Original Article

THE INTERNATIONAL JOURNAL of EVIDENCE & PROOF

The International Journal of Evidence & Proof
1–8

DOI: 10.1177/1365712718807226
journals.sagepub.com/home/epj
SAGE

Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos

Marie-Helen Maras
John Jay College of Criminal Justice, New York, USA

Alex Alexandrou
John Jay College of Criminal Justice, New York, USA

Abstract

Deepfake videos are the product of artificial intelligence or machine-learning applications that merge, combine, replace and superimpose images and video clips onto a video, creating a fake video that appears authentic. The main issue with Deepfake videos is that anyone can produce *explicit content* without the consent of those involved. While some of these videos are humorous and benign, the majority of them are pornographic. The faces of celebrities and other well-known (and lesser-known) individuals have been superimposed on the bodies of porn stars. The existence of this technology erodes trust in video evidence and adversely affects its probative value in court. This article describes the current and future capabilities of this technology, stresses the need to plan for its treatment as evidence in court, and draws attention to its current and future impact on the authentication process of video evidence in courts. Ultimately, as the technology improves, parallel technologies will need to be developed and utilised to identify and expose fake videos.

Keywords

Artificial intelligence, authentication, Deepfake, probative value, video evidence

Introduction

Deepfake videos provide the ability to swap one person's face onto another in a video clip or an image. The technology that creates these videos is designed to continuously improve its performance. Specifically, the algorithm that creates the fake videos learns, and improves the videos by continuing to mimic

Corresponding author:
Marie-Helen Maras, John Jay College of Criminal Justice, 524 W. 59th Street, Haaren Hall, Room 43311, New York 10019-1093, USA.
E-mail: mmaras@jjay.cuny.edu

the individual's facial expressions, gestures, voice and variations, making them more and more realistic. When starting with sufficient video and audio of a person, the algorithm can not only create the fake video but can also make the person say things they have not actually said. Eventually, these videos will be indistinguishable to the naked eye from authentic videos.

Until now, this kind of computer imaging technology was accessible only to Hollywood's big-budget movies, and is known as 'computer-generated imagery' (CGI). However, as with other technologies, faster processors, high-performance graphics cards and smarter algorithms make the technology more accessible to users. Anyone can now download the Deepfake app and follow its video tutorial to create face-swap videos. The video clips, lifted from sources like interviews or publicity photos, can contain anyone's face, and may be turned into an X-rated video. This article examines the challenges Deepfakes pose to criminal justice agents and lawyers, and concludes by discussing the impact of this technology on the probative value of video evidence and the technology needed to identify fake videos and authenticate digital video evidence.

AI: What it is and what it can do (at the moment)

Deepfake videos are created using technology relying on artificial intelligence (AI), 'computational models of human behavior and thought processes that are designed to operate rationally and intelligently' (i.e., simulate human behaviour) (Maras, 2017: 7), and machine learning, 'a branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience . . . [and] carry out complex processes by learning from data, rather than following pre-programmed rules' (The Royal Society, 2017: 5). AIs continuously improve their performance by learning from experience and adjusting their behaviour based on prior performance and new inputs. AI technology is still in its infancy. Some examples of first-generation AI technology are Apple's Siri, Amazon's Alexa, Google's Nest learning thermostat and Pandora's automated music recommendation service. These software programs use machine-learning technology that identifies queries and requests using spoken human language and responds with answers from a database.

In 2015, Google released its own open source artificial intelligence tool, TensorFlow, for machine learning and image processing.[1] According to TensorFlow's website, Google's translation system and Gmail teams have used the tool to understand the context of the messages it receives, and to predict the replies that will be sent (called 'smart replies'). In the field of medicine, doctors have used TensorFlow to predict diabetic retinopathy, the most common cause of blindness among working-age adults with diabetes.[2] The Google team programmed TensorFlow to detect diabetic retinopathy by using a database of images characterised and sorted by ophthalmologists. The doctors can use the tool to look at a new image. TensorFlow compares the image to the images in its database and can predict whether or not the image depicts diabetic retinopathy. However, even though TensorFlow is a valuable tool for machine learning and image processing, the open source TensorFlow backend has been used in malicious ways to create Deepfake videos.

