DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection

Saniat Javid Sohrawardi[*]
Rochester Institute of Technology

Sovantharith Seng
Rochester Institute of Technology

Akash Chintha
Rochester Institute of Technology

Bao Thai
Rochester Institute of Technology

Andrea Hickerson
University of South Carolina

Raymond Ptucha
Rochester Institute of Technology

Matthew Wright
Rochester Institute of Technology

ABSTRACT

Although the concern over deliberately inaccurate news is not new in media, the emergence of *deepfakes*—manipulated audio and video generated using artificial intelligence—changes the landscape of the problem. As these manipulations become more convincing, they can be used to place public figures into manufactured scenarios, effectively making it appear that anybody could say anything. Even if the public does not believe these are real, it will generally make video evidence appear less reliable as a source of validation, such that people no long trust anything they see. This increases the pressure on trusted agents in the media to help validate video and audio for the general public. To support this, we propose to develop a robust and an intuitive system to help journalists detect deepfakes. This paper presents a study of the perceptions, current procedures, and expectations of journalists regarding such a tool. We then combine technical knowledge of media forensics and the findings of the study to design a system for detection of deepfake videos that is usable by, and useful for, journalists.

KEYWORDS

deepfake detection, news verification, video forensics, journalism

1 INTRODUCTION

Currently, about two-thirds of the US population consume their news through social media, such as Reddit, Twitter, and Facebook [20]. While this increases the penetration of news content, it also presents itself as a potent breeding ground for the proliferation of maliciously falsified information. Rapid improvements in artificial intelligence have led to more advanced methods of creating false information that could be used to more effectively and efficiently trick the public and journalists alike. One such advancement is the ability to generate *deepfakes*, videos that can make it appear that a well-known person said and did things that the person never took part in. These videos represent a serious problem for journalists, who must discern real videos from fakes and do so quickly to avoid the fakes from becoming perceived as real by the as-yet unsuspecting public. Unfortunately, the quality of deepfakes is quickly getting better to the point that even careful observation by an informed expert may not be enough to spot them.

Although several research groups have investigated automatic detection of deepfakes, there is currently no deployed tool that a journalist could turn to for determining whether a video is a fake or not. Our project seeks to develop such a tool, which we call DeFake. While it is possible to build a detection tool and simply put it out on the web for journalists or anyone else to use, our work takes a user-centered approach. In this paper, we present a user study to answer the following research questions:

(1) What type of videos should be the main focus of a detection platform?
(2) What are the performance expectations and requirements for the tool?
(3) What type of analyses can be useful, and what is the best way to present them?

We hope that answers could guide the engineering and interface design processes of deepfake detection tools for use by journalists.

2 LITERATURE REVIEW

Deepfake Generation

Deepfake manipulation techniques can be grouped into two categories: face swapping and puppet master.

Face Swapping. In a face swap video, a *source* face from one video is placed onto a *target* face in another video. Simple techniques, such as the approach by Kowalski [14], use traditional computer graphics techniques. They can operate

[*]Email at: saniat.s@mail.rit.edu

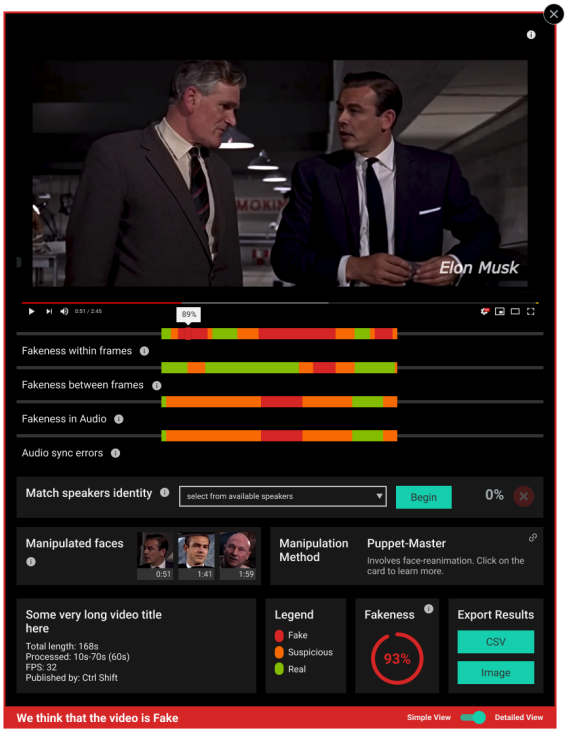

Figure 1: Detailed analysis interface showing 4 different analysis results over a timeline, unique faces with high fakeness levels, additional biometric analysis [3] and methodology detection.

in real time, but the results are far less polished than other, more complex methods. Better deepfakes can be generated by using recent advances in machine (deep) learning, namely Generative Adversarial Networks (GANs) [10]. The FakeApp tool, for example, learns the facial structures of both the source and target faces to create high-quality face swaps. The FakeApp has been taken down but remains available through online forums. There are also several alternatives on Github, such as faceswap-GAN [16] and DeepFaceLab [1]. Due to this relatively easy access, many videos using this method are available on the internet, mostly in form of either comedic face-swaps or fake celebrity pornography.

