Vulnerability of Face Recognition to Deep Morphing

Anonymous submission

Abstract

It is increasingly easy to automatically swap faces in images and video or morph two faces into one using generative adversarial networks (GANs). The high quality of the resulted deep-morph raises the question of how vulnerable the current face recognition systems are to such fake images and videos. It also calls for automated ways to detect these GAN-generated faces. In this paper, we present the publicly available dataset of the Deepfake videos with faces morphed with a GAN-based algorithm. To generate these videos, we used open source software based on GANs, and we emphasize that training and blending parameters can significantly impact the quality of the resulted videos. We show that the state of the art face recognition systems based on VGG and Facenet neural networks are vulnerable to the deep morph videos, with 85.62% and 95.00% false acceptance rates, respectively, which means methods for detecting these videos are necessary. We consider several baseline approaches for detecting deep morphs and find that the method based on visual quality metrics (often used in presentation attack detection domain) leads to the best performance with 8.97% equal error rate. Our experiments demonstrate that GAN-generated deep morph videos are challenging for both face recognition systems and existing detection methods, and the further development of deep morphing technologies will make it even more so.

1. Introduction

Recent advances in automated video and audio editing tools, generative adversarial networks (GANs), and social media allow the creation and the fast dissemination of high quality tampered video content. Such content already led to appearance of deliberate misinformation, coined ‘fake news’, which is impacting political landscapes of several countries [2]. A recent surge of videos (started as obscene) called Deepfakes[1], in which a neural network is used to train a model to replace faces with a likeness of someone else, are of a great public concern[2]. Accessible open source software and apps for such face swapping lead to large amounts of synthetically generated Deepfake videos appearing in social media and news, posing a significant technical challenge for detection and filtering of such content.

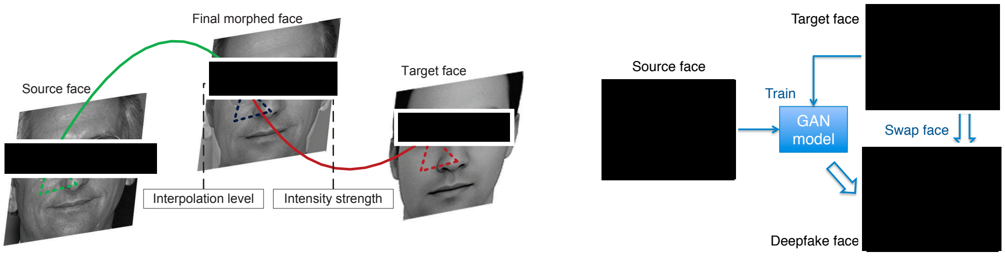
Although the original purpose of GAN-based Deepfake is to swap faces of two people in an image or a video, the resulted synthetic face is essentially a morph, i.e., a *deep morph*, of two original faces. The main difference from more traditional morphing techniques is that deep-morph can seamlessly mimic facial expression of the target person and, therefore, can also be successfully used to generate convincing fake videos of people talking and moving about. However, to understand how threatening such videos can be in the context of biometric security, we need to find out whether these deep-morphed videos pose a challenge to face recognition systems and whether they can be easily detected.

Traditional face morphing (Figure 1a illustrates the morphing process) has been shown to be challenging for face recognition systems [3, 16] and several detection methods has been proposed since [10, 18, 9]. For the GAN-based deep-morphing, until recently, most of the research was focusing on advancing the GAN-based face swapping [6, 8, 12, 14]. However, responding to the public demand to detect these synthetic faces, researchers started to work on databases and detection methods, including image and video data [15] generated with a previous generation of face swapping approach Face2Face [19] or videos collected using Snapchat[3] application [1]. Several methods for detection of Deepfakes have also been proposed [7, 21, 5].

[1] Open source: https://github.com/deepfakes/faceswap

[2] BBC (Feb 3, 2018): http://www.bbc.com/news/technology-42912529

[3] https://www.snapchat.com/


[redacted] [redacted] [redacted] [redacted] [redacted] [redacted]

(a) Morphing faces

(b) Generating Deepfake faces

Figure 1: Comparing morphing and GAN-based face swapping techniques.

