EEG-Features for Generalized Deepfake Detection

Arian Beckmann*[1] (arian.beckmann@hhi.fraunhofer.de)

Tilman Stephani*[2] (stephani@cbs.mpg.de)

Felix Klotzsche[2] (klotzsche@cbs.mpg.de)

Yonghao Chen[2] (cheny@cbs.mpg.de)

Simon M. Hofmann[2] (simon.hofmann@cbs.mpg.de)

Arno Villringer[2] (villringer@cbs.mpg.de)

Michael Gaebler[2] (gaebler@cbs.mpg.de)

Vadim Nikulin[2] (nikulin@cbs.mpg.de)

Sebastian Bosse[1] (sebastian.bosse@hhi.fraunhofer.de)

Peter Eisert[1,3] (peter.eisert@hhi.fraunhofer.de)

Anna Hilsmann[1] (anna.hilsmann@hhi.fraunhofer.de)

[1]Fraunhofer Heinrich-Hertz-Institute, Einsteinufer 37, 10587 Berlin, Germany

[2]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1A, 04103 Leipzig, Germany

[3]Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Abstract

Since the advent of Deepfakes in digital media, the development of robust and reliable detection mechanism is urgently called for. In this study, we explore a novel approach to Deepfake detection by utilizing electroencephalography (EEG) measured from the neural processing of a human participant who viewed and categorized Deepfake stimuli from the FaceForensics++ datset. These measurements serve as input features to a binary support vector classifier, trained to discriminate between real and manipulated facial images. We examine whether EEG data can inform Deepfake detection and also if it can provide a generalized representation capable of identifying Deepfakes beyond the training domain. Our preliminary results indicate that human neural processing signals can be successfully integrated into Deepfake detection frameworks and hint at the potential for a generalized neural representation of artifacts in computer generated faces. Moreover, our study provides next steps towards the understanding of how digital realism is embedded in the human cognitive system, possibly enabling the development of more realistic digital avatars in the future.

Keywords: EEG; Deepfake; perception; realism

Introduction

The pope wearing puffer jackets, Tom Cruise doing magic tricks on TikTok, Donald Trump being arrested by the police. Visual ”proof” of these situations went viral on social media – yet they never happened. Deepfake technology can readily delude a viewer’s beliefs about what a certain person says, does, and looks like. To maintain the delineation between truth and lie, it is hence of paramount importance for modern society to be able to identify and counteract such Deepfake technologies. A common approach for developing Deepfake detectors involves training convolutional neural networks (CNNs) to detect the spatio-temporal artifacts appearing in Deepfakes (Rössler et al., 2019; Wang, Bao, Zhou, Wang, & Li, 2023). Although they perform well on benchmark datasets, these detectors often struggle to generalize to new manipulation domains that omit unfamiliar artifacts (Beckmann, Hilsmann, & Eisert, 2023). Another possibility may be to add the human in the detection loop, and, more specifically, take advantage of the high-dimensional information contained in neurophysiological measurements of the perceptual and cognitive processing of Deepfake stimuli that leads or does not lead to the subjective percept of a fake human face. Previous work demonstrated that fake videos can be discriminated from genuine ones if the observer is familiar with at least one of the displayed persons (Tauscher, Castillo, Bosse, & Magnor, 2021). Moreover, (Moshel, Robinson, Carlson, & Grootswagers, 2022) demonstrated that GAN generated images can be decoded by people’s neural activity. In this proof-of-concept study, we test whether human electroencephalography (EEG) can inform the detection of deepfaked faces and whether it allows – in contrast to naively trained CNNs – to generalize across different Deepfake generation methods.

*Equal Contribution.

