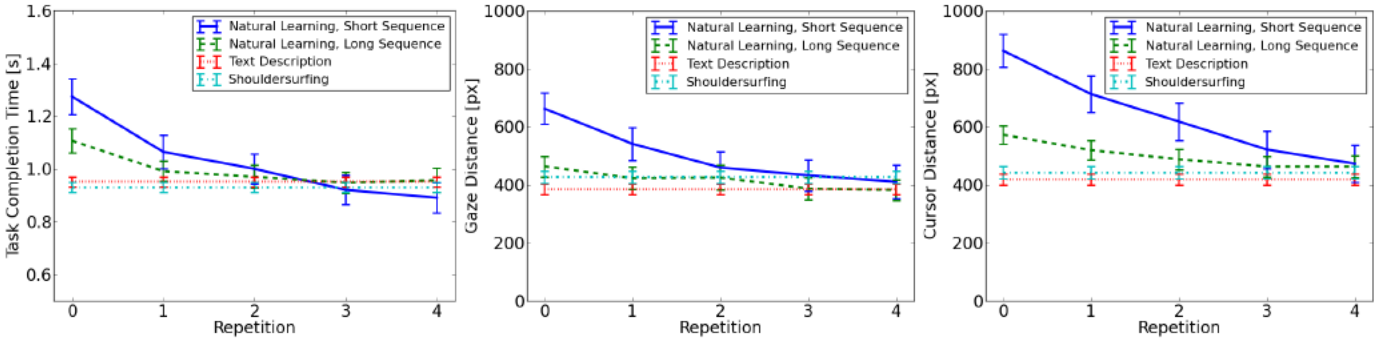
In order to overcome these sources of error we define a specific set of tasks that all users must complete. Our goal is to determine whether the users' eye movements are distinguishable, even if they are completing the same task the same way with the same knowledge. If this is indeed the case that means that there are *inherent* differences between users that can not be attributed to different ways of completing a single task. Nevertheless, as we do not make any assumptions about the experimental design when choosing the features the results are transferable to more general settings (e.g., web browsing or writing e-mails). We realize our design goals through a set of experiments.

B. Experiment Structure

We first introduce terminology to make it easier to refer to different parts of our interaction with test subjects, please 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 not only consistent 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 grey 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 grey 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.

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



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

Fig. 6: Changes of three different performance metrics caused by natural learning, text descriptions and shouldersurfing. When repeating the same sequence multiple times the performance increases steadily (natural learning), but the same level of performance is achieved through transferred knowledge (shouldersurfing and text descriptions). The error bars indicate the 95% confidence intervals.

version has 7 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 is facing. We assume that the subject's performance is affected by increasing task familiarity as well as memory-based learning effects when he 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 is accustomed to on a victim's workstation to cover his own tracks.

At each repetition the test subject is informed that the sequence will remain the same for the next iterations. Our hypothesis is that learning effects are more significant for shorter sequences. In order to test this we repeat the experiment 5 times with the same 5-dot sequence and 5 times with the same 7-dot 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.

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 give 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 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, i.e., 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 5-dot sequences and 3 times with different 7-dot sequences, to capture both simple and more complex tasks.

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

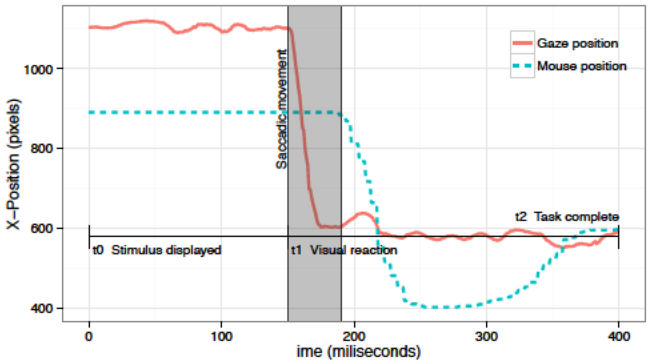


Fig. 7: At time $t0$ a new dot appears followed by a period of inactivity (reaction time) in which neither the gaze, nor the cursor move significantly. After about 150ms, at $t1$, a visual reaction in the form of a large saccade occurs (the gray area) and the gaze and cursor converges to the position of the stimulus.

where the adversary is able to observe the legitimate user while he performs his tasks, also known as "shouldersurfing". 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 himself.

