Robustness of Eye Movement Biometrics Against Varying Stimuli and Varying Trajectory Length

Christoph Schröder [*]
University of Bremen
Bremen, Germany
schroeder.c@cs.uni-bremen.de

Sahar Mahdie Klim Al Zaidawi [*]
University of Bremen
Bremen, Germany
saharmah@cs.uni-bremen.de

Martin H.U. Prinzler
University of Bremen
Bremen, Germany
martin.prinzler@cs.uni-bremen.de

Sebastian Maneth
University of Bremen
Bremen, Germany
maneth@cs.uni-bremen.de

Gabriel Zachmann
University of Bremen
Bremen, Germany
zach@cs.uni-bremen.de

ABSTRACT

Recent results suggest that biometric identification based on human's eye movement characteristics can be used for authentication. In this paper, we present three new methods and benchmark them against the state-of-the-art. The best of our new methods improves the state-of-the-art performance by 5.2 percentage points. Furthermore, we investigate some of the factors that affect the robustness of the recognition rate of different classifiers on gaze trajectories, such as the type of stimulus and the tracking trajectory length. We find that the state-of-the-art method only works well when using the same stimulus for testing that was used for training. By contrast, our novel method more than doubles the identification accuracy for these transfer cases. Furthermore, we find that with only 90 seconds of eye tracking data, 86.7 % accuracy can be achieved.

Author Keywords

eye tracking; gaze detection; eye movement biometrics

CCS Concepts

•**Security and privacy** → *Biometrics;* •**Human-centered computing** → *User models;* •**Computing methodologies** → *Classification and regression trees;*

INTRODUCTION

It has been observed that eye movements can be used as biometrics, i.e., as a way to identify a person within a larger pool of persons. This identification enables not only access control to secret information but also to tailor the user experience for each user. Especially in scenarios where the user wears an HMD, and therefore only the user's eyes are visible to cameras, gaze tracking biometrics can deliver a continuous identification. Principled research on this topic started about 15 years ago with the seminal paper by Kasprowski and Ober [17]. Since then, vast improvements on the accuracy of eye movement biometrics have been achieved, which have been facilitated by the continuous creation of high-quality datasets (many of which are publicly available), and by the application of current methods from statistics and machine learning.

In order to better compare the various existing methods for eye movement biometrics, a competition series has been set up in 2012 [16]. The most recent competition is *BioEye 2015* [23], which evaluated competitors using two different datasets: in the TEX dataset, participants read a complex poem presented on a monitor; in the RAN dataset, participants observed a randomly moving dot on the screen, see Fig. 1 for two examples. Competitors could train their models on subsets of the datasets; then, during the actual competition, their models/methods were evaluated on hitherto unseen samples (but the same stimulus). We feel it is important to note that, during the evaluation, they also knew the dataset to which each given test sample belonged to (TEX or RAN). Thus, predictions are, strictly speaking, *stimulus-dependent* (aka. *task-dependent* [19, 5]).

It should be noted that the two types of stimuli used in the contest, TEX, and RAN, are *very* different in nature. TEX relies on reading a poem, i.e., on a highly cognitive activity, while RAN relies more on a "hunter's task" of following a target (see also Friedman et al. [7] for this and related aspects). In our experience, classifiers that are trained on eye tracking trajectories obtained with one specific task (e.g., reading text) perform much worse when classifying trajectories obtained with a very different task (e.g., following a random dot). This kind of application of classifiers is called *stimulus-independent*. We will call such scenarios *strongly task-independent* classification.

[*] These two authors contributed equally to this work

CHI '20, April 25–30, 2020, Honolulu, HI, USA.
Copyright 2020 Association of Computing Machinery.
ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376534

By contrast, when training is performed on eye tracking data obtained from participants watching several images, and the evaluation is done using similar, but *different* images, we call this *weakly task-independent* classification.

