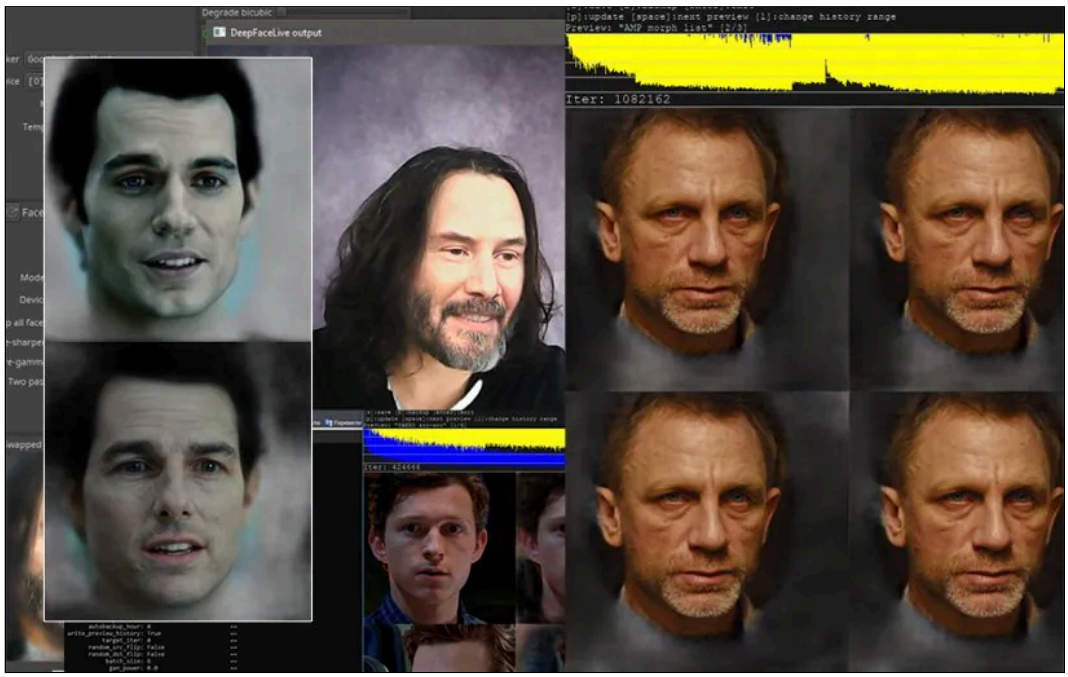
Deepfake Detection Based on Original Human Biometric Traits

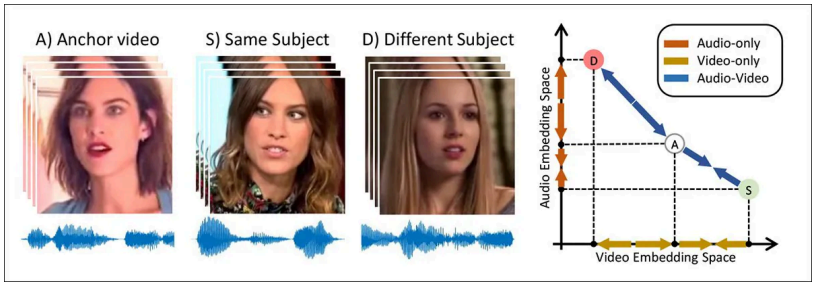
Originally published at https://www.unite.ai/deepfake-detection-based-on-original-human-biometric-traits/



Images produced by deepfakers at the DeepFaceLab Discord Channel

A new paper from researchers in Italy and Germany proposes a method to detect deepfake videos based on biometric face and voice behavior, rather than artifacts created by face synthesis systems, expensive watermarking solutions, or other more unwieldy approaches.