The face-swap technology (Deepfake)

The Deepfake program uses Google's image search, explores social media websites, and then, on its own, enters data to replace the faces in videos almost flawlessly. The program does not need any human supervision after the initial machine learning process. Instead, the algorithm continues to improve the process autonomously. Anyone can create pornographic videos starring celebrities, politicians, friends or enemies. Some of the celebrities who were victims of Deepfake videos include Aubrey Plaza, Daisy Ridley, Gal Gadot-Varsano, Natalie Portman, Scarlett Johansson, Meghan Markle and Taylor Swift

1. For information about TensorFlow, see: www.tensorflow.org (accessed 9 September 2018).
2. For information about diabetic eye disease, see: https://nei.nih.gov/health/diabetic/retinopathy (accessed 9 September 2018).

(Cuthbertson, 2018). Even former US first lady Michelle Obama was the target of a Deepfake video on Reddit (Farokhmanesh, 2018). Her face was positioned over the body of a pornography film actress whose facial structure was similar to hers.

In a society where information is consumed and reproduced quickly through social media and other forums, Deepfakes can have detrimental effects on those who are targeted in the videos. These videos often remain online for long periods and may be transferred to different forums when removed; sometimes they even reappear on the same forum. These videos can be used for revenge porn (sharing of X-rated videos or images of a person without their permission; the preferred term is 'image-based abuse'), bullying, video evidence, political sabotage, propaganda, blackmail and even fake news, which consists of methodical disinformation and propaganda that distorts actual news and facts by replacing knowledge with false images and information. While there are many basic face-morphing software programs on the market, the introduction of artificial intelligence in these programs ultimately enables the creation of more seamless videos. Put simply, when AI technology is used in the future, it may be impossible to determine that the video is fake. Even today, if this technology is paired with lower-quality videos (such as CCTV footage), the videos may be difficult to distinguish as fakes.

Beyond the creator of Deepfakes, other digital media manipulation tools exist, and are currently being developed, that can modify the images of their users. Adobe, for example, is planning to incorporate AI into its apps, to, among other things, automatically tweak selfies (Tiffany, 2017). Other software, such as Scene Stitch, created by Adobe, uses AI to alter images, enabling users to remove unwanted parts of an image. The company is also working on incorporating AI into a video editing program, which will enable users to identify an object within the frames of a video and remove it (Vincent, 2017). What is more, companies such as Pinscreen have created programs, not yet ready for public release, that enable users to create digital avatars of themselves (audio and visual representations) (Pierson, 2018). Coupled with other technologies, such as Adobe's prototype voice manipulation software (VoCo; the so-called 'Photoshop for Voice'), users will be able to create fake audio recordings, based on existing audio recordings. Eventually, individuals will easily be able to fabricate information, manipulate existing data and spread misinformation and fake audio and video recordings (Dockrill, 2016). Another company, Lyrebird,[3] has provided an audio version of TensorFlow that allows anyone to imitate a person's voice and say anything at all. In the demo page of their website,[4] the company provides the voices of Barack Obama, and Donald Trump, saying things they never actually said, which sound incredibly realistic. Nonetheless, as with TensorFlow, positive uses of the software exist, for example as a tool to help people who have lost their voices due to illness.

Probative value lost

People tend to believe what they see (Parry, 2009). For this reason, images and other forms of digital media are often accepted at face value. Digital images and videos are a powerful form of persuasion on a fact of a matter being asserted (Sherwin et al., 2006: 241–242). Visual representations of an alleged fact can be more convincing than words (Sherwin et al., 2006: 241–242). Visual presentations also inspire greater confidence in what is being conveyed visually than what is being expressed verbally (Sherwin et al., 2006: 244). Ultimately, images and videos have the ability to convince users, irrespective of whether the videos and images have been fabricated (Moonkin, 1998; Porter and Kennedy, 2012).

