DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation

Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, Aythami Morales
Biometrics and Data Pattern Analytics Lab - BiDA Lab
Universidad Autonoma de Madrid
{javier.hernandezo, ruben.tolosana, julian.fierrez, aythami.morales}@uam.es

Abstract

This work introduces a novel DeepFake detection framework based on physiological measurement. In particular, we consider information related to the heart rate using remote photoplethysmography (rPPG). rPPG methods analyze video sequences looking for subtle color changes in the human skin, revealing the presence of human blood under the tissues. In this work we investigate to what extent rPPG is useful for the detection of DeepFake videos. The proposed fake detector named DeepFakesON-Phys uses a Convolutional Attention Network (CAN), which extracts spatial and temporal information from video frames, analyzing and combining both sources to better detect fake videos. DeepFakesON-Phys has been experimentally evaluated using the latest public databases in the field: Celeb-DF and DFDC. The results achieved, above 98% AUC (Area Under the Curve) on both databases, outperform the state of the art and prove the success of fake detectors based on physiological measurement to detect the latest DeepFake videos.

Introduction

DeepFakes have become a great public concern recently (Citron 2019; Cellan-Jones 2019). The very popular term “DeepFake” is usually referred to a deep learning based technique able to create fake videos by swapping the face of a person by the face of another person. This type of digital manipulation is also known in the literature as Identity Swap, and it is moving forward very fast (Tolosana et al. 2020b).

Currently, most face manipulations are based on popular machine learning techniques such as AutoEncoders (AE) (Kingma and Welling 2013) and Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), achieving in general very realistic visual results, specially in the latest generation of public DeepFakes (Tolosana et al. 2020a), and the present trends (Karras et al. 2020). However, and despite the impressive visual results, are current face manipulations also considering the physiological aspects of the human being in the synthesis process?

Physiological measurement has provided very valuable information to many different tasks such as e-learning (Hernandez-Ortega et al. 2020a), health care (McDuff et al. 2015), human-computer interaction (Tan and Nijholt 2010), and security (Marcel et al. 2019), among many other tasks.

In physical face attacks, a.k.a. Presentation Attacks (PAs), real subjects are often impersonated using artifacts such as photographs, videos, and masks (Marcel et al. 2019). Face recognition systems are known to be vulnerable against these attacks unless proper detection methods are implemented (Galbally, Marcel, and Fierrez 2014; Hernandez-Ortega et al. 2019). Some of these detection methods are based on liveness detection by using information such as eye blinking or natural facial micro-expressions (Bharadwaj et al. 2013). Specifically for detecting 3D mask impersonation, which is one of the most challenging type of attacks, detecting pulse from face videos using remote photoplethysmography (rPPG) has shown to be an effective countermeasure (Hernandez-Ortega et al. 2018). When applying this technique to a video sequence with a fake face, the estimated heart rate signal is significantly different to the heart rate extracted from a real face (Erdogmus and Marcel 2014).

Seeing the good results achieved by rPPG techniques when dealing with physical 3D face mask attacks, and since DeepFakes are digital manipulations somehow similar to them, in this work we hypothesize that fake detectors based on physiological measurement can also be used against DeepFakes after adapting them properly. DeepFake generation methods have historically tried to mimic the visual appearance of genuine faces. However, to the best of our knowledge, they do not emulate the physiology of human beings, e.g., heart rate, blood oxygenation, or breath rate, so estimating that type of signals from the video could be a powerful tool for the detection of DeepFakes.

The **novelty of this work consists in using rPPG features previously learned for the task of heart rate estimation and adapting them for the detection of DeepFakes by means of a knowledge-transfer process, thus obtaining a novel fake detector based on physiological measurement named DeepFakesON-Phys.** In particular, the information related to the heart rate is considered to decide whether a video is real or fake. Our physiological detector intends to be a robust solution to the weaknesses of most state-of-the-art DeepFake detectors based on the visual features existing in fake videos (Matern, Riess, and Stamminger 2019; Agarwal

