Using Benford's Law for Deepfake Detection

Miguel Leão
miguel.leao@isr.uc.pt
Nuno Gonçalves
nunogon@deec.uc.pt

Institute of Systems and Robotics
University of Coimbra
Portugal

Abstract

The developments in home computers, united with the thousands upon thousands of images/videos of individuals present on the Internet, allowed for the proliferation of deepfaked media affecting the lives of private individuals and the dangerous spread of misinformation. There's need for a consistent method of detection of the tampered media as to impede these negative aspects. This work investigates the use of Benford's Law to detect deepfaked material by analyzing the frequency domain of bonafide facial images and the deepfakes, searching for a significant deviation in their value distribution that may be used to distinguish between the two. Through the approaches presented in this paper, this deviation was not found.

1 Introduction

The name "Deepfake" comes from the joining of the "fake" image of an identity swap, using deep learning techniques. While they can be presented as a benign or beneficial technology, being used only in academic purposes or in entertainment, they also pose very real threats. The use of deepfakes for the creation of fake scenarios to spread misinformation, disrupt the flow of information through media and personally affect individuals as already been proven to be a very real problem[1], and as such there needs to be countermeasures in place as to conserve the viability of news cycles.

Benford's Law [2] has been applied to many sets of data since it was proposed in 1938 ranging from voting data to image forensics (further explained in section 2 of this paper). It is in the interest of image forensics that this paper presents the possibility of a different approach to deepfake detection by leveraging the compliance to Benford's Law of bonafide face images. Inspired by previous work where morphed fingerprints were distinguishable from bonafide fingerprints since only the bonafide images followed Benford's distribution. The hypothesis is that the deepfake images won't comply and as such the analysis of the Discrete Cosine Transform (DCT) or other transforms of the video frames or images could be an alternative for deepfake detection.

2 Literature Review

Deepfakes are usually created through auto encoders or GAN's and have been detected through several approaches. For the first generation of deepfakes (related to the quality of the created image and the available technology at the time), most detection techniques searched for inconsistencies in the data. Either in face movement, head pose, eye blinking, smiles, lighting and shadows or other aspects that result in an "uncanny" video/image [1], or in artifacts related to the video/image editing like smear frames or different image noise between segments of the media [13]. With the development of technology and the coming of newer, higher quality, deepfakes, these detection methods require innovation, shifting from forensic approaches to machine learning solutions.

Benford's Law [2] or the law of anomalous numbers is an empirical law, that states that the leading digit $d \in (1,2,...,9)$ will occur in a set of data with a probability of:

$$P(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(1+\frac{1}{d}) \tag{1}$$

It as been applied to many different sets of data and has been shown to detect fraudulent manipulation of election and financial data. The variety of data that complies with Benford's Law eventually reached digital images with Jolion in [12], showing that the magnitude of the gradient of an image follows the law, followed by Acebo and Sbert showing the same with the light intensities in natural images [5] and Fu et al. [7] developing a generalized Benford's Law which they applied to JPEG compression, in order to study how their discoveries may be used in image forensics.

Biometrics also seem to follow Benford's Law with facial images, fingerprints [16] and even human speech patterns [9], continuing the trend that legitimate the inference that the data follows the suggested distribution. It stands to question that tampered biometrics would not follow the same distribution as the legitimate images. With the Generalized Benford's Law described by Fu et al., Iorliam and Caleb [10] leverage the law to distinguish between optically acquired, artificially printed contactless acquired latent and synthetically generated fingerprints by studying the divergence from Benford's distribution for each category. Satapathy et al. [16] follow the same approach to detect morphed (or double-identity) fingerprints.

3 Approach

The approach to the hypothesis presented in this paper is simple: confirming that images of bonafide faces follow Benford's Law as described by Iorliam et al. [11] and then checking if the deepfake images have a significant deviation as to be detected.

3.1 "Reading" the image

The analysis of the image isn't as straight forward as could be expected. Pérez-González et al. [15] state that a grey-level image won't respect Benford's Law but the DCT of an image does indeed comply. While their work and the previously referenced ones work with 8×8 block-DCT, Pérez-González observed that different block sizes result in similar, positive, results. It then stands to believe that a direct 2D-DCT of the full face gray-scale image is adequate to conduct this work.

While the standard Benford's Law applies to the first significant digit, Hill proposes a generalization for digits beyond the first [8] with the probability that a digit d is encountered as the n-th digit being:

$$\sum_{k=10^{n-2}}^{10^{n-1}-1} log_{10}(1+\frac{1}{10k+d}) \tag{2}$$

It is then a matter of counting the occurrences of the digits of interest and measuring the quality of the fit using the χ^2 divergence:

$$\chi^2 = \sum_{i=N}^{9} \frac{(p_i - b_i)^2}{b_i} \tag{3}$$

where N is equal to 1 for the standard Benford's Law and equal to 0 for Hill's generalization, p_i is the observed probability of the n-th significant digit and b_i the predicted probability.

3.2 Image Acquisition

The images used in this work were sourced from the VGGFace2 dataset [4] and the Deepfake Detection Challenge (DFDC) Dataset [6].

VGGFace2 contains 3.31 million images of 9131 different identities, collected from Google Image Search of public figures. It represents a wide range of several factors relevant to the individuals themselves such as age or ethnicity but also of poses across the images. From this dataset a number of deepfaked images were created using Faceswap. On average, each identity will have 362.6 images, which is a low number for deepfake creation, resulting in low quality deepfakes easily recognized by human observers.