While some applications of eye tracking biometrics will be in weakly task-independent settings, we believe that many applications will present strongly task-independent settings, yet the identification has to be as frictionless as possible. Therefore, stimulus independence is an important property.

Stimulus-independence in general machine learning is difficult to achieve and active reasearch [15]. In gaze biometrics, though, both training and testing stimuli are not fed directly into the classifier. Instead, the user observes the stimulus, and the algorithm works on the gaze trajectory as a common modality. One question we address in this work is how much this abstraction of the stimulus helps overcome the mentioned difficulties.

To our knowledge, the degree to which classification performance depends on the degree of task-independence (weak or strong) has not been investigated yet. Also, other effects important for real-world applications, such as trajectory length, have not been investigated much.

Our main contributions in this paper are:

- We present two extensions of the method by George and Routray [10], which is, to our knowledge, the best classifier, at least for weakly task-independent scenarios. One extension uses more features; the other one uses a different classifier. In total, we evaluate and compare four different methods in this paper.

- To the best of our knowledge, we are the first to compare the stimulus-agnostic performance of gaze biometrics methods (i.e., different training/testing stimulus types), which is important for potential real-world application of eye tracking biometrics.

- As a third major contribution, we analyze the effect of different tracking lengths on classification performance.

- We make an exact re-implementation of the method by George and Routray and all our methods publicly available as python module (https://cgvr.cs.uni-bremen.de/research/smida_ml/).

RELATED WORK

There is a large body of literature on eye movement biometrics, see, e.g., recent surveys [8, 6, 23, 9], which give a good overview. Here, we restrict our attention to work that is directly related or comparable to our approach.

Each of the seven participants of the BioEye 2015 competition computes a set of features from the eye tracking trajectories and then use some statistical or machine learning technique to carry out the classification (see Table 5 in [23]). Indeed, the winners of the competition, George and Routray, use the largest number of features when compared to the other participants. Kinnunen et al. [19] study task-independent person authentication using eye movement signals. Their approach is inspired by their earlier work on text-independent speaker recognition [18]. They first train a Gaussian mixture model (GMM), called the "universal background model," from a very large set of samples. Each person is then modeled by a GMM that is an interpolation between the universal background model and the observed user data. Their experimental results are rather preliminary with error rates between 29 % and 47 % (for a dataset with 17 users). Darwish [5] also investigates task-independent biometric identification based on eye movements (see also [4]). They present two different images to 17 participants (with no specific task), each image for 20 seconds. Using random forests over a small set of features they obtain a precision of 37 %. Combining the eye movement data with iris features, they obtain a precision of 79 %.

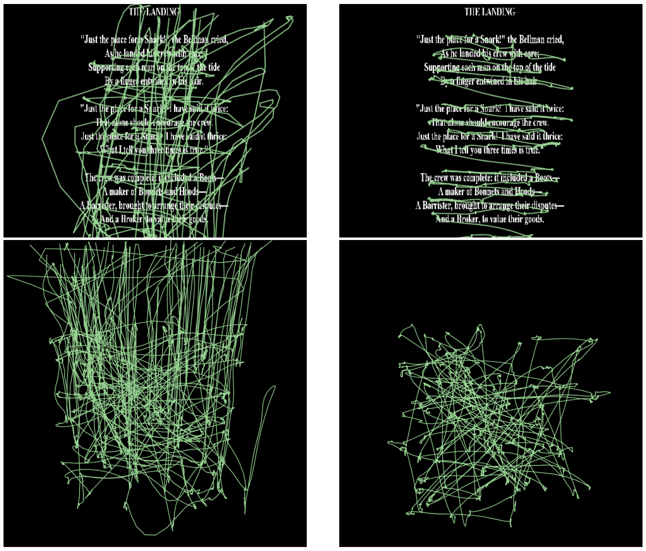

Figure 1: Participants 80 (left column) and 307 (right) reading a text from the TEX dataset (top row) and looking at random dots from the RAN dataset (bottom row).