C. Design Validation

We can only draw conclusions from our results if we can be sure that the test subjects actually use the information they are given (i.e., text descriptions or shouldersurfing). We hypothesize that task performance will improve during the natural learning that occurs when the same dot sequence is repeated and that advance knowledge will yield similar benefits to natural learning. In order to measure task performance we define three metrics: (1) *Task Completion Time* is the time it takes a test subject to complete one sequence of the experiment; (2) *Gaze Distance*, the distance between a subject's

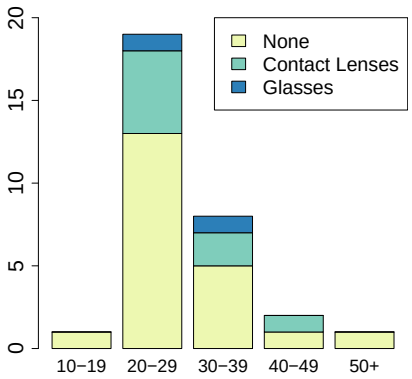


Fig. 8: Participant age distribution in decades. Out of 30 participants 2 are wearing glasses and 9 are wearing contact lenses.

Fig. 9: 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.

gaze position and the position of the stimulus (the dot), right before it is displayed; and (3) *Cursor Distance*, the distance between the cursor location and the position of the stimulus, right before it is displayed.

Figure 6 shows the results of our validation. As we do not perform repetitions with identical sequences for Experiment 2 and 3 (text descriptions and shouldersurfing), the figure shows the average over all sequences. We see that, as the number of repetitions go up, the average performance for Experiment 1 (natural learning) improves. The two knowledge transfer mechanisms cause the subjects to perform similarly or even better than through several repetitions of natural learning. We therefore conclude that our test subjects do benefit from the information in the same way that an attacker might.

D. Feature Stability Over Time

For eyetracking to be a useful defence 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 IV (Table I). In this section we present the main reasons why time stability is a challenging problem:

a) 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.

b) 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 II for details).

c) 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 [28]. As these values rely on the calibration of the gazetracker, they may change slightly across different sessions.

The changes described above 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 VI 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 session we are able to capture, and adjust for, both long-term and short-term feature decay.

E. Participant Recruitment

Our data is collected from 30 participants, recruited from the general public, 20 male and 10 female. The age distribution, as well as whether the subjects are wearing glasses or contact lenses, is given in Figure 8. The experiments are conducted with the approval of the ethics committee of the University of Oxford, reference SSD/CUREC1/13-064.

F. Experimental Setup

Figure 9 shows our experimental setup. We use an SMI RED500 eyetracking 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.

VI. Results and Analysis

In this section we will describe our classifier candidates and explain how the classification of raw samples can be extended to allow user authentication. We will discuss the impact that the feature selection and the time over which the data was collected have on the classifier performance. Finally we will give insights on how different parameters of our system can be chosen to reflect different security requirements.

A. Classifier Development

We measure the performance of the k-nearest-neighbors (knn) and Support Vector Machine (SVM) classifiers. 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 function. For all three kernels we varied the soft margin constant C in powers of ten between 1 and 10000. 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=10000 and γ=0.001 for the SVM.

