Scalable, Person-Agnostic Deepfake Detection In the Wild

Iacopo Masi, Kenneth Nadeau, Yash Doshi, Joe Mathai, Wael AbdAlmgaeed
USC Information Sciences Institute
Marina del Rey, CA, USA
{iacopo,knadeau,ydoshi,jmathai,wamageed}@isi.edu

Demo of Our DeepFake Detection System · Video Presentation

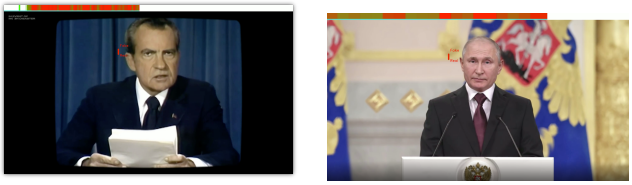
***Abstract*—We introduce a web service that helps fighting visual misinformation known as "deepfakes". A deepfake is a video containing a talking head, artificially manipulated in a hyper-realistic way using powerful AI tools. The rise of deepfakes calls for media forensics solutions that work reliably on videos and produce a low rate of false alarms at the video level. We present a web service offering a new way to assess if a face video has been manipulated. The web service relies on an engine for deepfake detection following our current research direction on video-based face manipulation detection. The research direction paves the way for achieving scalable, person-agnostic deepfake detection in the wild.**

The Deepfake Detection Web Service allows the user to upload a short video. The video will be processed in background using our deepfake detection engine. The user will be then notified, being able to review the detection output superimposed over the original video. The Deepfake Detection Web Service keeps track of a history of previously processed videos so that they can be easily inspected if need be. It also offers a user management system allowing users to privately inspect their own videos.

I. Introduction

The massive number of images and videos uploaded to the Internet every day, through social networks, presents both a challenge and opportunity to users and the society. The rapidly declining cost of mobile devices has allowed each person to hold a powerful computer in the palm of the hand and gain access to a variety of multimedia sharing applications. As a result, the frantic sharing of multimedia content skyrocketed in the last decades, in all forms including video, images and text, enhanced with augmented reality (AR) or computer vision (CV) effects.

On one hand, social networks and multimedia content improve human connectivity and information sharing. On the other hand, *visual* misinformation and technology-facilitated manipulations have dramatically increased on social networks and the Internet [1]. Nowadays, democratized artificial intelligence (AI) made it very easy to produce highly realistic face swaps with a few clicks, giving the ability to non-experts to synthesize content with "Hollywood-like" quality by simply using off-the-shelf applications and open-source tools [6]. The technology has been quickly developed to process videos, transferring the identity of a subject from a *source* video into a *target* video.



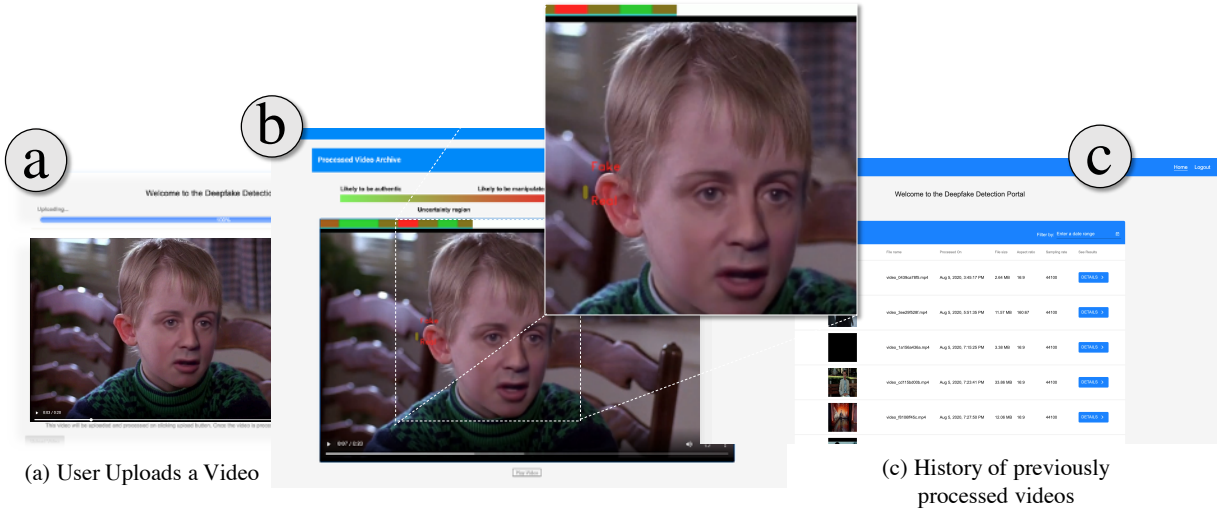
(a) In the event of moon disaster (b) A fake Putin

Fig. 1: **Deepfake Detection Results in the Wild**. Our webservice was able to successfully detect two very recent deepfakes videos created by (a) MIT Center for Advanced Virtuality (b) RepresentUs.

More recently, face swapping has been superseded by deepfakes in which the original face is replaced with a victim's face with the intent of presenting the victim to be saying or doing something he/she never said or did. The fake video is usually very realistic so that the viewer believes that the swapped subject is the actual acting person in the video. Deepfakes have marked a clear separation from the previous method of generating synthetic face swaps in the sense that prior face swapping [6] was only confined to a single image, obviously without audio. On the contrary, deepfakes revamped this idea, enabling the creation of fake "talking faces" in a video, augmented with audio for a fully believable entertainment experience. Many companies, such as Synthesia [2] and others, are using deepfake technology for entertainment and consumer purposes, especially with the increasingly improving capabilities.

Although deepfakes were initially mainly used for entertainment purposes, they have become a dangerous tool for technology-facilitated abuse with applications to nonconsensual pornography, revenge porn and defamation of actors or political targets [5]. Rapid sharing of deepfakes on the Internet may lead to a serious threat to society, heading to a new perception that *seeing is no longer believing* [1].

To respond to the rise of deepfakes, we present a web service for deepfake detection that leverages a scalable and person-agnostic deepfake detection engine backed by our current research direction on video-based face manipulation detection [8], [4]. Our solution is orthogonal to others [3] that



(a) User Uploads a Video

(b) Video is ready to be inspected

(c) History of previously processed videos

Fig. 2: **Flow of the DeepFake Detection demo.** (a) The user uploads a video containing a face (b) The Deepfake Detection engine works in background and the user is later on notified. Once this happens, the user can inspect the video for suspicious segments. (c) The UI offers a history of previously processed videos.

build person-specific models, by nature less prone to scale. Although our engine has been trained only on the modest-sized FaceForensics++ dataset [7], it achieves reasonable results on out of distribution (OOD) videos sampled from the Web. We hope this research direction could pave the way for achieving scalable, person-agnostic deepfake detection in the wild.

Fig. 1 shows a few examples of misleading deepfakes that have recently circulated on the Web to spark critical awareness among the public opinion. Our webservice was able to detect them as anomalous when compared to the real faces that the system was trained on [7].

In the following sections, we explain how our demo works and we point out conclusions and future directions.

II. Deepfake Detection Demo

Our demo employs a Web-based User Interface (UI) that offers the user the possibility to log in to the Deepfake Detection System, upload new videos, analyze the results, and keep an history of previously processed videos. The demo comprises of three main steps that are summarized in Fig. 2. First and foremost, after logging in, the user has the possibility to upload a video to be processed by the system. Once the video is uploaded—step a) in Fig. 2—the deepfake detection engine based on the technology described in [4] will assess each video segmented. A two-branch recurrent network will predict if each video segment deviates from "real" sequences used for training. Once this is done, the backend produces a new video with the superimposed detection result. In terms of speed, without any other video in the queue, a ballpark estimate is that we can consider a `x2.5` or `x3` time multiplier over the video length. E.g. if the video is processed immediately, it takes about seven minutes to get the result for a three minute video, comprising the time for uploading it. As soon as the video is ready, the user will receive a notification to the main UI, in case the web browser is still open. More importantly, the user will receive an additional notification by email. The email will inform that the submitted video is ready for inspection. The user can hence click on the provided link in the email for inspecting the video (see point b) in Fig. 2).

The DeepFake Demo superimposes a progress bar over the original video. The bar shows the likelihood for the video segment to be manipulated. The likelihood is represented with a gradient from green (no manipulation) to red (predicted to be tampered). The white color means no face was detected. The small turquoise bar at the bottom of the progress bar shows the overall progress of the video. In addition to this global feedback bar, the user can check which face is currently being processed: beside each detected face, the system shows a small likelihood bar. This vertical bar too is colored with the same gradient as explained before. Finally, videos uploaded in the past by users are not lost, on the contrary, the demo offers a way to keep track of previously processed video through a sortable history with different metadata of the video—point c) in Fig. 2.

Our demo is provided with a user management system that allows an administrative user to invite regular users to submit videos. This enforces that users will be accessing only their own videos.

III. Conclusions

We presented a demo of a web service that allows users to perform video-based deepfake detection. In the near future, the engine needs to be augmented with a multi face tracking system. In the long term, we plan to augment our demo with an explainability mechanism that enables the user to understand which facial features are used to assess its authenticity.

REFERENCES

[1] CNN - business - when seeing is no longer believing inside the pentagon's race against deepfake videos. https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/. 1

[2] Synthesia - AI Driven Video Generation. https://www.synthesia.io. Accessed: 2020-04-03. 1

[3] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, June 2019. 1

[4] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch Recurrent Network for Isolating Deepfakes in Videos. In *ECCV*, 2020. 1, 2

[5] Edvinas Meskys, Aidas Liaudanskas, Julija Kalpokiene, and Paulius Jurcys. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 15(1):24–31, 01 2020. 1

[6] Yuval Nirkin, Iacopo Masi, Anh Tran, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *AFGR*, 2018. 1

[7] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2

[8] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *CVPR Workshops*, pages 80–87, 2019. 1