Biometric Identification and Presentation-Attack Detection using Micro- and Macro-Movements of the Eyes

Silvia Makowski, Lena A. Jäger, Paul Prasse, Tobias Scheffer
Department of Computer Science, University of Potsdam
August-Bebel-Str. 89, 14482 Potsdam, Germany
{silvia.makowski, lena.jaeger, paul.prasse, tobias.scheffer}@uni-potsdam.de

Abstract

We study involuntary micro-movements of both eyes, in addition to saccadic macro-movements, as biometric characteristic. We develop a deep convolutional neural network that processes binocular oculomotoric signals and identifies the viewer. In order to be able to detect presentation attacks, we develop a model in which the movements are a response to a controlled stimulus. The model detects replay attacks by processing both the controlled but randomized stimulus and the ocular response to this stimulus. We acquire eye movement data from 150 participants, with 4 sessions per participant. We observe that the model detects replay attacks reliably; compared to prior work, the model attains substantially lower error rates.

1. Introduction

No single biometric characteristic that is known today is by itself sufficiently reliable for all biometric applications, unique, collectible, convenient, and universally available. For instance, while identification based on fingerprint and iris tend to be more accurate than facial identification, a good-quality fingerprint cannot be obtained for approximately 2-4% of the population due to degradation of the fingerprints from manual labor or hand-related disabilities, while long eyelashes, small eye apertures, cosmetic contact lenses, and conditions including glaucoma and cataract prevent the collection of good-quality images of the iris for an estimated 7% of the population [20]. It is therefore desirable to expand the space of biometric characteristics that can be used by themselves or as part of multimodal biometric identification systems. National population registers can serve as an illustrating example of an application in which multiple modalities are necessary for a biometric system to meet required false-acceptance, false-rejection, and failure-to-enroll rates across a large and diverse population.

At the same time, no universally reliable method for detection of presentation attacks exists, due to both the adversarial nature of the problem and the unbounded space of possible presentation attack instruments. Especially artefact-detection approaches are vulnerable to the development of new and unforseen presentation attack instruments. Challenge-response approaches can determine whether a presentation exhibits liveness properties. However, if the response requires a voluntary user action, the detection of presentation attacks is in conflict with a convenient user experience. If, in addition, the expected response can be derived easily from the challenge, the presentation attack can incorporate an automated or manually controlled response to an observed challenge. As an example application that calls for high resilience against presentation attacks with unforseen attack instruments, consider physical access control to high-security facilities.

It has long been known that the way we move our eyes in response to a given stimulus is highly individual [37] and more recent psychological research has shown that these individual characteristics are reliable over time [2]. Hence, it has been proposed to use eye movements as a behavioral biometric characteristic [24, 3].

Human eye movements alternate between *fixations* of around 250 ms during which the eye gaze is maintained on a location from which visual input is obtained and *saccades* of around 50 ms which are fast relocation movements that can reach up to $500^\circ/s$ and during which visual input is supressed. Moreover, three types of involuntary micro-movements always occur during attempted fixations which, among other functions, prevent visual fading of the fixated image. *Drift* movements are very slow movements of around 0.1-0.4$^\circ/s$ away from the center of a fixation which are superimposed by high-frequency, low-amplitude tremor of around 40-100 Hz whose velocity can reach up to $0.3^\circ/s$. *Microsaccades* are occasional small saccades that can reach velocities of up to $120^\circ/s$ and, among other functions, bring back the eye gaze to the intended center of a fixation after a drift movement has occurred [34, 35, 36, 39, 18].

Prior work on biometric identification using eye move-

ments extracts fixations and saccades from the eye tracking signal and measures the values of engineered explicit features such as fixation durations and saccadic amplitudes or velocities. Any information contained in the fixational micro-movements is discarded. Since saccades and fixations occur at a low frequency, a critical limitation of these approaches is that long eye gaze sequences of more than one minute [33] need to be observed before the system can reliably identify a user. The additional information contained in the high-frequency micro-movements bear the potential of considerably speeding up the identification. Recently, a neural network has been studied that processes a raw monocular eye tracking signal measured during reading [19]. This approach does not rely on any prior detection of specific types of macro- or micro-movements. In order to detect replay attacks, we develop a model in which the eye movements are the ocular response to a challenge in the form of a controlled stimulus. In this setting, however, the identification task becomes more challenging as fixation durations and saccade amplitudes are largely determined by the stimulus, and their distributional properties are less likely to vary across individuals.

This paper presents a number of contributions. We develop a deep convolutional neural network (CNN) that (a) processes binocular eye tracking signals while (b) the eye is responding to stimuli in the form of jumping dots on a screen. In addition to the eye-movement signals, the network processes the relative positions of the stimuli, enabling it to (c) detect replay attacks. Individual characteristics of eye movements correlate stonger within a session than across multiple sessions [2]. Therefore, we (d) experimentally study a setting in which enrollment and application data are collected on different days.

The remainder of this paper is structured as follows. Section 2 reviews prior work. Section 3 lays out the problem setting and Section 4 presents a convolutional network architecture for oculomotoric biometric identification with integrated liveness detection. Section 5 presents the data collection and the evaluation of the proposed method. Section 6 concludes.

2. Prior Work

In their seminal work, Kasprowski and Ober [24] and Bednarik *et al.* [3] presented proofs of concept showing that eye movements can be exploited as a behavioral biometric characteristic. Within the following decade, spurred by a series of competitions [23, 22, 28], the analysis of eye movements as a biometric feature has attracted increasing attention. Most existing methods extract explicit features from the eye gaze signal. The raw eye tracking signal is first preprocessed into fixations and saccades, and subsequently different sets of fixational (e.g., duration) and saccadic features (e.g., amplitude, velocity, acceleration) are derived. Some approaches [3, 10, 38] also use pupil size as input feature, which is standardly recorded by video-based eye trackers, but which has been demonstrated to be vulnerable to spoofing attacks [15]. Methods that operate on a binocular eye tracking signal use the difference between the eye gaze positions as additional feature [45, 5].