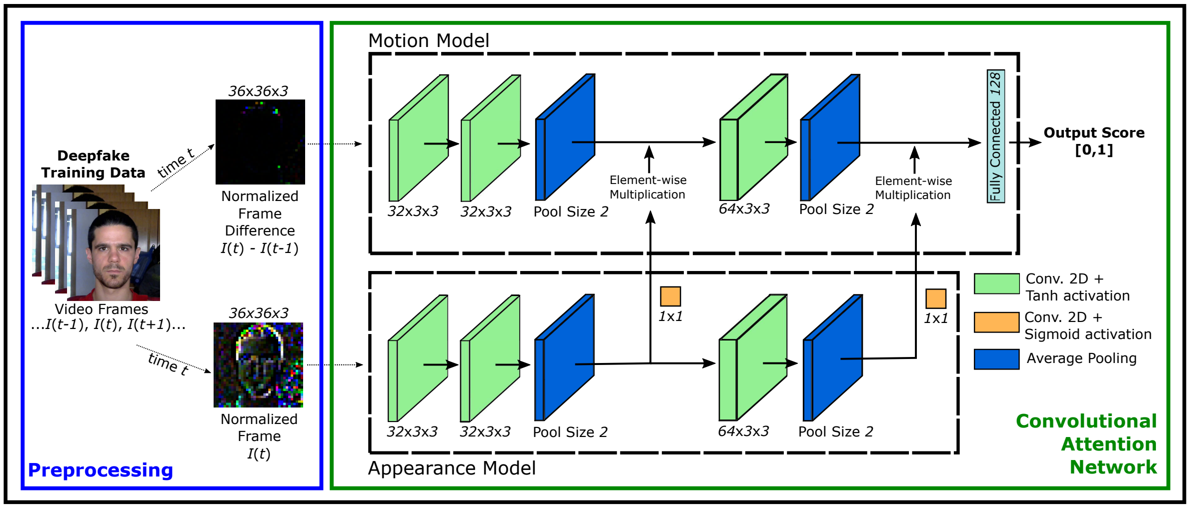

Figure 1: **DeepFakesON-Phys architecture**. It comprises two stages: *i)* a preprocessing step to normalize the video frames, and *ii)* a Convolutional Attention Network composed of Motion and Appearance Models to better detect fake videos.

and Farid 2019) and also on the artifacts/fingerprints inserted during the synthesis process (Neves et al. 2020), which are highly dependent on a specific fake manipulation technique.

DeepFakesON-Phys is based on DeepPhys (Chen and McDuff 2018), a deep learning model trained for heart rate estimation from face videos based on rPPG. DeepPhys showed high accuracy even when dealing with challenging conditions such as heterogeneous illumination or low resolution, outperforming classic hand-crafted approaches. We used the architecture of DeepPhys, but making changes to make it suitable for DeepFake detection. We initialized the weights of the layers of DeepFakesON-Phys with the ones from DeepPhys (meant for heart rate estimation based on rPPG) and we adapted them to the new task using fine-tuning. This process allowed us to train our detector without the need of a high number of samples (compared to training it from scratch). Fine-tuning also helped us to obtain a model that detects DeepFakes by looking to rPPG related features from the images in the face videos.

In this context, the main contributions of our work are:

- **An in-depth literature review of DeepFake detection** approaches with special emphasis to physiological techniques, including the key aspects of the detection systems, the databases used, and the main results achieved.
- **An approach based on physiological measurement to detect DeepFake videos: DeepFakesON-Phys**[1]**.** Fig. 1 graphically summarizes the proposed fake detection approach based on the original architecture DeepPhys (Chen and McDuff 2018), a Convolutional Attention Network (CAN) composed of two parallel Convolutional Neural Networks (CNN) able to extract spatial and temporal information from video frames. This architecture is adapted for the detection of DeepFake videos by means of a knowledge-transfer process.
- **A thorough experimental assessment of the proposed DeepFakesON-Phys**, considering the latest public databases of the 2$^{\text{nd}}$ DeepFake generation such as Celeb-DF v2 and DFDC Preview. DeepFakesON-Phys achieves high-accuracy results, outperforming the state of the art. In addition, the results achieved prove that current face manipulation techniques do not pay attention to the heart-rate-related physiological information of the human being when synthesizing fake videos.

The remainder of the paper is organized as follows. **Related Works** summarizes previous studies focused on the detection of DeepFakes. **Proposed Method: DeepFakesON-Phys** describes the proposed DeepFakesON-Phys fake detection approach. **Databases** summarizes all databases considered in the experimental framework of this study. **Experiments** describes the experimental protocol and the results achieved in comparison with the state of the art. Finally, **Conclusions** draws the final conclusions and points out future research lines.

Related Works

Different approaches have been proposed in the literature to detect DeepFake videos. Table 1 shows a comparison of the most relevant approaches in the area, paying special attention to the fake detectors based on physiological measurement. For each study we include information related to the method, classifiers, best performance, and databases for research. It is important to remark that in some cases, different evaluation metrics are considered, e.g., Area Under the Curve (AUC) and Equal Error Rate (EER), which complicate the comparison among studies. Finally, the results highlighted in *italics* indicate the generalization ability of the detectors against unseen databases, i.e., those databases were not considered for training. Most of these results are extracted from (Li et al. 2020).

The first studies in the area focused on the visual arti-

[1] https://github.com/BiDAlab/DeepFakesON-Phys

Table 1: **Comparison of different state-of-the-art fake detectors.** Results in *italics* indicate the generalization capacity of the detectors against unseen databases. FF++ = FaceForensics++, AUC = Area Under the Curve, Acc. = Accuracy, EER = Equal Error Rate.

