Detecting Real-Time Deep-Fake Videos Using Active Illumination

Candice R. Gerstner
National Security Agency
Fort George G. Meade MD, USA
crgerst@uwe.nsa.gov

Hany Farid
University of California, Berkeley
Berkeley CA, USA
hfarid@berkeley.edu

Abstract

While many have grown suspicious of viral images and videos found online, there is a general sense that we can and should trust that the person on the other end of our video-conferencing call is who it purports to be. The real-time creation of sophisticated deep fakes, however, is making it more difficult to trust even live video calls. Detecting deep fakes in real time introduces new challenges as compared to off-line forensic analyses. We describe a technique for detecting, in real-time, deep-fake videos transmitted over a live video-conferencing application. This technique leverages the fact that a video call typically places a user in front of a light source (the computer display) which can be manipulated to induce a controlled change in the appearance of the user's face. Deviations of the expected change in appearance over time can be measured in real time and used to verify the authenticity of a video-call participant.

1. Introduction

In early 2020, a United Arab Emirates' bank was swindled out of $35 million (USD). The bank teller was convinced to transfer the funds after receiving a phone call from the purported director of a company whom the bank manager knew and with whom he had previously done business. The voice on the other end of the phone instructed the bank manager to transfer the funds as part of a corporate acquisition. Because the request was consistent with previously received emails describing the acquisition, and because the purported director's voice was familiar to him, the bank manager transferred the funds. It was later revealed that the voice was that of an AI-synthesized voice made to sound like the director.

This was not the first time AI-synthesized content was used to steal large sums of money. In 2019, a United Kingdom based company suffered a similar fate when an imposter used an AI-synthesized voice to steal $243,000 (USD) in a similar type of scam.

These two incidents are almost certainly the canaries in the coal mine. As AI-synthesized audio and video continue to improve in quality and accessibility, it is reasonable to predict that these technologies will continue to be used to commit a range of small- to large-scale frauds, among other potentially nefarious uses.

Figure 1. A computer display acts as an area light source that can be controlled in real time to induce an authenticating pattern on the face of a video-call participant. This simple example shows the impact of switching between viewing a dark (left) and bright (right) browser window.

All forms of deep fakes pose potential threats from non-consensual sexual imagery to fraud, and disinformation campaigns. The creation of real-time deep fakes, however, poses unique threats because of the general sense of trust surrounding a live video or phone call, and the challenge of detecting deep fakes in real time, as a call is unfolding.

Over the past two pandemic years, we have grown accustomed to video calls replacing previously in-person meetings and phone calls. Although not yet perfected, deep fakes can be synthesized in real time and piped through a virtual camera (e.g., github.com/alievk/avatarify-python and github.com/iperov/DeepFacelive), meaning that it will become increasingly more difficult to distinguish a real person from an AI-synthesized person at the other end of a video call.

One approach to detecting deep-fake video calls is to employ any of a plethora of passive deep-fake forensic techniques (see Section 2). Most of these approaches, however,

struggle to run in real time, and most struggle to achieve the levels of accuracy that would be needed to be incorporated into a video-conferencing application.

In contrast to passive forensic techniques, active forensic techniques (a.k.a., control-capture [2, 25]) focus on authenticating content at the point of recording. By extracting a compact, digital signature at the point of recording and packaging this signature alongside the recording, audio, images, and videos can be efficiently and accurately authenticated.

Motivated by the reliability of active forensic approaches and the unique constrained environment afforded by a video-conferencing call, we describe an active approach for detecting real-time, deep-fake video calls. In particular, instead of explicitly trying to distinguish an authentic video from a deep-fake video, we authenticate videos by projecting a distinct illumination pattern onto the face of each call participant. This pattern can be induced by a call participant displaying the temporally varying pattern on a shared screen, or directly integrated into the video-call client. In either case, no specialized imaging or lighting hardware is required.

Through large-scale simulations, we evaluate the reliability of this approach under a range of imaging scenarios, and validate this approach in a variety of real-world settings. We begin by framing our technique within previous work.

2. Related Work

We provide an overview on the state of the art in creating and detecting deep-fake videos, with an emphasis on real-time deep fakes.

2.1. Creation

Since they splashed onto the scene in 2017 with full force (an earlier incarnation dates to a decade earlier [9]), AI-synthesized content – so-called deep fakes – have continued their rapid trajectory of increased sophistication, realism [26], and accessibility. This includes images of fully fabricated people [14, 15], audio recordings mimicking another voice [29], and videos of people saying anything the creator wants them to say [35].

Within this broad range of different types of AI-synthesized content, so-called puppet-master deep fakes (e.g., [13, 32]) are particularly intriguing for their power to create a deep fake in real time from a single source image. In particular, starting with a single image of a person (the puppet), and a recorded or live video of another person (the puppet master), a video of the puppet is synthesized to mimic the expressions, mouth movement, and head movements of the puppet master. The resulting synthesized video can then be piped into a live video call through a virtual camera. Although not yet perfected, this type of puppet-master deep fake holds the potential to deceive someone into believing they are talking with anyone that a fraudster wants to impersonate [7]. In contrast, this same technology holds the potential to significantly reduce the bandwidth necessary for a video call [1].

As compared to a puppet-master deep fake, a face-swap deep fake (`github.com/deepfakes/faceswap`, `github.com/shaoanlu/faceswap-GAN`) replaces the face – from eyebrows to chin and cheek to cheek – of the impersonator with that of another. The viral deep-fake Tom Cruise videos (`www.tiktok.com/@deeptomcruise`) are a particularly compelling example of this type of deep fake. While the creation of deep Tom Cruise is the result of a talented impersonator and a highly-skilled, special-effects artist, open-source software for creating real-time, face-swap deep fakes are emerging on the scene. DeepFaceLive (`github.com/iperov/DeepFacelive`), for example, allows the creator to swap their face with a celebrity in real time and, according to their documentation, also incorporates a color transfer that maps the creator's environmental lighting onto the deep fake.

2.2. Detection

Forensic techniques for detecting deep fakes can be broadly categorized into low- and high-level approaches. Low-level techniques detect pixel-level, synthesis artifacts, ranging from general artifacts [23, 36, 38–40], to warping artifacts [21], and blending artifacts [19]. High-level techniques focus on semantically meaningful features, including inconsistencies in eye blinks [20], head-pose [37], physiological signals [10], mouth shape and movement [5], and distinct mannerisms [3, 6].

While some of these techniques might be applicable to detecting real-time deep-fake videos, in our view, a class of particularly promising approaches takes advantage of the unique physical constraints of a live video call: call participants are in front of a light source (the computer display) that can be actively adjusted in real time to induce specific lighting patterns on a user's face. The consistency of a call participant's appearance under this induced lighting can then be used to verify their liveness and physical presence in front of the camera. We posit that this approach will be effective because either the deep-fake video simply fails to transfer the active illumination, or there is a temporal delay in transferring the active illumination. We will show that both of these scenarios are easily detected.

Exploiting the computer display as an active light source has previously been leveraged as an inexpensive light stage in which, after recording a user's appearance under a time-varying illumination pattern, her face can be synthetically re-lit under an arbitrary lighting environment [30]. This type of active approach has also previously been explored particularly for the purpose of thwarting playback or re-broadcast attacks [4]. FaceRevelio [11], for example, uses