The various methods that operate on saccadic and fixational features can be broadly classified into i) aggregational methods, ii) statistical methods and iii) generative models. Aggregational methods classify scanpaths after aggregating the features over the relevant time window [17, 5, 44]. Statistical methods compute the similarity between two eye gaze sequences by performing statistical tests on the distributions of the extracted features [41, 16, 43]. Generative approaches use hidden Markov models [48, 47] or other kinds of graphical models [31, 1, 33] to simulate a user's eye movement behavior.

DeepEyedentification [19] is the first method that does not operate on engineered features but rather computes a latent representation of an eye gaze sequence by training a deep convolutional network on the raw eye tracking signal. This approach has been demonstrated to outperform the best existing aggregational, statistical and generative approaches by one order of magnitude in terms of classification accuracy and by two orders of magnitude in terms of the duration of the eye tracking recording needed to identify a user.

The various approaches further differ with respect to the stimulus type they operate on; a static cross [3], various kinds of images [7] including faces [41, 14, 4], text [17, 1, 33, 19] or dynamical stimuli such as jumping points [21, 25, 42, 5, 45] or video sequences [26] have been used. Whereas most of the methods use the same stimulus for training, enrollment and testing—which makes them vulnerable to replay attacks—Kinnunen *et al.* are the first to identify users on a novel stimulus, and subsequent approaches identify readers on novel text [17, 1, 33, 19].

Prior work on presentation-attack detection in the context of gaze-based identification [27, 28] assumes that an attacker generates artificial eye movements, based on a model of a target individual's gaze patterns. The proposed methods use a classifier to discriminate *bona fide* from generated eye movements using the same engineered features that are used for identification. This approach relies on imperfections of the gaze model and cannot detect an attacker who replays actual eye movements that were recorded from the target individual. Approaches to presentation-attack detection that detect artefacts of specific presentation-attack instruments have been studied widely for other biometrics; for instance, for iris recognition. Work of Raja *et al.* [40] exploits phase information which is indicative of presentations on smartphone or table screens.

The analysis of eye movements has been combined with knowledge-based authentification procedures such as en-

tering a password with the eye gaze [32, 30, 8, 9, 46, 6], or with other behavioral biometrics such as keystroke dynamics [44] or physiological biometric methods such as iris scanning [29].

3. Problem Setting

We study the problems of biometric *identification*, *identity verification*, and *presentation-attack detection*. The system observes a sequence of yaw and pitch gaze angles of the left and right eye which is observed by means of an eye tracker. This sequence is the ocular response to a known visual stimulus which in our study has the form of a sequence of five dots that are positioned randomly on a screen and displayed for between 250 and 1000 ms each.

In the *verification* setting, the model compares the gaze sequence that is observed to one or more *enrollment sequences* of a single user. If the value of a similarity metric, defined between the application sequence and one of the enrollment sequences, exceeds a threshold, then the user is accepted and the user's presumed identity verified; otherwise, the user is rejected and classified as an impostor. The algorithm performance can be characterized by a *false-match rate* (FMR, fraction of impostors among all accepted users) and a *false non-match rate* (FNMR, fraction of falsely rejected users among all rejected users). By changing the decision threshold, one can observe a *detection error tradeoff curve (DET curve)*. The *equal error rate* (EER) is the point on this curve for which FMR equals FNMR.

In the *identification* setting, the gaze sequence that is observed at application time is compared to one or more *enrollment sequences* of *multiple* enrolled users. In case of a positive identification, the outcome is the matched identity; otherwise, the user is classified as *impostor*. The DET curve characterizes the trade-offs between *false-positive identification-error rate* (FPIR) and *false-negative identification-error rate* (FNIR) for enrolled users; here, false positive identifications can be impostors or enrolled users who are mistaken for different enrolled users.

In some approaches, the similarity metric is defined as a metric on a vector of engineered features which are extracted from the gaze sequence [17, 43]. In our approach, we measure the cosine similarity between *neural embeddings* of gaze sequences. This embedding is trained on a separate set of training users which is disjoint from the users that are encountered at application time. The neural network is trained such that the embedding is similar for all gaze sequences of a particular user but different for gaze sequences of distinct users.

The *presentation-attack detection* problem is to detect whether the observed gaze sequence is presented with the goal of interfering with the biometric system. We study the case of a complete artificial replay attack by an adversary who can observe both the size of the display on which the stimulus is presented and the duration for which each stimulus is displayed. The adversary does not, however, have advance information about the randomized positions of the five dots; therefore, they are limited to replaying a gaze sequence for a random stimulus with the same display size and display duration. We measure the DET curve between the *attack-presentation classification-error rate* (APCER)—the proportion of attack presentations incorrectly classified as *bona fide* presentations—and the *bona-fide presentation-classification error rate* (BPCER)—the proportion of *bona fide* presentations that are misclassified as attack.

As an example presentation-attack instrument for this type of attack, an attacker may record eye movements of the target person unnoticed by means of a remote eye tracker. The attacker may then create a 3D printed replica of the target's eye whose orientation is controlled by servomotors and programmed to replay the recorded gaze pattern. To prevent presentation-attack detection by detection of artefacts in the camera image, the presentation may include an artificial facial mask and hair.

Note that presentation attacks by lifeless humans are not possible due to the lack of eye movements, and that humans cannot be altered to exhibit another person's patterns of ocular micromovements. In a nonconformant presentation, the gaze patterns would be absent while a conformant *zero-effort* presentation attack by a human impostor would require a false match or false-positive identification, respectively, to be successful.