Study	Method	Classifiers	Best Performance	Databases
(Matern, Riess, and Stamminger 2019)	Visual Features	Logistic Regression MLP	AUC = 85.1%	Own
			AUC = 78.0%	*FF++ / DFD*
			AUC = 66.2%	*DFDC Preview*
			AUC = 55.1%	*Celeb-DF*
(Li and Lyu 2019; Li et al. 2020)	Face Warping Features	CNN	AUC = 97.7%	UADFV
			AUC = 93.0%	*FF++ / DFD*
			AUC = 75.5%	*DFDC Preview*
			AUC = 64.6%	*Celeb-DF*
(Rössler et al. 2019)	Mesoscopic Features Steganalysis Features Deep Learning Features	CNN	Acc. ≃ 94.0% Acc. ≃ 98.0% Acc. ≃ 100.0%	FF++ (DeepFake, LQ) FF++ (DeepFake, HQ) FF++ (DeepFake, RAW)
			Acc. ≃ 93.0% Acc. ≃ 97.0% Acc. ≃ 99.0%	FF++ (FaceSwap, LQ) FF++ (FaceSwap, HQ) FF++ (FaceSwap, RAW)
(Nguyen, Yamagishi, and Echizen 2019)	Deep Learning Features	Capsule Networks	*AUC = 61.3%*	*UADFV*
			AUC = 96.6%	*FF++ / DFD*
			AUC = 53.3%	*DFDC Preview*
			AUC = 57.5%	*Celeb-DF*

(Dang et al. 2020)	Deep Learning Features	CNN + Attention Mechanism	AUC = 99.4%
			EER = 3.1%
				DFFD

(Dolhansky et al. 2019)	Deep Learning Features	CNN	Precision = 93.0%
			Recall = 8.4%
				DFDC Preview

(Sabir et al. 2019)	Image + Temporal Features	CNN + RNN	AUC = 96.9%
			AUC = 96.3%
				FF++ (DeepFake, LQ)
				FF++ (FaceSwap, LQ)

(Tolosana et al. 2020a)	Facial Regions Features	CNN	AUC = 100.0%	UADFV
			AUC = 99.5%	FF++ (FaceSwap, HQ)
			AUC = 91.1%	DFDC Preview
			AUC = 83.6%	Celeb-DF
(Conotter et al. 2014)	Physiological Features	-	Acc. = 100%	Own
(Li, Chang, and Lyu 2018)	Physiological Features	LRCN	AUC = 99.0%	UADFV
(Agarwal and Farid 2019)	Physiological Features	SVM	AUC = 96.3%	Own (FaceSwap, HQ)

(Ciftci, Demir, and Yin 2020)	Physiological Features	SVM/CNN	Acc. = 94.9%
			Acc. = 91.5%
				FF++ (DeepFakes)
				Celeb-DF

(Jung, Kim, and Kim 2020)	Physiological Features	Distance	Acc. = 87.5%	Own
(Qi et al. 2020)	Physiological Features	CNN + Attention Mechanism	Acc. = 100.0%	FF++ (FaceSwap)
			Acc. = 100.0%	FF++ (DeepFake)
			Acc. = 64.1%	*DFDC Preview*
DeepFakesON-Phys [Ours]	**Physiological Features**	**CAN**	**AUC = 99.9%**	**Celeb-DF v2**
			AUC = 98.2%	**DFDC Preview**

facts existed in the 1st generation of fake videos. The authors of (Matern, Riess, and Stamminger 2019) proposed fake detectors based on simple visual artifacts such as eye colour, missing reflections, and missing details in the teeth areas, achieving a final 85.1% AUC.

Approaches based on the detection of the face warping artifacts have also been studied in the literature. For example, (Li and Lyu 2019; Li et al. 2020) proposed detection systems based on CNN in order to detect the presence of such artifacts from the face and the surrounding areas, being one of the most robust detection approaches against unseen face manipulations.

Undoubtedly, fake detectors based on pure deep learning features are the most popular ones: feeding the networks with as many real/fake videos as possible and letting the networks to automatically extract the discriminative features. In general, these fake detectors have achieved very good results using popular network architectures such as Xception (Rössler et al. 2019; Dolhansky et al. 2019), novel ones such as Capsule Networks (Nguyen, Yamagishi, and Echizen 2019), and novel training techniques based on attention mechanisms (Dang et al. 2020).

Fake detectors based on the image and temporal discrepancies across frames have also been proposed in the literature. (Sabir et al. 2019) proposed a Recurrent Convolutional Network similar to (Güera and Delp 2018), trained end-to-end instead of using a pre-trained model. Their proposed detection approach was tested using FaceForensics++ database (Rössler et al. 2019), achieving AUC results above 96%.

Although most approaches are based on the detection of fake videos using the whole face, in (Tolosana et al. 2020a) the authors evaluated the discriminative power of each facial region using state-of-the-art network architectures, achieving interesting results on DeepFake databases of the 1st and 2nd generations.