Thanks to digital imaging technology and easy access to applications from Adobe, Apple, Google and Microsoft, it has never been so easy to mislead the eye with tampered images or video. Although in the past manipulation of images was easy to detect (Bianchini and Bass, 1998: 309–310; Paul, 2000: B10; 2006: 46), this is not the case today. In fact, technology that has been around for quite some time can

3. The Beta version allows anyone to create their digital voice with only one minute of audio. Available at: https://lyrebird.ai/ (accessed 9 September 2018).
4. Demo of the voices of Barack Obama and Donald Trump. Available at: https://lyrebird.ai/vocal-avatar.

subtly alter images and in some cases create close-to-seamless significant alterations to these images (Bianchini and Bass, 1998: 306–307; House of Lords, 1998) As the AI technology that manipulates videos continues to learn and evolve, users will reach a point where they are unable to discern whether a video is fake. In view of that, AI-manipulated digital videos may eventually have little (if any) probative value in courts.

Authenticating digital images and videos

A review of US case law reveals limits on when digital images and videos can be introduced in a court of law as evidence. In *Nooner* v *State of Arkansas*,[5] a murder was partially captured on video, and stills of the video were introduced as evidence of the crime. The court ruled that the evidence was admissible as long as the reliability of the digital evidence could be verified. Image and video evidence may be submitted under the silent witness theory, which holds that an image or video may be introduced as evidence without the need of a witness to verify its authenticity if it has been established that the manner in which it was produced was reliable.[6] In addition to reliability, in *United States* v *Beeler*[7] and *Dolan* v *State of Florida*,[8] the courts held that digital images are admissible as long as their trustworthiness, accuracy, and authenticity can be established. The question that follows is: how are these established?

Before this question can be answered, the process of creating a digital image needs to be explored. The digital camera consists of an image plate or digital sensor that collects information through the lens aperture (the opening within a lens, from which light goes into the digital sensor). The sensor contains about a million tiny detecting sensor elements (sensel). The sensel capture the intensity of the light on the sensor and allocate a colour filter which segregates light at a specific frequency, and is either is Red, Green or Blue (RGB). The RGB mimics the human eye, and is based on the human perception of colours (trichromacy). To construct an image, the camera's processor uses the data in the image to manage the levels of colour (RGB) at each pixel. The digital image is made of pixels that are too small for the human eyes to see; therefore, the brain interprets the light emitted from the pixels as a single image. Similar to the digital image principle recording in a digital sensor, a digital video is made up of individual frames. Each frame represents an individual image. For example, at 30 frames per second (fps), every second of digital video shows 30 still images. Digital images may come in formats like BMP, GIF, PNG, JPEG, TIFF and RAW. The formats enable compression (images can be large or small, depending on the distribution of an image) and differ in the number of colours they contain. When a digital image has higher resolution, the file is larger and contains more information and detail.

Photographic images have been manipulated and modified long before digital image technology was invented (Fineman, 2012). Prior to digital image technology, the forensic investigator had to compare the actual photograph with the negative to verify the authenticity of an image. In digital imaging, RAW file format is the counterpart of a film negative. The RAW file format is produced by a digital camera and is the original or unaffected pixel information of an image. The post-processing (i.e., the alteration) of a produced digital image leaves traces (or footprints). For example, using an audit trail and history log in Adobe Photoshop CS can reveal whether valid techniques were used, and how each technique altered the image. The methods used to detect alterations to images involve the recovery of these footprints and their use to reconstruct processing actions post image production (a form of 'reverse engineering').

Digital image forensics has largely focused on detecting low-level alterations in images, such as dropping or duplicating a frame or frames and/or regions (Gironi et al., 2014; Li et al., 2007; Long et al., 2017; Mahalakshmi et al., 2012; Wang and Farid, 2007), splicing and copy-pasting a part or parts of the original image and placing them in other areas (i.e., a copy-move manipulation) (Bestagini et al., 2013;

5. *Nooner* v *State of Arkansas*, 907 S.W.2d 677 (1995).
6. *United States* v *McMahon*, 2008 CCA LEXIS 87 (N-M.C.C.A. 2008); *United States* v *Harris*, 55 M.J. 433 (C.A.A.F. 2001).
7. *United States* v *Beeler*, 62 F Supp. 2d 136 (July 1, 1999, United States District Court, D. Maine).
8. *Dolan* v *State of Florida*, 743 S. 2d 544 (July 21, 1999, Court of Appeal of Florida, Fourth District).