Pfeuffer et al. [22] compare different behavioral features to identify users in VR. For eye tracking they differentiate between 18 users. They use 8 angular features with a Support Vector Machine and a Random Forests classifier. Their view based features that include eye tracking achieve an accuracy of 24 %.

APPLIED METHODS

We use the Velocity Threshold (I-VT) algorithm [24] to segment each eye movement recording into a sequence of fixations and saccades, which is used in many works, for instance, in [12, 13, 20]. In order to replicate the results of George and Routray [10], we use the version of I-VT from their paper, and the same parameters as mentioned in their paper.

Radial Basis Function Networks

Here, we give a short recap of the Radial Basis Function Networks [3], which is used by George and Routray [10]. The proposed network consists of a fully connected, hidden layer (after the input layer) with $C \cdot K$ neurons, where C is the number of classes and K is a hyperparameter. A so-called radial basis function activates each neuron in the hidden layer:

$$\varphi(x) = e^{-\beta\|x-\mu\|^2} \quad \text{with} \quad \beta = \frac{1}{2\sigma^2}$$

where μ and σ are found for all neurons with a training procedure as follows. Their class labels partition all training data X. For each subset X_c, $c =$ class ID, we generate K seed vectors $\mu_{c,i}$, $i = 1,\ldots,K$, using k-means clustering. These cluster centers are the mean of all elements from X_c that belong to the corresponding cluster. Then, $\sigma_{c,i}$ is computed as the mean Euclidean distance of all elements in each cluster i to $\mu_{c,i}$. We weigh the output vector from the hidden layer $\varphi(x)$ with length $K \cdot C$ in a fully connected way to produce a prediction vector y of length of C. The weights matrix W has a dimension of $K \cdot C \times C$. During the training, W can be learned either by gradient descent or using the Moore-Penrose pseudoinverse to minimize the least squares error between the output vector and the one-hot encoding of the training labels. Before the actual training, we perform normalization on all the features: for each feature, we compute the mean and standard deviation over the training dataset; then, we subtract the mean from that feature and divide by the stddev in the whole dataset. This ensures that each feature's values are in the same range and, therefore, contributed equally to the classification.

Random Decision Forests

Random Decision Forests are a powerful and universal classification method [2], combining multiple decision trees as weak classifiers and use bagging [1] and random subspaces [11] for training. The randomness avoids the training to get stuck at a local minimum, which improves the predictive accuracy and controls over-fitting.

State-of-the-Art Eye Movement Biometrics

To the best of our knowledge, the method proposed by George and Routray [10] is currently the best performing method for eye movement biometrics. Their method ("RBFN") works as follows: Gaze trajectories are encoded as a list of x and y gaze angles. They use the I-VT algorithm to split each trajectory into sequences of saccades and fixations. For saccades and fixations, they extract different features respectively and train two independent classifiers. For these classifiers, each saccade or fixation is one training sample with the trajectory's participant as the label. After training, for the class prediction of a trajectory, they split the trajectory and obtain the corresponding features as it was done during the training. The classifiers produce two lists of class probabilities, which they average into one vector with one prediction probability per class. From this vector, they choose the class with the maximum probability as the prediction for the trajectory.

As classifiers they use radial base function networks. They set $K = 32$ and estimate the weight matrix W by calculating the Moore-Penrose pseudoinverse as in the original paper.

NEW METHODS

In our work, we propose a new method ("RDF") that is similar to RBFN. Instead of radial base function networks we use Random Decision Forest. From the randomness, we expect RDF to generalize better than RBFN. Further, Random Decision Forests are inherently parallel and therefore fast to train, they have an excellent baseline performance without hyperparameter tuning, and their successful use in all kinds of application domains attests a powerful performance. We use precisely the same features as for RBFN and the same protocol for evaluation. While Random Decision Forests depend on several parameters, such as the number of trees, the number of features considered at each split, the maximum tree depth, and others, we found by preliminary experiments that most of the default parameters from the *scikit-learn* software package [21] work very well (no limit to the maximum tree depth, at least two samples per split, consider $\sqrt{F}$ features at each split, where F is the length of the feature vector). The only parameter we choose ourselves is the number of trees, which we set to 400 for all our experiments.

