Oculomotoric Biometric Identification under the Influence of Alcohol and Fatigue

Silvia Makowski[1], Paul Prasse[1], Lena A. Jäger[1,2], and Tobias Scheffer[1]

[1]Department of Computer Science, University of Potsdam, Germany
[2]Department of Computational Linguistics, University of Zurich, Switzerland
{silvia.makowski, paul.prasse, lena.jaeger, tobias.scheffer}@uni-potsdam.de

Abstract

Patterns of micro- and macro-movements of the eyes are highly individual and can serve as a biometric characteristic. It is also known that both alcohol inebriation and fatigue can reduce saccadic velocity and accuracy. This prompts the question of whether changes of gaze patterns caused by alcohol consumption and fatigue impact the accuracy of oculomotoric biometric identification. We collect an eye tracking data set from 66 participants in sober, fatigued and alcohol-intoxicated states. We find that after enrollment in a rested and sober state, identity verification based on a deep neural embedding of gaze sequences is significantly less accurate when probe sequences are taken in either an inebriated or a fatigued state. Moreover, we find that fatigue and intoxication appear to randomize gaze patterns: when the model is fine-tuned for invariance with respect to inebriation and fatigue, and even when it is trained exclusively on inebriated training person, the model still performs significantly better for sober than for sleep-deprived or intoxicated subjects.

1. Introduction

Human eye movements alternate between *saccades*—fast relocation movements of around 50 ms—and *fixations* of around 250 ms during which visual input is obtained. Moreover, high-frequency involuntary micro-movements occur during attempted fixations which, among other functions, prevent visual fading of the fixated image; these fixational micro-movements are termed *drift*, *tremor*, and *microsaccades* [15, 16, 17, 19, 8]. It has long been known that patterns of eye movements are highly individual [18], and 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 [10, 3].

Early work on oculomotoric biometric identification extracts fixations and saccades from an eye-tracking signal and measures the values of engineered explicit features, such as fixation durations and saccadic amplitudes and velocities. Since these approaches only process information contained in the low-frequency macro-movements, they require long eye gaze sequences of more than one minute [12] for an identification.

The *DeepEyedentification* method [9, 13, 14] that uses deep convolutional neural networks to process the raw angular velocities, by contrast, is able to exploit patterns in both micro- and macro-movement which reduces the time to identification by an order of magnitude. Involuntary eye movements can also be cross-checked against a controlled stimulus, which is a major obstacle for any potential presentation-attack instrument [14].

Psychological research has found that the saccadic accuracy and peak saccadic velocity can be negatively impacted by fatigue [6, 7]. In a driving-simulator experiment, saccadic duration has been found to increase, saccadic speed to decrease, and their standard deviation to increase with increasing fatigue [21]. Similarly, the acute consumption of alcohol has been observed to impair saccadic latency, velocity, and accuracy [11, 20]. This raises the questions whether fatigue or alcohol consumption would "break" oculomotoric biometric systems and, if this turns out to be the case, if oculomotoric identification can be made robust against the mental state of subjects.

This paper investigates the robustness of oculomotoric biometric identification with respect to fatigue and acute alcohol consumption. To this end, we collect eye-gaze data of users in sleep-deprived and intoxicated states, in addition to the baseline state. We fine-tune the DeepEyedentification model for invariance against fatigue and intoxication on persons recorded in multiple states, and train a version

of the model exclusively using inebriated training persons. We investigate the performance of all the models for probe sequences recorded in a fatigued state and after alcohol consumption and investigate the reasons for deviations from the performance in the baseline state. As an additional contribution, we release the new *Potsdam Binge / JuDo* data set of gaze data of sleep-deprived and inebriated subjects to the research community.

The rest of this paper is structured as follows. Section 2 lays out the problem setting, and Section 3 summarizes the DeepEyedentification [14] network. Section 4 details our data collection, experimental setting, and training and fine-tuning procedures used. Section 5 presents our results, and Section 6 concludes.

2. Problem Setting

We will study the problem of oculomotoric biometric *identity verification*. The input to each system is given as a sequence of eye gaze yaw and pitch angles of the left and right eye over an observation period.

In a biometric identity verification scenario, each user first enrolls with one or more *enrollment gaze sequences*. In this study, we assume that users are enrolled in a *baseline state*—neither fatigued nor under the influence of alcohol. At application time, a *probe sequence* is compared to the enrollment sequences of the presumed enrolled identity by means of a parametric *similarity metric*. In case the similarity exceeds a decision threshold, the claimed identity is verified; otherwise, the user is classified as *impostor*.

The performance of identity verification models 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 trade-off curve (DET curve)*. The *equal error rate* (EER) is the point on this curve for which FMR equals FNMR.