Methods

Stimuli For our stimulus set, we utilize the FaceForensics++ (Rössler et al., 2019) benchmark dataset as it contains forged facial videos originating from various different manipulation methods. We use stimuli of two different fake methods, ”Deepfakes” (DF)[2] and ”FaceSwap” (FS), along with their respective original counterparts. We chose these two fake methods as they produce different characteristic artifacts that are not difficult to identify. Per category, 500 videos were selected of which we randomly selected 8 (16) frames as fake (real) images, in total resulting in 16,000 images with balanced fake/real labels. Note that we selected the videos such that, per category, 360 videos belong to the training set and 70 videos to the validation and testing sets respectively, as specified in Rössler et al. (2019).

Experimental procedure The images were presented in random order on a computer screen to a human observer (co-author, male, 22 years), while measuring EEG. Each image was centrally presented for 350 ms, followed by a blank screen with a fixation target for 350 ms. Subsequently, in a subset of the trials, the participant was tasked to indicate via button press whether he perceived the stimulus as real or fake (within 1000 ms). The tasks appear at random, in 12.5% of the trials, as to avoid the participant expecting the task. In the remaining 87.5% of trials, the experiment continued with stimulus presentation of the following trial. The whole experiment consisted of 160 blocks with 100 trials each, amounting to a total duration of around 4 hours measurement time.

EEG setup and preprocessing EEG data were recorded from 63 Ag/AgCl electrodes at a sampling frequency of 1000 Hz with a NeurOne Tesla EEG system (Bittium, Oulu, Finland). A built-in band-pass filter between 0.16 and 250 Hz was used. Electrodes were placed according to the international 10-10 system, mounted in an elastic cap (EasyCap, Hersching, Germany). FCz served as reference and CPz as ground electrode. During offline processing, the EEG data were bandpass filtered between 0.5 and and 40 Hz, re-referenced to an average reference, and ICA served to remove eye blink, eye movement, and heart artifacts. Subsequently, the data were cut into epochs from -300 to +700 ms relative to stimulus onset, including a baseline correction from -200 to 0 ms. Epochs with values exceeding +-400 µV at any electrode were excluded from further analysis (n=4).

Deepfake classification We construct the following experiment to analyze whether the recorded EEG data can inform Deepfake detection, particularly assessing the potential of a generalized artifact representation: For each video in each category, we average over all recorded trials to obtain a denoised sample. Thus, we are left with 1500 denoised samples, distributed evenly across the three categories – ”Deepfakes” (DF), ”Faceswap” (FS) and ”real”. Note that for denoising real samples, we use 8 instead of 16 recorded trials to ensure a similar signal quality between real and fake samples. Then, we form training, validation and testing sets according to Rössler et al. (2019). We process the data in two different variations, resulting from an extensive ablation study. We denote these variants by **V1** and **V2** respectively. Both variations ignore the first 300 ms pre stimulus onset and, ultimately, merge the spatial and temporal dimensions before applying dimensionality reduction. For **V1**, we use all remaining data and reduce its dimensionality via PCA with 64 components. Concerning **V2**, we split the data with respect to the remaining 700 ms along the spatial and temporal dimensions into chunks of length 100 ms per electrode, resulting in 441 chunks (63 electrodes X 7 100 ms intervals). For each chunk, we train a separate binary support vector classifier (SVC) to discriminate between neural signals representing real or deepfaked stimuli. Subsequently, we evaluate the classifiers on the validation sets of the respective chunks. The top 100 chunks by validation F1-score were selected, consolidated and further reduced by ICA with 128 components.

The following training and evaluation process is performed separately for data pre-processed according to both **V1** and **V2**: We train an SVC (with default parameters in scikit-learn) to discern real and fakes on training data containing DF and ”real” and evaluate it on the respective test set. Moreover, to test out-of-domain detection performance, we evaluate the classifier on the testing subset of FS. Likewise, we perform the experiment including FS instead of DF in the training set, while still testing on both fake subsets. The results of our experiments are shown in Table 1. DF→FS refers to the case

[2] The general term Deepfake originates from this seminal forgery method, to avoid confusion we refer to it solely as DF.