Additionally, we propose two further methods ("RBFN-all" and "RDF-all"). These work, as described above, with the only difference that more features are added. Those features are mentioned by George and Routray [10], but not used (they are certain statistical features indicated with an "N" in Table 2 of that paper). We question the reason for this omission and thus evaluate both classifiers with all features as well. We suspect that the individual features do not add much to the classification accuracy, whereas the combination of all features can increase the performance.

EXPERIMENTS

In this section, we describe our two experiments. First, we replicate the results by George and Routray [10], where they identify individuals from the BioEye 2015 dataset. Here, we extend their results by changing their feature selection and their classifier. We show how the variants perform when stimuli are varied, i.e., we train on TEX and test on RAN, and vice versa. Second, on the MIT dataset, we analyze how the classification accuracy depends on the number of training data as well as the amount of testing data used.

Datasets

The first dataset in our experiments is from the *BioEye* 2015 competition [23][1]. The second dataset is the *MIT dataset* [14][2] The BioEye 2015 dataset contains data for two different stimuli, obtained from 153 participants, whose task was (1) to read a poem, and (2) to observe a randomly moving dot. For stimulus 1, there are two 60 second recordings per participant. For stimulus 2, there are again two recordings, each of length 100 seconds. All sessions were recorded with an EyeLink-1000 eye-tracker (1000 Hz) but provided with 250 Hz. The participants comprise of males and females between the ages of 18 and 46. During the recordings, each subject was positioned at a distance of 550 mm from a computer screen of size 474 x 297 mm and screen resolution of 1680 x 1050 pixels. The head of the subjects was stabilized with the help of a chin rest to mitigate the potential eye tracking artifacts caused by significant head movements.

[1] The data was kindly provided to us by Oleg Komogortsev.

[2] As of this writing, it can be obtained from http://saliency.mit.edu/datasets.html.

The MIT dataset has 15 participants (male and female, aged 18 to 35). Each participant was recorded when looking at an image for 3 seconds, for a total of 1003 images per participant (with a 1-second gray screen between images). The images where views on a 19-inch computer screen (with resolution 1280 x 1024) in a dark room and using a chin rest, situated at a distance of approximately two feet from the screen. An ETL 400 ISCAN eye-tracker was used (240 Hz).

Metric

In our evaluation, we report classification accuracy (the number of correct classifications divided by test cases). Accuracy allows easy comparison to the baseline of change. It is valid in all our test cases, as in all our evaluations, we have balanced class distributions. Further, it allows us to directly compare our results to not only state of the art but all relevant methods mentioned in related work. When we compare accuracies, we report the absolute difference in percentage points (pp.). The absolute difference is more intuitive and makes it easier to compare gains in different ranges. For example, we report that the accuracy between 5 % and 20 % has increased by 15 pp. in contrast to 400 %.

Experiments on the BioEye 2015 Dataset

In our first experiment, we replicate the results from George and Routray [10]. As the original code was not published and is no longer available from the authors, we re-implement the RBFN classifier.

From the BioEye dataset, we use one of the two provided sessions from each participant for training and the other one for evaluation. Thereby, we follow the procedure described by George and Routray [10] and avoid biasing the evaluation with data from the same sample. From our tests, we determined that George and Routray must have used session 2 for training and session 1 for testing.

Further, we repeat the same evaluation with our RDF method. We hypothesize that RDF performs equally well or better than RBFN. We base this assumption on the fact that Random Decision Forests are generally scale-invariant and, due to their randomness, generalize well.

The competition only evaluates the algorithm on one kind of stimulus at a time. We add another evaluation where we train both, RBFN and RDF classifiers on the poem and evaluate them on the random dot test data and vice versa. Our question is, whether the classifier learns task-specific or independent features.