Finally, we pay special attention to the fake detectors based on physiological information. The eye blinking rate was studied in (Li, Chang, and Lyu 2018; Jung, Kim, and Kim 2020). (Li, Chang, and Lyu 2018) proposed Long-Term Recurrent Convolutional Networks (LRCN) to capture the temporal dependencies existed in human eye blinking. Their method was evaluated on the UADFV database, achieving a final 99.0% AUC. More recently, (Jung, Kim, and Kim

2020) proposed a different approach named DeepVision. They fused the Fast-HyperFace (Ranjan, Patel, and Chellappa 2017) and EAR (Soukupova and Cech 2016) algorithms to track the blinking, achieving an accuracy of 87.5% over an in-house database.

Fake detectors based on the analysis of the way we speak were studied in (Agarwal and Farid 2019), focusing on the distinct facial expressions and movements. These features were considered in combination with Support Vector Machines (SVM), achieving a 96.3% AUC over their own database.

Finally, fake detection methods based on the heart rate have been also studied in the literature. One of the first studies in this regard was (Conotter et al. 2014) where the authors preliminary evaluated the potential of blood flow changes in the face to distinguish between computer generated and real videos. Their proposed approach was evaluated using 12 videos (6 real and fake videos each), concluding that it is possible to use this metric to detect computer generated videos.

Changes in the blood flow have also been studied in (Ciftci, Demir, and Yin 2020; Qi et al. 2020) using DeepFake videos. In (Ciftci, Demir, and Yin 2020), the authors considered rPPG techniques to extract robust biological features. Classifiers based on SVM and CNN were analyzed, achieving final accuracies of 94.9% and 91.5% for the DeepFakes videos of FaceForensics++ and Celeb-DF, respectively.

Recently, in (Qi et al. 2020) a more sophisticated fake detector named DeepRhythm was presented. This approach was also based on features extracted using rPPG techniques. DeepRhythm was enhanced through two modules: *i)* motion-magnified spatial-temporal representation, and *ii)* dual-spatial-temporal attention. These modules were incorporated in order to provide a better adaptation to dynamically changing faces and various fake types. In general, good results with accuracies of 100% were achieved on FaceForensics++ database. However, this method suffers from a demanding preprocessing stage, needing a precise detection of 81 facial landmarks and the use of a color magnification algorithm prior to fake detection. Also, poor results were achieved on databases of the 2$^{\text{nd}}$ generation such as the DFDC Preview (Acc. = 64.1%).

In the present work, in addition to the proposal of a different DeepFake detection architecture, we enhance previous approaches, e.g. (Qi et al. 2020), by keeping the preprocessing stage as light and robust as possible, only composed of a face detector and frame normalization. To provide an overall picture, we include in Table 1 the results achieved with our proposed DeepFakesON-Phys in comparison with key related works, which shows that we outperform the state of the art on Celeb-DF v2 and DFDC Preview databases.

Proposed Method: DeepFakesON-Phys

Fig. 1 graphically summarizes the architecture of DeepFakesON-Phys, the proposed fake detector based on heart rate estimation. We hypothesize that rPPG methods should obtain significantly different results when trying to estimate the subjacent heart rate from a video containing a real face, compared with a fake face. Since the changes in color and illumination due to oxygen concentration are subtle and invisible to the human eye, we think that most of the existing DeepFake manipulation methods do not consider the physiological aspects of the human being yet.

The initial architecture of DeepFakesON-Phys is based on the DeepPhys model described in (Chen and McDuff 2018), whose objective was to estimate the human heart rate using facial video sequences. The model is based on deep learning and was designed to extract spatio-temporal information from videos mimicking the behavior of traditional handcrafted rPPG techniques. Features are extracted through the color changes in users' faces that are caused by the variation of oxygen concentration in the blood. Signal processing methods are also used for isolating the color changes caused by blood from other changes that may be caused by factors such as external illumination, noise, etc.

As can be seen in Fig. 1, after the first preprocessing stage, the Convolutional Attention Network (CAN) is composed of two different CNN branches:

- **Motion Model**: it is designed to detect changes between consecutive frames, i.e., performing a short-time analysis of the video for detecting fakes. To accomplish this task, the input at a time *t* consists of a frame computed as the normalized difference of the current frame $I(t)$ and the previous one $I(t-1)$.
- **Appearance Model**: it focuses on the analysis of the static information on each video frame. It has the target of providing the Motion Model with information about which points of the current frame may contain the most relevant information for detecting DeepFakes, i.e., a batch of attention masks that are shared at different layers of the CNN. The input of this branch at time t is the raw frame of the video $I(t)$, normalized to zero mean and unitary standard deviation.

The attention masks coming from the Appearance Model are shared with the Motion Model at two different points of the CAN. Finally, the output layer of the Motion Model is also the final output of the entire CAN.

In the original architecture (Chen and McDuff 2018), the output stage consisted of a regression layer for estimating the time derivative of the subject's heart rate. In our case, as we do not aim to estimate the pulse of the subject, but the presence of a fake face, we change the final regression layer to a classification layer, using a sigmoid activation function for obtaining a final score in the [0,1] range for each instant *t* of the video, related to the probability of the face being real.