In contrast, the DFDC dataset applied eight different deepfake creation techniques to commissioned videos of individuals. The total 48,190

[1] See for example:
https://www.bbc.com/news/entertainment-arts-65854112
https://www.bbc.com/news/technology-60780142
https://www.bbc.com/news/av/technology-55431004

bonafide videos of 3,426 subjects resulted in 104,500 fake videos with varying degrees of deepfake quality.

4 Experiments and Results

The previously described approach yielded the results presented in table 1, with similar results being obtained when analyzing patches of the images relating to the area of the eyes, nose and mouth.

Table 1: χ^2 divergences between the first, second, and third significant digits and the expected distribution according to Hill's Generalized Benford's Law.

	1st digit	2nd digit	3rd digit
VGG2 bonafide	0.0033	0.0004	0.0003
VGG2 deepfake	0.0033	0.0004	0.0003
DFCD bonafide	0.0085	0.0002	0.00004
DFDC deepfake	0.0092	0.0003	0.00004

Various combinations of frequency domains such as Discrete Fourier Transform (DFT) or Discrete Wavelet Transform (DWT), and correlation metrics were used, and like [17] an extended vector of first significant digits across all transforms was tested with Pearson's correlation coefficient. These results showed no difference between unadulterated material and deepfake material, concluding with this approach not being usable for deepfake detection.

5 Discussion and Conclusion

The results presented do not give merit to the proposed approach. Wang et al. [18] had previously raised concerns with the use of Benford's Law in image forensics, showing that both "natural" landscape pictures and "unnatural" artistic renditions of landscapes followed Benford's Law closely when treated with common image processing techniques, not allowing for a discrimination to be made.

In contrast, Bonettini et al. [3] showed that GAN generated images could be detected with Benford's law, explaining the slight discrepancy for the DFDC data which used GAN methods for some of their deepfakes (note that the dataset wasn't entirely used, the discrepancy showed could increase or decrease if more of the data is used, depending on the distribution of methods used), while there is no discrepancy between the bonafides and deepfakes of the VGGFace2 dataset since all deepfakes were created using auto-encoders. They do however recognize some Convolutional Neural Networks seem to produce images that are harder to recognize as GAN generated.

There are still further developments that can be made to study if and how Benford's Law may be used for deepfake detection, but from the results that were already achieved it is possible to conclude that the use of Hill's [8] generalization mat be dropped since the χ^2 divergences obtained for the second and third significant digits are overall very small and vary little between themselves, giving little discriminatory value.

While the proposed approach proved not viable for deepfake detection, it is not correct to assume that "Benford's Law cannot be used for deepfake detection". There are multiple paths that can be taken in an attempt to turn this idea into a viable approach.

- Analyzing not only the eyes, nose and mouth patches, but instead following the approach of Nirkin et al. [14] where the focus is the "border" created by the deepfaked face and how it interacts in the context of the image;
- There are many more deepfakes available for study, and many deepfakes creation methods not approached in this work. There is always the possibility that what does not work for deepfakes created using method A might work for the ones created with method B.

References

[1] Shruti Agarwal and Hany Farid. Detecting deep-fake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 981–989, June 2021.

[2] Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938. ISSN 0003049X.

[3] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro. On the use of benford's law to detect gan-generated images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5495–5502, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[5] Esteve del Acebo and Mateu Sbert. Benford's law for natural and synthetic images. In *International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, 2005.

[6] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *ArXiv*, abs/2006.07397, 2020.

[7] Dongdong Fu, Yun Qing Shi, and Wei Su. A generalized benford's law for jpeg coefficients and its applications in image forensics. In *Electronic imaging*, 2007.

[8] Theodore P. Hill. The significant-digit phenomenon. *American Mathematical Monthly*, 102:322–327, 1995.

[9] Leo Hsu and Visar Berisha. Does human speech follow benford's law? In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[10] Aamo Iorliam and Shangbum F. Caleb. On the use of benford's law to detect jpeg biometric data tampering. *Journal of Information Security*, pages 240–256, 2017.

[11] Aamo Iorliam, Anthony Tung Shuen Ho, Norman Poh, and Yun Qing Shi. Do biometric images follow benford's law? *2nd International Workshop on Biometrics and Forensics*, pages 1–6, 2014.

[12] Jean-Michel Jolion. Images and benford's law. In *Journal of Mathematical Imaging and Vision*, volume 14, pages 73–81, 2001.

[13] Sohail Ahmed Khan and Hang Dai. Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1821–1828, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517.

[14] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105, 2018.

[15] Fernando Pérez-González, Greg L. Heileman, and Chaouki T. Abdallah. Benford's law in image processing. In *2007 IEEE International Conference on Image Processing*, volume 1, pages 405–408, 2007.

[16] Govind Satapathy, Gaurab Bhattacharya, Niladri Bihari Puhan, and A.T.S. Ho. Generalized benford's law for fake fingerprint detection. *2020 IEEE Applied Signal Processing Conference (ASPCON)*, pages 242–246, 2020.

[17] Domonkos Varga. Benford's law and perceptual features for face image quality assessment. *Signals*, 4(4):859–876, 2023. ISSN 2624-6120.

[18] Jingwei Wang, Byung-Ho Cha, Seong-Ho Cho, and C.-C. Jay Kuo. Understanding benford's law and its vulnerability in image forensics. In *2009 IEEE International Conference on Multimedia and Expo*, pages 1568–1571, 2009.