Experiments on the MIT Dataset

For real-world applications of gaze biometrics, it is relevant to estimate the amount of trajectory data that is needed per user for a reliable identification. With our second experiment, we investigate how many segments of an eye tracking trajectory is needed to identify individual participants. We consider both the amount of training data and how much test data is required. Since the BioEye dataset does not contain enough data to vary the amount of training and test data in a meaningful way, we use the much larger MIT dataset for this experiment. With more than 50 minutes of gaze data per participant, we vary the amount of data used for training and testing as follows. We will call the tracking data that was obtained while a participant was looking at one image a *sample*. To vary the amount of training data, we use segments from multiple samples of each participant. In the evaluation, we average the class probability predictions from multiple test samples.

	Ran	Tex	Tex⇒Ran	Ran⇒Tex
RDF	84.3	81.7	14.4	5.2
RBFN	88.9	85.0	5.2	4.6
RDF-all	90.9	85.6	23.5	7.8
RBFN-all	94.1	90.8	11.8	14.4

Table 1: Accuracy of different classifiers and different cases in percent. Columns with arrows (⇒) denote transfer cases; columns without arrows use the same dataset for training and testing.

We investigate the performance of all four different classifiers with a varying number of samples used for training. In the first scenario, we fix the number of testing samples per participant to 30. This is equal to 90 seconds of continuous trajectory and therefore is similar to the training data from the BioEye dataset (60 and 100 seconds for RAN and TEX, respectively). At the same time, we vary the number of training samples up to 700.

In the second scenario, we fix the number of training samples per participant. Again, we choose 30 samples for the same reason as above. Here, we limit the maximum number of test images to 300, as initial tests indicated not much difference when using more.

To test the robustness of our evaluation, we repeat all experiments 5 times on a random subset of the dataset and thereby get a mean accuracy as well as standard deviation. In all experiments, the hyperparameters of the classifiers were the same. Only for the tests with less than 30 training samples, we had to reduce parameter K for RBFN since the k-means clustering can only return as many clusters as there are training samples. We choose K equal to the number of training samples for less than 20 samples and $K = 32$ for 20 samples and more.

RESULTS

Task-Independence

First, we verify our re-implementation of the RBFN algorithm from George and Routray [10] with the same configuration and on the same part of the BioEye dataset. We obtain exactly the same results as the original paper (see Section 5.3):[3] the deviation between ours and their results is at most 1 % for both datasets; for the TEX stimulus, our implementation achieves an accuracy of 84.97 %, while the authors report 85.62 %. On the RAN stimulus, we achieve 88.89 % accuracy while they report 89.54 %. We attribute the variation of less than 1 % to the different seeds used for the k-means clustering and numerical instabilities in the calculation of the pseudo-inverse.

[3] We can only compare our results to Table 5 in [10] and not the competition summary [23], as we do not have access to the evaluation data.