4. System and Network Architecture

This section derives the *DeepEyedentificationLive* system and the neural network that performs binocular oculomotoric biometric identification and liveness detection. An eye scanner records the user's eye gaze while a display (see Figure 1) shows a sequence of dots at random locations. The gaze sequence of absolute yaw x and pitch gaze angles y of the left l and right eye r recorded with sampling frequency ρ in Hz is transformed into sequences of yaw δ_i^x and pitch δ_i^y gaze velocities in $^\circ/s$ where $\delta_i^x = \frac{\rho}{2}(x_{i+1} - x_{i-1})$ and $\delta_i^y = \frac{\rho}{2}(y_{i+1} - y_{i-1})$. These four velocity sequences constitute four of the input channels into the network: $\langle \delta_1^{x,l}, \ldots, \delta_n^{x,l} \rangle$ is the sequence of yaw angular velocities of the left eye; $\langle \delta_1^{y,l}, \ldots, \delta_n^{y,l} \rangle$ is the sequence of pitch angular velocities; $\langle \delta_1^{x,r}, \ldots, \delta_n^{x,r} \rangle$ and $\langle \delta_1^{y,r}, \ldots, \delta_n^{y,r} \rangle$ are the corresponding yaw and pitch velocities of the right eye.

Since the velocity of saccadic and fixational eye movements occur at vastly different scales, global normalization would squash the slow fixational drift and tremor to near-zero and as a consequence much of the information in the eye tracking signal would be lost. The solution to this challenge is a model architecture with two independent convolutional subnets which observe the same input: a *fast subnet*

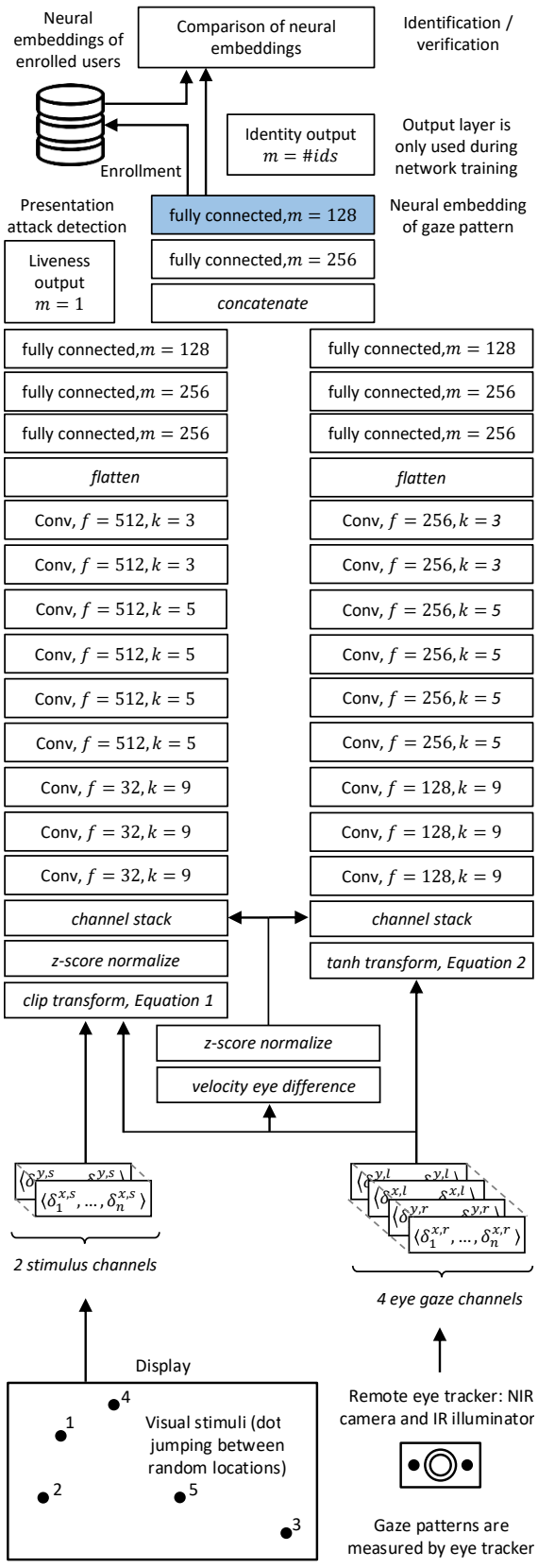


Figure 1. *DeepEyedentificationLive* architecture.

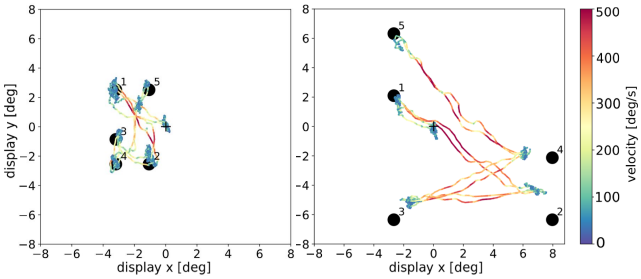


Figure 2. Exemplary eye traces for two different trial configurations: 500 ms stimulus display duration and small grid (*left*) and 250 ms display duration on big grid (*right*). The cross is displayed before the onset of the trial.

is designed to process the fast (micro-)saccadic eye movements and a *slow subnet* to processs the slow fixational eye movements (drift, tremor). The two subnets have the same type of layers except for a transformation layer that transforms the input to resolve the fast and the slow movements, respectively. For the fast subnet, absolute velocities below a minimal velocity ν_{min} are truncated and z-score normalization is applied (see Equation 1).

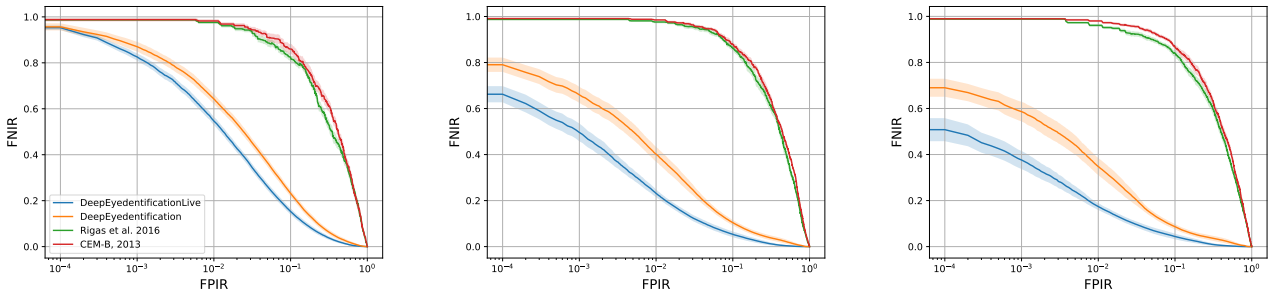
$$t_f(\delta_i^x, \delta_i^y) = \begin{cases} z(0) & \text{if } \sqrt{\delta_i^{x2} + \delta_i^{y2}} < \nu_{min} \\ (z(\delta_i^x), z(\delta_i^y)) & \text{otherwise} \end{cases} \quad (1)$$