Since the original DeepPhys model from (Chen and McDuff 2018) is not publicly available, instead of training a new CAN from scratch, we decided to initialize DeepFakesON-Phys with the weights from the model pretrained for heart rate estimation presented in (Hernandez-Ortega et al. 2020b), which is also an adaptation of DeepPhys but trained using the COHFACE database (Heusch, Anjos, and Marcel 2017). This model also showed to have high accuracy in the heart rate estimation task using real face videos, so our idea is to take benefit of that acquired knowledge to better train DeepFakesON-Phys through a proper

Table 2: **Identity swap publicly available databases** of the 2nd generation considered in our experimental framework.

	2nd Generation	
Database	**Real Videos**	**Fake Videos**
Celeb-DF v2 (Li et al. 2020)	590 (Youtube)	5,639 (DeepFake)
DFDC Preview (Dolhansky et al. 2019)	1,131 (Actors)	4,119 (Unknown)

fine-tuning process.

Once we initialized DeepFakesON-Phys with the mentioned weights, we freeze the weights of all the layers of the original CAN model apart from the new classification layer and the last fully-connected layer, and we retrain the model. Due to this fine-tuning process we take benefit of the weights learned for heart rate estimation, just adapting them for the DeepFake detection task. This way, we make sure that the weights of the convolutional layers remain looking for information relative to heart rate and the last layers learn how to use that information for detecting the existence of DeepFakes.

Databases

Two different public databases are considered in the experimental framework of this study. In particular, Celeb-DF v2 and DFDC Preview, the two most challenging DeepFake databases up to date. Their videos exhibit a large range of variations in aspects such as face sizes (in pixels), lighting conditions (i.e., day, night, etc.), backgrounds, different acquisition scenarios (i.e., indoors and outdoors), distances from the person to the camera, and pose variations, among others. These databases present enough images (fake and genuine) to fine-tune the original weights meant for heart rate estimation, obtaining new weights also based in rPPG features but adapted for DeepFake detection. Table 2 summarizes the main characteristics of the databases.

Celeb-DF v2

The aim of the Celeb-DF v2 database (Li et al. 2020) was to generate fake videos of better visual quality compared with the previous UADFV database. This database consists of 590 real videos extracted from Youtube, corresponding to celebrities with a diverse distribution in terms of gender, age, and ethnic group. Regarding fake videos, a total of 5,639 videos were created swapping faces using DeepFake technology. The final videos are in MPEG4.0 format.

DFDC Preview

The DFDC database (Dolhansky et al. 2019) is one of the latest public databases, released by Facebook in collaboration with other companies and academic institutions such as Microsoft, Amazon, and the MIT. In the present study we consider the DFDC Preview dataset consisting of 1,131 real videos from 66 paid actors, ensuring realistic variability in gender, skin tone, and age. It is important to remark that no publicly available data or data from social media sites were used to create this dataset, unlike other popular databases. Regarding fake videos, a total of 4,119 videos were created using two different unknown approaches for fakes generation. Fake videos were generated by swapping subjects with similar appearances, i.e., similar facial attributes such as skin tone, facial hair, glasses, etc. After a given pairwise model was trained on two identities, the identities were swapped onto the other's videos.

Experiments

Experimental Protocol

Celeb-DF v2 and DFDC Preview databases have been divided into non-overlapping datasets, development and evaluation. It is important to remark that each dataset comprises videos from different identities (both real and fake), unlike some previous studies. This aspect is very important in order to perform a fair evaluation and predict the generalization ability of the fake detection systems against unseen identities. Also, it is important to remark that the evaluation is carried out at frame level as in most previous studies (Tolosana et al. 2020b), not video level, using the popular AUC and accuracy metrics.

For the Celeb-DF v2 database, we consider real/fake videos of 40 and 19 different identities for the development and evaluation datasets respectively, whereas for the DFDC Preview database, we follow the same experimental protocol proposed in (Dolhansky et al. 2019) as the authors already considered this concern.

Fake Detection Results: DeepFakesON-Phys

This section evaluates the ability of DeepFakesON-Phys to detect the most challenging DeepFake videos of the 2nd generation. Table 3 shows the fake detection performance results achieved in terms of AUC and accuracy over the final evaluation datasets of Celeb-DF v2 and DFDC Preview. It is important to highlight that a separate fake detector is trained for each database.

In general, very good results are achieved in both DeepFake databases. For the Celeb-DF v2 database, DeepFakesON-Phys achieves an accuracy of 98.7% and an AUC of 99.9%. Regarding the DFDC Preview database, the results achieved are 94.4% accuracy and 98.2% AUC, similar ones to the obtained for the Celeb-DF database.

Observing the results, it seems clear that the fake detectors have learnt to distinguish the spatio-temporal differences between the real/fake faces of Celeb-DF v2 and DFDC Preview databases. Since all the convolutional layers of the proposed fake detector are frozen (the network was originally initialized with the weights from the model trained to predict the heart rate (Hernandez-Ortega et al. 2020b)), and we only train the last fully-connected layers, we can conclude that the proposed detection approach based on physiological measurement is successfully using pulse-related features for distinguishing between real and fake faces. These results prove that current face manipulation techniques do not pay attention to the heart-rate-related physiological information of the human being when synthesizing fake videos.

