https://doi.org/10.2352/EI.2024.36.4.MWSF-334

Contribution of residual signals to the detection of face swapping in deepfake videos

Paul Tessé, Christophe Charrier, Emmanuel Giguet; Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC; Caen, France

Abstract

The impressive rise of Deep Learning and, more specifically, the discovery of generative adversarial networks has revolutionised the world of Deepfake. The forgeries are becoming more and more realistic, and consequently harder to detect. Attesting whether a video content is authentic is increasingly sensitive. Furthermore, free access to forgery technologies is dramatically increasing and very worrying. Numerous methods have been proposed to detect these deepfakes and it is difficult to know which detection methods are still accurate regarding the recent advances. Therefore, an approach for face swapping detection in videos, based on residual signal analysis is presented in this paper.

Key words

Deepfake videos, Face swapping, Residual signals, Digital forensics, Deep Learning.

Introduction

We live in a hyper-connected society, with billions of pieces of data in transit every day. Unfortunately, these days we can no longer be sure that this data is reliable and risk-free. This is especially true when we are talking about videos and images, which represent a phenomenal amount of data distributed and shared on a massive scale. The rapid growth in Deep Learning schemes has led to the emergence of numerous efficient models for generating false images or videos. As a matter of fact, while these models are becoming increasingly powerful, they are also becoming more and more accessible to the public thanks to the internet and social media. The current risk is that we will eventually no longer be able to distinguish the real from the fake, hence an inability to trust these ubiquitous media that are images and videos.

As a consequence, many researchers have studied deepfake detection, proposing methods mainly based on the use of Deep Learning models. However, although these models presented as state-of-the-art models show very good performance, the durability of these models and their ability to generalize to any database are lacking. Nevertheless, the main drawback of these models lies in the fact that they are black boxes. In fact, it is not really possible to provide a justification for the pronounced verdicts, which makes these detectors unusable. It is for all these reasons we have investigate this highly hot and urgent topic.

Objective(s)

As explained previously, the problem is vast, which is why we have set ourselves a number of objectives for this work. The main objective is to develop a model that takes a video as input and returns a verdict on its authenticity as output. The problem is therefore characterized as a binary classification problem where the classes are ”authentic” and ”forged”. The secondary objectives considered in this study are the following:

- as the model must be able to be used to assist the judicial system, particular attention must be paid to the explainability of the results;
- the model must work without any reference to pronounce its diagnosis;
- the model must be as robust and generalizable as possible;
- we will focus on the detection of face swapping, and a face detection mechanism is required in order to target the area to be studied;
- the model must work for videos of variable length with the fake part that can appear at any position or moment.

Eventually, the aim is to achieve the best possible trade-off between efficiency and explainability. The aim is not to come up with a perfect solution, but rather a proof of concept to determine whether or not the proposed approach is viable.

Method

State-of-the-art methods

In order to propose a suitable alternative, we took a look at state-of-the-art deepfake detection methods. As introduced earlier, the best performing state-of-the-art models are those based on the usage of Deep Learning. In [1], the authors present a deep architecture split into a feature extractor part and a classifier part. The feature extractor uses convolutional networks to extract spatial features from images, while a LSTM is used to extract temporal features. Regarding the classifier, the authors use a likelihood estimation in order to predict the probability of a video to be real or fake.

The authors reported very good results with this method. However, these results remain poorly explainable. As a matter of fact, it is difficult to give a sense to the verdict since it is based on features we do not understand. This is the reason why we looked for an alternative to this feature extraction system. In [2], the author lists many different methods used in forensics.

Many methods seek to analyze what are known as residual signals. These are characteristics intrinsic to an image that are generated during the acquisition process and that can be altered during the face swapping process. As videos are a succession of images, we can analyze these characteristics frame by frame. These signals, which are invisible to the naked eye and look like a hidden signature, are varied and can be explained, enabling us to extract features that can be explained and thus, used for prediction.

The residual errors investigated

In this study we focused on two residual signals in particular: 1) Image quality assessemnt and 2) frequency spectrum of images.