B. From Classification to Authentication

After completing the training phase and generating a template for each user the authentication system decides continuously whether a new sample belongs to the currently logged in user. This decision can be either based on a single sample or combine multiple samples. Combining multiple samples will increase the accuracy of the decision but also introduces a delay before an imposter can be detected. As eyetracking 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 parametrized through the number of samples n that are used for the decision and the threshold t which defines how many of these samples must support the current user. This procedure is described in Algorithm 1. A discussion of the effects of both parameters will be given in the next section.

Algorithm 1 The authentication algorithm accepts the current user if at least t out of the last n classifications match his user ID. This allows us to control the trade-off between the FAR and the FRR.

```
Input: t,n,uid
classifications ← []
loop
    s ← collect_sample()
    classifications ← classifications + classify(s)
    window ← last n classifications
    accepted ← all uid ∈ window where count(uid)≥ t
    if uid ∈ accepted then
        accept sample
    else
        reject sample
    end if
end loop
```

C. Results and Discussion

In order to ensure a high statistical robustness we split the datasets into training and test sets using 5-fold stratified cross validation, resulting in 80% of the data being used for training and 20% for testing. The following results reflect the average of the 5 folds. The dataset contains data from all experiments. The second and third session form the inter-session dataset, the first and second are combined for the 2-weeks dataset. We consider all of our subjects as potential imposters of every other subject. This realistically reflects an insider threat scenario in which every person enrolled in the system could be a potential attacker. We use two performance metrics: The equal error rate (EER) and the minimal and maximal class distance (d_{min} and d_{max}). The equal error rate is the rate at which the false accept rate (FAR) and false reject rate (FRR) are equal and is a good measure to compare different classifiers. The class distance measures the distance between the template of a user and the most successful out of the 29 imposters and gives insights about the *distribution* of false classifications. We derive the class distance d_c for each user c from the confusion matrix cm as follows:

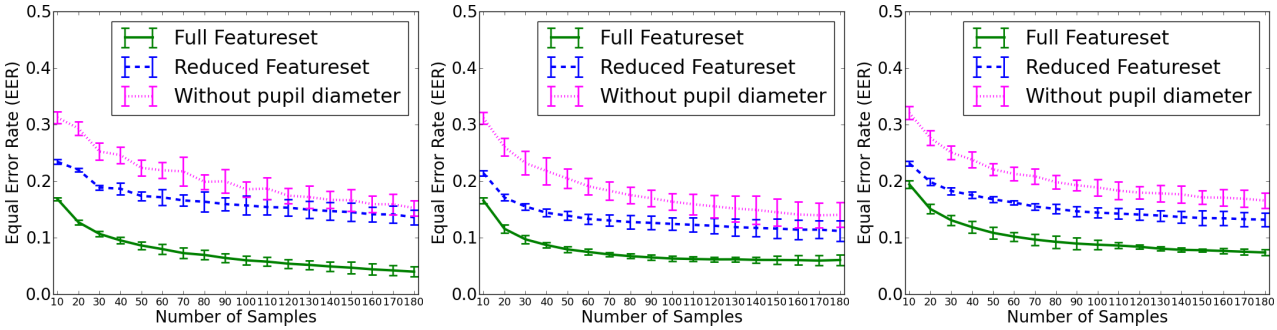
$$d_c = \min_{i \neq c} \frac{cm[c,c]}{cm[c,i]}$$

A class distance lower than 1 means that the best attacker is more likely to be accepted than the legitimate user, a high class distance means that the user is harder to impersonate. As only the *best* out of the 29 imposters is considered this is an extremely conservative metric. The equal error rate is computed using the authentication algorithm described in Section VI-B. As the parameter that controls the trade-off between the false accept rate and the false reject rate (the threshold parameter t) is an integer we report the average of the FAR and the FRR for the value of t for which they are closest.