In this paper, we focus on evaluating the vulnerability of face recognition systems to Deepfake videos where real faces are replaced by GAN-generated images trained on the faces of two people. The resulted synthetic face is essentially a deep morph of two people. The database was created using the open source software with cyclic GAN model[4] (see Figure 1b for illustration), which is developed from the original autoencoder-based Deepfake algorithm[1]. We manually selected 16 similar looking pairs of people from publicly available VidTIMIT database[5]. For each of 32 subjects, we trained two different models (see Figure 2 for examples), referred to in the paper as the low quality (LQ) model, with 64×64 input/output size, and the high quality (HQ) model, with 128×128 size. Since there are 10 videos per person in VidTIMIT database, we generated 320 videos corresponding to each version, resulting in total 620 videos with faces swapped. For the audio, we kept the original audio track of each video, i.e., no manipulation was done to the audio channel.

We assess the vulnerability of face recognition to deep morph videos using two state of the art systems: based on VGG [13] and Facenet[6] [17] neural networks. For detection of the deep morphs, we applied several baseline methods from presentation attack detection domain, by treating deep morph videos as digital presentation attacks [1], including simple principal component analysis (PCA) and linear discriminant analysis (LDA) approaches, and the approach based on image quality metrics (IQM) and support vector machine (SVM) [4, 20].

To allow researchers to verify, reproduce, and extend our work, we provide the database YYY [7], face recognition and deep morph detection systems with corresponding scores as an open source Python package [8].

[redacted]

[redacted]

(g) Original 1 (h) Original 2 (i) LQ swap 1 → 2 (j) HQ swap 1 → 2 (k) LQ swap 2 → 1 (l) HQ swap 2 → 1

Figure 2: Screenshot of the original videos from VidTIMIT database and low (LQ) and high quality (HQ) deep morphs.

[4] https://github.com/shaoanlu/faceswap-GAN
[5] http://conradsanderson.id.au/vidtimit/
[6] https://github.com/davidsandberg/facenet
[7] http://anonymized/for/review
[8] http://anonymized/for/review

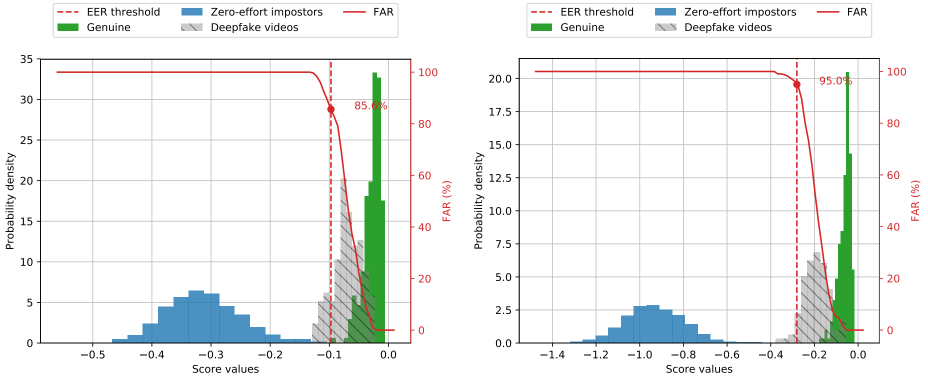
2. Database of deep morph videos

As the original data, we took video from VidTIMIT database[5]. The database contains 10 videos for each of 43 subjects, which were shot in controlled environment with people facing camera and reciting predetermined short phrases. From these 43 subject, we manually selected 16 pairs in such a way that subjects in the same pair have similar prominent visual features, e.g., mustaches or hair styles. Using GAN-based algorithm based on the available code[4], for each pair of subjects, we generated videos where their faces are replaced by a GAN-generated deep morphs (see the example screenshots in Figure 2).

