4U-01

Preliminary User Evaluation of Deepfake Detection System in Criminal Justice Facial Evidence Verification

Ebrima Hydara, Masato Kikuchi, Tadachika Ozono†
Nagoya Institute of Technology†

1 Introduction

The use of manipulated images or videos as evidence in proceedings can have serious consequences, such as wrongful convictions or the dismissal of valid charges. Conventional methods [1] for verifying media are becoming less effective against the sophisticated techniques employed in creating deepfakes. To address this issue, our study focuses on evaluating a deepfake detection system designed specifically for verifying facial evidence. The evaluation measured the system prediction performance but also assesses the confidence it instills in users who must rely on its judgments. This paper shows that the preliminary user evaluation of our system for deepfake detection in criminal justice can be effective by using the system alongside human perception.

2 Deepfake Detection System

The originality of our work [2] is the introduction of a deepfake detection system using the Vision Transformer model specifically trained for our task, to compliment the current facial evidence verification methods by verifying the veracity of the evidence. We incorporated forensic oriented techniques such as confidence threshold, frame extraction, frame timestamps, and heatmaps in a criminal justice reporting scheme.

We implemented and evaluated 4 functions: confidence threshold to filter the relevant frames based on the confidence level of the prediction result from the model, frame extraction to extract user-targeted frames, frame timestamps to track frames, and heatmaps to visualize manipulations for the forensic analysts. We assume that the combination of these functions with the current methods will address the deepfake threat to facial evidence in criminal investigations. The system is designed to assist forensic analysts in a try and error mode by using confidence threshold to get only the video frames returned with the highest level of prediction confidence. This will help investigation to ascertain the veracity of the evidence.

3 Experiment Setup

Our research study combined automated deepfake detection with human evaluation methods focusing on two objectives. These objectives aimed to assess the usefulness of the deepfake detection system and its interaction with users. We aimed to determine how effective the system was in helping users differentiate between manipulated videos and measure user confidence in the system judgments.

We designed the experiment with the above goals in mind to test the technical efficiency of the system and understand how users perceived its results and functionalities. For this purpose,

† Email: ebrima@ozlab.org

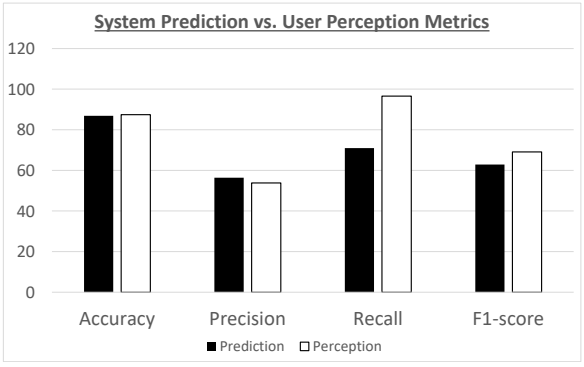

Fig.1 Collaborative performance evaluation

we created a dataset consisting of 200 videos that included both real and deepfake content. This dataset was carefully selected to simulate real-world scenarios.

A diverse group of 10 individuals interacted with the system as part of our study. Each participant was assigned 20 videos to authenticate which were randomly selected to avoid any bias based on sequence. The participants assessments were recorded using a spreadsheet for further analysis.

4 Experiment Results

The system performed exceptionally well aligning closely with user perceptions and demonstrating its ability to resonate with human evaluators. While users were less precise than the system, there is room for improvement in ensuring better alignment between user judgment and the system output. Interestingly, users tended to be more cautious when labeling videos as deepfakes compared to the system, indicating an approach when dealing with potential deepfake content.

When comparing F1 score, as shown in Fig.1, it became evident that users struck a balance between recall and precision by leveraging their intuition to navigate the complexities of deepfake detection. This highlights the importance of combining algorithmic analysis with human intuition. A hybrid approach that is crucial for maintaining integrity in criminal justice proceedings.

User confidence levels varied across accuracy, precision, recall, and F1 score measurements. Generally speaking, there was a sense of confidence in the accuracy and precision of the system. However, confidence levels were more diverse regarding recall performance suggesting some concerns about the ability of the system to comprehensively detect all instances of interest. The variation in trust levels shows a need to better educate users on deepfake detection and understand the nuances involved.

5 Conclusion

Our investigation has shown that combining artificial intelligence with human judgment shows great potential for the criminal justice system in detecting deepfakes. The ability of the system to align with user perceptions makes it more likely to be adopted in real-world settings.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Numbers JP19K12266, JP22K18006.

References

[1] Bacci, N., Davimes, J. G., Steyn, M., Briers, N.: Forensic Facial Comparison: Current Status, Limitations, and Future Directions. Biology, 10(12), 1269, pp. 1–26 (2021).

[2] Hydara, E., Kikuchi, M., and Ozono, T. (2023): Deepfake Detection System for Facial Evidence Verification in Criminal Justice and its Legal and Ethical Implications. ISDA2023, Springer, 10p. (to appear)