In this study, we use a state-of-the-art neural-network model [14] in which the similarity metric is the cosine similarity between *neural embeddings* of gaze sequences. The embedding function 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.

We will compare the cases of a probe sequence that is observed (a) in the baseline state, (b) in a state of fatigue induced by prolonged sleep deprivation, and (c) under the influence of alcohol.

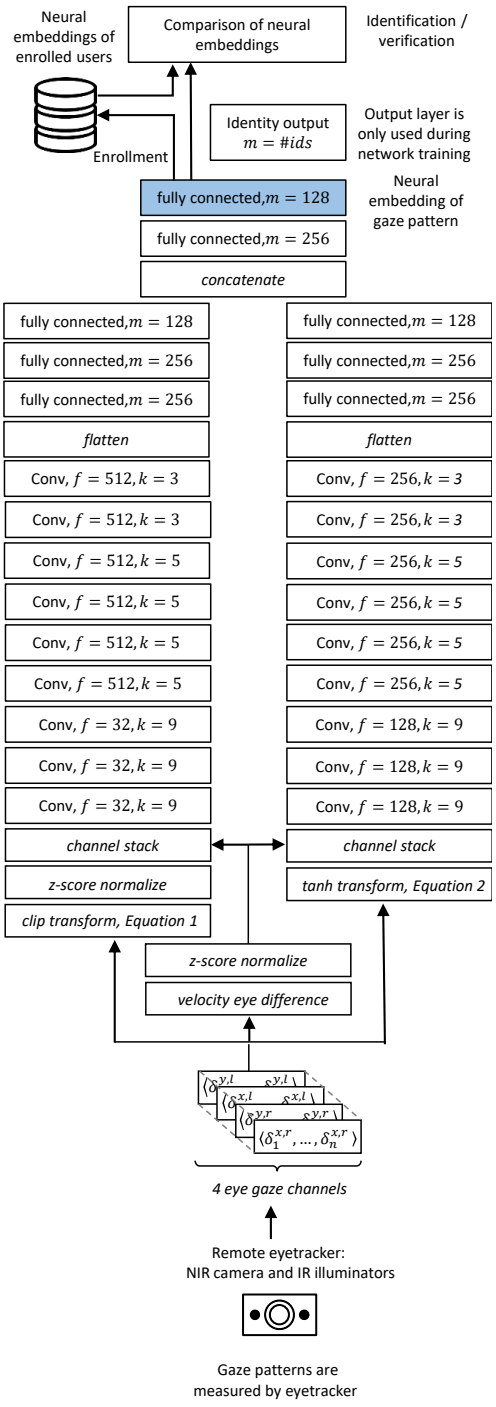


Figure 1: Binocular version of the DeepEyedentification architecture. Figure adapted from [13].

3. System and Network Architecture

This section provides an overview of the binocular version of the *DeepEyedentification* architecture [14]. An eye

tracker records binocular gaze sequences of absolute yaw x and pitch gaze angles y of the left l and right eye r at a sampling frequency of ρ, measured in Hz. The binocular DeepEyedentification network (see Figure 1) receives yaw and pitch gaze velocities δ_i^x and δ_i^y in $^\circ/s$ as input which are computed from the recorded gaze sequence as $\delta_i^x = \frac{\rho}{2}(x_{i+1} - x_{i-1})$ and $\delta_i^y = \frac{\rho}{2}(y_{i+1} - y_{i-1})$ for the left and the right eye, respectively. This results in a total of four input channels, namely the sequence of yaw $\langle \delta_1^{x,l}, \ldots, \delta_n^{x,l} \rangle$ and pitch angular velocities of the left eye $\langle \delta_1^{y,l}, \ldots, \delta_n^{y,l} \rangle$ and the corresponding yaw $\langle \delta_1^{x,r}, \ldots, \delta_n^{x,r} \rangle$ and pitch angular velocities of the right eye $\langle \delta_1^{y,r}, \ldots, \delta_n^{y,r} \rangle$. The network processes input sequences of 1,000 time steps corresponding to 1 s of 1000 Hz eye tracking recording.

The key feature of the network's architecture is that the input channels are duplicated and directed into two separate convolutional subnets. The *fast subnet* is designed to process the high angular velocities of (micro-) saccadic eye movements whereas the *slow subnet* is designed to process the slow fixational eye movements (drift and tremor). Each of the subnets is preceded by a transformation layer that applies a transformation to the input to resolve the fast saccadic and slow fixational eye movements, respectively. For the fast subnet, saccadic eye movements are resolved by applying a clipping function that truncates velocities below a threshold ν_{min} to zero and a subsequent z-score normalization (see Equation 1). Based on hyperparameter tuning within a range of psychologically plausible parameters on two independent data sets [9, 13], the velocity threshold ν_{min} is set to $40^\circ/s$.