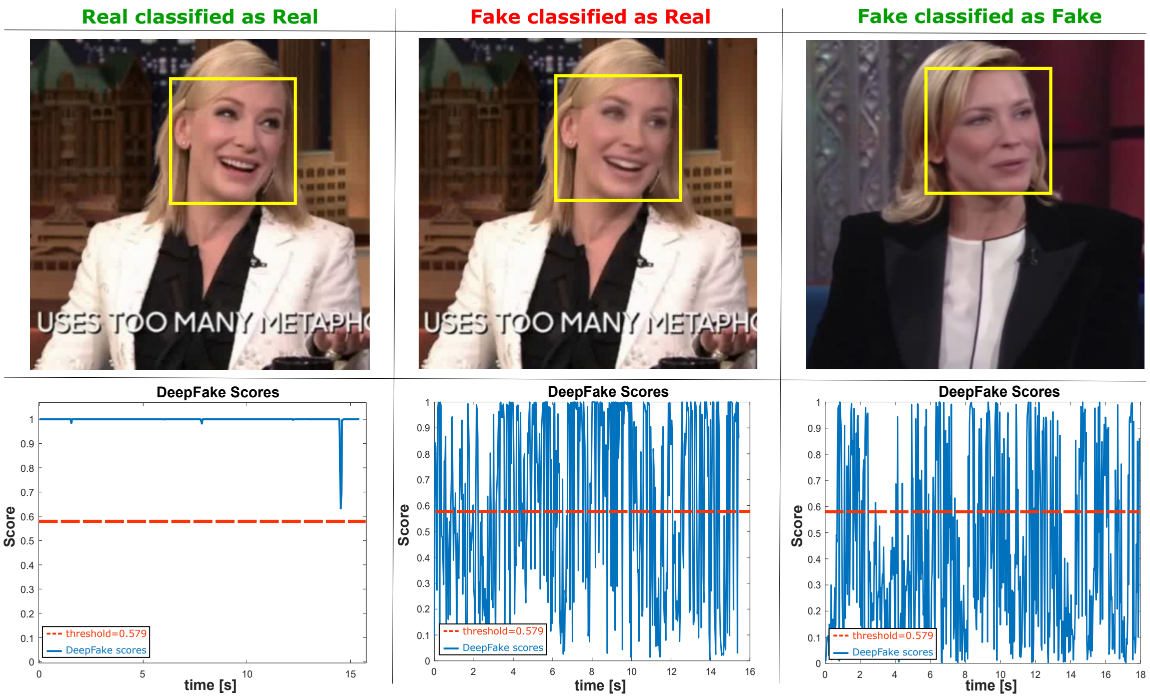

Figure 2: **Examples of successful and failed DeepFake detections**. Top: sample frames of evaluated videos. Bottom: score distribution for each sample video. For the fake video misclassified as containing a real face, the DeepFake detection scores present a higher mean compared to the case of the fake video correctly classified as a fake.

Table 3: **Fake detection performance** results in terms of AUC and Accuracy over the final evaluation datasets.

Database	AUC Results (%)	Acc. Results (%)
Celeb-DF v2	99.9	98.7
DFDC Preview	98.2	94.4

Fig. 2 shows some examples of successful and failed detections when evaluating the proposed approach with real/fake faces of Celeb-DF v2. In particular, all the failures correspond to fake faces generated from a particular video, misclassifying them as real faces. Fig. 2 shows a frame from the original real video (top-left), one from a misclassified fake video generated using that scenario (top-middle), and another from a fake video correctly classified as fake and generated using the same real and fake identities but from other source videos (top-right). The detection threshold is the same for all the testing databases and videos, and it has been selected to maximize the accuracy in the evaluation.

Looking at the score distributions along time of the three examples (Fig. 2, bottom), it can be seen that for the real face video (left) the scores are 1 for most of the time and always over the detection threshold. However, for the fake videos considered (middle and right), the score changes constantly, making the score of some fake frames to cross the detection threshold and consequently misclassifying them as real. Nevertheless, it is important to remark that these mistakes only happen if we analyze the results at frame level (traditional approach followed in the literature (Tolosana et al. 2020b)). In case we consider an evaluation at video level, DeepFakesON-Phys would be able to detect fake videos by integrating the temporal information available in short-time segments, e.g., in a similar way as described in (Hernandez-Ortega et al. 2018) for continuous face anti-spoofing.

We believe that the failures produced in this particular case are propitiated by the interferences of external illumination. rPPG methods that use handcrafted features are usually fragile against external artificial illumination in the frequency and power ranges of normal human heart rate, making difficult to distinguish those illumination changes from the color changes caused by blood perfusion. Anyway, the proposed physiological approach presented in this work is more robust to this kind of illumination perturbations than hand-crafted methods, thanks to the fact that the training process is data-driven, making possible to identify those interferences by using their presence in the training data.