The results of our analysis are listed in Table II. The SVM outperforms the knn classifier for almost every combination of featureset and dataset. While the training phase is much slower for the SVM the classification decision for a new sample is virtually instantaneous, therefore this does not constitute a serious limitation. When using the full feature set the best performance is achieved with the intra-session dataset. The EER increases from 3.98% to 6.05% when using the inter-session dataset. This transition reflects degradation effects caused by technical artifacts (e.g., different calibration accuracies) across the two sessions. The performance takes another drop to 7.37% when considering data collected over two weeks. Given the behavioral nature of our feature set these changes are to be expected as behavior is usually less stable than physical characteristics. The fact that the EER is very good but the minimal class distance is low suggests that our classifier

			Full			Reduced			Without Pupil Diameter		
Dataset	**Subjects**	**Classifier**	**EER**	d_{min}	d_{max}	**EER**	d_{min}	d_{max}	**EER**	d_{min}	d_{max}
Intra-Session	30	knn	7.07%	0.37	5.71	13.92%	0.18	3.6	19.05%	0.25	1.98
Intra-Session	30	SVM	3.98%	0.52	3.61	13.6%	0.14	4.77	15.25%	0.22	3.44
Inter-Session	20	knn	8.86%	0.81	2.76	10.87%	0.60	2.86	16.58%	0.76	2.86
Inter-Session	20	SVM	6.05%	1.08	4.07	11.17%	0.51	3.00	14.03%	0.54	3.05
2-weeks	20	knn	9.27%	0.4	7.28	13.83%	0.46	5.15	21.32%	0.31	1.92
2-weeks	20	SVM	7.37%	0.49	4.72	13.18%	0.41	4.34	16.56%	0.45	2.46

TABLE II: Classifier Performance on different datasets and different sets of features. The reduced feature set includes the ten features selected by the mRMR algorithm (see Table I). The equal error rate was calculated using Algorithm 1 with 180 samples. d_{min} and d_{max} refer to the maximal and minimal relative difference between any user and the most successful imposter.



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

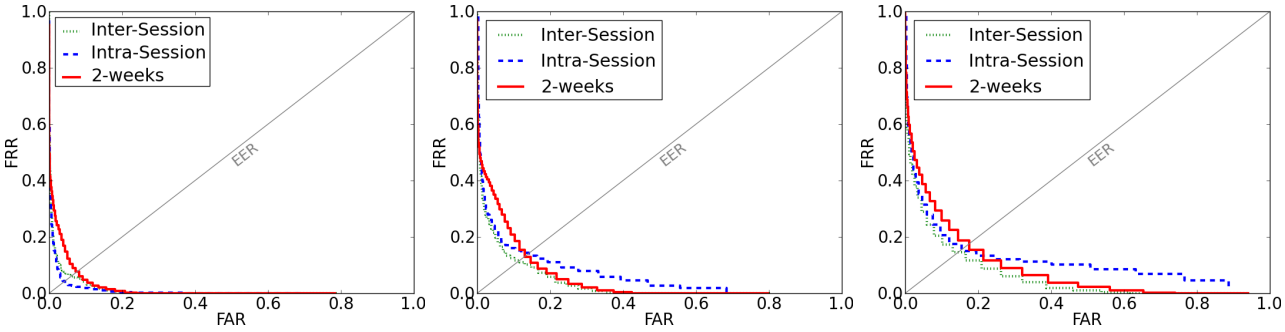
Fig. 10: Average Equal Error Rates obtained through 5-fold stratified cross validation on three different datasets using the SVM classifier. The error bars indicate 95% confidence intervals.

performs extremely well for most users but that the templates of few users are too similar to allow reliable distinction. In order to mitigate this problem it would be possible to determine the closeness of templates directly after the training phase, after obtaining the class distance for each user pair it is then possible to give security guarantees for each user. Users whose templates are not distinctive enough within the target population can then be authenticated with an alternative mechanism.

When using the reduced featureset described in Section IV the error rates increase significantly. The magnitude of this change is surprising, as the features that were removed to form this set exhibit either low distinctiveness or high correlation with other features. Nevertheless, this difference in error rates confirms that even features that carry little information on their own help to correctly classify samples near the decision boundary. This suggests that it won't be sufficient for an attacker to emulate a few distinctive features when using the full feature set. As the complexity of an imitation attack grows rapidly with the number of features that have to be emulated this underlines the resilience of our system against such attacks.

