Detection of Deepfake Video Manipulation

Conference Paper · August 2018

CITATIONS
0

READS
1,727

3 authors, including:

Andrea Macarulla Rodriguez
Netherlands Forensic Institute
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Zeno Geradts
Netherlands Forensic Institute
88 PUBLICATIONS **680** CITATIONS
SEE PROFILE

Detection of Deepfake Video Manipulation

Marissa Koopman, Andrea Macarulla Rodriguez, Zeno Geradts

University of Amsterdam & Netherlands Forensic Institute

Abstract

The Deepfake algorithm allows a user to switch the face of one actor in a video with the face of a different actor in a photorealistic manner. This poses forensic challenges with regards to the reliability of video evidence. To contribute to a solution, photo response non uniformity (PRNU) analysis is tested for its effectiveness at detecting Deepfake video manipulation. The PRNU analysis shows a significant difference in mean normalised cross correlation scores between authentic videos and Deepfakes.

Keywords: Video Manipulation, Digital Forensics, PRNU, Neural Network, Deepfake.

1 Introduction

Photographic and video evidence are commonly used in the courtroom and police investigations, and are seen as reliable types of evidence. With the advances in video editing techniques however, video evidence is becoming potentially unreliable. It is probable that, in the near future, it will be required that video evidence needs to be examined for traces of tampering before being deemed admissible to court.

A new video manipulation technique known as Deepfake has established itself online over the last few months. Deepfake manipulation allows a user to replace the face of an actor in a video with the face of a second actor, provided that enough images (several hundred to thousands) are available of both actors. These videos are known as 'Deepfakes'. Deepfakes quickly gained notoriety in the media due to their application to porn videos, where the faces of famous actresses and politicians were 'Deepfaked' into existing porn videos on websites such as Reddit and PornHub [Matsakis, 2018].

What distinguishes Deepfakes from other video manipulation techniques are, firstly, its potential for photorealistic results; with enough images of both actors, and enough computer training time, the resulting videos can be extremely convincing. Secondly, the availability of the technique to laypersons. An app named FakeApp was quickly released on Reddit, which is essentially a guided user interface wrapped around the Deepfake algorithm, allowing users with limited knowledge of programming and machine learning to create Deepfakes. Several other versions followed such as OpenFaceSwap [Anonymous, 2018].

The combination of photorealistic results and ease of use poses a unique forensic challenge. It becomes increasingly feasible that every day videographic evidence has been manipulated, creating an increased need for verified authentication methods to detect the Deepfake manipulation. This is an especially sensitive and urgent problem in the current 'fake news' era, going beyond law enforcement and becoming relevant also to journalists, video hosting websites, and social media users. Due to this, authentication methods which are approachable and usable for a wide and private audience are ideal.

Considering the above, this paper explores the use of photo response non uniformity (PRNU) analysis applied to Deepfakes to assess the method's accuracy and ease of use in detecting the Deepfake manipulation. The PRNU pattern of a digital image is a noise pattern created by small factory defects in the light sensitive sensors of a digital camera [Lukas et al., 2006]. This noise pattern is highly individualising, and is often referred to as

 ISBN 978-0-9934207-3-3

the fingerprint of the digital image [Rosenfeld and Sencar, 2009]. PRNU analysis is considered a method of interest because it is expected that the manipulation of the facial area will affect the local PRNU pattern in the video frames. Furthermore, it is widely used in image forensics and is as such familiar to the experts working in the field.

2 Current authentication efforts

No academic papers could be found on the detection of Deepfakes, although efforts to detect and remove them are being made at websites such as Gyfcat [Matsakis, 2018]. Gyfcat attempts to use artificial intelligence and facial recognition software to spot inconsistencies in the rendering of the facial area of an uploaded video. When a video has been flagged as suspicious, a second program masks the facial area and checks whether a video with the same body and background has been uploaded before. If such a video is found, but the faces of the original and the newly uploaded video do not match, then the software concludes that the new video has been manipulated [Matsakis, 2018].

