Advancing DeepFake Detection

Guide: Dr. Rajiv Ratn Shah
Indraprastha Institute of Information Technology (IIITD)

Shubham Attri* Trilok Singh* Adish Jain*
Dhruv Garg* Archit garg* Raj Gupta*

1. Problem Identification:

Deepfakes, synthetic media generated using deep learning techniques, have rapidly advanced from basic face swaps to highly realistic manipulated video and audio. Initially limited to fake celebrity pornographic images, deepfakes now have concerning implications for identity theft, fraud, and disinformation. Recent research has shown deepfake video generation advancing to even 96% fool rate on face authentication systems. This exponential improvement, combined with the accessibility of open-source deepfake tools, signals an urgent need to bolster detection efforts.

2. Importance of the Problem:

As the deep fake generation becomes more accessible, instances of identity theft and fraud through the usage of deep fakes are likely to rise. Cybercriminals may leverage deepfakes to impersonate individuals and gain access to sensitive personal or financial data. Deepfakes also threaten to undermine public trust and exacerbate the disinformation crisis. The potential societal impacts highlight the need for robust deepFake detection, particularly focusing on forensics identity verification use cases.

3. Related Work:

Prior deepfake detection research has focused heavily on analyzing facial cues and artifacts in images and videos. However, recent advancements in generative models can synthesize realistic faces that bypass traditional biometrics. Some recent works have begun exploring supplementary modalities like audio, speech patterns, and micro-expressions to improve detection. Our work aims to build on these multimodal approaches, focusing on identity verification as the use case.

4. Novelty of the Proposed Idea:

We propose a novel multi-modal framework leveraging image, frame, video cues for detecting deepfake identity theft attempts. Unlike previous works, our method uniquely integrates signed biometric modalities found in the Forensic system to enable matching against enrolled identities.
The fusion of explicit biometric signals with learned deepfake artifacts provides a robust identity verification approach.

5. Techniques/Algorithms:

Forensic Analysis:

Compression Artifacts: Analyze video frames for compression artifacts. Deepfake generation may introduce different compression patterns compared to genuine videos.

Noise Patterns: Examine noise patterns in the image, as deepfake algorithms may produce unnatural noise that differs from real-world video recording.

Temporal Analysis:

Frame Consistency: Deepfake videos may exhibit inconsistencies between frames. Analyze the temporal consistency of facial features, shadows, and other elements throughout the video.

Lip Sync Analysis: Evaluate lip sync accuracy by comparing lip movements with corresponding audio. Deepfake lip syncing may not perfectly match the spoken

6. Evaluation Methodology:

We evaluate our framework on a proprietary dataset containing deepfake impersonation attempts against enrolled individuals. Metrics include biometric matching accuracy, presentation attack detection error rates, and computational efficiency to validate real-time identity verification viability. We also measure performance gains over individual modalities to demonstrate the value of our multi-modal fusion approach.

7. Potential Contributions:

- Development of a novel deepfake detection model integrating advanced neural network architectures.
- Enhancement of detection accuracy by capturing intricate patterns and temporal dependencies in manipulated media content.
- Robustness against evolving deepfake generation techniques, providing a more sustainable solution.
- Empirical validation through extensive experimentation on diverse datasets, demonstrating the practical utility of the proposed method.

In summary, this research endeavors to address the escalating challenges posed by deepfake technology through the development of an innovative detection method. By introducing a hybrid neural network architecture, our approach aims to push the boundaries of current deepfake detection capabilities, contributing to the ongoing efforts to mitigate the potential harms associated with manipulated media content with infrastructure to deployable defense mechanisms against credential compromise attempts using AI-generated media.

Literature Review:

Forensic Analysis and Artifacts:

Several studies focused on analyzing compression artifacts and inconsistencies introduced during the deepfake generation process. Researchers explored how deepfake videos might exhibit different patterns of noise, artifacts, and distortions compared to authentic content.

Facial Feature Analysis:

A common theme in the literature involves the analysis of facial features for detecting deepfakes. This includes examining discrepancies in facial landmarks, expressions, and eye movements. Machine learning models, especially those based on convolutional neural networks (CNNs), were often employed for facial feature analysis. But Among all the studies they lack the accuracy and precision of detection.

Literature Review references:

https://paperswithcode.com/paper/faceforensics-learning-to-detect-manipulated
https://paperswithcode.com/paper/taming-transformers-for-high-resolution-image
https://paperswithcode.com/paper/celeb-df-a-new-dataset-for-deepfake-forensics
https://paperswithcode.com/paper/unmasking-deepfakes-with-simple-features
https://paperswithcode.com/paper/combining-efficientnet-and-vision
https://paperswithcode.com/paper/video-face-manipulation-detection-through
https://paperswithcode.com/paper/cross-forgery-analysis-of-vision-transformers
https://paperswithcode.com/paper/undercover-deepfakes-detecting-fake-segments