Considering the distinctiveness of the pupil diameter features (see Table I) it is not surprising that removing them from the feature set has a significant impact on our performance metrics. However, the changes of the error rates caused by increasing time distance is less pronounced for this feature set. This suggests that a lot of the degradation observed when using the full feature set was caused by changes in the pupil diameter features.

Using our classifier in conjunction with the algorithm from Section VI-B continuous authentication of users is possible. However, there are still some design decisions to be made. While the EER is a good measure to compare classifiers it is rarely useful in a real-world environment. In an office environment the FRR should usually be extremely low in order to avoid a high number of false positives. The ROC curve in Figure 11 shows that a FRR of 0 is possible when using the full feature set, in order to achieve this a FAR of 19.2% has to be taken into account. While this may seem like a prohibitively high number it is important to remember that our system does not make a one-time decision but authenticates users *continuously*. Conversely, a higher FRR may be acceptable in a high-security context if it ensures the quick detection of an attacker. Another parameter that directly impacts the detection speed is the number of samples used for the authentication decision. Figure 10 shows the effect of this number on the EER. Increasing the number of samples up to 40 rapidly decreases the EER, after that diminishing returns are observed. If the quick detection of an imposter is important the smallest number that still yields acceptable error rates should be chosen. It is noteworthy that our biometric provides a much higher and more constant sampling rate than those relying on active user behavior (e.g., typing or mouse movements). Using our sampling rate of 4Hz even the highest number of 180 samples will correspond



(a) Full feature set (b) Reduced feature set (c) without-pupil feature set

Fig. 11: The ROC curve shows the tradeoff between the false accept rate (FAR) and the false reject rate (FRR) depending on the threshold parameter t for the SVM classifier.

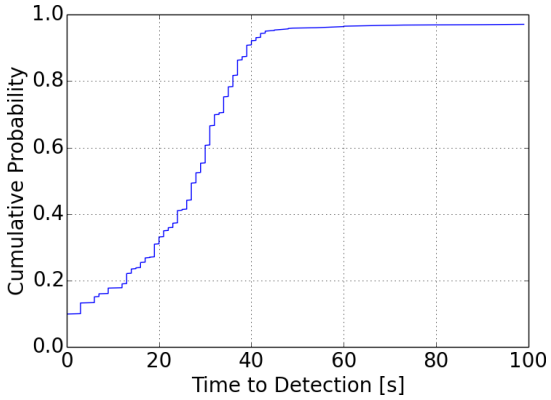


Fig. 12: The ECDF plot shows that 92.2% of all attackers are detected within the first 40 seconds of using the system. The system failed to detect 2.76% of attackers as their biometric templates are very close to that of a legitimate user.

to only 45 seconds. Figure 12 shows that most attackers can be detected even before this 45-second mark, as the number of the attacker's samples in the sliding window gradually increases. While most attackers are detected quickly (92.2% within 40 seconds) the system fails to detect 2.76% of attackers within the scope of our data (i.e., the system exhibits systematic false negatives). These false negatives occur when the biometric templates of two users are very close. This problem could be dealt with by using a second biometric (see Section VII) that is likely to be independent from eye movements. A framework that allows combining several biometrics is described in [29].

These results are very encouraging and significantly outperform related work both in terms of error rates and universality (see Section VII for details). Our solution allows a fine-grained trade-off between classification speed, accuracy, detection time and resistance to imitation attacks. The time stability of our features makes it possible to use old templates for an extended period without having to frequently retrain the classifier (which would require extensive effort). As blindness is the only known condition that prevents reliable eye tracking this makes our biometric an excellent step towards universal and transparent continuous authentication.