$$t_f(\delta_i^x, \delta_i^y) = \begin{cases} z(0) & \text{if } \sqrt{\delta_i^{x\,2} + \delta_i^{y\,2}} < \nu_{min} \\ (z(\delta_i^x), z(\delta_i^y)) & \text{otherwise} \end{cases} \quad (1)$$

The slow fixational eye movements are resolved by applying a sigmoidal function that stretches the slow velocities of drift and tremor approximately within the interval between -0.5 and $+0.5$ and squashes the (micro-) saccadic velocities to the interval between -0.5 and -1 or $+0.5$ and $+1$, depending on their direction (see Equation 2). Independent hyperparameter optimization on two data sets showed that an appropriate value for the scaling factor c of Equation 2 is 0.02 [9, 13].

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

Since binocular alignment is an informative individual characteristics, the four untransformed input velocity channels are also fed into a subtraction layer which 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 $\langle \delta_1^{y,r} - \delta_1^{y,l}, \ldots, \delta_n^{y,r} - \delta_n^{y,l} \rangle$. After each of the transformation layers, a stacking layer is inserted which stacks these additional two channels with the input of each of the two subnets.

The two subnets share the same number and type of layers. Each of the subnets consists of a series of one-dimensional convolutional layers, where the convolutions are applied to the six input channels over the temporal axis. The number of filters and kernel size of the convolutional layers (f and k in Figure 1), as well as the number of units of the subsequent fully connected layers (m in Figure 1), are allowed to differ between the two subnets. For our experiments, we use established hyperparameter [13] (see Figure 1). After each of the convolutional and fully connected layers, batch normalization and ReLU activation is applied. An average pooling layer with pooling size 2 and stride size 1 is inserted after each convolutional layer.

For training, a softmax output layer with one unit for each user in the training data is added. Using categorical cross-entropy as loss function, the network is trained to predict a viewer's identity from an eye tracking sequence. Once training is completed, the softmax output layer is removed and the activation of the last fully connected layer (highlighted in blue in Figure 1) is used as neural feature embedding of an input gaze sequence. At application time, the similarity between an enrolment and a test sequence is computed as the cosine similarity of their neural embeddings, averaged over all input windows of 1,000 ms.

4. Methods

This section reports on data collection, training and fine-tuning of the models, and experiments.

4.1. Data Collection

We collect the *Potsdam Binge / JuDo* data set[1] of binocular eye movement data (horizontal and vertical gaze coordinates) from 66 subjects, aged 18 to 48 years, with a mean age of 24. Participants have given their written informed consent and the study has been approved by the ethics committee of the University of Potsdam.

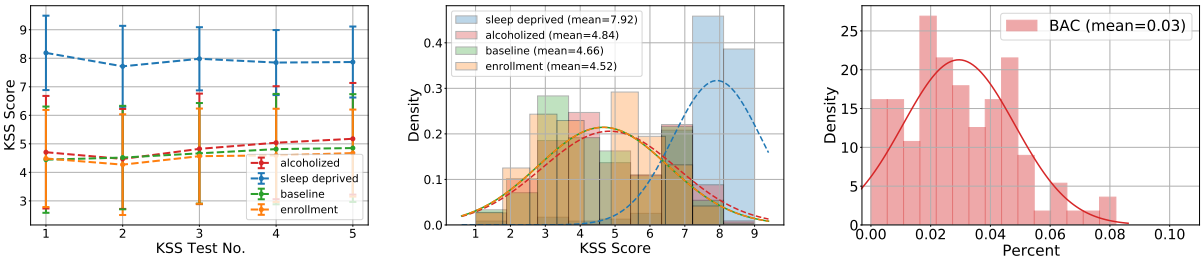
4.1.1 Experimental Design

Participants are recorded over four experimental sessions. We ensure that any two consecutive sessions are separated by a time lag of at least one week and we randomly vary the order of experimental conditions across participants.

1. For a *sleep-deprived* session, we ask participants to refrain from sleeping within 24 hours prior the starting time. We do not monitor participants for compliance.

2. In an *alcohol* session, participants imbibe a beverage with 26 ml of alcohol at the start of the session; this

[1] https://osf.io/drn4x/



(a) Development of the KSS score over time.

(b) Distribution of KSS scores across participants.

(c) Distribution of the BAC across participants in the alcohol session.

Figure 2: Descriptive statistics for the *Potsdam Binge / JuDo* data set.

dose results in a blood alcohol concentration of below 0.05% for most individuals.