For each pair of subjects, we have trained two different GAN models and generated two versions of the deep morphs:

1. The low quality (LQ) model has input and output image (facial regions only) of size 64×64. About 200 frames from the videos of each subject were used for training and the frames were extracted at 4 fps from the original videos. The training was done for $10'000$ iterations and took about 4 hours per model on Tesla P40 GPU.
2. The high quality (HQ) model has input/output image size of 128×128. About 400 frames extracted at 8 fps from videos were used for training, which was done for $20'000$ iterations (about 12 hours on Tesla P40 GPU).

Also, different blending techniques were used when generating deep morph videos using different models. With LQ model, for each frame from an input video, generator of the GAN model was applied on the face region to generate the fake counterpart. Then a facial mask was detected using a CNN-based face segmentation algorithm proposed in [12]. Using this mask, the generated fake face was blended with the face in the target video. For HQ model, the blending was done based on facial landmarks (detected with publicly available MTCNN model [22]) alignment between generated fake face and the original face in the target video. Finally, histogram normalization was applied to the blended result to adjust for the lighting conditions, which makes the result more realistic (see Figure 2).



(a) VGG-based face recognition

(b) FaceNet-based face recognition

Figure 3: Histograms show the vulnerability of VGG and Facenet based face recognition to high quality deep morphs.

2.1. Evaluation protocol

When evaluating vulnerability of face recognition, for the *licit* scenario without the deep morph videos, we used the original VidTIMIT[5] videos for the 32 subjects for which we have generated corresponding deep morph videos. In this scenario, we used 2 videos of the subject for enrollment and the other 8 videos as probes, for which we computed the verification scores.

From the scores, for each possible threshold θ, we computed commonly used metrics for evaluation of classification systems: false acceptance rate (FAR) and false reject rate (FRR). Threshold at which these FAR and FRR are equal leads to an equal error rate (EER), which is commonly used as a single value metric of the system performance.

To evaluate vulnerability of face recognition, in *tampered* scenario, we use deep morph videos (10 for each of 32 subjects) as probes and compute the corresponding scores using the enrollment model from the *licit* scenario. To understand if face recognition perceives deep morphs to be similar to the genuine original videos, we report the FAR metric computed using EER threshold θ from *licit* scenario. If FAR value for deep morph videos is significantly higher than the one computed in *licit* scenario, it means the face recognition system cannot distinguish synthetic videos from originals and is therefore vulnerable to deep morphs.

Table 1: Baseline detection systems for low (LQ) and high quality (HQ) deep morph videos. EER and FRR when FAR equal to 10% are computed on Test set.

Database	**Detection system**	**EER (%)**	**FRR@FAR10% (%)**
LQ deep morph	Pixels+PCA+LDA	39.48	78.10
	IQM+PCA+LDA	20.52	66.67
	IQM+SVM	3.33	0.95
HQ deep morph	IQM+SVM	8.97	9.05

When evaluating deep morph detection, we consider it as a binary classification problem and evaluate the ability of detection approaches to distinguish original videos from deep morph videos. All videos in the dataset, including genuine and fake parts, were split into training (*Train*) and evaluation (*Test*) subsets. To avoid bias during training and testing, we arranged that the same subject would not appear in both sets. We did not introduce a development set, which is typically used to tune hyper parameters such as threshold, because the dataset is not large enough. Therefore, for deep morph detection system, we report the EER and the FRR (using the threshold when $FAR = 10\%$) values on the *Test* set.

3. Vulnerability of face recognition

We used publicly available pre-trained VGG and Facenet architectures for face recognition. We used the *fc7* and *bottleneck* layers of these networks, respectively, as features and used cosine distance as a classifier. For a given test face, the confidence score of whether it belongs to a pre-enrolled model of a person is the cosine distance between the average feature vector, i.e., model, and the features vector of a test face. Both of these systems are state of the art recognition systems with VGG of 98.95% [13] and Facenet of 99.63% [17] accuracies on labeled faces in the wild (LFW) dataset.

We conducted the vulnerability analysis of VGG and Facenet-based face recognition systems on low quality (LQ) and high quality (HQ) face swaps in VidTIMIT[5] database. In a *licit* scenario when only original videos are present, both systems performed very well, with EER value of 0.03% for VGG and 0.00% for Facenet-based system. Using the EER threshold from *licit* scenario, we computed FAR value for the scenario when deep morph videos are used as probes. In this case, for VGG the FAR is 88.75% on LQ deep morphs and 85.62% on HQ deep morphs, and for Facenet the FAR is 94.38% and 95.00% on LQ and HQ deep morphs respectively. To illustrate this vulnerability, we plot the score histograms for high quality deep morph videos in Figure 3. The histograms show a considerable overlap between deep morph and genuine scores with clear separation from the zero-effort impostor scores (the probes from *licit* scenario).

From the results, it is clear that both VGG and Facenet based systems cannot effectively distinguish GAN-generated synthetic faces from the original ones. The fact that more advanced Facenet system is more vulnerable is also consistent with the findings about presentation attacks [11].

4. Detection of deep morph videos

We considered several baseline deep morph detection systems:

- *Pixels+PCA+LDA*: use raw faces as features with PCA-LDA classifier, with 99% retained variance resulting in 446 dimensions of transform matrix.
- *IQM+PCA+LDA*: IQM features with PCA-LDA classifier with 95% retained variance resulting in 2 dimensions of transform matrix.
- *IQM+SVM*: IQM features with SVM classifier, each video has an averaged score from 20 frames.

The systems based on image quality measures (IQM) are borrowed from the domain of presentation (including replay attacks) attack detection, where such systems have shown good performance [4, 20]. As IQM feature vector, we used 129 measures of image quality, which include such measures like signal to noise ratio, specularity, bluriness, etc., by combining the features from [4] and [20].

The results for all detection systems are presented in Table 1. The results demonstrate that the IQM+SVM system has a reasonably high accuracy of detecting deep morph videos, although videos generated with HQ model pose a more serious challenge. It means that a more advanced techniques for face swapping will be even more challenging to detect.

5. Conclusion

In this paper, we demonstrated that state of the art VGG and Facenet-based face recognition algorithms are vulnerable to the deep morph morphed videos and fail to distinguish such videos from the original ones with up to 95.00% equal error rate. We also evaluated several baseline detection algorithms and found that the techniques based on image quality measures with SVM classifier can detect HQ deep morph videos with 8.97% equal error rate.

However, the continued advancements in development of GAN-generated faces will result in more challenging videos, which will be harder to detect by the existing algorithms. Therefore, new databases and new more generic detection methods need to be developed in the future.

References

[1] 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)*, pages 659–665, Oct 2017.

[2] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[3] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics (BTAS)*, pages 1–7, Sep. 2014.

[4] J. Galbally and S. Marcel. Face anti-spoofing based on general image quality assessment. In *International Conference on Pattern Recognition*, pages 1173–1178, Aug 2014.

[5] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Nov 2018.

[6] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017.

[7] P. Korshunov and S. Marcel. Vulnerability assessment and detection of Deepfake videos. In *International Conference on Biometrics (ICB 2019)*, Crete, Greece, June 2019.

[8] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, Oct 2017.

[9] R. S. S. Kramer, M. O. Mireku, T. R. Flack, and K. L. Ritchie. Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, 4(1):28, Jul 2019.

[10] A. Makrushin, T. Neubert, and J. Dittmann. Automatic generation and detection of visually faultless facial morphs. In *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, pages 39–50. INSTICC, SciTePress, 2017.

[11] A. Mohammadi, S. Bhattacharjee, and S. Marcel. Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biometrics*, 7(1):15–26, 2018.

[12] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni. On face segmentation, face swapping, and face perception. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 98–105, May 2018.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[14] H. X. Pham, Y. Wang, and V. Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *arXiv.org*, 2018.

[15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv.org*, 2018.

[16] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, Feb. 2019.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[18] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert. Detection of face morphing attacks by deep learning. In C. Kraetzer, Y.-Q. Shi, J. Dittmann, and H. J. Kim, editors, *Digital Forensics and Watermarking*, pages 107–120, Cham, 2017. Springer International Publishing.

[19] J. Thies, M. Zollhfer, M. Stamminger, C. Theobalt, and M. Niener. Face2Face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, June 2016.

[20] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, April 2015.

[21] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, May 2019.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.