VII. Related Work

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. The authors of [30] 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 can not 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 [31], [32]. 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 [33] or more complex 3D video models [34].

Behavioral biometrics are typically less susceptible to these kinds of replication attacks, but their performance with regard to false accept rates (FAR) and false reject rates (FRR) 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 has been proposed in 1980 and exploits distinctive keystroke patterns [35]. 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 into the typing process. Additionally templates are usually tied to keyboard layouts and even physical devices. As the identifying features are conceptually simple this type of identification can be imitated. The authors of [36] designed a software that facilitates imitation attacks by providing positive and negative

feedback depending on the difference between the attacker's and the user's patterns. Two recent comprehensive surveys of keystroke dynamics can be found in [37], [38].

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 [39]. The best accuracy has been reported with a FAR of 0.36% and a FRR of 0% [40]. 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 [41].

Given the increasing share of smartphones and tablets keyboard and mouse are no longer used ubiquitously. A recent study reported an equal error rate of 2-3% when identifying subjects across sessions based on their stroke patterns on a smartphone touchscreen [25]. A similar approach that also tests the resistance to imitation attacks is described in [42]. 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. The authors of [43] 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. However, the number of samples collected is extremely low. It is unclear whether the accuracy stays at these levels when subjects are monitored continuously. Similar work that uses bioimpedance as a biometric reports a recognition rate of 90%, but requires augmentation with hand geometry [44]. Furthermore, the scope of the study was limited to a family-size study with up to 5 subjects.

Eye movements have previously been studied as an input channel that is resistant to shouldersurfing attacks. These systems still rely on a conventional PIN, a password or a passphrase. The authors of [45] 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 shouldersurfing attacks to 55% with an average input time of 5.3 seconds [46].

Our work is perhaps most closely related to [47]. The authors use a Tobii X120 gazetracker with a sampling rate of 120Hz to capture a subject's eye movements while he is watching a movie and use short-term eye gaze direction to construct feature vectors which are modeled using Gaussian mixtures. Depending on the amount of training data an equal error rate 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. attempts to distinguish individuals by the way they look at different images [22]. However, their approach is not suitable for task-independent identification and they do not state to what degree these patterns change over time, especially given the static nature of the pictures. Using density and duration of fixations as their main features they report an EER of 27.06%. Liang et al. measure the eye's tracking behaviour when a moving stimulus is displayed [23]. They use the acceleration of eye movements while the subjects are pursuing a moving shape as input to both Support Vector Machines (SVM) and a Back-Propagation neural network. In an experiment with five subjects they achieve an identification accuracy of 82%. However, their design requires the display of specific stimuli and can not be adapted to general tasks or continuous authentication. Furthermore they do not evaluate the time stability of the user templates.

VIII. Conclusion

In this work we have contributed a set of 21 discriminative features based on a person's eye movement patterns. The usefulness of these features is not limited to our design, they can be used with a wide set of general tasks like web browsing or writing e-mails. 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. The data shows that eye movements, specifically the features we have presented, provide a rich source of distinguishing information. Using data from a single session we achieve an equal error rate of 3.98%. While the effects of task familiarity are measurable, it does not allow the attacker to circumvent our authentication system. In order to test the time stability of our features we performed two repetitions of the experiments, two weeks apart. Our results indicate that users can be authenticated reliably over the entire period. The universal nature of eye movements and the low error rates make this biometric an excellent primitive, on which to build other continuous authentication mechanisms.

Acknowledgement

We would like to thank our shepherd Gianluca Stringhini and the anonymous reviewers for their invaluable feedback. This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/M50659X/1].

References

[1] I. G. Group, "Eye tracking and gaze interaction," www.gazegroup.org.

[2] A. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2007, vol. 373.

[3] B. Cassin, M. L. Rubin, and S. Solomon, *Dictionary of eye terminology*. Wiley Online Library, 1984.