For the slow subnet, a sigmoidal function is applied such that the slow fixational movements (drift and tremor) are stretched within the interval between -0.5 and $+0.5$ whereas the fast microsaccades and saccades are squashed to values between -0.5 and -1 or $+0.5$ and $+1$ (see Equation 2). Based on the psychological literature [18] and prior tuning, the threshold ν_{min} of Equation 1 is set to $40°/s$ and the scaling factor c of Equation 2 to 0.02.

$$t_s(\delta_i^x, \delta_i^y) = (\tanh(c\delta_i^x), \tanh(c\delta_i^y)) \quad (2)$$

The original input velocities are also fed into a subtraction layer that computes the yaw $\langle \delta_1^{x,r} - \delta_1^{x,l}, \ldots, \delta_n^{x,r} - \delta_n^{x,l} \rangle$ and pitch velocity differences between the two eyes $\langle \delta_1^{y,r} - \delta_1^{y,l}, \ldots, \delta_n^{y,r} - \delta_n^{y,l} \rangle$. These two channels are then stacked with each of the outputs of the transformation layers.

The network additionally processes the positions of the visual stimuli to which the gaze sequence is the oculomotoric response. In our experiments, dots are displayed at five random positions in each trial, each dot is displayed for a time of Δ. The stimuli are represented as offsets in x and y direction to the previous stimulus: $\langle \delta_1^{x,s}, \ldots, \delta_n^{y,s} \rangle$ and $\langle \delta_1^{y,s}, \ldots, \delta_n^{y,s} \rangle$, where each $\delta_i^s \Delta$ is the offset in degrees between the stimulus displayed at time i and the previously displayed stimulus. Note that when a stimulus is displayed from time t to time t' and the user's eye gaze moves from the previous stimulus to exactly the new stimulus within this



(a) 1 sec (b) 5 secs (c) 10 secs

Figure 3. False-Negative Identification-Error Rate (FNIR) over False-Positive Identification-Error Rate (FPIR) in the idenification setting. Colored bands show the standard error.

Table 1. Metrics for identification. Values marked "*" are significantly better ($p < 0.05$) than the next-best value.

	DeepEyedentificationLive	**DeepEyedentification**	**Rigas et al. 2016**	**CEM-B, 2013**
EER @ 1 s	**0.1246 ± 0.0204***	0.1549 ± 0.0219	0.4314 ± 0.044	0.4577 ± 0.048
EER @ 5 s	**0.072 ± 0.0242***	0.1024 ± 0.0223	0.4522 ± 0.0355	0.4706 ± 0.0248
EER @ 10 s	**0.0638 ± 0.0242***	0.0917 ± 0.0225	0.4456 ± 0.0301	0.4695 ± 0.0342
EER @ 60 s	**0.0554 ± 0.0268***	0.0774 ± 0.0253	0.479 ± 0.0903	0.4758 ± 0.0649
FNIR @ FPIR 10^{-2}**@1** s	**0.5471 ± 0.0566***	0.6433 ± 0.0684	0.9811 ± 0.0193	0.9922 ± 0.0123
FNIR @ FPIR 10^{-2}**@5** s	**0.2316 ± 0.054***	0.4019 ± 0.1017	0.9851 ± 0.0119	0.9879 ± 0.0088
FNIR @ FPIR 10^{-2}**@10** s	**0.1742 ± 0.0477***	0.3477 ± 0.1182	0.9717 ± 0.0162	0.9826 ± 0.016
FNIR @ FPIR 10^{-2}**@60** s	**0.1186 ± 0.0562**	0.3112 ± 0.1495	0.9709 ± 0.0272	0.9829 ± 0.0181

interval, then, both for the left and right eye, it holds that

$$\sum_{i=t}^{t'} \delta_i^x - \sum_{i=t}^{t'} \delta_i^{x,s} = \sum_{i=t}^{t'} \delta_i^y - \sum_{i=t}^{t'} \delta_i^{y,s} = 0. \qquad (3)$$

The network processes the input in one-dimensional convolutions over time. This is in analogy to other CNN architectures that process time-sequential data—for instance, speech recognition—and in contrast to image-processing CNNs that use two-dimensional convolutions. The six input sequences, x and y are processed as as six channels, in analogy to the red, green and blue channels by image processing CNNs. Convolutional neural networks require an input sequence of fixed width. Therefore, the network processes the gaze sequences in windows of 1,000 time steps, which corresponds to one second of input data. All input sequences are therefore split into subsequences of 1,000 ms, and the results—the similarity metric for identification and verification, and the activation of the output unit for liveness detection—are averaged across all subsequences of each gaze sequence.