Such a method is uniquely suited to a website such as Gyfcat, where millions of videos are uploaded and there are vast databases of reference videos. The method may not be as applicable to forensic cases, where for instance some CCTV footage from a store robbery could be Deepfaked. Nor would it detect Deepfakes which do not have an original version stored in the databases. As such, techniques which do not rely on vast databases are needed.

3 Methods

3.1 Dataset

The dataset consists of ten authentic, unmanipulated videos between 20 and 40 seconds in length, and of 16 Deepfakes made by the researcher. The videos are made by a Canon PowerShot SX210 IS, and are in .MOV format, with a resolution of 1280x720 pixels. The Deepfakes are made using the ten authentic videos captured by the Canon PowerShot SX210 IS, and using the Deepfake GUI OpenFaceSwap [Anonymous, 2018]. Three different actors are used interchangeably to create the authentic videos, as well as the Deepfakes.

3.2 PRNU analysis

The videos are turned into a series of frames as PNGs with the use of the software 'FFmpeg' [FFmpeg Developers, 2018], named sequentially, and kept in labelled folders. In order to increase the significance of the expected change in PRNU pattern in the facial area of the frame, the frames will be cropped to frame the face, also with the use of 'FFmpeg'. Each frame of a video is cropped by the exact same pixels in order to leave the portion of the PRNU pattern which is examined consistent between each cropped frame. An example of how the frames are cropped can be found in figure 1.

The frames are then sequentially divided into eight groups of equal size, and an average PRNU pattern is created for each group using the second order (FSTV) method [Baar et al., 2012] with the software 'PRNUCompare' [Ministry of Security and Justice, 2013]. These eight PRNU patterns are then compared to one another, and normalised cross correlation scores are returned. The variations in correlation scores and the average correlation score for each video are calculated. A Welch's t-test will be applied to the results in order to assess the statistical significance between the results for Deepfakes and for authentic videos [Welch, 1947].

 ISBN 978-0-9934207-3-3

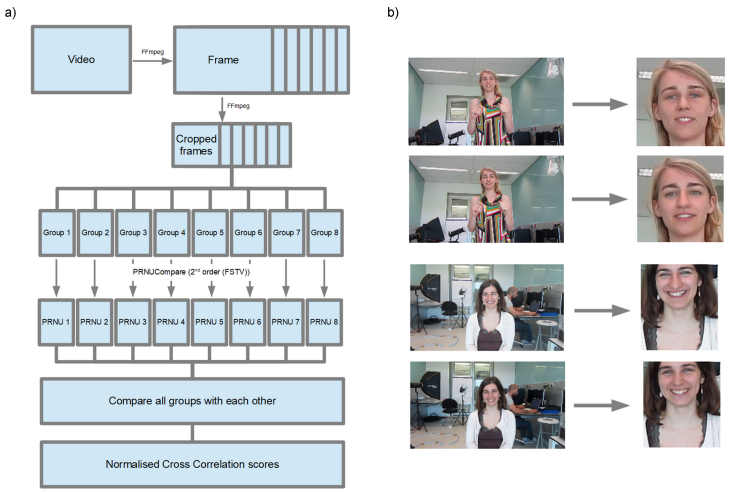

Figure 1: a) Frames are extracted from the video and cropped to contain the questioned face. The cropped frames are split evenly and sequentially over eight groups. An average PRNU pattern is calculated for each group. The PRNU pattern of each group is then compared to the PRNU patterns of the other seven groups. Normalised cross correlation scores are calculated for each comparison. b) Frames are extracted from the video, and cropped down to the exact same pixels which contain the questioned face.

4 Results

The mean normalised cross correlation scores per video and the variance in normalised cross correlation scores per video are calculated. The results are illustrated in figure 2.

