WHITEPAPER

Deep Dive into Deepfakes

Mitigating the growing threat to biometric security posed by fake digital imagery and injection attacks

IDR&D

Introduction: Secure biometric authentication relies on liveness detection

Biometrics leverage a person's unique physical or behavioral characteristics such as their face or voice to identify and authenticate them remotely using their devices. The security and convenience challenges that come with passwords make biometrics an ideal alternative. While passwords demonstrate a secret that you *know*—and can be forgotten, shared, stolen, or phished—biometrics demonstrate something unique that you inherently are.

Biometrics rely on confirmation that the user is live and present when capturing their biometrics. Without countermeasures, bad actors can conduct successful attacks using non-live biometric imagery; for example, they can "spoof" facial recognition algorithms with photos or videos that impersonate their potential victim analogously to stealing their password. Liveness detection mitigates the risk by detecting such attacks.

Deepfakes are rendered digital imagery–like special effects in a movie–that allow a fraudster to create realistic videos and use them to spoof biometric security. Concerns about fraud are growing as the technology to create deepfakes becomes increasingly accessible. The threat of deepfakes is substantial. In 2023, 20% of successful account takeover attacks will leverage deepfakes (**Gartner**).

This paper introduces various types of deepfakes, how they are used in attacks to defeat biometric security, and how the threat is mitigated.

Generating fraudulent imagery: deepfakes and cheapfakes

Deepfakes are a rapidly evolving type of synthetic media in which artificial intelligence (AI) is used to generate realistic-looking digital imagery that depicts existing people doing or saying things they never actually did or said, or even fabricating people that don't actually exist.

Deepfakes are being used for a variety of nefarious purposes, including spreading misinformation, but a growing application of deepfakes is to spoof facial recognition algorithms. **According to Idiap Research Institute**, 95% of facial recognition systems are unable to detect deepfakes.

Unlike photos and screen replays that are "presented" to the camera on a physical medium (a presentation attack), deepfakes are natively digital, created using deep neural network-based machine learning. Realistic deepfakes can be "injected" into a device by way of hardware and software hacks in a way that bypasses the camera (an injection attack), which makes them difficult for presentation attack detection.

© 2023 ID R&D

There are three main types of a facial deepfake highlighted by experts:

Face swapping is one of the most straightforward deepfake techniques, as there are numerous out-of-the-box solutions to generate facial deepfakes. Face swapping websites and apps can "insert" the face of any real person into a video.

Face synthesis involves creating a highly realistic photo or video of a real or non-existent face. They are produced with the help of Generative Adversarial Network (GAN). Face synthesis is a sophisticated process and requires a potential attacker to have a high level of expertise.

Altered facial expression or face manipulation are able to add more realism to a falsified image or video. Face manipulation is also based on GAN usage and can manipulate a face in different ways: changing facial expressions, age, gender, hair, and eye color, etc. By utilizing a system of discriminators, generators, and target domain labels, the network can apply any emotion to a target's face.