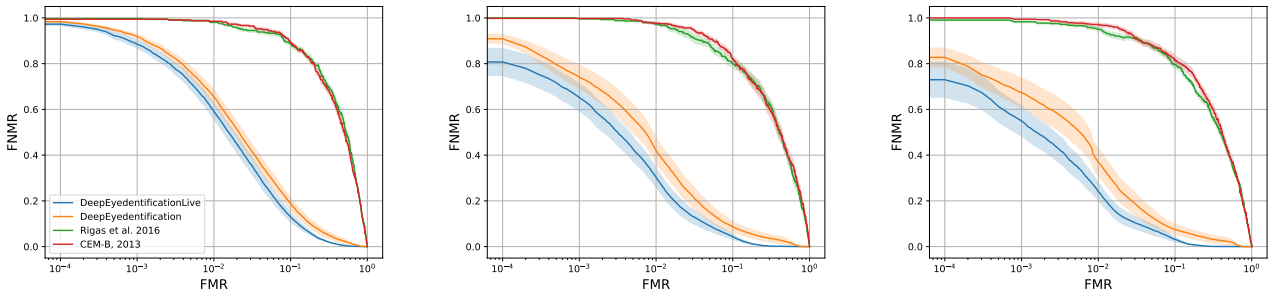
Parameter f in Figure 1 shows the number of filters, k specifies the kernel size of convolutions. Parameter m characterizes the number of units of fully connected layers. Convolutional and fully connected layers are all batch normalized and followed by a ReLu activation function. Each convolutional layer is followed by an average pooling layer with pooling size 2 and stride of 1. The network has a sigmoid layer for the liveness ouput.

For the purpose of training the network, a softmax output layer with one unit for each training user is added. The network is then trained on gaze sequences of training users. For half of the training sequences, the correct stimulus is presented to the network as input and the target liveliness output is +1. In the remaining cases, a random stimulus with the same display size and display duration is chosen and the target liveness output value is -1. After training, the identification softmax layer is discarded and the embedding layer provides the neural feature embedding of each gaze sequence. Because the network has learned to identify the training users based on the activations of the embedding units, the embedding distills signals that vary across individuals and are indicative of the viewer's identity.

The similarity metric between enrollment and application sequence is the given by the cosine similarity, averaged over all input windows of 1,000 ms. The similarity value between an application sequence and a user is the maximum similarity over that user's enrollment sequences. For the process of enrollment, the neural embedding of one or several gaze sequences are determined and stored in a database.

5. Experiments

This section reports on data collection and our comparison of DeepEyedentificationLive and reference methods.



(a) 1 sec (b) 5 secs (c) 10 secs

Figure 4. False Non Match Rate (FNMR) over False Match Rate (FMR) in the verification setting. Colored bands show the standard error.

Table 2. Metrics for verification setting. Values marked "*" are significantly better ($p < 0.05$) than the next-best value.

	DeepEyedentificationLive	**DeepEyedentification**	**Rigas et al. 2016**	**CEM-B, 2013**
EER @ 1 s	**0.1087 ± 0.0301**	0.1373 ± 0.0432	0.5187 ± 0.0482	0.4996 ± 0.0398
EER @ 5 s	**0.0599 ± 0.0322**	0.0837 ± 0.0444	0.4527 ± 0.0591	0.4556 ± 0.065
EER @ 10 s	**0.0517 ± 0.0324**	0.0694 ± 0.0446	0.4535 ± 0.0292	0.4642 ± 0.0522
EER @ 60 s	**0.0384 ± 0.0334**	0.0587 ± 0.0505	0.3094 ± 0.077	0.319 ± 0.0528
FNMR @ FMR 10^{-2}**@1** s	**0.5928 ± 0.1246***	0.6555 ± 0.1118	0.994 ± 0.0119	0.9942 ± 0.0116
FNMR @ FMR 10^{-2}**@5** s	**0.3035 ± 0.1471***	0.4206 ± 0.1891	0.9818 ± 0.0245	0.9907 ± 0.0114
FNMR @ FMR 10^{-2}**@10** s	**0.2406 ± 0.1374***	0.3701 ± 0.2017	0.9585 ± 0.0308	0.9789 ± 0.0278
FNMR @ FMR 10^{-2}**@60** s	**0.2774 ± 0.2705***	0.3018 ± 0.2268	0.9226 ± 0.0757	0.9338 ± 0.0627

5.1. Data Collection

We collect the *JuDo1000 data set*[1] of binocular eye movement data (horizontal and vertical gaze coordinates) from 150 participants (18 to 46 years old, mean age 24 years), each of whom participate in four experimental sessions with a lag of at least one week between any two sessions. Eye movements are recorded using an Eyelink Portable Duo eye tracker (tripod mounted camera) at a sampling frequency of 1,000 Hz. Participants are seated in front of a 38×30 cm (1280×1024 px) computer monitor at a height adjustable table with their head stabilized by a chin- and forehead rest. In each session, participants are presented with a total of 108 experimental trials in which a black dot with a diameter of 0.59 cm (20 px) appears consecutively at 5 random positions on a white background.

The duration for which each dot is displayed is varied between 250, 500 and 1000 ms with a fixed value within each trial; the size of the screen area in which the dots appear is varied between 7.6×14.0 cm, 11.4×17.0 cm, and 19.0×23.0 cm around the center of the monitor with a fixed area within each trial. The combination of display duration and areas results in nine *trial configurations*. Figure 2 shows example eye-movement traces of the left and right eye for different display duration and areas.

[1]The *JuDo1000 data set* is accessible at https://osf.io/5zpvk/.

Table 3. Parameter space used for random grid search: kernel size k and number of filters f of all convolutional layers and number of units m of all fully connected layers.

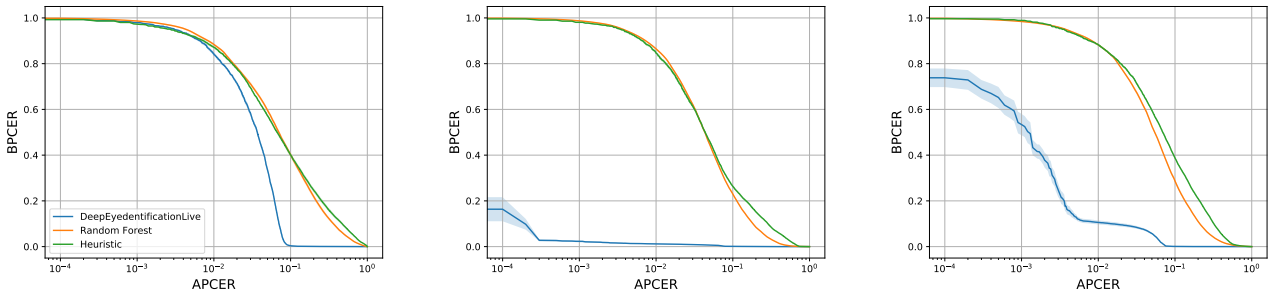
Parameter	Search space
k	$\{3, 5, 7, 9\}$
f	$\{32, 64, 128, 256, 512\}$
m	$\{64, 128, 256, 512\}$

5.2. Hyperparameter Tuning

We tune the hyperparameters with a random grid search in the parameter search space shown in Table 3. As validation data we use one hold-out trial from each configuration per session, which is removed from the training and testing data of the final model. We constrain kernel sizes and number of filters of the convolutional layers to be identical within layers 1-3, 4-7 and 8-9 of both subnets. Kernel sizes are furthermore constrained to be smaller or equal and filter sizes greater or equal to the previous layer block. Figure 1 shows the best parameter configuration.