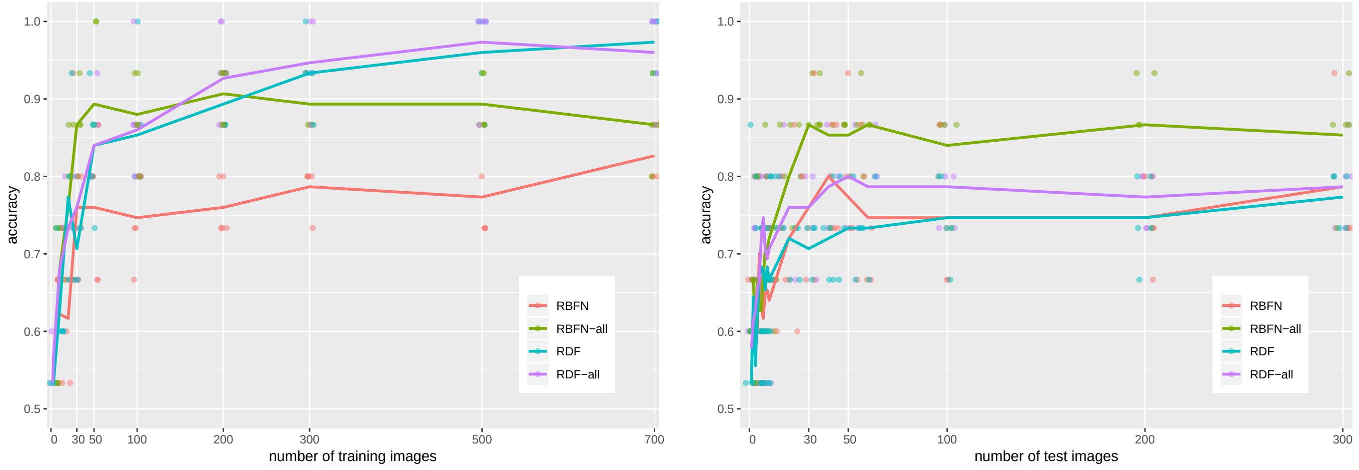

Figure 2: Accuracy of different methods, by number of training images (left) and by the number of testing images (right). The number of testing or training images is fixed to 30, respectively. The lines mark the methods' mean values.

We then replace the RBFN by an RDF, while using exactly the same features in the BioEye dataset. This causes the accuracy to drop by 4.6 pp. for RAN, and by 3.3 pp. for TEX.

In contrast to George and Routray [10], we find that the performance of both RBFN and RDF with the complete feature set (*RBFN-all* and *RDF-all* in Section 5.3) are better than with the restricted feature set used in the original paper. On average, the accuracy of RBFN increases by 5.6 pp. and for RDF by 5.2 pp.

For all previous experiments, the training and testing stimulus were the same, either both text or both random dots. When we train on the trajectories from the TEX stimulus and evaluate the performance on RAN, the accuracy drops significantly. The baseline method drops from 85 % to 5.2 %, and RDF from 81.7 % to 14.4 %. When we train on RAN and evaluate on TEX, the accuracy drops even more, from 88.9 % to 4.6 % for RBFN and from 84.3 % to 5.2 % for RDF. RDF has the highest performance in this setting with 23.5 % accuracy, where it is trained on TEX and evaluated on RAN. Training on RAN and evaluating on TEX works better with RBFN where it achieves 14.4 % accuracy compared to 7.8 % with RDF.

Relationship between trajectory length and robustness

When we vary the number of training samples but fix the number of samples for testing to 30, we see that accuracy increases in general (see Fig. 2, upper). With 30 training images, RBFN with all features performs best with an average accuracy and standard derivation of 86.67 %(4.71) over the 5 runs with randomly selected sample subsets. With 300 training samples, RDF, both with the full feature set as well as with the smaller subset, starts to outperform both RBFN methods. Here, RDF with all features has an accuracy of 94.67 %(5.58) in our experiments.

Similarly to the previous experiment, when we use an increasing number of samples for testing, but fix the number of training samples to 30, we see that the performance of all methods increases (see Fig. 2, lower). For all methods except RBFN with the limited features, the performance does not increase with more than 40 test samples.

While we only show plots for 30 training and testing samples, we also analyzed the other combinations. The general form of the curves is the same – only the accuracy increase with more training samples.

DISCUSSION

Our results show that none of the tested methods is strongly task-independent. Learning on one type of stimulus does not lead to a reasonable identification rate on a different type of stimulus. Remarkably, the transfer (if any) is highly non-symmetric for RDF: For instance, when training on TEX and evaluating on RAN, the classification performance is three times as high as vice versa. By contrast, the transfer performance of RBFN is nearly symmetric, albeit pretty bad too. This finding could suggest further avenues for research on the transfer performance of classifiers.

Our findings also suggest that random stimuli lead to more user-specific eye movements, which results in higher accuracy. In all our tests, when training on RAN and evaluating on RAN the classification performance is better than when training on TEX and evaluating on TEX.