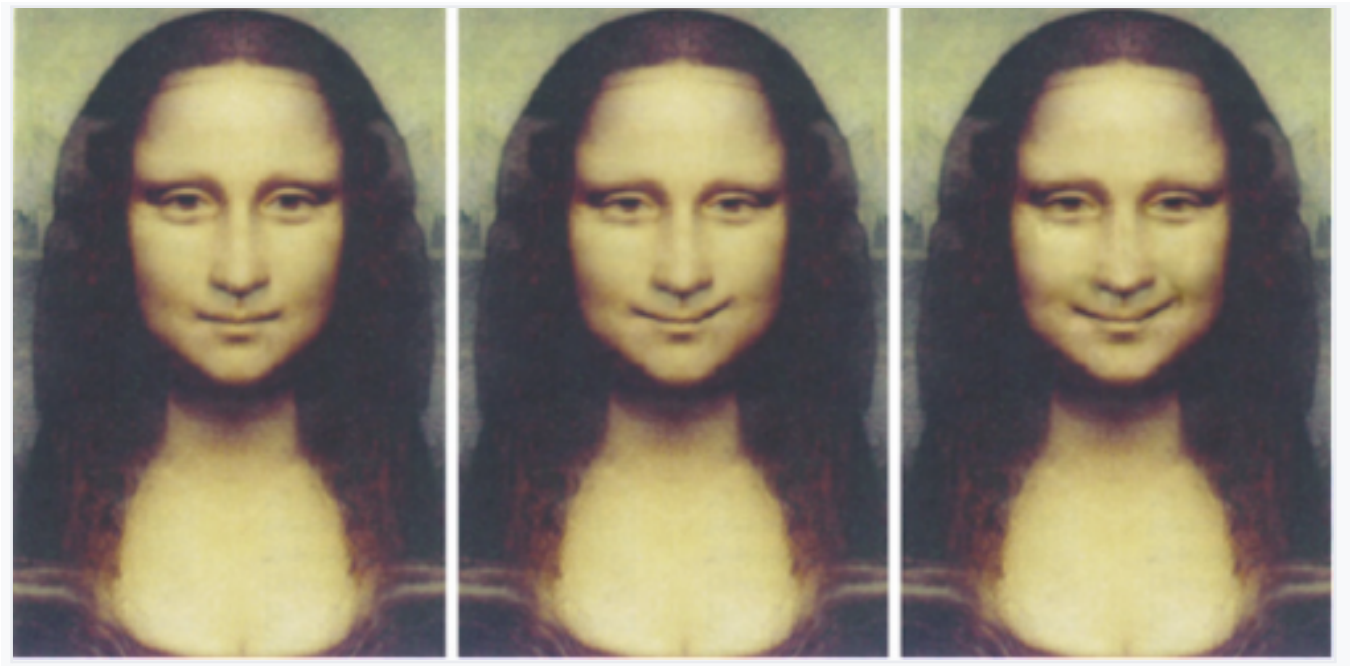


Figure: Altered facial expression technique applied to the Mona Lisa

Cheapfakes are a class of fake media that is easy and quick to produce, especially for amateur users. Compared to traditional deepfakes, cheapfakes do not require sophisticated coding, manual neural network tuning, or post-production skills. Experts highlight that cheapfakes can be dangerous because they take little effort and cost to fabricate.

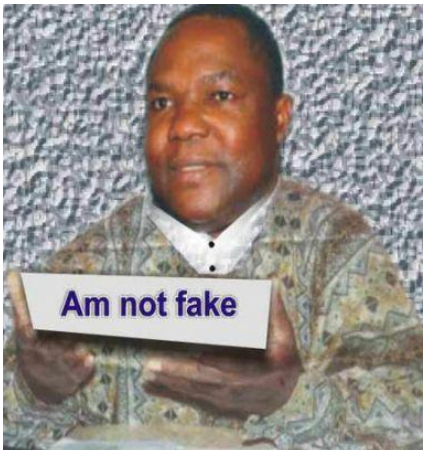


Figure: A cheapfake used by Nigerian email scammers

© 2023 ID R&D

Delivery of fraudulent imagery: presentation attacks and injection attacks

There are two classes of attacks on biometric systems that aim to spoof biometrics by defeating liveness detection: presentation attacks and injection attacks.

Presentation attacks involve "presenting" non-live imagery to the camera during the capture process and representing it as live using physical media, such as paper photos and digital screens as outlined in the intro. More information about presentation attacks and the presentation attack detection techniques applied to defeating them can be found here.

Injection attacks involve hardware or software hacks that enable a fraudster to "inject" digital imagery in such a way that bypasses a proper capture process with the primary camera. Liveness detection techniques that assume a proper capture process can be defeated with this type of attack. A fraudster is essentially able to represent their injected deepfake video as imagery captured in front of the camera when it is not.