3. For each of two *baseline sessions*, participants are asked to appear well rested and sober before the session.

We adopt the same experimental design that was used by Makowski *et al.* [14] to collect the *JuDo1000* data set: In each session, participants are presented with a total of 144 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 the dot is displayed is varied between 250 ms, 500 ms, 1,000 ms and 1,500 ms with a fixed value within each trial; the size of the screen area in which the dot appears is varied between 76 by 140 mm, 114 by 170 mm, and 190 by 230 mm around the center of the monitor with a fixed area within each trial. The combination of display duration and areas results in twelve trial configurations.

Eye movements are recorded using a tripod-mounted Eyelink Portable Duo eye tracker at a sampling frequency of 1,000 Hz. Participants are seated in front of a 380 by 300 mm (1280 by 1024 px) computer monitor at a height adjustable table with their head stabilized by a chin- and forehead rest.

Participants self-report their perceived level of fatigue before the recording, three times during the recording and after the experiment on the Karolinska sleepiness scale (KSS) [1] from being "1–extremely alert" to "9–very sleepy, great effort to keep alert, fighting sleep". This results in five KSS scores per session. In the alcohol session the blood alcohol concentration (BAC) is estimated with a breathalyzer (Dräger Alcotest 5820) before and after the recording.

4.1.2 Quantitative Data Analysis

Figure 2a shows the mean KSS scores across the session and Figure 2b the distribution of scores over subjects. We can see that during the baseline and alcohol session the scores are mostly in the medium range ("neither very alert nor fatigued") due to the repetitiveness of the task whereas the sleep-deprived subjects are mostly drowsy or fighting sleep. Figure 2c shows the distribution of BAC in the alcohol session. Most probes are taken in the BAC range of up to 0.05% with some participants reaching up to 0.08%.

4.2. Model Training and Fine Tuning

We train several models on combinations of the *JuDo1000* data set [14] of 150 subjects recorded over four experimental sessions in the baseline condition and the *Potsdam Binge / JuDo* data set of sleep-deprived and intoxicated subjects. In all cases, we use the hyper-parameters reported by Makowski *et al.* [14].

We split the *Potsdam Binge / JuDo* data into 22 test persons who have completed all four sessions, and 35 training persons who have completed at least one baseline session and the sleep-deprived session, the alcohol session, or both. The resulting training portion of the *Potsdam Binge / JuDo* data comprises 95 recording sessions: 27 sleep sessions, 25 alcohol sessions and 43 baseline sessions; 25 training persons were recorded over three sessions, 10 training persons were recorded over two sessions. The rationale of fine-tuning on persons that have completed at least sessions in two different states is that it forces the embedding layer towards invariance with respect to these states, since these training persons have to be recognized in either state.

Our first model is trained only on the *JuDo1000* data set of 150 subjects using the and training protocol of Makowski *et al.* [14]. This model is trained with a softmax layer that distinguishes between 150 training subjects which is discarded after the training; the final embedding layer below this softmax layer is used to calculate the cosine similarity.

Table 1: Comparison of EER between baseline, fatigued, and intoxicated probe sequences after enrollment on baseline sequences from another session; verification with one enrolled user and 4 impostors. Mean EER $\pm$ standard deviation of 5 times resampling 5 users out of 22 users. A star (*) indicates a significantly ($p < 0.05$) higher EER compared to the sober probe. A cross ($^+$) indicates a significantly ($p < 0.05$) higher EER compared to the sober probe without fine tuning. Bold font indicates the lowest value in a column.