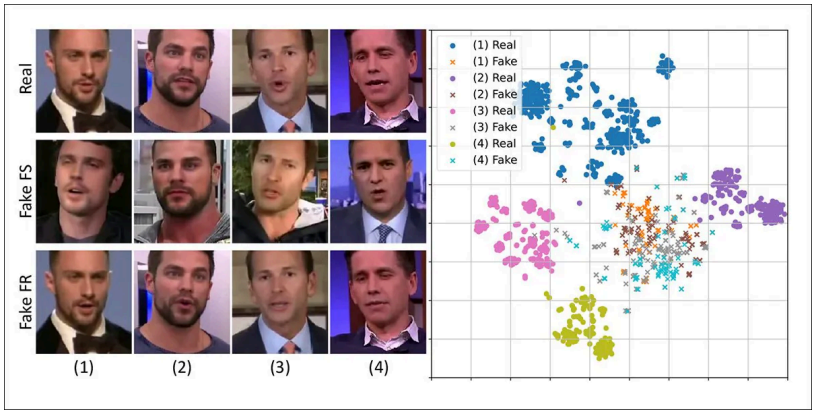
The framework requires an input of 10 or more varied, non-fake videos of the subject. However, it does not require to be specifically trained, retrained or augmented on per-case videos, as its incorporated model has already abstracted the likely vector distances between real and fake videos in a broadly applicable manner.



Contrastive learning underpins the approach of POI-Forensics. Vectors derived from source material on a per-case basis are compared to the same vectors in a p and traits drawn from both video and audio components of the potentially faked footage. Source: https://arxiv.org/pdf/2204.03083.pdf

Titled *POI-Forensics*, the approach relies on movement and audio cues unique to the real individual being deepfaked.

Though such a system could allow completely automated, 'pre-rendered' authentication frameworks for celebrities, politicians, YouTube influencers, and other people for whom a great deal of video material is readily available, it could also be adapted into a framework where ordinary victims of deepfake technologies could potentially have a platform to prove the inauthenticity of attacks against them.



Visualizations of extracted features from genuine and faked videos across four subjects in POI-Forensics, via the t-SNE framework.

The authors claim that POI-Forensics achieves a new state of the art in deepfake detection. Across a variety of common datasets in this field, the framework is reported to achieve an improvement in AUC scores of 3%, 10%, and 7% for high quality, low quality and 'attacked' videos, respectively. The researchers promise to release the code shortly.

	AUC/ACC	pDFDC	DF-TIMIT	FakeAVCel.	KoDF	AVG
High quality	Seferbekov	85.5 / 72.0	95.7 / 87.6	98.6 / 95.0	88.4 / 79.2	92.0 / 83.5
	FTCN	72.3 / 63.9	100.0 / 87.4	84.0 / 64.9	76.5 / 63.0	83.2 / 69.8
	LipForensics	68.7 / 60.0	98.8 / 78.0	97.6 / 83.3	92.9 / 56.1	89.5 / 69.3
	ID-Reveal	91.3 / 80.4	99.0 / 92.8	70.2 / 60.3	87.6 / 63.7	87.0 / 74.3
	POI-Forensics	95.2 / 86.7	99.2 / 85.7	94.1 / 86.6	89.9 / 81.1	**94.6**/ **85.0**
Low quality	Seferbekov	62.6 / 54.0	81.8 / 71.4	61.7 / 58.5	79.6 / 55.9	71.4 / 59.9
	FTCN	51.3 / 50.0	78.3 / 58.9	37.6 / 42.1	68.4 / 58.6	58.9 / 52.4
	LipForensics	46.5 / 45.8	70.9 / 62.1	58.3 / 53.8	85.7 / 52.3	65.3 / 53.5
	ID-Reveal	86.5 / 73.9	93.6 / 75.5	70.8 / 58.9	85.0 / 63.8	84.0 / 68.0
	POI-Forensics	93.9 / 82.0	98.2 / 76.3	94.4 / 88.7	89.0 / 81.5	**93.9**/ **82.1**

POI-Forensics' performance against rival SOTA frameworks pDFDC, DeepFakeTIMIT, FakeAVCelebV2, and KoDF. Training in each case was performed on FaceForensics++ and the authors' own ID-Reveal on VoxCeleb2. Results include high and low quality videos.

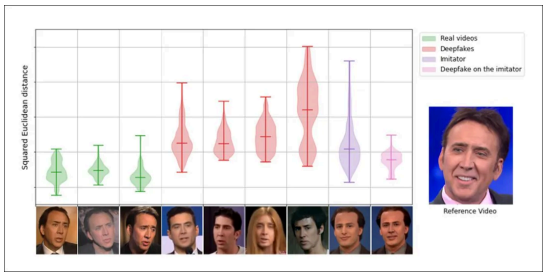
The authors state:

> *'Training is carried out exclusively on real talking-face videos, thus the detector does not depend on any specific manipulation method and yields the highest generalization ability. In addition, our method can detect both single-modality (audio-only, video-only) and multi-modality (audio-video) attacks, and is robust to low-quality or corrupted videos by building only on high-level semantic features.'*

The new paper, which incorporates elements of some of the authors' vision-based *ID-Reveal* project of 2021, is titled *Audio-Visual Person-of-Interest DeepFake Detection*, and is a joint effort between the University of Federico II in Naples and the Technical University of Munich.

The Deepfake Arms Race

To defeat a detection system of this nature, deepfake and human synthesis systems would require the capability to at least simulate visual and audio biometric cues from the intended target of the synthesis – technology which is many years away, and likely to remain in the purview of costly and proprietary closed systems developed by VFX companies, which will have the advantage of the cooperation and participation of the intended targets (or their estates, in the case of simulation of deceased people).



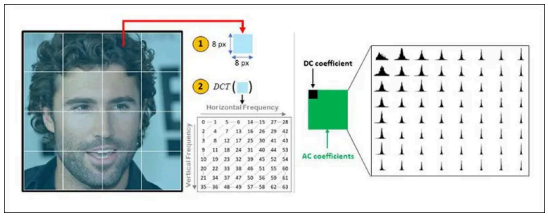
The authors' previous approach, ID-Reveal, concentrated entirely on visual information. Source: https://arxiv.org/pdf/2012.02512.pdf

Successful and popular deepfake methods such as FaceSwap and DeepFaceLab/Live currently have zero capacity to create such granular biometric approximations, relying at best on talented impersonators on whom the faked identity is imposed, and much more commonly on apposite in-the-wild footage of 'similar' people. Nor does the structure of the core 2017 code, which has little modularity and which remains the upstream source for DFL and FaceSwap, make adding this kind of functionality feasible.

These two dominant deepfake packages are based on autoencoders. Alternative human synthesis methods can use a Generative Adversarial Network (GAN) or Neural Radiance Field (NeRF) approach to recreating human identity; but both these lines of research have years of work ahead even to produce fully photorealistic human video.

With the exception of audio (faked voices), biometric simulation is very far down the list of challenges facing human image synthesis. In any case, reproducing the timbre and other qualities of the human voice does not reproduce its eccentricities and 'tells', or the way that the real subject uses semantic construction. Therefore even the perfection of AI-generated voice simulation does not solve the potential firewall of biometric authenticity.

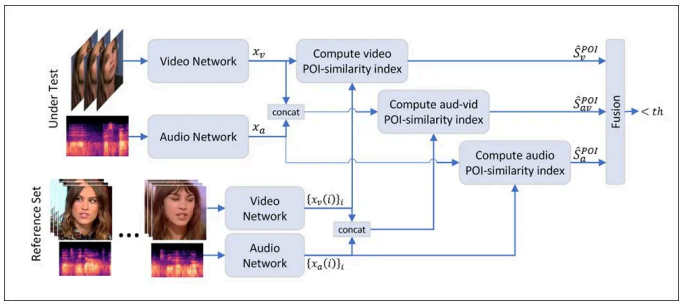
At Arxiv alone, several deepfake detection strategies and innovations are released each week. Recent approaches have hinged on Voice-Face Homogeneity, Local Binary Pattern Histogram (FF-LBPH), human perception of audio deepfakes, analyzing face borders, accounting for video degradation, and 'Forensic Ballistics' – among many others.



Segmented histogram analysis is among the latest techniques offered to improve deepfake detection. Source: https://arxiv.org/pdf/2203.09928.pdf

Approach, Data and Architecture

POI-Forensics takes a multi-modal approach to identity verification, leveraging soft biometrics based on visual and audio cues. The framework features separate audio and video networks, which ultimately derive characteristic vector data that can be compared to the same extracted features in a potential deepfake video under study.