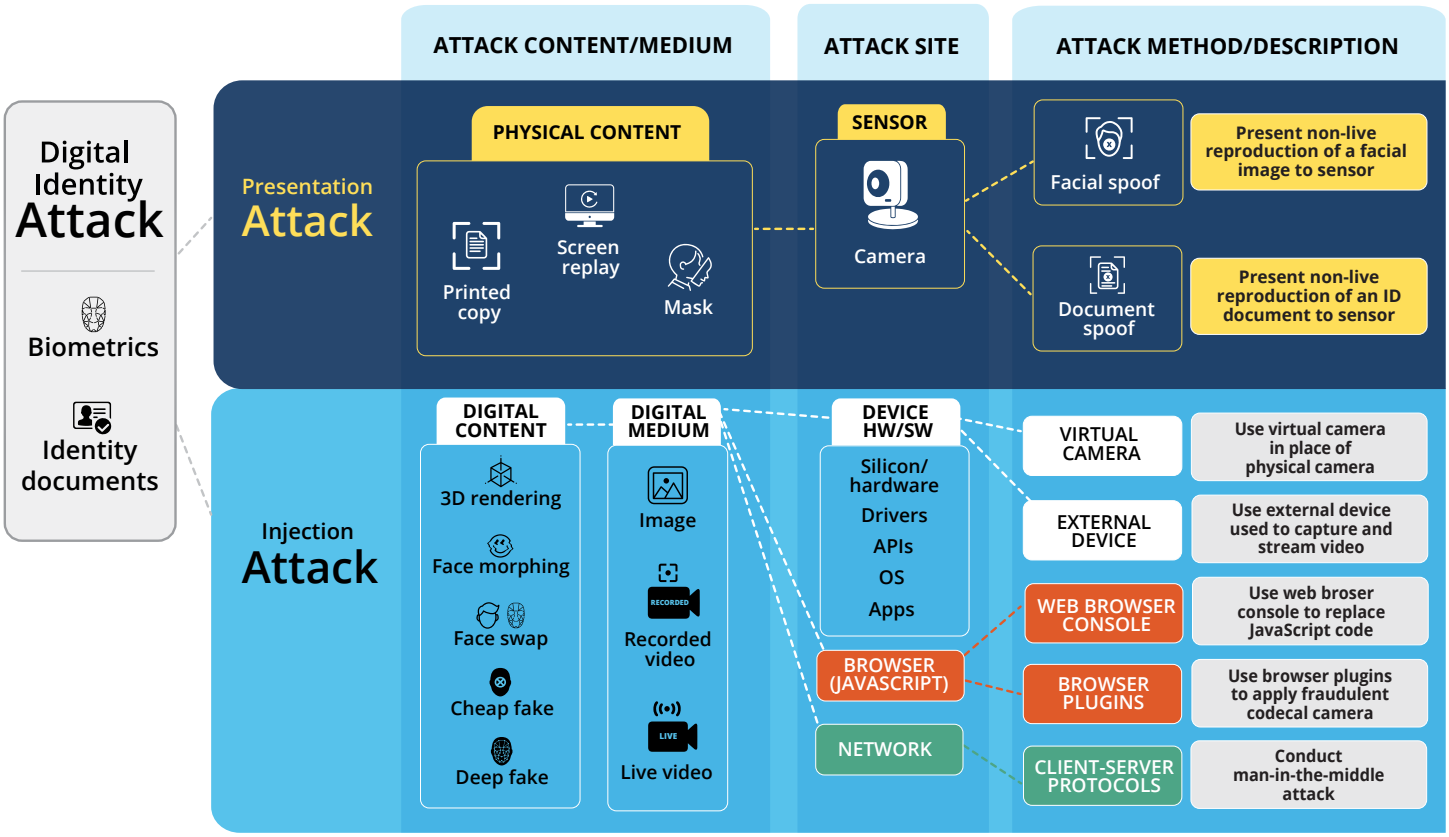


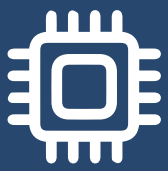
Figure: a map of attack content, sites, and methods

© 2023 ID R&D

Consider an analogous attack where a satellite transmission of a live speech by a political leader is intercepted and replaced with a digital deepfake video of them saying something else. The audience believes they are watching a live speech from their leader, when in fact they are watching a recorded speech from a deepfake saying something different. Injection attack detection helps mitigate this threat in the realm of biometric security.

