Interpretable deepfake detection using facial dynamics

Project examining facial behavioural dynamics as interpretable signals for deepfake detection, and analysing discourse to evaluate UK policing and policy responses to deepfake-related harms.

Key details

Lead institution	P-ACE LAB (University of Leicester, Aston University and University of Birmingham)
Principal researcher(s)	Timothy Murphy tim.murphy@bristol.ac.uk
Police region	West Midlands
Collaboration and partnership	• Dr Hélio Cuve, University of Bristol • Prof Jennifer Cook, University of Birmingham • Centre for National Training and Research Excellence in Understanding Behaviour (CENTRE-UB) • National Police Chiefs' Council (NPCC)
Level of research	PhD
Project start date	October 2024
Date due for completion	September 2028

Research context

Deepfakes, synthetic media generated using artificial intelligence (AI) that can convincingly depict individuals saying or doing things they never did, pose growing challenges for law enforcement,

criminal justice and public safety. Advances in generative AI have lowered barriers to creation, with documented uses including non-consensual intimate imagery (NCII), fraud, political disinformation and the fabrication of evidential material (Chesney and Citron, 2019).

Existing automated detection approaches predominantly treat deepfake identification as a binary classification problem using deep neural networks trained on low-level visual artefacts. These methods tend to lack interpretability, where outputs cannot be easily explained or scrutinised, and generalise poorly across different generative systems (Tolosana and others, 2020). This limits their use in forensic and evidential contexts where transparent, accountable outputs are required.

This PhD investigates facial behavioural dynamics as an alternative basis for interpretable deepfake detection. The main hypothesis is that current generative AI systems fail to faithfully reproduce the fine-grained temporal patterns of genuine facial movement, leaving detectable behavioural traces. Analysing these patterns may yield detection methods whose outputs are grounded in human facial behaviour rather than opaque visual features.

This project has two overarching aims:

- develop and validate an interpretable machine learning pipeline for deepfake detection based on Facial Action Units (standardised codings of facial muscle movements (Ekman and Friesen, 1978) and temporal pattern analysis
- map how deepfake-related harms are represented in UK media discourse, providing empirical context for understanding the real-world landscape within which detection research operates

Research methodology

The project uses a mixed-methods design integrating machine learning with computational discourse analysis.

Detection pipeline

Facial Action Unit (AU) features are extracted from video using OpenFace (Baltrusaitis and others, 2018), an open-source toolkit based on the Facial Action Coding System (FACS). These are decomposed using Non-Negative Matrix Factorisation (NMF), a dimensionality reduction technique that identifies interpretable underlying components of facial expression. Temporal features are then derived using the Complexity Measures and Features for Time Series Classification (CMFTS)

framework (Baldán and Benítez, 2023), capturing dynamic patterns in facial movement across multiple time scales.

To identify features predictive of deepfakes, a feature selection process (Kursa and Rudnicki, 2010) is used, and the identified important features are used for classification. A variety of machine learning classifiers are used, ranging from ensemble decision trees, to logistic regression. Traditional machine learning methods are favoured over deep learning counterparts for their inherent interpretability. Random Forests support feature importance analysis directly, aiding transparency and explainability.

The pipeline is further validated on emotion classification across multiple datasets, establishing its generalisability as a facial dynamics analysis tool beyond deepfake detection.

Discourse analysis

A corpus of over 500 UK news articles referencing deepfakes was collected from nine major outlets covering 2017 to 2025. Topic modelling is conducted using BERTopic (Grootendorst, 2022), a transformer-based method using contextual language embeddings to identify coherent themes. Sentiment analysis is additionally planned to characterise the affective dimensions of public discourse. Temporal analysis tracks how topic prevalence shifts across the collection period. Understanding how deepfake-related harms are framed in public discourse provides empirical grounding for evaluating policing and policy responses.

Interim reports or publications

Murphy TJ, Cook J and Cuve HCJ. (2026). Interpretable facial dynamics as behavioral and perceptual traces of deepfakes. arXiv.

Research participation

This project does not involve direct recruitment of participants. The technical detection studies draw on existing secondary video datasets comprising adult participants engaged in naturalistic conversation and emotion elicitation tasks. Any other participant data is drawn from previously conducted studies.

References

Baldán FJ and Benítez JM. (2023). Complexity measures and features for times series classification. Expert Systems with Applications, volume 213, article 119227

Baltrusaiti T and others. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), pages 59–66

Citron D and Chesney R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, volume 107

Ekman P and Friesen WV (1978). 'Facial Action Coding System: A Technique for the Measurement of Facial Movement'. Consulting Psychologists Press

Grootendorst M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (Version 1). arXiv

Kursa MB and Rudnicki WR. (2010). Feature selection with the Boruta package. Journal of Statistical Software, volume 36, issue 11

Tolosana R and others. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. arXiv