SUBJECTIVE AND OBJECTIVE EVALUATION OF DEEPFAKE VIDEOS

Pavel Korshunov and Sébastien Marcel

Idiap Research Institute, Martigny, Switzerland
{pavel.korshunov,sebastien.marcel}@idiap.ch

ABSTRACT

Practically anyone can now generate a realistic looking deepfake video. It is clear that the online prevalence of such fake videos will erode the societal trust in video evidence even further. To counter the looming threat, many methods to detect deepfakes were recently proposed by the research community. However, it is still unclear how realistic deepfake videos are for an average person and whether the algorithms are significantly better than humans at detecting them. Therefore, this paper, presents a subjective study, which, using 60 naïve subjects, evaluates how hard it is for humans to see if a video is a deepfake or not. For the study, 120 videos (60 deepfakes and 60 originals) were manually selected from the Facebook database used in Kaggle's Deepfake Detection Challenge 2020. The results of the subjective evaluation were compared with two state of the art deepfake detection methods, based on Xception and EfficientNet (B4 variant) neural network models pre-trained on two other public databases: Google and Jiqsaw subset from FaceForensics++ and Celeb-DF v2 dataset. The experiments demonstrate that while the human perception is very different from the perception of a machine, both successfully but in different ways are fooled by deepfakes. Specifically, algorithms struggle to detect the deepfake videos that humans find to be very easy to spot.

Index Terms— Deepfake videos, subjective evaluation, deepfake detection

1. INTRODUCTION

Autoencoders and generative adversarial networks (GANs) significantly improved the quality and realism of the automated image generation and face swapping, leading to the deepfake phenomena. Many are starting to believe that the proverb 'seeing is believing' is starting to loose its meaning when it comes to digital video [1]. This public unease prompted researchers to propose various datasets of deepfakes and methods to detect them. Some of the latest approaches demonstrate encouraging accuracy, especially, if they are trained and evaluated on the same datasets.

Many databases with deepfake videos were created to help develop and train deepfake detection methods. One of the first freely available database was based on VidTIMIT [2], followed by the FaceForensics database, with 1000 deepfakes generated from Youtube videos, which later was extended to FaceForensics++ with a larger set of high resolution videos provided by Google and Jiqsaw [3]. Another recently proposed 5000 videos-large database of deepfakes generated from Youtube videos is Celeb-DF v2 [4]. But the most extensive and the largest database to date with more than $100K$ videos (80% of which are deepfakes) is the dataset from Facebook [5], which was used in Deepfake Detection Challenge 2020 hosted by Kaggle[1] (see face examples in Figure 1).

These datasets were generated using either the popular open source code[2] (e.g., DeepfakeTIMIT [2], FaceForensics++ [3], and Celeb-DF [4]) or the variety of different GAN-based methods used in datasets by Google (specific information is not available) and Facebook (the list of methods can be found in [5]). The recent surge in the number of publicly available deepfake databases allowed researchers to train and test detection approaches based on very deep neural networks, such as Xception [3], capsules networks [6], ResNet-50 [7], and EfficientNet [8], which were shown to outperform the methods based on shallow CNNs, facial physical characteristics [9, 10, 11], or distortion features [12, 13, 14].

However, despite the public and media concern about deepfake videos and the surge of automated methods for their detection, little is known about how 'good' the deepfakes actually are at 'fooling' the human perception. It is commonly assumed that deepfakes are very realistic, because of some manually retouched video examples available on Youtube. However, there is a lack of scientific studies on how true it is for automatically generated deepfakes and whether they can pose a threat to the human video perception. One relavant study [3] evaluated 60 still images (30 of them were deepfakes) in a subjective study claiming that 80% of the deepfake images were successfully recognized as fake, while another [15] focused on a crowdsourcing-based methodology for evaluating synthetic images generated with GANs, but no deepfake videos were considered in any of the studies.