Christlein et al., 2012; D'Amiano et al., 2018; Hashmia et al., 2014; Mahalakshmi et al., 2012). The various image forensics methods, however, focus 'on specific image features which oftentimes disappear with post-processing' (Rossler et al., 2018). Most 'forensic analysis of video content proves to be harder with respect to the analysis of still images since video data are practically always available in compressed formats and several times a high compression factor is used to store it. Strong compression ratios may cancel or fatally compromise the existing footprints so that the processing history is, entirely or in part, no longer recoverable' (Milani et al., 2012). These challenges are more pronounced today as videos uploaded and distributed online in general, and on social media platforms in particular, are compressed and/or of low resolution (Rossler et al., 2018). These challenges also serve as obstacles to the verification of the trustworthiness, accuracy and authenticity of digital images and videos.

To admit digital evidence, images and videos, Rule 901(b) (9) of the US Federal Rules of Evidence further requires the introduction of 'evidence describing a process or system and showing that it produces an accurate result'. This requires the testimony of someone with technical, scientific or specialised knowledge of the issue so the person can explain why the evidence is valid and reliable. In addition, the ability to convince that digital evidence is worthy depends on the qualifications and competence of the forensic expert. Under Rule 702 of the US Federal Rules of Evidence, 'a witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if . . . the expert's scientific, technical, or other specialised knowledge will help the trier of fact to understand the evidence or to determine a fact in issue.' The qualifications of the digital media forensics expert who authenticates digital media may be questioned. Usually, a digital media forensics expert has a computer or law enforcement background with experience in collecting, examining, preserving and analysing digital information, and finally presenting this evidence in a courtroom (Maras, 2014). During this process, the expert must provide a detailed description of the steps taken throughout the digital media forensics process, what was uncovered and the conclusions reached, based on the results of the process and the evidence revealed. Machine-learning algorithms and applications have been developed to verify the integrity of a digital image and have provided reliable results (Zhou et al, 2018; Farid, 2008) However, the use of artificial intelligence and machine-learning applications could pose problems in the reporting and presentation phases of the digital media forensics process because experts may not be able to explain how these results were obtained. Therefore, experts in the field of image and video forensics analysis must also, in the near future, be well-versed in machine learning and artificial intelligence as applied to the field of media forensics to explain why the results obtained using them are valid and reliable.

Because of the existence of digital manipulation technologies, criminal convictions of perpetrators based on images and videos will be at risk if criminal justice agents and legal professionals do not consider the vulnerability of images and videos to manipulation. These individuals should be cognisant of how easily fabricated videos could produce false evidence and lead to wrongful convictions. Digital recordings present unique authentication issues, and their original unaltered videos may or may not be available. Because images and videos that have been authenticated but are not what they purport to be could be mistakenly admitted as evidence, criminal justice agents and legal professionals should exercise caution when introducing this type of evidence in court. Corroborating evidence is needed to show that the images and videos were not modified or manipulated. Historically, digital media has been introduced to corroborate eyewitness testimony. Presently, and more so in the future, eyewitness testimony will be mandated to corroborate images and videos. Nevertheless, as machine learning and AI technology advances, the testimony of digital media forensic experts may not be enough to authenticate evidence, because even expert witnesses may not be able to discern the modifications made to digital videos.

What the future holds: Using AI to detect AI video manipulation

Deepfake videos are relative newcomers to digital media forensics. Videos must be authenticated before they are introduced as evidence in a court of law; however, this process is complicated by the existence of machine-learning algorithms and AI. While not yet operating at their full potential, machine-learning

algorithms and AI are designed to constantly improve their performance. As such, it is only a matter of time before fake videos are so convincing that they are difficult to identify as fakes.