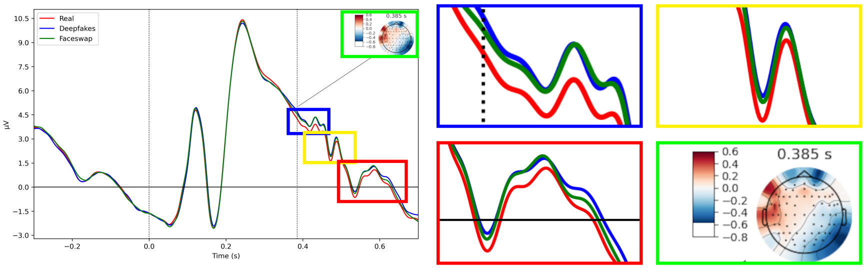

Figure 1: Mean EEG responses with respect to the three stimuli classes for electrode PO8 as well as the topography (across all electrodes) of the difference between fake and real images at 385 ms after stimulus onset (green box).

Table 1: Macro F1-Scores for both variations on multiple train-test splits. Bold numbers highlight out-of-domain testing.

Variation	DF→ DF	DF→ FS	FS→ DF	FS→ FS
V1	0.62	**0.58**	**0.59**	0.61
V2	0.61	**0.58**	**0.61**	0.56

in which the train set contains DF and ”real” and the model is evaluated on the testing sets corresponding to FS and ”real”. The other columns follow the same logic.

Results and Discussion

The left-hand side of Figure 1 shows the EEG responses averaged over the respective classes for electrode PO8. The magnified regions on the right-hand side show a significant difference between the responses to the real images and their manipulated counterparts (confirmed by cluster-based permutation testing). Additionally, the green box displays the topography of the difference in the responses to faked (DF and FS) and real images at 385 ms after stimulus onset. These descriptive results tentatively demonstrate that neural processing may contain a generalized representation of artificiality with respect to computer-generated faces. This interpretation obtains further support from the decoding results depicted in Table 1. As can be seen in the third and fourth columns, the classifier is able to produce above chance level performance when confronted with fakes not seen during training. We check the significance of these results by permutation testing against chance-level with 10,000 repetitions for $p = 0.05$. The resulting p-values are $.0309$ and $.0128$ for **V1**, as well as $.0277$ and $.0036$ for **V2** (same order as shown in Table 1). Nonetheless, to further support our hypothesis, we aim to perform more experiments with a wider variety of Deepfakes and more participants in our future work.

Conclusion

In this pilot experiment, we not only demonstrated that features derived from EEG recordings can be used to detect Deepfakes, but also that these features can be utilized for out-of-domain fake detection – hinting at the potential for a generalized representation of artifacts or uncanny content within neural processing signals. For subsequent experiments, we plan to include more high-quality images and to manually add a variety of artifacts, to enable more control and a broader analysis.

Acknowledgments

This work has received funding through the Max Planck-Fraunhofer collaboration project NeuroHum.

References

Beckmann, A., Hilsmann, A., & Eisert, P. (2023). Fooling state-of-the-art deepfake detection with high-quality deepfakes. In *Proceedings of the 2023 acm workshop on information hiding and multimedia security* (p. 175–180). New York, USA. doi: 10.1145/3577163.3595106

Moshel, M. L., Robinson, A. K., Carlson, T. A., & Grootswagers, T. (2022). Are you for real? decoding realistic ai-generated faces from neural activity. In *Vision res. 2022 oct;199:108079.*

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *International conference on computer vision (iccv).*

Tauscher, J.-P., Castillo, S., Bosse, S., & Magnor, M. (2021). Eeg-based analysis of the impact of familiarity in the perception of deepfake videos. In *2021 ieee international conference on image processing (icip)* (p. 160-164). doi: 10.1109/ICIP42928.2021.9506082

Wang, Z., Bao, J., Zhou, W., Wang, W., & Li, H. (2023, June). Altfreezing for more general video face forgery detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 4129-4138).