As presentation attack detection has become more effective, digital injection attacks are becoming a more prevalent attack vector. The injection of fake imagery can take place at any of several attack sites within a device using either hardware or software hacks.

Examples of injection attacks are:



Hardware attacks

Use of an external camera in place of the primary device camera

Interception and replacement of signals at the chip and pin level of the device

Software attacks

Use of a virtual camera in place of the primary device camera

Interception and replacement of signals at the driver, API, and OS layer

Browser console and plugin hacks

Network attacks

Interception and replacement of signals at the network layer (e.g. man-in-the-middle attack)

Mitigating the risk: detecting attacks

Presentation attack detection (PAD) is specifically designed to detect fraudulent content presented to the camera; the presentation attacks as described above. There are multiple approaches that fall into categories of either “passive”--designed to be transparent and eliminate impact on user experience–and “active”, which rely on interaction with the user to confirm liveness.

Unlike the *physical* media used for presentation attacks, such as paper photos and screens, deepfakes and related *digitally* rendered content pose a different kind of threat. While deepfakes can be used in screen replay presentation attacks, these can be mitigated with PAD. But they can also be used in injection attacks that, as described above, bypass the camera and therefore warrant different detection methods.

© 2023 ID R&D

With the evolution of deepfakes, countermeasures to combat it have advanced as well. There are numerous methods depending on the data available for analysis:

1. **Multi-modal analysis.** AI analyzes lip movement and mouth shapes to detect discrepancies. If there are any discrepancies detected between them and the words pronounced by the deepfake — even if they are minute — the system will detect the video as a deepfake automatically.

2. **Watermarking and digital signature techniques.** These methods require calculating a hash or using a technique that embeds a hidden, digital signal or code within a video that can be used to identify the video's origin.

3. **Video analysis.** This includes a wide spectrum of different approaches based on machine learning and deep neural networks:

    a. *Motion analysis:* These methods use motion analysis algorithms to detect inconsistencies in the way objects move in the video, such as the lack of motion blur.

    b. *Frame-by-frame analysis:* These methods analyze each frame of the video individually and look for inconsistencies, such as changes in lighting, shading, or texture, which may indicate that the video has been manipulated.

    c. *Image artifact detectors:* These methods use machine learning algorithms specifically designed to detect deepfakes. They can be trained to recognize the specific artifacts that are characteristic of deepfake videos, such as non-symmetry of the face, teeth, or background noise.

4. **Digital injection attack detection.** These methods are based on detecting the bypass of the normal capture process from a proper camera as described above. They may collect additional information from the operating system, browser, and whatever software and hardware is used for video capturing. A wide range of information is collected and compared with that known to be a proper capture and known to be an injection attack.

Summing up: a comprehensive approach to liveness includes detection of presentation and injection attacks

Ultimately, the importance of protecting biometric security from deepfakes and injection attacks cannot be overstated, given the rapid advancement and proliferation of deepfake rendering software. Deepfakes have the potential to seriously compromise biometric authentication, and it is crucial that organizations take steps to protect against them. By using presentation attack detection in concert with injection attack detection, organizations can better ensure the integrity of their systems.

Source
www.antispoofing.org

© 2023 ID R&D

About ID R&D

ID R&D is an award-winning biometrics company on a mission to deliver the next generation of "zero-effort" frictionless authentication and fraud prevention. With extensive expertise in the science of biometrics and the industry's leading R&D team, we deliver a new breed of AI-driven voice, face and behavioral biometrics, and unique face, voice, and document liveness detection capabilities.

1350 Broadway, Suite 605
New York, NY 10018 USA
info@idrnd.net

www.idrnd.ai

ALL RIGHTS RESERVED IDR&D 2023