Research in this field, such as research on face manipulation detection, is sparse, and so too is the availability of datasets that could be used to assist in the detection of face manipulation and other alterations in videos. In 2018, researchers created a dataset including 'about half a million edited images (from over 1000 videos)', and published their research, which showed that the machine-learning algorithm (XceptionNet) they developed could distinguish between face swaps and videos without this modification (Rossler et al., 2018). The same algorithm, however, can also be used to make the face swaps more seamless, making the detection of fake videos even more difficult (Emerging Technology from the arXiv, 2018). Realising the scope and magnitude of the problem of face-swapping technology, Deepfakes and other fake videos, the United States Defense Advanced Research Projects Agency (DARPA) launched a major research initiative in media forensics (named Media Forensics or MediFor for short)[9] to facilitate the creation of technology that automatically detects and assesses the integrity of digital visual media (i.e., determines whether they are unadultered images or fakes). These technologies, however, apart from XceptionNet, have not yet been developed and tested.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Bestagini P, Milani S, Tagliasacchi M and Tubaro S (2013) Local tampering detection in video sequences. IEEE International Workshop on Multimedia Signal Processing, Pula, Italy, 30 September—2 October 2013, pp. 488–493.

Bianchini VE and Bass H (1999) A paradigm for the authentication of photographic evidence in the Digital Age. *Thomas Jefferson Law Review* 20: 303–322.

Christlein V, Riess C, Jordan J and Angelopoulou E (2012) An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security* 7(6): 1841–1854.

Cuthbertson A (2018) What is Deepfake porn? AI brings face-swapping to a disturbing new level. *Newsweek*, 8 February. Available at: www.newsweek.com/what-deepfake-porn-ai-brings-face-swapping-disturbing-new-level-801328 (accessed 9 September 2018).

D'Amiano L, Cozzolino D, Poggi G and Verdoliva L (2018) A patchmatch-based dense-field algorithm for video copy-move detection and localization. IEEE Transactions on Circuits and Systems for Video Technology. Early access. DOI: 10.1109/TCSVT.2018.2804768.

Dockrill P (2016) Adobe's new 'Photoshop for voice' app lets you put words into people's mouths. *Science Alert*, 11 November 11. Available at: www.sciencealert.com/adobe-s-new-photoshop-for-voice-app-lets-you-put-words-in-people-s-mouths (accessed 9 September 2018).

Emerging Technology from the arXiv (2018) This algorithm automatically spots 'face swaps' in videos. *MIT Technology Review*, April 10. Available at: www.technologyreview.com/s/610784/this-algorithm-automatically-spots-face-swaps-in-videos/ (accessed 9 September 2018).

Farid H (2008) Digital image forensics. *Scientific American* 298(6): 66–71.

9. For information about Media Forensics (MediFor) see: www.darpa.mil/program/media-forensics (accessed 9 September).

Farokhmanesh N (2018) Deepfakes are disappearing from parts of the web, but they are not going away. The Verge, 9 February. Available at: www.theverge.com/2018/2/9/16986602/deepfakes-banned-reddit-ai-faceswap-porn (accessed 9 September 2018).

Fineman M (2012) *Faking It: Manipulated Photography before Photoshop*. New York: Metropolitan Museum of Art.

Gironi A, Fontani M, Bianchi T and Piva A Barni M (2014) A video forensic technique for detection frame deletion and insertion. IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014, pp. 6226–6230.

Hashmia MF, Anandb V and Keskarc AG (2014) Copy-move image forgery detection using an efficient and robust method combining un-decimated wavelet transform and scale invariant feature transform. *AASRI Procedia* 9: 84–91.

House of Lords (1998) Science and Technology—Fifth Report. Select Committee on Science and Technology Committee Reports, Session 1997–98. Available at: https://publications.parliament.uk/pa/ld199798/ldselect/ldsctech/064v/st0503.htm (accessed 9 September 2018).