This work was funded by Hasler Foundation's VERIFAKE project and Swiss Center for Biometrics Research and Testing.

[1] https://www.kaggle.com/c/deepfake-detection-challenge
[2] https://github.com/deepfakes/faceswap

(f) Very easy (g) Easy (h) Moderate (i) Difficult (j) Very difficult

Fig. 1: Cropped faces from different categories of deepfake videos of Facebook database (top row) and the corresponding original versions (bottom row).

Therefore, in this paper, we conduct subjective assessments of deepfake videos, using QualityCrowd 2 [16], a specialized web-based framework for crowdsourcing experiments. The main point of the evaluation is to understand how easily an average human observer is fooled by different types of deepfakes. We manually selected 120 videos (60 original and 60 deepfakes) from more than $100K$ videos of Facebook dataset [5], because it contains large variety of deepfakes, ranging from obvious to realistically looking. Please note that the original ground truth provided in Facebook dataset contained only a fake/original label. Hence, we had to rely on our expertise to select a representative range of visually different videos. We defined five categories of deepfakes (12 videos in each) by judging them as 'very easy', 'easy', 'moderate', 'difficult', and 'very difficult' to spot as being fake (see examples in Figure 1). In total, 60 naïve human subjects (i.e., no prior specialized knowledge or work in the area), including PhD students, senior scientists, and admin staff, participated in the experiments with 20 subjective scores obtained for each video, which is higher than the recommended number of observers [17] to ensure the statistical significance of the evaluation results.

Understanding how well people recognize deepfake is important, but also is the understanding of how detection algorithms recognize them too. Policy decisions as well as people's perceptions are often based on the assumption that automated detection algorithms perceive videos in a way that is similar to humans [18], which can be dangerous when it comes to such impactful technology as deepfake detection.

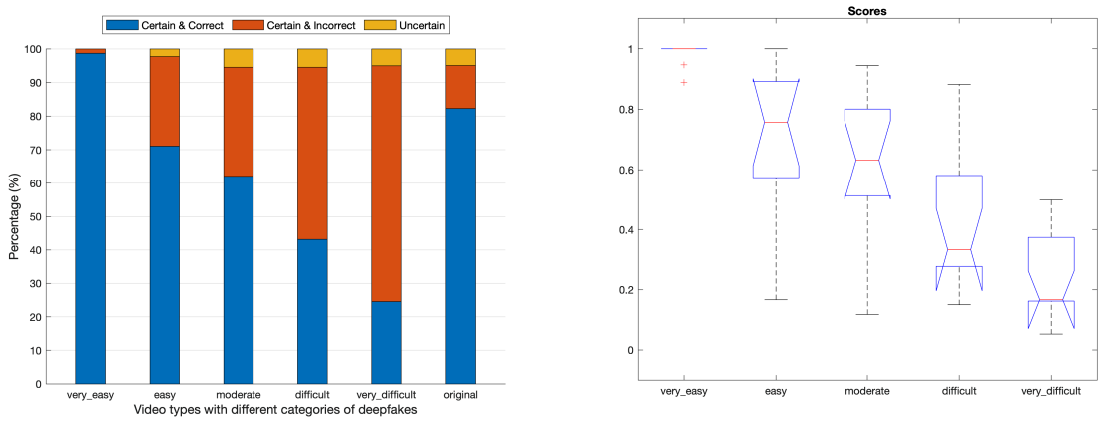
Therefore, in this paper, we also assess how two state of the art algorithms, based on Xception model [19] and EfficientNet variant B4 [8], both of which showed a great performance on several deepfake databases [3], pre-trained on two other databases from Google [3] (subset of FaceForensics++) and Celeb-DF [4], perform on the same videos and categories of deepfakes that we used in our subjective evaluation. This comparison provides a scientific insight into the differences between human and machine perception of deepfakes.

To allow researchers to verify and reproduce our work, we provide the pre-trained models, subjective scores, and the scripts used to analyze the data as an open source package[3].

2. DATA AND SUBJECTIVE EVALUATION

Since the resulted videos produced by automated deepfake generation algorithms vary drastically visually, depending on many factors (training data, the quality of the video for manipulation, and the algorithm itself), we cannot label all deepfakes into one visual category. Therefore, we have manually looked through thousands of videos from Facebook database[2] and pre-selected 60 deepfake videos, split into five categories depending of how clearly fake they look, with the corresponding 60 original videos (see examples in Figure 1). Please note that we have manually selected the videos based on our expertise and judgement, because the dataset annotations only contain the labels of whether a given video is fake or not, with no information about the method used to generate the deepfake. For the overview of methods used in this dataset and the details on how the original videos were captured, please refer to [5].

[3]https://gitlab.idiap.ch/bob/bob.paper.subjective-deepfakes



(a) Subjective answers

(b) ANOVA test

Fig. 2: Subjective answers and median values with error bars from ANOVA test for different deepfake categories.

The subjective evaluation was conducted using Quality-Crowd 2 framework [16] designed for crowdsourcing-based evaluations. This framework allows us to make sure subjects watch each video fully at least once and are not able to skip any questions. Prior to the evaluation itself, a display brightness test was performed using a method similar to that described in [20]. Since deepfake detection algorithms typically evaluate only the face regions cropped using a face detector, to have a comparable scenario, we have shown to the human subjects the cropped face regions next to the original video.

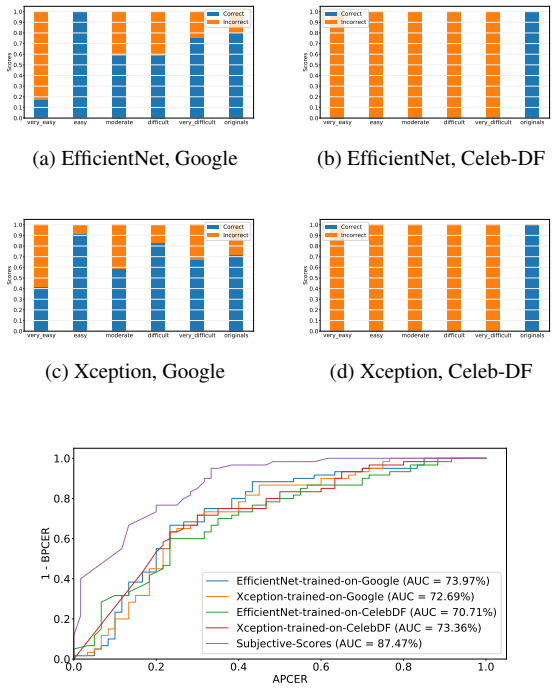
Each of the 60 naïve human subjects (unfamiliar with deepfakes beyond typical news) had to answer the question after watching a given video: "Is face of the person in the video real or fake?" with the following options: "Fake", "Real", and "I do not know." Prior to the evaluation, the explanation of the test was given to the subjects with several test video examples of different fake categories and real videos. The 120 videos were also split into random batches of 40 each to reduce the total evaluation time for one subject, so the average time per one evaluation was about 16 minutes, which is consistent with the standard recommendations [17].

Due to privacy concerns, we did not collect any personal information from our subjects such as age or gender. Also, the licensing conditions of Facebook database[2] restricted the evaluation to the premises of Idiap research institute, which signed the license agreement not do distribute the data outside. Therefore, the subjects consisted of PhD students, scientists, administration, and management of Idiap. The major shortcoming of the crowdsourcing-based subjective experiments is the inability to supervise participants behavior. Therefore, to evaluate the 'trustworthiness' of the subjects, we used the 'honeypot' method [20, 21] to filter out scores from people who did not pay attention, ending up with 18.66 answers per video on average, which is higher than the recommended number by TU-R BT.500-13 standard [17].