5.3. Identification and Identity Verification

As reference methods for identification and identity verification, we use the DeepEyedentification network [19]—which differs from DeepEyedentificationLive in that it can only process monocular data and lacks presentation-attack detection—and the statistical models of Rigas *et al.* [43] and CEM-B [16] which have been observed to outperform



(a) 1.25s (b) 2.5s (c) 5s

Figure 5. Presentation-attack detection with one trial as input at test time for different display durations (250ms, 500ms, 1000ms) of the five dots. Colored bands show the standard error. Time in seconds denotes resulting trial length.

Table 4. Presentation-attack detection with one trial as input at test time. Table shows EER for different display durations (250ms, 500ms, 1000ms) of the five dots. Time in seconds denotes resulting trial length.

	DeepEye.Live	**Random Forest**	**Heuristic**
EER	**0.0407 ± 0.0039**	0.1713 ± 0.005	0.2024 ± 0.0086
EER @ $1.25s$	**0.0732 ± 0.0028**	0.2075 ± 0.0135	0.2248 ± 0.0162
EER @ $2.5s$	**0.0106 ± 0.0039**	0.1455 ± 0.0051	0.1749 ± 0.0085
EER @ $5.0s$	**0.0513 ± 0.0039**	0.1596 ± 0.0058	0.2046 ± 0.0062

aggregational approaches. We do not compare any generative models specifically designed for eye movement analysis during reading [31, 1, 33] or free viewing during a visual task [48, 47]. For the DeepEyedentification network we apply the same hyperparameter tuning as for DeepEyedentificationLive. The statistical reference methods do not have any hyperparameters. As a fusion metric, we use the simple mean metrics which Rigas *et al.* [43] use in their main experiments for both their own model and its predecessor CEM-B. For these methods, we preprocess the data using a velocity-based saccade detection alogorithm [12, 11, 13].

The verification (identification) of an enrolled user counts as true match (true-positive identification) if the cosine similarity between any input window from the test sequence and any window from an enrollment sequences exceeds the recognition threshold; otherwise, it counts as a false non-match (false-negative identification). A false match (false-positive identification) occurs when the cosine similarity of a test window and an enrollment window of an enrolled user and an impostor exceeds the threshold; otherwise, a true non-match (true-negative identification) occurs.

To evaluate both settings, we resample 10 times from the data set—in each iteration, we randomly select a training population of 125 users to train an embedding. From the 25 remaining users, we select 20 users as enrolled users and 5 users as impostors for the identification setting. In the verification setting, we select one enrolled user and 24 impostors. In all settings, enrollment and test data are also split across the four recording sessions to avoid session bias. For each enrolled user, we select 9 trials with unique trial configurations, drawn from 3 training sessions as enrollment. At application time, we use recordings from a different test session to calculate test embeddings. The same evaluation procedure is applied for the reference methods.

Figure 3 and Table 1 shows the identification performance of DeepEyedetificationLive and the baseline models for 20 enrolled users, averaged across all trial configurations. Here, the EER decreases from 0.1246 for one second of input data at test time to 0.0554 after 60 seconds. DeepEyedentificationLive obtains a lower FNIR than DeepEyedentification for every FPIR and every trial duration; differences are significant based on a paired t-test with $p < 0.05$. The performance gap between DeepIdentificationLive on one hand and CEM-B and Rigas *et al.* on the other hand is dramatic. It should be noted that both baselines use distributional properties of fixation durations and saccadic properties which here are largely dominated by the controlled stimuli. In the verification setting, shown averaged across all trial configurations in Figure 4 and Table 2, there is only one identity each imposter can be confused with. Here, the EER ranges from 0.1087 for one second to 0.0384 for 60 seconds of input at test time. Again, the performance gap to Cem-B and the method of Rigas *et al.* is dramatic.

5.4. Presentation-Attack Detection

We simulate an attacker who has observed the number of stimuli, display duration, and the size of the display area, and who is able to record and replay, without detectable imperfections, a gaze sequence of the target individual for this exact configuration. A possible instrument for this type of attack is a pair of 3D-printed replicas of a target person's eyes that are controlled by servomotors and programmed to execute a gaze sequence that has been recorded from the target person by a remote eye tracker. We generate examples of attacks by pairing a test gaze sequence with a stimulus se-

quence for which the positions have been randomly drawn with the same display duration and area. For each *bone fide* presentation in the data, we create one attack presentation.

We compare DeepEyedentificationLive against a simple *heuristic* and a *Random Forest* model with engineered features. The heuristic is based on Equation 3; it measures, how well the fixation sequence matches the sequence of stimuli by computing the average, over four stimulus relocations, of the differences between the aggregate gaze movements during presentation of the stimulus and the offset between current and last stimulus. Based on a threshold, a pair of fixation sequence and stimuli is classified as *attack* or *bona fide*. The *Random Forest* baseline uses five the four differences between aggregate eye movements and stimulus offsets during presentation of one of the stimuli and the average of these values as features.