The results indicate that there is no correlation between the authenticity of the video and the variance in correlation scores. There does appear to be a correlation between the mean correlation scores and the authenticity of the video, where on average original videos have higher mean normalised cross correlation scores compared to the Deepfakes. The difference in the distribution of mean normalised cross correlation scores is statistically significant, with a p-value of $5.21 * 10^{-5}$.

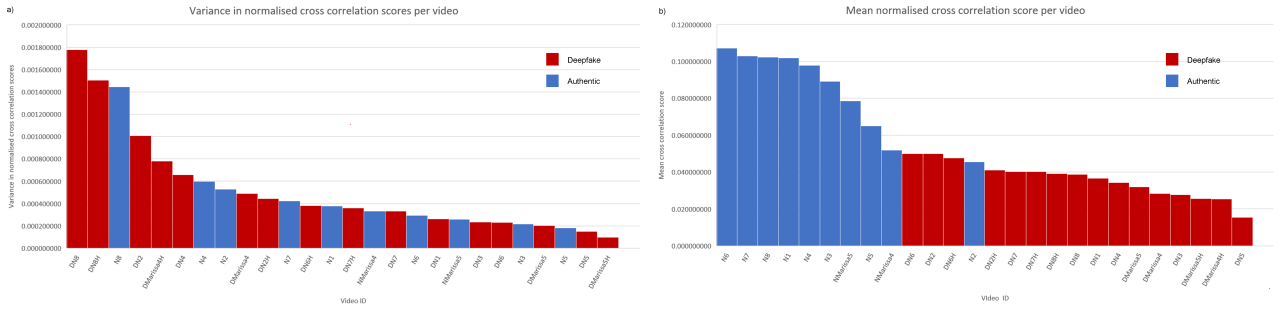

Figure 2: a) The average variation in correlation scores per authentic and per Deepfake video. b) The average correlation score per authentic and per Deepfake video.

5 Conclusion

It appears that the mean normalised cross correlation score can be used to distinguish Deepfakes from authentic videos. The dataset is too small to formulate guidelines for likelihood ratios, as is desired in forensic sciences. However, a cut-off value of 0.05 results in a 3.8% false positive rate, and a 0% false negative rate within our

 ISBN 978-0-9934207-3-3

dataset. As such, PRNU analysis may be suitable for the detection of Deepfakes. Before such an application can be advised however, further research must be done with larger datasets in order to confirm the correlation and to formulate reliable likelihood ratios.

Acknowledgments

I would like to thank Hugo Dictus for his continued support and advice throughout the project.

References

[Anonymous, 2018] Anonymous (2018). Openfaceswap: A deepfakes gui. https://www.deepfakes.club/openfaceswap-deepfakes-software/.

[Baar et al., 2012] Baar, T., van Houten, W., and Geradts, Z. (2012). Camera identification by grouping images from database, based on shared noise patterns. *arXiv preprint arXiv:1207.2641*.

[FFmpeg Developers, 2018] FFmpeg Developers (2018). ffmpeg tool (version 4.0 "wu"). http://ffmpeg.org/.

[Lukas et al., 2006] Lukas, J., Fridrich, J., and Goljan, M. (2006). Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214.

[Matsakis, 2018] Matsakis, L. (2018). Artificial intelligence is now fighting fake porn. https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/.

[Ministry of Security and Justice, 2013] Ministry of Security and Justice (2013). Finding the link between camera and image. camera individualisation with prnu compare. professional from the netherlands forensic institute. https://www.forensicinstitute.nl/binaries/forensicinstitute/documents/publications/2017/03/06/brochure-prnu-compare-professional/brochure-nfi-prnu-compare-professional_tcm36-21580.pdf.

[Rosenfeld and Sencar, 2009] Rosenfeld, K. and Sencar, H. T. (2009). A study of the robustness of prnu-based camera identification. In *Media Forensics and Security*, volume 7254, page 72540M. International Society for Optics and Photonics.

[Welch, 1947] Welch, B. L. (1947). The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

 ISBN 978-0-9934207-3-3