Li G, Qiong W, Tu D and Sun S (2007) A sorted neighbourhood approach for detecting duplicated regions in image forgeries based on DWT and SVD. In: *Proceedings of IEEE International Conference on Multimedia and Expo*, Beijing, China, 2–5 July 2007, pp. 1750–1753.

Long C, Smith E, Basharat A and Hoogs A (2017) A C3D-based convolutional neural network for frame dropping detection in a single video shot. IEEE Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 21–26 July, 2017, pp. 1898–1906.

Mahalakshmi DS, Vijayalakshmi K and Priyadharsini S (2012) Digital image forgery detection and estimation by exploring basic image manipulations. *Digital Investigation* 8(3): 215–225.

Maras M-H (2014) *Computer Forensics: Cybercriminals, Laws, and Evidence*, 2nd ed. Burlington, MA: Jones and Bartlett.

Maras M-H (2017) Social media platforms: Targeting the 'found space' of terrorists. *Journal of Internet Law* 21(2): 3–9.

Milani S, Fontani M, Bestagini P, Barni M, Piva A, Tagliasacchi M and Tubaro S (2012) An overview on video forensics. APSIPA Transactions on Signals and Information Processing, Los Angeles, USA, 3–6 December, 2012, pp. 1–18.

Moonkin JL (1998) The image of truth: Photographic evidence and the power of analogy. *Yale Journal of Law and Humanities* 10(1): 1–74. Available at: http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1181&context=yjlh (accessed 9 September 2018).

Paul G (2000) Fabrication of evidence: A click away. *The National Law Journal*, 21 February.

Paul G (2006) The authenticity crisis in real evidence. *Practical Litigator* November: 45–52. Available at: www.lrrc.com/files/uploads/documents/Paul_TheAuthenticityCrisisInRealEvidence_Practicallitigator_2004.pdf (accessed 9 September 2018).

Parry ZB (2009) Digital manipulation of photographic evidence: Defrauding the courts one thousand words at a time. *Journal of Law, Technology & Policy* 81: 176.

Pierson D (2018) Fake videos are on the rise. As they become more realistic, seeing shouldn't always be believing. Los Angeles Times, 19 February 19. Available at: www.latimes.com/business/technology/la-fi-tn-fake-videos-20180219-story.html (accessed 9 September 2018).

Porter G and Kennedy M (2012) Photographic truth and evidence. *Australian Journal of Forensic Sciences* 44(2): 183–192.

Rossler A, Cozzolino D, Verdoliva L, Reiss C, Thies J and Niebner M (2018) FaceForensics: A-large-scale video dataset for forgery detection in human faces. Available at: https://arxiv.org/pdf/1803.09179.pdf (accessed 9 September 2018).

Sherwin RK, Feingenson N and Spiesel C (2006) Law in the Digital Age: How visual communication technologies are transforming the practice, theory, and teaching of law. *Boston University Journal of Science and Technology Law* 12(2): 227–270.

The Royal Society (2017) *Machine Learning: The Power And Promise Of Computers That Learn By Example*. London: The Royal Society. Available at: https://royalsociety.org/∼/media/policy/projects/machine-learning/publications/machine-learning-report.pdf (accessed 9 September 2018).

Tiffany K (2017) Adobe tries to make selfies less embarrassing using AI and machine learning. The Verge, 6 April. Available at: www.theverge.com/tldr/2017/4/6/15209202/adobe-sensei-selfie-improving-ai-machine-learning (accessed 9 September 2018).

Vincent J (2017) Adobe's prototype AI tools let you instantly edit photos and videos. *The Verge*, 24 October 24. Available at: www.theverge.com/2017/10/24/16533374/ai-fake-images-videos-edit-adobe-sensei (accessed 9 September 2018).

Wang W and Farid H (2007) Exposing digital forgeries in interlaced and deinterlaced video. *IEEE Transactions on Information Forensics and Security* 2(3): 438–449.

Zhou P, Han X, Morariu VI and Davis LS (2018) Learning rich features for image manipulation detection. arXiv preprint arXiv:1805.04953.