Prior work on presentation-attack detection for gaze-based identification assumes that the presentation is generated with imperfections in the distributional features of the user's fixation durations and amplitude and velocity features of saccades [27, 28]. In our setting, the attacker replays a recorded gaze sequence of the target user. Therefore, this known approach cannot distinguish these replay attacks from *bona fide* presentations and we do not include it in the experimental comparison. Reference methods that detect specific artefacts that are indicative of a particular presentation-attack instrument are generally ineffective against different attack instruments. For instance, prior work that exploits phase information which is indicative of smartphone screens [40], or other methods that detect video presentations by detecting mobile devices cannot detect a presentation attack using 3D-printed eyeball replicas. We therefore do not believe that including these methods in our experimental comparison would provide new insights.

We evaluate by random resampling from the data set. We split the data across users, with 125 users used for training and 25 users for testing. At test time, a decision is made based on input data of one trial, whose length depends on the display duration of the experimental configuration. To investigate the influence of display durations on presentation-attack detection performance, we also evaluate the models on three subsets of the data (using only trials with 250 ms, 500 ms or 1000 ms display duration in training and test respectively). As Figure 5 and Table 4 show, we attain the lowest EER of 0.0106, when only using data from trials with 500 ms stimulus display duration and an EER of 0.0407 when using all experimental configurations. The performance gap between DeepEyedentificationLive and both baseline methods is dramatic.

6. Conclusion

We have developed DeepEyedentificationLive, a convolutional network for oculomotoric biometric identification that processes both the controlled stimulus and the binocular response in the form of sequences of gaze velocities. The model determines an embedding of gaze sequences and simultaneously performs presentation-attack detection.

Our model of an attacker who is informed about the display duration and area and can replay gaze sequences of the target individual without imperfection is arguably the most challenging attacker model. We conclude that five stimuli displayed for 500 ms each are the best configuration for presentation-attack detection under investigation. Presentation-attack detection can be integrated without any additional sensory cost or secondary tasks. The ocular response to the visual stimuli does not require the users' attention.

Most existing methods [16, 43] are based on the analysis of a scanpath during free viewing. During a liveness detection challenge, the trajectory of the scanpath as well as the fixation durations are largely determined by the stimulus. We have demonstrated that DeepEyedentificationLive is able to identify users even in this setting.

We conclude that DeepEyedentificationLive outperforms its monocular predecessor DeepEyedentification consistently and significantly for identification and for low FMR values for verification. Both DeepEyedentificationLive and DeepEyedentification dramatically outperform reference methods that extract explicit saccadic and fixational features. By processing the low-velocity, high-frequency micro-movements in a separate sub-network, DeepEyedentificationLive is able to automatically identify micro-movement features that vary across individuals. From our experiments we can conclude that a stimulus display durations of 250 ms works well for identification but is too fast for liveness detection; a display duration of 500 ms is the overall best configuration under investigation. DeepEyedentificationLive only receives gaze velocities as input and hence is insensitive to user-specific offsets that otherwise would have to be compensated by calibration.

We conclude that using eye movements bears high potential for applications that require a fast and unobtrusive identification. Eye movements are a necessary prerequisite of vision, and are therefore available for a large fraction of the population. Eye movements are orthogonal to established biometric features, but could potentially be measured using the same infrared sensors that can also be used for iris scans or face recognition. Hence, it might complement these technologies in a multimodal biometric system could be more robust to colored contact lenses and small eye apertures—which may prevent the use of iris scans—and niqabs and masks, which pose problems for facial identification.

Acknowledgments

This work was partially funded by the German Science Foundation under grant SFB1294.

References

[1] A. Abdelwahab, R. Kliegl, and N. Landwehr. A semiparametric model for Bayesian reader identification. In *EMNLP 2016*, pages 585–594, 2016. 2, 7

[2] G. Bargary, J. M. Bosten, P. T. Goodbourn, A. J. Lawrance-Owen, R. E. Hogg, and J. Mollon. Individual differences in human eye movements: An oculomotor signature? *Vision Research*, 141:157–169, 2017. 1, 2

[3] R. Bednarik, T. Kinnunen, A. Mihaila, and P. Fränti. Eye-movements as a biometric. In *SCIA 2005*, pages 780–789, 2005. 1, 2

[4] V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio. GANT: Gaze analysis technique for human identification. *Pattern Recognition*, 48:1027–1038, 2015. 2

[5] N. Cuong, V. Dinh, and L. S. T. Ho. Mel-frequency cepstral coefficients for eye movement identification. In *ICTAI 2012*, pages 253–260, 2012. 2

[6] D. Cymek, A. Venjakob, S. Ruff, O.-M. Lutz, S. Hofmann, and M. Roetting. Entering PIN codes by smooth pursuit eye movements. *Journal of Eye Movement Research*, 7:1–11, 2014. 2

[7] A. Darwish and M. Pasquier. Biometric identification using the dynamic features of the eyes. In *BTAS 2013*, pages 1–6, 2013. 2

[8] A. De Luca, R. Weiss, H. Hußmann, and X. An. Eyepass – eye-stroke authentication for public terminals. In *CHI EA 2008*, pages 3003–3008, 2007. 2

[9] P. Dunphy, A. Fitch, and P. Olivier. Gaze-contingent passwords at the ATM. In *COGAIN 2008*, pages 59–62, 2008. 2

[10] S. Eberz, K. Rasmussen, V. Lenders, and I. Martinovic. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *NDSS*, 2015. 2

[11] R. Engbert. Microsaccades: A microcosm for research on oculomotor control, attention, and visual perception. *Progress in Brain Research*, 154:177–192, 2006. 7

[12] R. Engbert and R. Kliegl. Microsaccades uncover the orientation of covert attention. *Vision Research*, 43:1035–1045, 2003. 7

[13] R. Engbert and K. Mergenthaler. Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences of the U.S.A.*, 103:7192–7197, 2006. 7

[14] C. Galdi, M. Nappi, D. Riccio, V. Cantoni, and M. Porta. A new gaze analysis based softbiometric. In *MCPR 2013*, pages 136–144, 2013. 2

[15] I. Griswold-Steiner, Z. Fyke, M. Ahmed, and A. Serwadda. Morph-a-dope: Using pupil manipulation to spoof eye movement biometrics. In *UEMCON 2018*, pages 543–552, 2018. 2

[16] C. Holland and O. Komogortsev. Complex eye movement pattern biometrics: Analyzing fixations and saccades. In *ICB 2013*, 2013. 2, 7, 8

[17] C. Holland and O. V. Komogortsev. Biometric identification via eye movement scanpaths in reading. In *IJCB 2011*, pages 1–8, 2011. 2, 3

[18] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, Oxford, 2011. 1, 4

[19] L. A. Jäger, S. Makowski, P. Prasse, S. Liehr, M. Seidler, and T. Scheffer. Deep Eyedentification: Biometric identification using micro-movements of the eye. In *ECML PKDD 2019*, 2020. 2, 6

[20] A. K. Jain. Biometric recognition: overview and recent advances. In *Iberoamerican Congress on Pattern Recognition*, pages 13–19. Springer, 2007. 1

[21] P. Kasprowski. *Human identification using eye movements*. PhD thesis, Silesian Unversity of Technology, 2004. 2

[22] P. Kasprowski and K. Harkeżlak. The second eye movements verification and identification competition. In *Proceedings of the International Joint Conference on Biometrics*, 2014. 2

[23] P. Kasprowski, O. V. Komogortsev, and A. Karpov. First eye movement verification and identification competition at BTAS 2012. In *BTAS 2012*, pages 195–202, 2012. 2

[24] P. Kasprowski and J. Ober. Eye movements in biometrics. In *International Workshop on Biometric Authentication*, pages 248–258, 2004. 1, 2

[25] P. Kasprowski and J. Ober. Enhancing eye-movement-based biometric identification method by using voting classifiers. In *SPIE 5779: Biometric Technology for Human Identification II*, pages 314–323, 2005. 2

[26] T. Kinnunen, F. Sedlak, and R. Bednarik. Towards task-independent person authentication using eye movement signals. In *ETRA 2010*, pages 187–190, 2010. 2

[27] O. V. Komogortsev, A. Karpov, and C. Holland. CUE: Counterfeitresistant usable eye-based authentication via oculomotor plant characteristics and complex eye movement patterns. In *SPIE Defence Security and Sensing Conference on Biometric Technology for Human Identification IX*, pages 1–10, 2012. 2, 8

[28] O. V. Komogortsev, A. Karpov, and C. D. Holland. Attack of mechanical replicas: Liveness detection with eye movements. *IEEE Transactions on Information Forensics and Security*, 10(4):716–725, 2015. 2, 8

[29] O. V. Komogortsev, A. Karpov, C. D. Holland, and H. P. Proença. Multimodal ocular biometrics approach: A feasibility study. In *BTAS 2012*, pages 209–216, 2012. 2

[30] M. Kumar, T. Garfinkel, D. Boneh, and T. Winograd. Reducing shoulder-surfing by using gaze-based password entry. In *SOUPS 2007*, pages 13–19, 2007. 2

[31] N. Landwehr, S. Arzt, T. Scheffer, and R. Kliegl. A model of individual differences in gaze control during reading. In *EMNLP 2014*, pages 1810–1815, 2014. 2, 7

[32] A. Maeder, C. Fookes, and S. Sridharan. Gaze based user authentication for personal computer applications. In *2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 727–730, 2004. 2

[33] S. Makowski, L. A. Jäger, A. Abdelwahab, N. Landwehr, and T. Scheffer. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *ECML PKDD 2018*, pages 209–225, 2019. 2, 7

[34] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5:229–240, 2004. 1

[35] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and T. A. Dyar. Microsaccades counteract visual fading during fixation. *Neuron*, 49:297–305, 2006. 1

[36] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and D. H. Hubel. Microsaccades: A neurophysiological analysis. *Trends in Neurosciences*, 32:463–475, 2009. 1

[37] D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971. 1

[38] N. Nugrahaningsih and M. Porta. Pupil size as a biometric trait. In *Revised Selected Papers of International Workshop on Biometric Authentication*, pages 222–233, 2014. 2

[39] J. Otero-Millan, X. G. Troncoso, S. L. Macknik, I. Serrano-Pedraza, and S. Martinez-Conde. Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *Journal of Vision*, 8(14):21–21, 2008. 1

[40] K. B. Raja, R. Raghavendra, and C. Busch. Video presentation attack detection in visible spectrum iris recognition using magnified phase information. *IEEE Transactions on Information Forensics and Security*, 10(10):2048–2056, 2015. 2, 8

[41] I. Rigas, G. Economou, and S. Fotopoulos. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recogntion Letters*, 33:786–792, 2012. 2

[42] I. Rigas, G. Economou, and S. Fotopoulos. Human eye movements as a trait for biometrical identification. In *BTAS 2012*, pages 217–222, 2012. 2

[43] I. Rigas, O. Komogortsev, and R. Shadmehr. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transactions on Applied Perception*, 13(2):6, 2016. 2, 3, 6, 7, 8

[44] D. L. Silver and A. Biggs. Keystroke and eye-tracking biometrics for user identification. In *ICAI 2006*, volume 2, pages 344–348, 2006. 2

[45] N. Srivastava, U. Agrawal, S. Roy, and U. S. Tiwary. Human identification using linear multiclass SVM and eye movement biometrics. In *IC3 2015*, pages 365–369, 2015. 2

[46] J. Weaver, K. Mock, and B. Hoanca. Gaze-based password authentication through automatic clustering of gaze points. In *SMC 2011*, pages 2749–2754, 2011. 2

[47] H. Yoon, T. Carmichael, and G. Tourassi. Temporal stability of visual search-driven biometrics. In *SPIE Medical Imaging: Image Perception, Obserformance, and Technology Assessment*, 2015. 2, 7

[48] H.-J. Yoon, T. R. Carmichael, and G. Tourassi. Gaze as a biometric. In *SPIE Medical Imaging Conference: Image Perception, Observer Performance, and Technology Assessment*, 2014. 2, 7