Regarding the transfer case, our initial assumption was that the RAN⇒TEX performance would be better than the TEX⇒RAN performance, because the simple RAN case performed better than the simple TEX case. Surprisingly, when training on TEX and evaluating on RAN, the performance is better than the other way around, at least when using the RDF classifiers.

In all our experiments, the models using all features clearly win, both for weakly and strongly task-independent classifications. This is surprising since George and Routray [10] performed a special feature selection step. Also, using 100-

fold cross-validation on a subset of 50 participants, we find that the standard deviation of all methods varies by 0.76 pp. on average. Thus, our findings suggest that while adding single features does not contribute much to the performance of the model, the combination of several weak features improves the accuracy a lot.

Regarding the trajectory length needed for gaze based user identification, our results suggest that only 120 seconds of trajectory data for testing is necessary for the maximum identification rate. At the same time, more training data generally increases the accuracy. Even though the RBFN methods work well with fewer training samples, their performance is worse than the RDF methods with more than 200 training samples. Even with only 90 seconds of trajectory for training and evaluation, the accuracy already is at 86.7 %. This short application time enables the use of gaze biometrics in continuous identification scenarios.

In all non-transfer cases, the RBFN classifier performs slightly better than RDF. For the transfer cases, there is no clear winner; in two scenarios, RDF performs better than RDF and vice versa in the other two scenarios. However, the mean accuracy of RDF overall transfer cases is 3.8 pp. higher than the mean accuracy of RBFN. We found anecdotal evidence that small perform gains could be achieved by more extensive hyperparameter tuning.

CONCLUSION

In this work, we have investigated the performance of three new methods and compared their performance to the state-of-the-art regarding their robustness against varying stimuli and trajectory length.

The best of our new methods improves the state-of-the-art performance by 5.2 pp. for a common dataset with equal training and evaluation tasks. Furthermore, we evaluate and compare the four methods on a popular dataset that has never been used for gaze biometrics before. We find that when training and evaluation are done weakly task-independent on this dataset, our method achieves 86.7 % accuracy with only 30+30 samples (trainging+testing), each 3 seconds duration. With 300 training samples, our methods can achieve even 94.7 %.

None of the tested methods is capable of strong task-independent user identification. Especially, our results suggest that transfer learning is highly non-symmetric. Training on text and evaluating on random dots performs three times better than the other way around.

Finally, we make all our source code, including our re-implementation of the method by George and Routray [10], publicly available: https://cgvr.cs.uni-bremen.de/research/smida_ml/.

In the future, we believe more work is needed on methods that perform well with strongly task-independent settings. Especially the difference in the performance of the different classifiers we observed seems to be a promising start. Task-independence could also be combined with multi task learning similar to the approach from Kaiser et al. [15]. Further, for security and authentication applications, future methods should investigate possibilities to reject participants that are unknown to the classifier.

ACKNOWLEDGMENTS

We thank George and Routray and, in particular, Anjith George for feedback wrt. our implementation. We thank Oleg Komogortsev for providing the BioEye 2015 dataset.

REFERENCES

[1] Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140. DOI: http://dx.doi.org/10.1007/BF00058655

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. DOI: http://dx.doi.org/10.1023/A:1010933404324

[3] David S. Broomhead and David Lowe. 1988. Multivariable Functional Interpolation and Adaptive Networks. *Complex Systems* 2, 3 (1988).

[4] Ali Darwish and Michel Pasquier. 2013. Biometric identification using the dynamic features of the eyes. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*. 1–6. DOI:http://dx.doi.org/10.1109/BTAS.2013.6712724

[5] Ali Alhaj Darwish. 2013. *Biometric Identification Based on Eye Movements and Iris Features Using Task-Driven and Task-Independent Stimuli*. Master's thesis. American University of Sharjah.