Comparison with the State of the Art

Finally, we compared in Table 4 the results achieved in the present work with other state-of-the-art DeepFake detection approaches: head pose variations (Yang, Li, and Lyu 2019), face warping artifacts (Li et al. 2020), mesoscopic features (Afchar et al. 2018), pure deep learning features (Dang et al. 2020; Tolosana et al. 2020a), and physiological features (Qi et al. 2020; Ciftci, Demir, and Yin 2020). The best results achieved for each database are remarked in **bold**. Results in *italics* indicate that the evaluated database was not used for training. Some of these results are extracted

Table 4: **Comparison of different state-of-the-art fake detectors with our proposed DeepFakesON-Phys.** The best results achieved for each database are remarked in **bold**. Results in *italics* indicate that the evaluated database (Celeb-DF or DFDC) was not used for training.

Study	**Method**	**Classifiers**	**AUC Results (%)**	
			Celeb-DF (Li et al. 2020)	**DFDC** (Dolhansky et al. 2019)
(Yang, Li, and Lyu 2019)	Head Pose Features	SVM	*54.6*	*55.9*
(Li et al. 2020)	Face Warping Features	CNN	*64.6*	*75.5*
(Afchar et al. 2018)	Mesoscopic Features	CNN	*54.8*	*75.3*
(Dang et al. 2020)	Deep Learning Features	CNN + Attention Mechanism	*71.2*	-
(Tolosana et al. 2020a)	Deep Learning Features	CNN	83.6	91.1
(Qi et al. 2020)	Physiological Features	CNN + Attention Mechanism	-	*Acc. = 64.1*
(Ciftci, Demir, and Yin 2020)	Physiological Features	SVM/CNN	Acc. = 91.5	-
DeepFakesON-Phys [Ours]	**Physiological Features**	**CNN + Attention Mechanism**	**AUC = 99.9** **Acc. = 98.7**	**AUC = 98.2** **Acc. = 94.4**

from (Li et al. 2020).

Note that the comparison in Table 4 is not always under the same datasets and protocols, therefore it must be interpreted with care. Despite of that, it is patent that the proposed DeepFakesON-Phys has achieved state-of-the-art results in both Celeb-DF and DFDC Preview databases. In particular, it has further outperformed popular fake detectors based on pure deep learning approaches such as Xception and Capsule Networks (Tolosana et al. 2020a) and also other recent physiological approaches based on SVM/CNN (Ciftci, Demir, and Yin 2020).

Conclusions

This work has evaluated the potential of physiological measurement to detect DeepFake videos. In particular, we have proposed a novel DeepFake detector named DeepFakesON-Phys based on a Convolutional Attention Network (CAN) originally trained for heart rate estimation using remote photoplethysmography (rPPG). The proposed CAN approach consists of two parallel CNN networks that extract and share temporal and spatial information from video frames.

DeepFakesON-Phys has been evaluated using Celeb-DF v2 and DFDC Preview databases, two of the latest and most challenging DeepFake video databases. Regarding the experimental protocol, each database was divided into development and evaluation datasets, considering different identities in each dataset in order to perform a fair evaluation of the technology.

The soundness and competitiveness of DeepFakesON-Phys has been proven by the very good results achieved, AUC values of 99.9% and 98.2% for the Celeb-DF and DFDC databases, respectively. These results have outperformed other state-of-the-art fake detectors based on face warping and pure deep learning features, among others. Finally, the experimental results of this study reveal that current face manipulation techniques do not pay attention to the heart-rate-related or blood-related physiological information.

Immediate work may consist in replicating the state of the art DeepFake works and training them with the same databases than the ones used to train DeepFakesON-Phys in order to make a fair comparison of accuracy, and showing the actual performance of our method. Another future work will be oriented to the analysis of the robustness of the proposed fake detection approach against face manipulations unseen during the training process (Tolosana et al. 2020b), temporal integration of frame data (Hernandez-Ortega et al. 2018), and the application of the proposed physiological approach to other face manipulation techniques such as face morphing (Raja and *et al.* 2020).

Acknowledgments

This work has been supported by projects: IDEA-FAST (IMI2-2018-15-two-stage-853981), PRIMA (ITN-2019-860315), TRESPASS-ETN (ITN-2019-860813), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), and edBB (Universidad Autonoma de Madrid, UAM). J. H.-O. is supported by a PhD fellowship from UAM. R. T. is supported by a Postdoctoral fellowship from CAM/FSE.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *Proc. IEEE Int. Workshop on Information Forensics and Security*.

Agarwal, S.; and Farid, H. 2019. Protecting World Leaders Against Deep Fakes. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*.

Bharadwaj, S.; Dhamecha, T. I.; Vatsa, M.; and Singh, R. 2013. Computationally Efficient Face Spoofing Detection with Motion Magnification. In *Proc. IEEE/CVF Conf. on Comp. Vision and Pattern Recognition Workshops*.

Cellan-Jones, R. 2019. Deepfake Videos Double in Nine Months. URL https://www.bbc.com/news/technology-49961089.