3. SUBJECTIVE EVALUATION RESULTS

For each deepfake or original video, we computed the proportions of answers that were 'certain & correct', 'certain & incorrect', and 'uncertain' (the answer was 'I do not know'). We have averaged these scores across videos in each category and the results are shown in Figure 2(a). From the figure, we can note the low number of uncertain answers, which means human subjects tend to be sure when it comes to judging the realism of a video. It also means people can be easily fooled by a good quality deepfake video, since only in 24.5% cases the videos from 'very difficult' category were correctly perceived as fakes. In the scenario when deepfakes are distributed via social media, we can expect the number of people noticing them to be significantly lower. Also, it is interesting to note that even videos from 'easy' category were not as easy to spot (71.1% correct answers) compared to the original videos (82.2%). Overall, we can see that people are only 'good' at recognizing either very obvious examples of deepfakes or real videos.

To check whether the difference between videos from the five deepfake categories is statistically significant based on the subjective scores, we performed ANOVA test with the corresponding box plot shown in Figure 2(b). The scores were computed for each video (and per category when applicable) by averaging the answers from all corresponding observers. For each correct answer the score is 1 and for both wrong and uncertain answers the score is 0. The red lines in Figure 2(b) correspond to median values. The *p-value* of ANOVA test is below 4.7×10^{-11}, which means the deepfake categories are significantly different on average. However, the ANOVA test also shows that 'easy', 'moderate', and 'difficult' categories have large score variations and overlaps, which means some of the videos from the neighboring categories are perceived similarly.



(a) EfficientNet, Google

(b) EfficientNet, Celeb-DF

(c) Xception, Google

(d) Xception, Celeb-DF

(e) ROC curves of algorithms and humans

Fig. 3: The detection accuracy of Xception and EfficientNet models pre-trained on Google and Celeb-DF databases (threshold at APCER=10%) for video categories from subjective test set and the corresponding ROC plots (subfig (e)).

4. EVALUATION OF ALGORITHMS

As part of our objective evaluation, we took two state of the art deepfake detection algorithms: based on Xception model [19] and EfficientNet variant B4 [8] shown to be performing well on different datasets and benchmarks [3]. We pre-trained (using image augmentations and data balancing) these models for 20 epochs each on the Google's set from FaceForensics++ [3] and Celeb-DF [4] databases to demonstrate the impact of different training conditions on the evaluation results. If evaluated on the test sets of the same databases they were trained on, both Xception and EfficientNet models demonstrate high accutacy with the area under the curve (AUC) metric equal to almost 100% in all cases.

We evaluated these neural network models on the 120 videos we used in the subjective test. Since these videos come from Facebook database, they can be considered as unseen data, which is still an obstacle for many neural network based systems, as they do not generalize well [22]. To compute the performance accuracy, we need to select the threshold. We chose the threshold corresponding to the attack presentation classification error rate (APCER) of 10%, selected on the development set of the respective database. APCER measures the proportion of attacks, deepfakes in our case, that are incorrectly classified as bona fide or original videos. The counterpart metric is bona fide presentation classification error rate (BPCER), which measures the proportion of incorrectly classified original videos. Both metrics are defined in ISO standard [23].

Figure 3 show the evaluation results of pre-trained Xception and EfficientNet models with the threshold chosen at APCER=10% on the videos from the subjective test. The results are shown separately for each deepfake category to be visually comparable to the subjective results in Figure 2(a). Similarly, the blue bar corresponds to the percent of correctly detected videos in the given category and the orange correspond to the percent of incorrectly detected. Noticeably, the results for algorithms are very different from the subjective results in Figure 2(a). The accuracy of the algorithms have no correlation to the visual appearance of deepfakes. It is evident that algorithms struggle the most with the deepfake videos that were easy for human subjects. Of course, the choice of threshold and the training data impact the resulted accuracy. Still, in practical scenario, one cannot assume the way an algorithm works is at all related to a human perception.