[4] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and T. A. Dyar, "Microsaccades counteract visual fading during fixation," *Neuron*, vol. 49, no. 2, pp. 297–305, 2006.

[5] R. Abadi and E. Gowen, "Characteristics of saccadic intrusions," *Vision research*, vol. 44, no. 23, pp. 2675–2690, 2004.

[6] A. Jones, R. Friedland, B. Koss, L. Stark, and B. Thompkins-Ober, "Saccadic intrusions in alzheimer-type dementia," *Journal of neurology*, vol. 229, no. 3, pp. 189–194, 1983.

[7] B. A. Clementz, J. A. Sweeney, M. Hirt, and G. Haas, "Pursuit gain and saccadic intrusions in first-degree relatives of probands with schizophrenia." *Journal of abnormal psychology*, vol. 99, no. 4, p. 327, 1990.

[8] K. Rayner, C. M. Rotello, A. J. Stewart, J. Keir, and S. A. Duffy, "Integrating text and pictorial information: eye movements when looking at print advertisements." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 219, 2001.

[9] M. Wedel and R. Pieters, "Eye fixations on advertisements and memory for brands: A model and findings," *Marketing science*, vol. 19, no. 4, pp. 297–312, 2000.

[10] R. J. Jacob, "Eye tracking in advanced interface design," *Virtual environments and advanced interface design*, pp. 258–288, 1995.

[11] W. L. Ottati, J. C. Hickox, and J. Richter, "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, no. 1. SAGE Publications, 1999, pp. 66–70.

[12] D. Tock and I. Craw, "Tracking and measuring drivers' eyes," *Image and Vision Computing*, vol. 14, no. 8, pp. 541–547, 1996.

[13] T. Ito, S. Mita, K. Kozuka, T. Nakano, and S. Yamamoto, "Driver blink measurement by the motion picture processing and its application to drowsiness detection," in *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*. IEEE, 2002, pp. 168–173.

[14] M. S. Devi and P. R. Bajaj, "Driver fatigue detection based on eye tracking," in *Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on*. IEEE, 2008, pp. 649–652.

[15] C. MacLachlan and H. C. Howland, "Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years," *Ophthalmic and Physiological Optics*, vol. 22, no. 3, pp. 175–182, 2002.

[16] D. Kahneman and J. Beatty, "Pupil diameter and load on memory." *Science*, 1966.

[17] S. Taptagaporn and S. Saito, "How display polarity and lighting conditions affect the pupil size of vdt operators," *Ergonomics*, vol. 33, no. 2, pp. 201–208, 1990.

[18] D. R. Jasinski, J. S. Pevnick, and J. D. Griffith, "Human pharmacology and abuse potential of the analgesic buprenorphine: a potential agent for treating narcotic addiction," *Archives of General Psychiatry*, vol. 35, no. 4, p. 501, 1978.

[19] (2011) Cybersecurity watch survey. [Online]. Available: http://resources.sei.cmu.edu/asset_files/Presentation/2011_017_001_54029.pdf

[20] Michelle and E. Kowalski, "Insider Threat Study: Computer System Sabotage in Critical Infrastructure Sectors," May 2005. [Online]. Available: http://www.cert.org/archive/pdf/insidercross051105.pdf

[21] M. Kandias, A. Mylonas, N. Virvilis, M. Theoharidou, and D. Gritzalis, "An insider threat prediction model," in *Trust, Privacy and Security in Digital Business*. Springer, 2010, pp. 26–37.

[22] V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio, "Gant: Gaze analysis technique for human identification," *Pattern Recognition*, 2014.

[23] Z. Liang, F. Tan, and Z. Chi, "Video-based biometric identification using eye tracking technique," in *Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 728–733.

[24] Z. M. Hafed and J. J. Clark, "Microsaccades as an overt measure of covert attention shifts," *Vision research*, vol. 42, no. 22, pp. 2533–2545, 2002.

[25] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Transactions on Information Forensics and Security*, 2012.