Chen, W.; and McDuff, D. 2018. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *Proc. European Conf. on Computer Vision*, 349–365.

Ciftci, U. A.; Demir, I.; and Yin, L. 2020. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Trans. on Pattern Analysis and Machine Intelligence* .

Citron, D. 2019. How DeepFake Undermine Truth and Threaten Democracy. URL https://www.ted.com.

Conotter, V.; Bodnari, E.; Boato, G.; and Farid, H. 2014. Physiologically-Based Detection of Comp. Generated Faces in Video. In *Proc. IEEE Int. Conf. on Image Processing*.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. 2020. On the Detection of Digital Face Manipulation. In *Proc. IEEE/CVF Conf. on Comp. Vision and Pattern Recognition*.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv preprint:1910.08854* .

Erdogmus, N.; and Marcel, S. 2014. Spoofing Face Recognition with 3D Masks. *IEEE Transactions on Information Forensics and Security* 9(7): 1084–1097.

Galbally, J.; Marcel, S.; and Fierrez, J. 2014. Biometric Anti-Spoofing Methods: A Survey in Face Recognition. *IEEE Access* 2: 1530–1552.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proc. Advances in Neural Information Processing Systems*.

Güera, D.; and Delp, E. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *Proc. Int. Conf. on Advanced Video and Signal Based Surveillance*.

Hernandez-Ortega, J.; Daza, R.; Morales, A.; Fierrez, J.; and Tolosana, R. 2020a. Heart Rate Estimation from Face Videos for Student Assessment: Experiments on edBB. In *Proc. IEEE Comp. Software and Applications Conf.*

Hernandez-Ortega, J.; Fierrez, J.; Morales, A.; and Diaz, D. 2020b. A Comparative Evaluation of Heart Rate Estimation Methods using Face Videos. In *Proc. IEEE Intl. Workshop on Medical Computing*.

Hernandez-Ortega, J.; Fierrez, J.; Morales, A.; and Galbally, J. 2019. Introduction to Face Presentation Attack Detection. In *Handbook of Biometric Anti-Spoofing*, 187–206. Springer.

Hernandez-Ortega, J.; Fierrez, J.; Morales, A.; and Tome, P. 2018. Time Analysis of Pulse-Based Face Anti-Spoofing in Visible and NIR. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition Workshops*.

Heusch, G.; Anjos, A.; and Marcel, S. 2017. A reproducible study on remote heart rate measurement. *arXiv preprint:1709.00962* .

Jung, T.; Kim, S.; and Kim, K. 2020. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* 8: 83144–83154.

Karras, T.; et al. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. IEEE/CVF Conf. on Comp. Vision and Patter Recognition*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. In *Proc. Int. Conf. on Learning Represent.*

Li, Y.; Chang, M.; and Lyu, S. 2018. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. In *Proc. IEEE Int. Work. Information Forensics and Security*.

Li, Y.; and Lyu, S. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *Proc. IEEE/CVF Conf. on Comp. Vision and Pattern Recognition Workshops*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proc. IEEE/CVF Conf. on Comp. Vision and Pattern Recognition*.

Marcel, S.; Nixon, M.; Fierrez, J.; and Evans, N. 2019. *Handbook of Biometric Anti-Spoofing (2nd Edition)*.

Matern, F.; Riess, C.; and Stamminger, M. 2019. Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations. In *Proc. IEEE Winter App. of Comp. Vision Workshops*.

McDuff, D. J.; Estepp, J. R.; Piasecki, A. M.; and Blackford, E. B. 2015. A Survey of Remote Optical Photoplethysmographic Imaging Methods. In *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*.

Neves, J.; et al. 2020. GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing* 14(5): 1038–1048.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Use of a Capsule Network to Detect Fake Images and Videos. *arXiv preprint:1910.12467* .

Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; and Zhao, J. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. *arXiv preprint:2006.07634* .

Raja, K.; and *et al.* 2020. Morphing Attack Detection - Database, Evaluation Platform and Benchmarking. *IEEE Transactions on Information Forensics and Security.* .

Ranjan, R.; Patel, V. M.; and Chellappa, R. 2017. Hyperface: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 41(1): 121–135.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proc. IEEE/CVF Int. Conf. on Comp. Vision*.

Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; and Natarajan, P. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *Proc. IEEE/CVF Conf. on Comp. Vision and Pattern Recognition Workshops*.

Soukupova, T.; and Cech, J. 2016. Real-Time Eye Blink Detection Using Facial Landmarks. In *Proc. Comp. Vision Winter Workshop*.

Tan, D.; and Nijholt, A. 2010. Brain-Computer Interfaces and Human-Computer Interaction. In *Brain-Computer Interfaces*, 3–19. Springer.

Tolosana, R.; Romero-Tapiador, S.; Fierrez, J.; and Vera-Rodriguez, R. 2020a. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. *Proc. International Conference on Pattern Recognition Workshops* .

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020b. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion* 64: 131–148.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*.