Image Quality Assessment

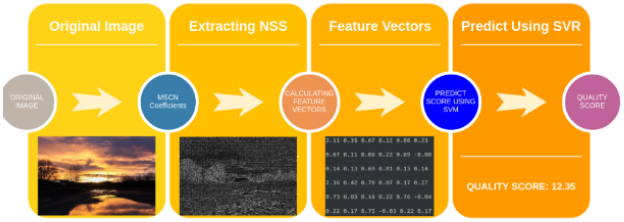

Figure 1. *Quality features extraction process*

The first residual signal we focus on is the assessment of image quality. Actually, it can be observed that image falsification processes tend to reduce the quality of the original images introducing artifacts. In [3], a detection method based on image quality analysis was presented. This work was then taken up more recently in [4]. The authors greatly increase the number of quality measurements carried out and test their regression model on state-of-the-art databases. The obtained results are convincing, which is why we looked for a method of quality estimation without reference. Among all no-reference Image Quality assessment schemes, Blind/Referenceless Image Spatial Quality Evaluator also known as BRISQUE [5] is selected due to its high correlation with human judgments. Introduced in 2012, the principle of this method is schematized in Fig. 1. The main idea is that the distribution of pixel intensities of natural images differs from that of distorted images. This difference in distributions is much more pronounced when we normalize pixel intensities and calculate the distribution over these normalized intensities. Finally, 18 elements are obtained. The image is downsized to half its original size and the same process is repeated to obtain 18 new numbers bringing the total to 36 features.

We were then able to compute 37 features relating to image quality without using a reference (36 features plus the predicted quality score).

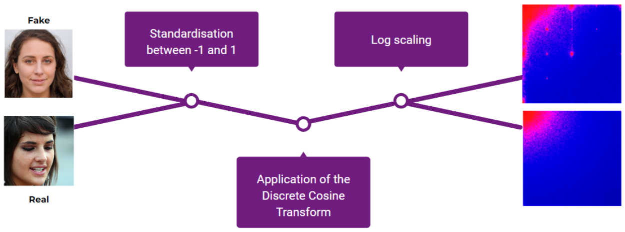
Fake
Real
Standardisation between -1 and 1
Log scaling
Application of the Discrete Cosine Transform

Figure 2. *Frequency spectrum analysis*

Frequency spectrum analysis

When processing image, it is common to move from the time domain to the frequency domain. In [6], the authors investigate the impact of the deepfake generation process on the frequency spectrum of images. Their results, illustrated by the Fig. 2, clearly indicate that deepfakes have a higher high-frequency intensity than real images and that it is possible to use this property to detect deepfakes. The origin of this phenomenon lies in the use of Generative Adversarial Networks (GANs), which are the keystone of modern deepfake models. These models are forced to use upsampling to generate images, which introduces more high frequencies due to interpolation. We therefore reproduced the pipeline used to obtain the frequency spectrum of the images and calculated the tenth quartile in order to quantify the increase in high frequencies.

The Proposed Architecture

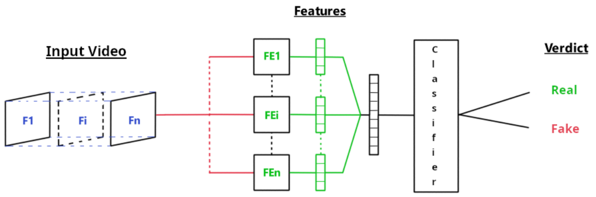

Figure 3. *Synopsis of the proposed architecture*

The two-part architecture

As alluded above, there are many other residual signals. We based our choice on the results obtained, the independence of the mechanism used for face swapping and the latest published results. We then implemented our two feature extractors presented earlier and devised an architecture adapted to combine both explicability and performance.

Based on the classic two-part architecture (extractor and classifier), we have devised the following architecture, as presented in Fig. 4:

- The input video is split into frames from which the faces are extracted.
- These faces are then analysed by our Features Extractors (FE), which generate a vector of explainable features.
- These explainable features are then concatenated and passed on to the deep classifier, which gives its verdict at a frame level.