[26] J. Dougherty, R. Kohavi, M. Sahami *et al.*, "Supervised and unsupervised discretization of continuous features," in *ICML*, 1995, pp. 194–202.

[27] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[28] K. Holmqvist, M. Nyström, and F. Mulvey, "Eye tracker data quality: what it is and how to measure it," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 45–52.

[29] H. Crawford, K. Renaud, and T. Storer, "A framework for continuous, transparent mobile device authentication," *Computers & Security*, vol. 39, pp. 127–136, 2013.

[30] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *Information Forensics and Security, IEEE Transactions on*, vol. 1, no. 2, pp. 125–143, 2006.

[31] C. Barral and A. Tria, "Fake fingers in fingerprint recognition: Glycerin supersedes gelatin," in *Formal to Practical Security*. Springer, 2009, pp. 57–69.

[32] F. Rieger. (2013) Chaos computer club breaks apple touchid. [Online]. Available: http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid/

[33] N. M. Duc and B. Q. Minh, "Your face is not your password face authentication bypassing lenovo–asus–toshiba," *Black Hat Briefings*, 2009.

[34] A. Boehm, D. Chen, M. Frank, L. Huang, C. Kuo, T. Lolic, I. Martinovic, and D. Song, "Safe: Secure authentication with face and eyes," in *IEEE PRISMS 2013*, June 2013.

[35] R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro, "Authentication by keystroke timing: Some preliminary results," DTIC Document, Tech. Rep., 1980.

[36] C. M. Tey, P. Gupta, and D. GAO, "I can be you: Questioning the use of keystroke dynamics as biometrics." The 20th Annual Network & Distributed System Security Symposium (NDSS 2013), 2013.

[37] D. Shanmugapriya and G. Padmavathi, "A survey of biometric keystroke dynamics: Approaches, security and challenges," *arXiv preprint arXiv:0910.0817*, 2009.

[38] S. P. Banerjee and D. L. Woodard, "Biometric authentication and identification using keystroke dynamics: A survey," *Journal of Pattern Recognition Research*, vol. 7, no. 1, pp. 116–139, 2012.

[39] K. Revett, H. Jahankhani, S. T. de Magalhães, and H. M. Santos, "A survey of user authentication based on mouse dynamics," in *Global E-Security*. Springer, 2008, pp. 210–219.

[40] Y. Nakkabi, I. Traoré, and A. A. E. Ahmed, "Improving mouse dynamics biometric performance using variance reduction via extractors with separate features," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 40, no. 6, pp. 1345–1353, 2010.

[41] Z. Jorgensen and T. Yu, "On mouse dynamics as a behavioral biometric for authentication," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*. ACM, 2011, pp. 476–482.

[42] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on smartphones via tapping behaviors," Tech. Rep. WM-CS-2012-06, Tech. Rep., 2012.

[43] K. B. Rasmussen, M. Roeschlin, I. Martinovic, and G. Tsudik, "Authentication using pulse-response biometrics," in *Proceedings of the 21st Network and Distributed System Security Symposium (NDSS 2014)*, 2014.

[44] C. Cornelius, J. Sorber, R. Peterson, J. Skinner, R. Halter, and D. Kotz, "Who wears me? bioimpedance as a passive biometric," in *Proc. 3rd USENIX Workshop on Health Security and Privacy*, 2012.

[45] M. Kumar, T. Garfinkel, D. Boneh, and T. Winograd, "Reducing shoulder-surfing by using gaze-based password entry," in *Proceedings of the 3rd symposium on Usable privacy and security*. ACM, 2007, pp. 13–19.

[46] A. De Luca, M. Denzel, and H. Hussmann, "Look into my eyes!: Can you guess my password?" in *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 2009, p. 7.

[47] T. Kinnunen, F. Sedlak, and R. Bednarik, "Towards task-independent person authentication using eye movement signals," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 187–190.