Puppet Master. In this group of methods, a *source* video (the puppet master in this analogy) drives the expressions and pose, which are placed on to a *target* video (the puppet). These techniques enable someone to create a source video of themselves in order to make the target, a public figure, say anything. Face2Face, developed by Thies et al. [25], was the first technique to enable transfer of facial expressions from a source video to a target video. A subset of those authors have since developed an additional method that achieves even more refined results based on a similar concept that also relies on facial landmarks [24]. While this method requires both source and target media to be provided as videos, there are alternative methods that are capable of working with less. Kim et al. [12] for example, created a method capable of forging a video from as little as one image of the target, while Suwajanakorn et al. [23] achieve impressive results with only audio input.

Deepfake Detection

Deepfake detection is a relatively nascent research space, starting in early 2018. There are two types of techniques.

Biological signals. Several works have looked at unusual behavior in deepfake videos, like lack of blinking [15], facial abnormalities [7], and movement anomalies [3]. These methods suffer from the potential for straightforward improvements in the video generation techniques, such as adding blinking.

Pixel level irregularities. There is a greater breadth of research that extracts faces and uses different forms of deep learning to target intra-frame [2, 7, 8, 17–19], or inter-frame inconsistencies [11, 21]. While many of these methods perform well on specific types of manipulations, they fall short at being able to generalize to multiple and unknown types of deepfakes, which is critical for open-world detection.

None of the techniques for deepfake detection have yet been developed into an actual tool to be used for detection in the real world. Nor have there been any studies, to our knowledge, on how to design such a tool for effective use by journalists.

Verification Tools

There is a range of tools, services, and methodologies available for journalists in their quest for information verification. As described by Brandtzaeg et al. [5, 6] tools that are frequently used by journalists for media verification are TinEye [1], Google Reverse Image search and InVid Project [2]. We next review some of the relevant work in the image and video verification space.

3 METHOD

We take a user-centered approach to developing the detection tool to help us orchestrate the design of the detection mechanisms as well as the interface. The research is based on a qualitative interview supplemented by either interactive prototypes or early iterations of the functional tool. We carried out interviews with 11 journalists involved in verification work as shown on Table 1

[1] www.tineye.com
[2] www.invid-project.eu

ID	Sex	Target Audience	Geographic Region
LNAM1	M	Local	North America
LNAM2	M	Local	North America
NNAM1	M	National	North America
NNAM2	M	National	North America
NNAM3	F	National	North America
NNAM4	M	National	North America
NNAM5	F	National	North America
LEE1	M	Local	Eastern Europe
NEU1	F	National	European Union
NEU2	F	National	European Union
NOC1	M	National	Oceania

Table 1: Characteristics of interview participants

Experimental Design

Recruitment. All participants were journalists involved in news verification, since they are more likely to interact with the tool. We used attended journalism conferences to get access to our target population and then used snowball sampling through personal contacts that we had developed.

Interviews. The interviews, each spanning about 40 minutes, were carried out in a conversational setting guided by a questionnaire split into three sections.

(1) **Current Process:** The discussions here would focus on the participants' responsibilities at their job and their current processes for verification of potentially manipulated information. We do not focus specifically on any type of media, however, we do ask them what triggers them to scrutinize information.
(2) **Deepfakes:** This section was focused on discussions about deepfakes. Here we talk about their past experiences about deepfakes, their attempts to verify the videos and the sources of these videos. We talk about their views on deepfakes in general and where they thought the general trend would lead us.
(3) **Detection Tool:** Here we tried to identify the needs and preferences of the participants for a deepfake detection tool. We briefly discuss their ideas on the visualization of the analysis and follow it up by showing prototypes for feedback. We allow them to interact with the provided interfaces asking them questions and assessing the intuitiveness of the interface. We also get their opinions on performance aspects of the tool.

Prototype Design

The initial interactive prototypes were designed and built in Figma,[3] to accelerate the ideation stage with quick sketches that we could evaluate internally and through the initial interviews. For the latter interviews we used a working iteration of the tool built using ReactJS and Python.

[3]www.figma.com

4 RESULTS

The interview sessions yielded several valuable themes and compelling findings that helped shape the initial versions of the tool and will further guide the iterative upgrades.

The Current State

Majority of the participants were very familiar with most types of disinformation and the several verification methods. Local news stations are less burdened to by news delivery speed and prefer to record their own content, generally due to lack of the budget for special verification teams. Participant LNAM1 stated:

> "*We don't go in and hire someone to come in and do a technical breakdown of the video. We don't have the resources or time to do that.*"

At the national level, the organizations have the ability to employ verification personnel that are more accustomed to fact-checking and the available tools.

The current verification process involves manual source and context verification for all media types. The only video verification tool that participants used was InVid, with some also using video metadata matching. For deepfake detection, participants would refer to observation, context irregularities and prior knowledge.

What triggers the verification process?

The participants would verify videos sourced from untrusted or bipartisan entities. Likewise, the theme and the political or emotional impact on specific groups is also an important factor for suspecting the video. Additionally, odd behavior for the people in the video, virality, and low quality are features that warrant more in-depth scrutiny.

What are the fears of deepfakes?

The primary fears from participants were national security and videos used for political smear campaigns aimed at highly influential entities. Some other worries were more personal: potential cyber-bullying and blackmail on a smaller scale. Most participants also admitted that it would be embarrassing for deepfakes to trickle down into the published media and possibly lead to the erosion of trust among the public. As participant NNAM1 mentioned while discussing the past media manipulation campaigns:

> "*We have seen gaslighting and erosion of trust in traditional media outlets for good reason. I mean, traditional media is in many times in the past shot themselves in the foot.*"

A deepfake itself would only be a story if it had made an impact, taking the narrative of either a debunking or focus on the subject within the videos.

Performance expectations and concerns

As one would expect, the journalists were more concerned with the accuracy of the tool and were willing to accept delays on the verdict. Participant NNAM2 was stated:

> "*Accuracy is the most important thing in journalism. Anyone who tells you otherwise is not a good journalist.*"

The majority of the participants agreed that they accept the tool being more aggressive at flagging the suspicious videos at the cost of false positives. One of the participants (NNAM1) went on to stress the importance of accuracy, stating that journalists might be caught off guard due to false sense of security.

> "*If the tool is wrong, I totally missed it. And it was a huge blind spot and I just opened myself up to it.*"

The participant also suggested that if the public finds inaccurate answers through the tool, it could spark arguments on social media. Alternatively, false-positive results might undermine people who recorded real events.

What analyses do the participants expect?

In this section, the participants shared their expectations of the analyses provided by the tool. The discussions here have the greatest impact in guiding the user interface. The requirement is best described by NNAM1:

> "*Give me tools that let me as a journalist keep doing my job, which is pick up the phone and call people or go knock on doors or you know, let me follow the trails.*"

Some of the participants would focus on the video file analyses: change in speech, frame-rate, unnatural physical behavior of objects, anomalous behavior of the subjects, location of the fakes on the timeline, localization of the manipulations and the types of deepfake. The participants mentioned that the reasoning behind the choices made by the tool would help them write their articles. The supplementary analyses mentioned by the participants had more to do with context rather than video analysis. The possible outputs they discussed included: social media activity on the video, the originated source and the reputation of the source accounts. Overall, the participants hoped for a broader and more explainable view of the detection. Quoting one of the participants, NNAM5:

> "*The more information you have, the more likely I am to be able to report it to my readers in a very comprehensive way.*"

As for the display of the output and the flexibility of it, the participants preferred to have a more standardized tool with more rigid options and visual analytics.

Access restrictions

The participants unanimously agreed to make the tool available to the public, as it would help with transparency of the reports. They were also contempt with access and brute force prevention mechanisms like logins, rate limiting, and captchas. Some did, however, request some trusted organizations to be white-listed.

5 DISCUSSION

The findings from the study helped us guide the design of both the user interface and the back-end.

Video Types

The initial detection schemes are focused on accurate detection of camera-facing single subject videos. The decision was guided by the preferences of journalists to focus on public figure addresses. The models should, however, be robust to different levels of compression, which is known to be able to obscure signs of manipulation. The tool should support different video sources, primarily YouTube, Facebook, Twitter and manual upload.

Performance

The video processing speeds faced less scrutiny as journalists are more focused on accuracy. Additionally, a higher recall is prioritized over precision, as participants preferred to avoid false negatives. In the long run a high recall performance would be advised for better verification efficiency.

Users could be prompted to select a subset of the video to save the processing speed and resources. This feature was tested on the participants using interactive prototypes and they all agreed that it was a sensible feature to have.

Analysis Methods

The common theme among all the participants was the request to provide explainable pieces of information that would help them make their own decisions. Moreover, the studies of adversarial machine learning suggest that one of the most viable defenses against adversarial examples is the inclusion of multiple different detection models [22]. Hence, as shown in Figure 1, we assess the past research and choose to include five distinct schemes to tackle detection: (1) Intra-frame detection [2, 7, 8, 17–19], (2) Inter-frame detection [11, 21], (3) Audio manipulation detection [4, 9, 21], (4) Audio-sync error detection [13] and (5) Soft-biometric-based speaker identity recognition [3]. The report also includes unique faces with the highest *fakeness* scores and the detected manipulation methodology family. Although past research suggests it is possible to generate masks showing manipulated areas in the frames [17, 19], we forego providing this information due to high infrastructure demands.

Cognitive Load

As all of the outlined methods can provide frame-by-frame predictions, we can align the detection timelines with the video timeline to show frame-level results. However, we precede the detailed results with a simplified analysis, showing only the verdict and the *fakeness* percentage of the overall video, to prevent initial distractions. The timeline and the verdict are both color-coded into three urgency levels: *fake*, *suspicious* and *real*.

Security

From the security perspective, it would be beneficial to add rate limitations and login barriers. Although, some participants suggested lifting the barriers for trusted organizations, the potential risk of phishing has to be weighed in. Overall, it is vital to prevent brute-force training of an attack model using these defenses.

6 CONCLUSION

This study explores the requirements of national as well as international journalists to formulate a tool for the detection of deepfake videos. We combine knowledge from previous research and our findings from the study to develop a system and interface design for the detection platform. Like any tool developed to solve a new problem, we believe that it is reasonable to require multiple iterations, with followup interviews before a universally acceptable form is achieved. Our findings show that the key requirement from a tool of this nature is explainability of the results. This work provides a well-laid foundation to provide the journalists with a tool that can seamlessly integrate itself into their current workflow.

Acknowledgements: This effort was funded in part by the Ethics and Governance of AI Initiative.

REFERENCES

[1] [n. d.]. iperov/DeepFaceLab. https://github.com/iperov/DeepFaceLab. Commit: adb9f3a346a4330858d064924eec5a194bfcdafb.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: A Compact Facial Video Forgery Detection Network. In *WIFS*.

[3] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *CVPR Workshops*.

[4] Moustafa Alzantot, Ziqi Wang, and Mani B Srivastava. 2019. Deep Residual Neural Networks for Audio Spoofing Detection. *arXiv preprint arXiv:1907.00501* (2019).

[5] Petter Bae Brandtzaeg, Asbjørn Følstad, and Maria Ángeles Chaparro Domínguez. 2018. How Journalists and Social Media Users Perceive Online Fact-checking and Verification Services. *Journalism Practice* (2018).

[6] Petter Bae Brandtzaeg, Marika Lüders, Jochen Spangenberg, Linda Rath-Wiggins, and Asbjørn Følstad. 2016. Emerging Journalistic Verification Practices Concerning Social Media. *Journalism Practice* (2016).

[7] Umur Aybars Ciftci and Ilke Demir. 2019. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *arXiv preprint arXiv:1901.02212* (2019).

[8] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. 2018. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510* (2018).

[9] Rohan Das, Jichen Yang, and Haizhou Li. 2019. Long Range Acoustic and Deep Features Perspective on ASVspoof 2019. https://doi.org/10.7488/ds/1994

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2672–2680.

[11] David Güera and Edward J Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *AVSS*.

[12] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics (TOG)* (2018).

[13] Pavel Korshunov and Sébastien Marcel. 2018. Speaker Inconsistency Detection in Tampered Video. In *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE.

[14] Marek Kowalski. [n. d.]. faceswap. https://github.com/MarekKowalski/FaceSwap.

[15] Yuezun Li, Ming-Ching Chang, Hany Farid, and Siwei Lyu. 2018. In ictu oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv preprint arXiv:1806.02877* (2018).

[16] Shao-An Lu. [n. d.]. faceswap-GAN. https://github.com/shaoanlu/faceswap-GAN. Commit: c563edc128e79c3b593da63825f0208acf7ea4d9.

[17] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. arXiv:cs.CV/1906.06876

[18] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2018. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. *arXiv preprint arXiv:1810.11215* (2018).

[19] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv preprint arXiv:1901.08971* (2019).

[20] Elisa Shearer and Katerina E Matsa. 2018. News Use Across Social Media Platforms. *Pew Research Center, Washington DC* (2018).

[21] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2019. Poster: Towards Robust Open-World Detection of Deepfakes. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM.

[22] Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. 2017. Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks. *arXiv preprint arXiv:1709.03423* (2017).

[23] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics (TOG)* (2017).

[24] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis using Neural Textures. *arXiv preprint arXiv:1904.12356* (2019).

[25] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time Face Capture and Reenactment of RGB Videos. In *CVPR*.