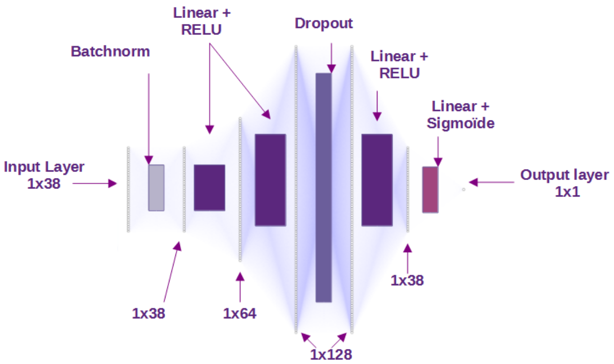

Figure 4. *The designed deep classifier*

The used deep classifier

As part of our experiments, we trained our classifier using a decreasing learning rate starting at 0.005. The number of epochs is 100 for a batch size of 1024. The loss used is a Binary Cross Entropy Logit function. Our classifier is mainly based on a succession of linear layers and RELU activations. Fig. 4 provides a detailed description of its characteristics.

More details about the proposed strategy will be given in the final paper.

Results

Exprimental setup

To evaluation the performance of the proposed method, we selected four datasets:

1. VidTIMIT dataset [7] which comprises of video and corresponding audio recordings of 43 people, reciting short sentences. This database will serve as real samples.
2. DeepFakeTIMIT dataset [4] which contains videos where faces are swapped using the open source GAN-based approach (adapted from https://github.com/shaoanlu/faceswap-GAN), which, in turn, was developed from the original autoencoder-based Deepfake algorithm. A total of 620 total videos with faces swapped is provided.
3. FF++ dataset [8] consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data has been sourced from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries.
4. Celeb-DF [9] is a large-scale challenging dataset for deepfake forensics. It includes 590 original videos collected from YouTube with subjects of different ages, ethnic groups and genders, and 5639 corresponding DeepFake videos.

From previously alluded databases, we generate one database containing 79 395 real and 85 886 fake extracted frames from

- 300 real videos randomly selected from the VidTIMIT database,
- 320 fake videos randomly selected from the DeepFake-TIMIT database,
- 200 real and 600 fake videos both randomly selected from the FF++ dataset,
- 50 real and 50 fake videos both randomly selected from the Celeb-DF databse.

These frames were then separated into 4 different sets: 1) training set (31,627 True frames and 34,826 False frames) , 2) validation set (13,474 True frames and 15,629 False frames), 3) test set (13,590 True frames and 14,466 False frames) and 4) generalization set (20,694 True frames and 20,905 False frames), used during the learning and evaluation process of the proposed classifier model. We isolated the Celeb-DF set in order to test the generalization capability of the designed classifier. The results were computed with a 5-fold process.

Set	Acccuracy	F1	AUC	Precision	Recall
Train	0.96	0.96	0.99	0.97	0.95
Validation	0.82	0.84	0.88	0.83	0.86
Test	0.84	0.85	0.89	0.83	0.87
Generalization	0.49	0.60	0.52	0.50	0.75

Classifier Training and Evaluation results

Performance evaluation

In order to evaluate the performance of the proposed scheme, five different measures have been used: 1) Accuracy, 2) Recall, 3) Precision, 4) F1-score and 5) AUC (Area Under the ROC Curve).

T_P, T_N, F_P, and F_N respectively represents true positive, true negative, false positive, and false negative The **Accuracy** is the fraction of predictions correctly identified by the model, ad is defined as:

$$\text{Accuracy} = \frac{T_P + F_N}{T_P + F_P + T_N + F_N} \tag{1}$$

The **Recall** is the percentage of positives well predicted by the model, defined as:

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{2}$$

The higher it is, the more the Machine Learning model maximizes the number of True Positives. When recall is high, this means it won't miss any positives. However, this gives no indication of its predictive quality on negatives.

The **Precision** is the number of positive predictions made defined as:

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{3}$$

The higher the precision, the more the Machine Learning model minimizes the number of False Positives. When precision is high, this means that the majority of the model's positive predictions are well-predicted positives.

The **F1-score** is an harmonic mean and provides a relatively accurate assessment of our model's performance. It is defined as

$$\text{F1-score} = 2 \times \frac{\text{Recall.Accuracy}}{\text{Recall + Accuracy}} \tag{4}$$

The higher the F1 Score, the better the model's performance.

The **AUC** computes the area under the ROC curve when plotting the precision versus the recall value. It represents the overall performance of the model, *i.e.*, the probability that a randomly chosen positive sample will be ranked higher by the model than a randomly chosen negative sample. A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5.

Results

Table 1 displays the obtained results. Whatever the considered measure, we obtained very good performance both in training and in validation and testing, although there was a drop in performance for the latter two, which is symptomatic of Deep Learning. The results for the generalization base (Celeb-DF) are much lower, which shows that our model is not yet sufficiently robust. These results are nevertheless satisfactory since Celeb-DF uses the latest deepfakes models, *i.e.*, unseen data, for which the quality is superior to those used in the three other sets. There is therefore a significant gap, which may explain this drop.

To face this drawback, the amount of data for the training process is increased by adding samples that are more difficult to diagnose from the DFDC database [10].

In order to assess the performance of the proposed strategy with state-of-the-art approaches is not obvious. Actually the performance assessment of existing methods is usually performed on different datasets and the used metrics are not common for all papers.

For example, in [11], Li *et al.* presented the performance of their proposed approach using the AUC value on the FF++ database. It shown an AUC value equal to 0.52. In [11], Ding *et al* used the accuracy value obtained from their proposed method, but on Chicago Face Dataset, that is not the one we used.

To conclude, the performance of our model is very encouraging. Furthermore, the proposed scheme can be trained in just a few minutes (less than 5 minutes on a Dell Laptop SP15 with a NVIDIA GeForce RTX 4070, 8 Go GDDR6) and is extremely light.

Conclusion

This work has enabled us to experiment with a hybrid approach between traditional forensic methods and state-of-the-art deepfake detection methods based on Deep Learning.

Our architecture is more explainable and therefore viable in an integration context. What's more, its good performance with just 38 features and little data seems to indicate that by digging deeper in this direction it would be possible to achieve very good results.

Finally, our architecture also has the advantage of being lightweight and quick to train and use, which is increasingly rare with the development of Deep Learning. It also makes it easy to integrate new feature extraction modules, which guarantees its durability.

Future works will investigate how new residual signals computed in spatial, frequency and color spaces may help to increase the performance of the approach.

References

[1] David Güera, Edward J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1-6, 2018

[2] Verdoliva, Luisa. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing. PP. 1-1. 10.1109/JSTSP.2020.3002101.

[3] J. Galbally and S. Marcel, "Face Anti-spoofing Based on General Image Quality Assessment," 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 2014, pp. 1173-1178, doi: 10.1109/ICPR.2014.211.

[4] Korshunov, Pavel and Marcel, Sébastien. (2018). DeepFakes: a New Threat to Face Recognition? Assessment and Detection.

[5] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

[6] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 304, 3247–3258.

[7] Sanderson, Conrad and Lovell, Brian. (2009). Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. LNCS. 5558. 10.1007/978-3-642-01793-3-21.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.

[9] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[10] Dolhansky, Brian and Bitton, Joanna and Pflaum, Ben and Lu, Jikuo and Howes, Russ and Wang, Menglin and Ferrer, Cristian. (2020). The DeepFake Detection Challenge (DFDC) Dataset. (arXiv:2006.07397)

[11] Li, L., Bao, J., Yang, H., Chen, D., Wen, F. (2020). Advancing high fidelity identity swapping for forgery detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5074-5083).

[12] Ding, X., Raziei, Z., Larson, E. C., Olinick, E. V., Krueger, P., Hahsler, M. (2020). Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security, 2020(1), 1-12.