			EER		
Training data	Fine-tuning	Probe duration	baseline	sleep deprived	intoxicated
JuDo (150 users; 600 sessions in total)	-	1 s	0.1148 ± 0.0377	$0.157 \pm 0.0397^*$	$0.1563 \pm 0.0557^*$
		5 s	$\mathbf{0.0577 \pm 0.0285}$	$\mathbf{0.1026 \pm 0.037}^*$	$\mathbf{0.121 \pm 0.053}^*$
		10 s	$\mathbf{0.0451 \pm 0.0272}$	$\mathbf{0.0898 \pm 0.0363}^*$	$\mathbf{0.1146 \pm 0.0542}^*$
		60 s	$\mathbf{0.0266 \pm 0.0272}$	$\mathbf{0.066 \pm 0.039}^*$	$0.109 \pm 0.0569^*$
JuDo (150 users; 600 sessions in total)	Embedding layer on Binge	1 s	$\mathbf{0.1121 \pm 0.0454}$	$\mathbf{0.155 \pm 0.0546}^{*+}$	$\mathbf{0.1509 \pm 0.0622}^{*+}$
		5 s	$0.083 \pm 0.0495^+$	$0.1291 \pm 0.0573^{*+}$	$0.1237 \pm 0.0642^{*+}$
		10 s	$0.0766 \pm 0.0521^+$	$0.1252 \pm 0.06^{*+}$	$0.1172 \pm 0.0645^{*+}$
		60 s	$0.0686 \pm 0.0588^+$	$0.1189 \pm 0.0675^{*+}$	$\mathbf{0.1083 \pm 0.0685}^{*+}$
JuDo (150 users; 600 sessions in total)	All layers on Binge	1 s	$0.1503 \pm 0.0441^+$	$0.2011 \pm 0.0479^+$	$0.1949 \pm 0.0661^+$
		5 s	$0.0987 \pm 0.0437^+$	$0.1642 \pm 0.0466^{*+}$	$0.1448 \pm 0.0709^+$
		10 s	$0.0881 \pm 0.0437^+$	$0.1591 \pm 0.0453^{*+}$	$0.1356 \pm 0.0693^+$
		60 s	$0.0746 \pm 0.0428^+$	$0.1543 \pm 0.0472^{*+}$	$0.1212 \pm 0.0637^+$
Binge (35 users; 95 sessions in total)	-	1 s	$0.165 \pm 0.0463^+$	$0.2081 \pm 0.0594^{*+}$	$0.2242 \pm 0.0547^{*+}$
		5 s	$0.119 \pm 0.0478^+$	$0.1798 \pm 0.0653^{*+}$	$0.1842 \pm 0.0576^{*+}$
		10 s	$0.1096 \pm 0.0496^+$	$0.1747 \pm 0.0681^{*+}$	$0.1766 \pm 0.0581^{*+}$
		60 s	$0.0969 \pm 0.0531^+$	$0.1696 \pm 0.0752^{*+}$	$0.1598 \pm 0.0636^{*+}$
JuDo (35 users, 95 sessions in total)	-	1 s	$0.155 \pm 0.0473^+$	$0.204 \pm 0.0572^{*+}$	$0.1966 \pm 0.0662^{*+}$
		5 s	$0.1043 \pm 0.0493^+$	$0.1638 \pm 0.0589^{*+}$	$0.1549 \pm 0.0555^{*+}$
		10 s	$0.0945 \pm 0.0492^+$	$0.1581 \pm 0.061^{*+}$	$0.1474 \pm 0.0535^{*+}$
		60s	$0.0805 \pm 0.0508^+$	$0.1518 \pm 0.0667^{*+}$	$0.1391 \pm 0.0505^{*+}$

We use a learning rate of 10^{-3} for both subnets and 10^{-4} for the remaining layers.

This first model also serves as starting point of a second and third fine-tuned second model. For these models, the softmax layer is replaced by a new softmax layer that distinguishes between the 35 training persons of the *Potsdam Binge / JuDo* data. For the second model, only the embedding and output layer are trained using a learning rate that starts with the terminal learning rate of pre-training on *JuDo1000* whereas all weights on lower layers are frozen.

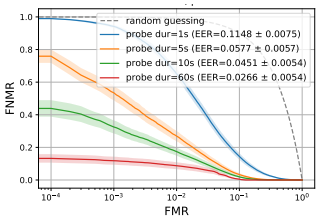
For the third model, all model parameters of the first model are fine-tuned on the training persons of *Potsdam Binge / JuDo* using the terminal learning rate of the pre-training process as starting point. For retraining the architecture on the Binge / JuDo data set, we use a learning rate of 0.001 to train the new softmax output layers of the subnets and 10^{-4} for the output layer of the joint architecture, whereas we use a learning rate of 10^{-5} for fine-tuning all layers.

The fourth model is trained *only* on the 35 training persons, with 95 sessions in total, of the *Potsdam Binge / JuDo* data set. Since each training person has been observed in the baseline state and in at least one of the sleep-deprived or intoxicated state, this model should exhibit the most homogeneous performance across states. In this experiment, we use a learning rate of 10^{-3} for both subnets and 10^{-4} for the remaining layers.

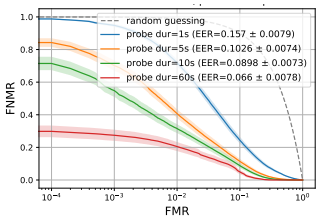
Due to the much smaller training population, this fourth model cannot be compared meaningfully to the model trained on 150 training subjects of the JuDo set. Therefore, we train a fifth and final model on a random subset of 35 JuDo training persons, 95 sessions in total. This model is identical to the first model in all aspects other than the size of the training population.

We use the Keras [4] and Tensorflow [5] libraries on an NVIDIA A100-SXM4-40GB GPU using the NVIDIA CUDA platform. All models are pre-trained and fine-tuned using the Adam optimizer. For all models we use a batch size of 64. For all models we use early stopping with a patience of 10 epochs for which we use 20% of each training person's data as validation data. All code can be found online.[2]

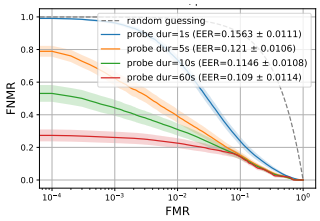
[2] https://osf.io/brzfn/



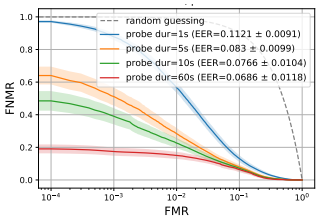
(a) Training on *JuDo1000*; enrollment: baseline session 1, probe: baseline session 2.



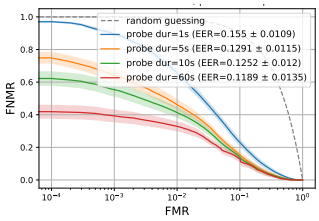
(b) Training on *JuDo1000*; enrollment: baseline session, probe: sleep-deprived session.



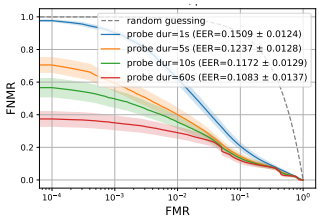
(c) Training on *JuDo1000*; enrollment: baseline session, probe: alcohol session.



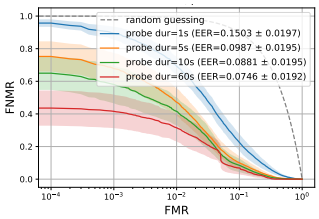
(d) Training on *JuDo1000*, fine-tuning of embedding layer on *Binge*; enrollment: baseline session 1, probe: baseline session 2.



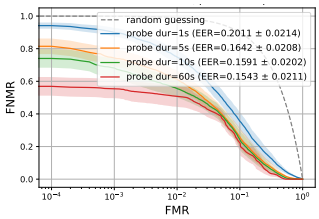
(e) Training on *JuDo1000*, fine-tuning of embedding layer on *Binge*; enrollment baseline session, probe: sleep-deprived session.



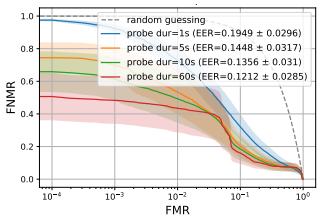
(f) Training on *JuDo1000*, fine-tuning of embedding layer on *Binge*; enrollment baseline session, probe: alcohol session.



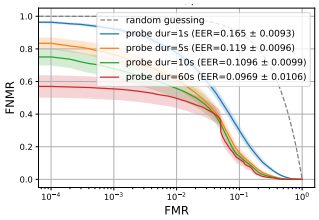
(g) Training on *JuDo1000*, fine-tuning of all layers on *Binge*; enrollment: baseline session 1, probe: baseline session 2.



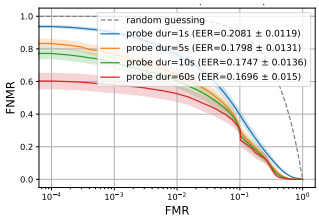
(h) Training on *JuDo1000*, fine-tuning of all layers on *Binge*; enrollment: baseline session, probe: sleep-deprived session.



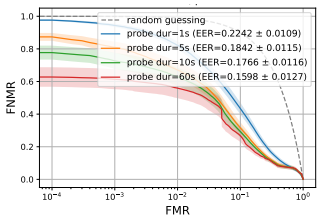
(i) Training on *JuDo1000*, fine-tuning of all layers on *Binge*; enrollment: baseline session, probe: alcohol session.



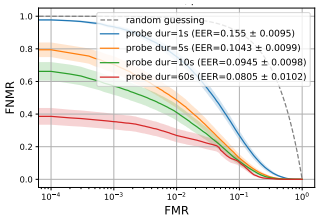
(j) Training on *Binge*; enrollment: baseline session 1, probe: baseline session 2.



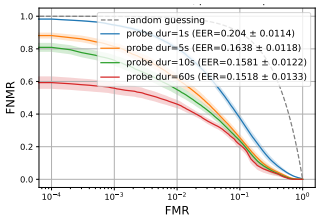
(k) Training on *Binge*; enrollment: baseline session, probe: sleep-deprived session.



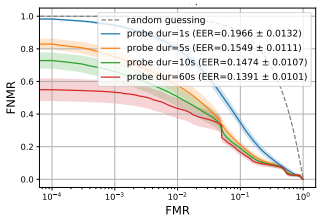
(l) Training on *Binge*; enrollment: baseline session, probe: alcohol session.



(m) Training on *JuDo (35 users)*; enrollment: baseline session 1, probe: baseline session 2.



(n) Training on *JuDo (35 users)*; enrollment: baseline session, probe: sleep-deprived session.



(o) Training on *JuDo (35 users)*; enrollment: baseline session, probe: alcohol session.

Figure 3: Verification performance. False Non Match Rate (FNMR) over False Match Rate (FMR). Shaded bands show the standard error.

Table 2: Comparison of FNMR at a FMR of 10^{-3} between baseline, fatigued, and intoxicated probe sequences after enrollment on baseline sequences from another session; verification with one enrolled user and 4 impostors. Mean FNMR $\pm$ standard deviation of 5 times resampling 5 users out of 22 users. A star (*) indicates a significantly ($p < 0.05$) higher FNMR compared to the sober probe. A cross ($^+$) indicates a significantly ($p < 0.05$) higher FNMR compared to the sober probe without fine tuning. Bold font indicates the lowest value in a column.

			FNMR@FMR 10^{-3}		
Training data	Fine-tuning	Probe duration	baseline	sleep deprived	intoxicated
JuDo (150 users; 600 sessions in total)	-	1 s	0.9422 ± 0.0448	0.9484 ± 0.0459	0.9635 ± 0.0453
		5 s	0.5356 ± 0.2279	$0.7033 \pm 0.1591^*$	0.6263 ± 0.2144
		10 s	$\mathbf{0.3277 \pm 0.203}$	$\mathbf{0.5526 \pm 0.1876}^*$	$\mathbf{0.4463 \pm 0.246}^*$
		60 s	$\mathbf{0.4227 \pm 0.409}$	$\mathbf{0.2753 \pm 0.1723}$	$\mathbf{0.3818 \pm 0.271}$
JuDo (150 users; 600 sessions in total)	Embedding layer on Binge	1 s	$\mathbf{0.8776 \pm 0.103}$	$\mathbf{0.897 \pm 0.0751}$	$\mathbf{0.9027 \pm 0.089}$
		5 s	$\mathbf{0.5131 \pm 0.275}$	$\mathbf{0.6441 \pm 0.1988}$	$\mathbf{0.5949 \pm 0.260}$
		10 s	0.3918 ± 0.2497	$0.5554 \pm 0.2147^{*+}$	$0.5006 \pm 0.2746^+$
		60 s	0.4937 ± 0.3929	0.3947 ± 0.1987	0.4714 ± 0.2716
JuDo (150 users; 600 sessions in total)	All layers on Binge	1 s	0.895 ± 0.0892	0.903 ± 0.0603	0.923 ± 0.054
		5 s	0.6806 ± 0.2139	$0.7438 \pm 0.1176^+$	$0.6994 \pm 0.2414^+$
		10 s	0.5845 ± 0.2571	0.683 ± 0.1326	0.6189 ± 0.2914
		60 s	0.7355 ± 0.2627	0.5553 ± 0.1309	0.4845 ± 0.3217
Binge (35 users; 95 sessions in total)	-	1 s	0.9227 ± 0.0618	0.8994 ± 0.0759	0.9445 ± 0.0503
		5 s	$0.7672 \pm 0.2224^+$	$0.767 \pm 0.1703^+$	$0.8108 \pm 0.1739^+$
		10 s	$0.6807 \pm 0.2967^+$	$0.716 \pm 0.193^+$	$0.731 \pm 0.2433^+$
		60 s	$0.6832 \pm 0.3301^+$	0.5797 ± 0.2558	$0.6231 \pm 0.312^+$
JuDo (35 users; 95 sessions in total)	-	1 s	0.9356 ± 0.0548	0.947 ± 0.035	0.9457 ± 0.0488
		5 s	$0.705 \pm 0.2521^+$	$0.8158 \pm 0.1144^{*+}$	$0.7532 \pm 0.2083^+$
		10 s	$0.5744 \pm 0.2723^+$	$0.7327 \pm 0.1592^{*+}$	$0.6645 \pm 0.2758^+$
		60 s	0.3545 ± 0.2404	$0.5609 \pm 0.1791^*$	$0.5994 \pm 0.3511^*$

5. Results

The embedding is evaluated on enrollment and probe sequences of 22 test subjects from the *Potsdam Binge / JuDo* data set who have completed the full four sessions. Enrollment is always performed on 20 randomly drawn trials from one of the baseline sessions, where each trial consists of gaze data for a stimulus of five random points. Depending on the display duration of each point, a trial is between 1.25 and 5 seconds long. Probe sequences are taken from the second baseline session, the alcohol session, or the sleep-deprived session.

Table 1 shows the equal error rates, Table 2 the FNMR at a FMR of 10^{-3} of all trained models. Figure 3 shows the corresponding DET curves. For most probe durations and probe session types, the model that was only trained on *JuDo1000* shows the lowest EER. This model has the lowest FNMR at a FMR of 10^{-3} in half of the cases. Fine-tuning of the embedding layer leads to slight, statistically insignificant improvements in one third of the cases with respect to EER and in half the cases with respect to FNMR at FMR of 10^{-3}.

Fine-tuning all layers and training only on *Binge* leads to higher EER and FNMR. For all probe durations and models, probe sequences taken in both the sleep-deprived and the intoxicated condition result in higher EER values than probe sequences in baseline conditions. Most differences are significant ($p < 0.05$) for EER and some also for FNMR, based on a two-tailed t-test.

Even the fourth model that was *only* trained on the 35 *Binge* training subjects performs significantly worse for probe sequences in sleep-deprived and intoxicated states than it does for sober probe sequences. Regarding the FNMR at FMR of 10^{-3}, the discrepancy between sober and sleep-deprived or intoxicated probe sequences is noticeably smaller for this fourth model, but only at the expense of generally higher FNMR values. Surprisingly, this fourth model does not show any advantage for probe sequences in any condition over the fifth model that was trained on an equally large population in only the sober state. EER and FNMR values appear to be marginally higher for the fourth than for the fifth model, but the differences are not significant.

6. Discussion

Eye movements are a novel and innovative biometric characteristic. Since eye movements are a response to a

stimulus that can be randomized, it would be extremely challenging to devise a presentation-attack instrument. Eye movements are also independent of other characteristics such as facial features, iris, or fingerprints, and therefore oculomotoric biometrics may prove to be a valuable factor in multi-modal biometric systems. Eye gaze is, for instance, unaffected by subjects wearing masks. But in order to gauge its potential for practical application, it is important to understand what the limiting factors of oculomotoric biometrics may be.

We find that a state-of-the-art oculomotoric biometric model that has been trained on subjects in a sober state performs significantly worse for probe sequences taken in a fatigued or alcohol-inebriated state. Fine-tuning the embedding layer on a small group of training subjects that have been recorded both sober and in fatigued or intoxicated states appears to result in (statistically insignificant) performance improvements for short probe sequences.

Training the model only on a small group of subjects that have been recorded in sober and fatigued or inebriated states mitigates the performance gap between sober, fatigued, and inebriated probe sequences somewhat, *but only* by deteriorating the performance for all probe conditions, compared to training on equally many sober training subjects.

It is known that both fatigue and alcohol reduce saccadic accuracy and peak saccadic velocity [6, 7, 11, 20]. Our interpretation of our experimental results therefore is that fatigue and alcohol intoxication add noise to eye movements that dilutes the generally highly individual gaze patterns. A model that has been trained on the Binge training users shows a small (statistically insignificant) performance degradation for probe sequences in any condition compared to a model that has been trained on an equally large population in a sober state only. This indicates that training data recorded in a sober state are "more valuable" to the network than training data recorded in a sleep-deprived or intoxicated state.

Our findings suggest that fine-tuning much larger fine-tuning population recorded in multiple states would somewhat narrow, but not close, the performance gap between sober probe sequences and fatigued or inebriated probe sequences on the other hand.

Acknowledgment

This work was partially funded by the German Federal Ministry of Education and Research under grant 01|S20043.

References

[1] T. Åkerstedt and M. Gillberg. Subjective and objective sleepiness in the active individual. *International journal of neuroscience*, 52(1-2):29–37, 1990.

[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.

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

[4] F. Chollet et al. Keras. `https://keras.io`, 2015.

[5] J. Dean, R. Monga, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[6] N. Galley. Saccadic eye movement velocity as an indicator of (de)activation. a review and some speculations. *Journal of Psychophysiology*, 3:229–244, 1989.

[7] K. Hirvonen, S. Puttonen, K. Gould, J. Korpela, V. F. Koefoed, and K. Müller. Improving the saccade peak velocity measurement for detecting fatigue. *Journal of neuroscience methods*, 187(2):199–206, 2010.

[8] 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.

[9] 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 *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science.*, pages 299–314. Springer, Cham, 2020.

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

[11] A. C. King and J. A. Byars. Alcohol-induced performance impairment in heavy episodic and light social drinkers. *Journal of Studies on Alcohol*, 65(1):27–36, 2004.

[12] 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 *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*, pages 209–225. Springer, Cham, 2019.

[13] S. Makowski, L. A. Jäger, P. Prasse, and T. Scheffer. Biometric identification and presentation-attack detection using micro-movements of the eyes. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2020.

[14] S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer. DeepEyedentificationLive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[15] 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.

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

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

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

[19] 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.

[20] W. Roberts, M. A. Miller, J. Weafer, and M. T. Fillmore. Heavy drinking and the role of inhibitory control of attention. *Experimental and Clinical Psychopharmacology*, 22(2):133, 2014.

[21] R. Schleicher, N. Galley, S. Briest, and L. Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51(7):982–1010, 2008.