If we ignore the threshold selection, we can plot receiver operating characteristic (ROC) curves for algorithms and human subjects for the same 120 videos, as shown in Figure 3(e) with corresponding AUC values. We can then notice that human subjects are significantly more accurate at detecting these videos with AUC value of 87.47%.

5. CONCLUSION

This paper presents the results of subjective and objective evaluations of different categories of deepfake videos, ranging from obviously fake to very realistic looking. The videos were manually pre-selected from Facebook database and evaluated by 60 human subjects and by two deepfake detection algorithms based on Xception and EfficientNet models, which were separately pre-trained on Google (from FaceForensics++) and Celeb-DF deepfake datasets.

The subjective evaluation demonstrated that people are confused by good quality deepfakes in 75.5% of cases. On the other hand, the algorithms have a totally different perception of deepfakes compared to human subjects. The algorithms struggle to detect many videos that look obviously fake to humans, while some of the algorithms (depending on the training data and the selected threshold) can accurately detect videos that are difficult for people.

This paper shows that the deepfakes are already at the level of realism that would confuse the majority of the public, especially when they are spread online. What is missing from this study is which artifacts or image regions impact the perception of subjects or algorithms and in which way. However, it is important not to confuse and not to anthropomorphize machine vision with human vision, because they are very different and are not related to each other.

6. REFERENCES

[1] Donie O'Sullivan, "When seeing is no longer believing," *CNN Business*, Oct 2019.

[2] Pavel Korshunov and Sébastien Marcel, "Vulnerability assessment and detection of Deepfake videos," in *International Conference on Biometrics (ICB 2019)*, Crete, Greece, June 2019.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[4] Y. Li, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," in *IEEE Conference on Computer Vision and Patten Recognition*, 2020.

[5] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, "The deepfake detection challenge dataset," *arXiv preprint*, June 2020.

[6] H.H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.

[7] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, "Cnn-generated images are surprisingly easy to spot...for now," in *CVPR*, 2020.

[8] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Güera, F. Zhu, and E. J. Delp, "Deepfakes detection with automatic face weighting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 2851–2859.

[9] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[10] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head pose," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8261–8265.

[11] S. Agarwal, T. El-Gaaly, H. Farid, and S.N. Lim, "Detecting deep-fake videos from appearance and behavior," *arXiv preprint*, 2020.

[12] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," in *IEEE International Conference on Signal and Image Processing (ICSIP)*, Aug 2017, pp. 15–19.

[13] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 659–665.

[14] Pavel Korshunov and Sebastien Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint*, 2018.

[15] Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Li Fei-Fei, and Michael S. Bernstein, "Hype: A benchmark for human eye perceptual evaluation of generative models," in *NeurIPS*, 2019, pp. 3444–3456.

[16] Christian Keimel, Julian Habigt, Clemens Horch, and Klaus Dieopold, "Qualitycrowd – a framework for crowd-based quality evaluation," in *Picture Coding Symposium (PCS)*, May 2012.

[17] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Jan. 2012.

[18] Elizabeth Fernandez, "AI Is Not Similar To Human Intelligence. Thinking So Could Be Dangerous," *Forbes*, Nov 2019.

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[20] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[21] Pavel Korshunov, Hiromi Nemoto, Athanassios Skodras, and Touradj Ebrahimi, "Crowdsourcing-based evaluation of privacy in HDR images," in *Optics, Photonics, and Digital Technologies for Multimedia Applications III*. 2014, vol. 9138, pp. 1 – 11, SPIE.

[22] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez, "Deepfakes evolution: Analysis of facial regions and fake detection performance," *arXiv preprint*, 2020.

[23] ISO/IEC JTC 1/SC 37 Biometrics, "DIS 30107-3:2016, information technology — biometrics presentation attack detection — part 3: Testing and reporting," American National Standards Institute, Oct. 2016.