[6] Nastaran Maus Esfahani. 2016. A Brief Review of Human Identification Using Eye Movement. *Journal of Pattern Recognition Research* 11, 1 (2016), 15–24. DOI: http://dx.doi.org/10.13140/RG.2.1.3466.3924

[7] Lee Friedman, Mark S. Nixon, and Oleg V. Komogortsev. 2017. Method to assess the temporal persistence of potential biometric features: Application to oculomotor, gait, face and brain structure databases. *PLOS ONE* 12, 6 (06 2017), 1–42. DOI: http://dx.doi.org/10.1371/journal.pone.0178501

[8] Chiara Galdi and Michele Nappi. 2019. *Eye Movement Analysis in Biometrics*. Springer Singapore, Singapore, 171–183. DOI: http://dx.doi.org/10.1007/978-981-13-1144-4_8

[9] Chiara Galdi, Michele Nappi, Daniel Riccio, and Harry Wechsler. 2016. Eye movement analysis for human authentication: a critical survey. *Pattern Recognition Letters* 84 (2016), 272–283. DOI: http://dx.doi.org/10.1016/j.patrec.2016.11.002

[10] Anjith George and Aurobinda Routray. 2016. A score level fusion method for eye movement biometrics. *Pattern Recognition Letters* 82 (2016), 207–215. DOI: http://dx.doi.org/10.1016/j.patrec.2015.11.020

[11] Tin Kam Ho. 1995. Random decision forests. In *Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I*. 278–282. DOI: http://dx.doi.org/10.1109/ICDAR.1995.598994

[12] Corey Holland and Oleg V. Komogortsev. 2011. Biometric identification via eye movement scanpaths in reading. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*. 1–8. DOI: http://dx.doi.org/10.1109/IJCB.2011.6117536

[13] Corey D. Holland and Oleg V. Komogortsev. 2012. Biometric verification via complex eye movements: The effects of environment and stimulus. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012*. 39–46. DOI: http://dx.doi.org/10.1109/BTAS.2012.6374556

[14] Tilke Judd, Krista A. Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. 2106–2113. DOI: http://dx.doi.org/10.1109/ICCV.2009.5459462

[15] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One Model To Learn Them All. *arXiv:1706.05137* (2017). http://arxiv.org/abs/1706.05137

[16] Pawel Kasprowski, Oleg V. Komogortsev, and Alex Karpov. 2012. First eye movement verification and identification competition at BTAS 2012. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012*. 195–202. DOI: http://dx.doi.org/10.1109/BTAS.2012.6374577

[17] Pawel Kasprowski and Józef Ober. 2004. Eye Movements in Biometrics. In *Biometric Authentication, ECCV 2004 International Workshop, BioAW 2004, Prague, Czech Republic, May 15, 2004, Proceedings*. 248–258. DOI: http://dx.doi.org/10.1007/978-3-540-25976-3_23

[18] Tomi Kinnunen and Haizhou Li. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* 52, 1 (2010), 12–40. DOI: http://dx.doi.org/10.1016/j.specom.2009.08.009

[19] Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. 2010. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA 2010, Austin, Texas, USA, March 22-24, 2010*. 187–190. DOI: http://dx.doi.org/10.1145/1743666.1743712

[20] Anneli Olsen and Ricardo Matos. 2012. Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In *Proceedings of the 2012 Symposium on Eye-Tracking Research and Applications, ETRA 2012, Santa Barbara, CA, USA, March 28-30, 2012*. 317–320. DOI:http://dx.doi.org/10.1145/2168556.2168625

[21] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[22] Ken Pfeuffer, Matthias J. Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 110, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300340

[23] Ioannis Rigas and Oleg V. Komogortsev. 2017. Current research in eye movement biometrics: An analysis based on BioEye 2015 competition. *Image Vision Comput.* 58 (2017), 129–141. DOI: http://dx.doi.org/10.1016/j.imavis.2016.03.014

[24] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2000, Palm Beach Gardens, Florida, USA, November 6-8, 2000*. 71–78. DOI:http://dx.doi.org/10.1145/355017.355028