The conceptual architecture of POI-Forensics.

Both separate (audio or video) and fusion analysis can be effected on target clips, arriving finally at a POI similarity index. The contrastive loss function employed is based on a 2021 academic collaboration between Google Research, Boston University, Snap Inc. , and MIT.

The base dataset was divided on a per-identity basis. 4608 identities were used for training, with 512 remaindered for validation. The 500 identities used in FakeAVCelebV2 (a testing candidate, see below) were excluded in order to obtain non-polarized results.

The two networks were trained for 12 epochs at an unusually large batch-size of 2304 batches per epoch, with each batch comprised of 8×8 video segments – 8 segments for 8 different identities. The Adam optimizer was used with decoupled weight decay at a learning rate of 10^{-4}, and a weight decay of 0.01.

Testing and Results

The deepfake datasets tested for the project were the preview DeepFake Detection Challenge dataset, which features face-swaps across 68 subjects, from which 44 identities were selected that have more than nine related videos, totaling 920 real videos and 2925 fake videos; DeepFake-TIMIT, a GAN-based dataset featuring 320 videos of 32 subjects, totaling 290 real videos and 580 fake videos of at least four seconds' duration; FakeAVCelebV2, comprising 500 real videos from Voxceleb2, and approximately 20,000 fake videos from various datasets, to which fake cloned audio was added with SV2TTS for compatibility; and KoDF, a Korean deepfake dataset with 403 identities faked through FaceSwap, DeepFaceLab, and FSGAN, as well as three First Order Motion Models (FOMM).

The latter also features audio-driven face synthesis ATFHP, and output from Wav2Lip, with the authors using a derived dataset featuring 276 real videos and 544 fake videos.

Metrics used included area under the receiver operating characteristic curve (AUC), and an approximated 10% 'false alarm rate', which would be problematic in frameworks that incorporate and train on fake data, but which concern is obviated by the fact that POI-Forensics takes only genuine video footage as its input.

The methods were tested against the Seferbekov deepfake detector, which achieved first place in the Kaggle Deepfake Detection Challenge; FTCN (Fully Temporal Convolution Network), a collaboration between China's Xiamen University and Microsoft Research Asia; LipForensics, a joint 2021 work between Imperial College London and Facebook; and ID-Reveal, a prior project of several of the new paper's researchers, which omits an audio aspect, and which uses 3D Morphable Models in combination with an adversarial game scenario to detect fake output.

In results (see earlier table above), POI-Forensics outperformed reference leader Seferbekov by 2.5% in AUC, and 1.5% in terms of accuracy. Performance was more competitive over other datasets at HQ.

However, the new approach demonstrated a notable lead over all competing reference methods for low-quality videos, which remain the likeliest scenario in which deepfakes are prone to fool casual viewers, based on 'real world' contexts.

The authors assert:

> *'Indeed, in this challenging scenario, only identity-based approaches keep providing a good performance, as they rely on high-level semantic features, quite robust to image impairments.'*

Considering that PIO-Forensics uses only real video as source material, the achievement is arguably magnified, and suggests that using the native biometric traits of potential deepfake victims is a worthwhile road forward to escaping the 'artifact cold war' between deepfake software and deepfake detection solutions.

In a final test, the researchers added adversarial noise to the input, a method that can reliably fool classifiers. The now venerable fast gradient sign method still proves particularly effective, in this regard.

Predictably, adversarial attack strategies dropped the success rate across all methods and datasets, with AUC descending in increments between 10% to 38%. However, only POI-Forensics, and the authors' earlier method ID-Reveal were able to maintain reasonable performance under this attack scenario, suggesting that the high-level features associated with soft biometrics are extraordinarily resistant to deepfake detection evasion.

The authors conclude:

> *'Overall, we believe our method is a first stepping stone; in particular, the use of higher-level semantic features is a promising future avenue for future research. In addition, the multimodal analysis could be further enriched by including more information from other domains such as textual data.'*

First published 8th April 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai