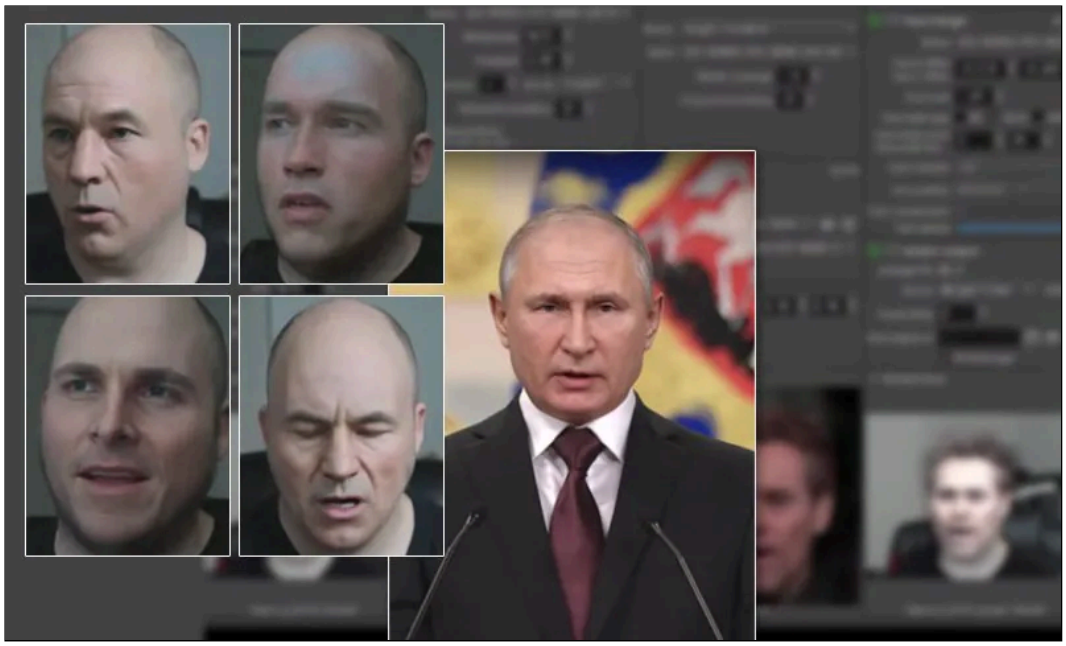
Detecting Deepfake Video Calls Through Monitor Illumination

Originally published at https://www.unite.ai/detecting-deepfake-video-calls-through-monitor-illumination/



A new collaboration between a researcher from the United States' National Security Agency (NSA) and the University of California at Berkeley offers a novel method for detecting deepfake content in a live video context – by observing the effect of monitor lighting on the appearance of the person at the other end of the video call.

CLICK TO VIEW ANIMATED GIF

Popular DeepFaceLive user Druuzil Tech & Games tries out his own Christian Bale DeepFaceLab model in a live session with his followers, while lighting sources change. Source: https://www.youtube.com/watch?v=XPQLDnogLKA

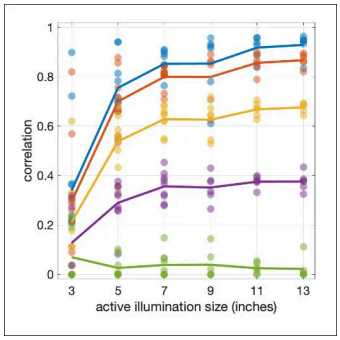
The system works by placing a graphic element on the user's screen that changes a narrow range of its color faster than a typical deepfake system can respond – even if, like real-time deepfake streaming implementation DeepFaceLive (pictured above), it has some capability of maintaining live color transfer, and accounting for ambient lighting.

The uniform color image displayed on the monitor of the person at the other end (i.e. the potential deepfake fraudster) cycles through a limited variation of hue-changes that are designed not to activate a webcam's automatic white balance and other *ad hoc* illumination compensation systems, which would compromise the method.

From the paper, an illustration of change in lighting conditions from the monitor in front of a user, which effectively operates as a diffuse 'area light'. Source: https://farid.berkeley.edu/downloads/publications/cvpr22a.pdf

The theory behind the approach is that live deepfake systems cannot respond in time to the changes depicted in the on-screen graphic, increasing the 'lag' of the deepfake effect at certain parts of the color spectrum, revealing its presence.

To be able to measure the reflected monitor light accurately, the system needs to account for and then discount the effect of general environmental lighting that is unrelated to light from the monitor. It is then able to distinguish shortfalls in the measurement of the active-illumination hue and the facial hue of users, representing a temporal shift of 1-4 frames' difference between each:



By limiting the hue variations in the on-screen 'detector' graphic, and ensuring that the user's webcam is not prompted to auto-adjust its capture settings by excessive changes in levels of monitor illumination, the researchers have been able to discern a tell-tale lag in the deepfake system's adjustment to the lighting changes.

The paper concludes:

> *'Because of the reasonable trust we place on live video calls, and the growing ubiquity of video calls in our personal and professional lives, we propose that techniques for authenticating video (and audio) calls will only grow in importance.'*

The study is titled *Detecting Real-Time Deep-Fake Videos Using Active Illumination*, and comes from Candice R. Gerstner, an applied research mathematician at the US Department of Defense, and Professor Hany Farid of Berkeley.

Erosion of Trust

The anti-deepfake research scene has pivoted notably in the last six months, away from general deepfake detection (i.e. targeting pre-recorded videos and pornographic content) and towards 'liveness' detection, in response to a growing wave of incidents of deepfake usage in video conference calls, and to the FBI's recent warning regarding the growing use of such technologies in applications for remote work.

Even where a video call transpires not to have been deepfaked, the increased opportunities for AI-driven video impersonators is beginning to generate paranoia.

The new paper states:

> *'The creation of real-time deep fakes [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.'*

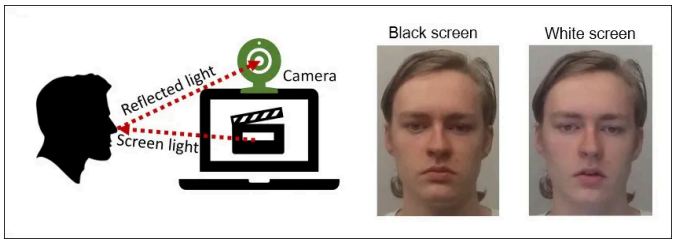
The research community has long since set itself the goal of finding infallible signs of deepfake content that can't easily be compensated for. Though the media has typically characterized this in terms of a technological war between security researchers and deepfake developers, most of the negations of early approaches (such as eye blink analysis, head pose discernment, and behavior analysis) have

occurred simply because the developers and users were trying to make more realistic deepfakes in general, rather than specifically addressing the latest 'tell' identified by the security community.

Throwing Light on Live Deepfake Video

Detecting deepfakes in live video environments carries the burden of accounting for poor video connections, which are very common in video-conferencing scenarios. Even without an intervening deepfake layer, video content may be subject to NASA-style lag, rendering artefacts, and other types of degradation in audio and video. These can serve to hide the rough edges in a live deepfaking architecture, both in terms of video and audio deepfakes.

The authors' new system improves upon the results and methods that feature in a 2020 publication from the Center for Networked Computing at Temple University in Philadelphia.



From the 2020 paper, we can observe the change in 'in-filled' facial illumination as the content of the user's screen changes. Source: https://cis.temple.edu/~jiewu/research/publications/Publication_files/FakeFace__ICDCS_2020.pdf

The difference in the new work is that it takes account of the way webcams respond to lighting changes. The authors explain:

> *'Because all modern webcams perform auto exposure, the type of high intensity active illumination [used in the prior work] is likely to trigger the camera's auto exposure which in turn will confound the recorded facial appearance. To avoid this, we employ an active illumination consisting of an isoluminant change in hue.*
>
> *'While this avoids the camera's auto exposure, it could trigger the camera's white balancing which would again confound the recorded facial appearance. To avoid this, we operate in a hue range that we empirically determined does not trigger white balancing.'*

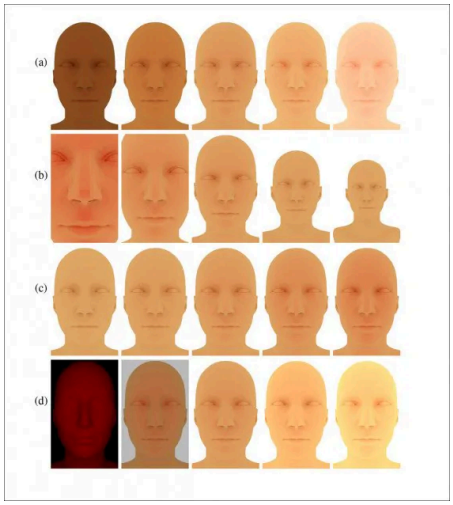
For this initiative, the authors also considered similar prior endeavors, such as LiveScreen, which forces an inconspicuous lighting pattern onto the end-user's monitor in an effort to reveal deepfake content.

Though that system achieved a 94.8% accuracy rate, the researchers conclude that the subtlety of the light patterns would make such a covert approach difficult to implement in brightly-lit environments, and instead propose that their own system, or one patterned along similar lines, could be incorporated publicly and by default into popular video-conferencing software:

> *'Our proposed intervention could either be realized by a call participant who simply shares her screen and displays the temporally varying pattern, or, ideally, it could be directly integrated into the video-call client.'*

Tests

The authors used a mixture of synthetic and real-world subjects to test their Dlib-driven deepfake detector. For the synthetic scenario, they used Mitsuba, a forward and inverse renderer from the Swiss Federal Institute of Technology at Lausanne.



Samples from the simulated environment tests, featuring varying skin tone, light source size, ambient light intensity, and proximity to camera.

The scene depicted includes a parametric CGI head captured from a virtual camera with a 90° field of view. The heads feature Lambertian reflectance and neutral skin tones, and are situated 2 feet in front of the virtual camera.

To test the framework across a range of possible skin tones and set-ups, the researchers ran a series of tests, varying diverse facets sequentially. The aspects changed included skin tone, proximity, and illumination light size.

The authors comment:

> *'In simulation, with our various assumptions satisfied, our proposed technique is highly robust to a broad range of imaging configurations.'*

For the real-world scenario, the researchers used 15 volunteers featuring a range of skin tones, in diverse environments. Each was subjected to two cycles of the restricted hue variation, under conditions where a 30Hz display refresh rate was synchronized to the webcam, meaning that the active illumination would only last for one second at a time. Results were broadly comparable with the synthetic tests, though correlations increased notably with greater illumination values.

Future Directions

The system, the researchers concede, does not account for typical facial occlusions, such as bangs, glasses, or facial hair. However, they note that masking of this kind can be added to later systems (through labeling and subsequent semantic segmentation), which could be trained to take values exclusively from perceived skin areas in the target subject.

The authors also suggest that a similar paradigm could be employed to detect deepfaked audio calls, and that the detecting sound necessary could be played in a frequency out of the normal human auditory range.

Perhaps most interestingly, the researchers also suggest that extending the evaluation area beyond the face in a richer capture framework could notably improve the possibility of deepfake detection*:

> *'A more sophisticated 3-D estimation of lighting would likely provide a richer appearance model which would be even more difficult for a forger to circumvent. While we focused only on the face, the computer display also illuminates the neck, upper body, and surrounding background, from which similar measurements could be made.*
>
> *'These additional measurements would force the forger to consider the entire 3-D scene, not just the face.'*

** My conversion of the authors' inline citations to hyperlinks.*

First published 6th July 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai