Chapter 12
DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame

Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales

Abstract This chapter describes a 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. This chapter explores to what extent rPPG is useful for the detection of DeepFake videos. We analyze the recent fake detector named DeepFakesON-Phys that is based on 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 v2 and DFDC. The results achieved for DeepFake detection based on a single frame are over 98% AUC (Area Under the Curve) on both databases, proving the success of fake detectors based on physiological measurement to detect the latest DeepFake videos. In this chapter, we also propose and study heuristical and statistical approaches for performing continuous DeepFake detection by combining scores from consecutive frames with low latency and high accuracy (100% on the Celeb-DF v2 evaluation dataset). We show that combining scores extracted from short-time video sequences can improve the discrimination power of DeepFakesON-Phys.

The present chapter is an updated adaptation of the conference paper [21].

J. Hernandez-Ortega (✉) · R. Tolosana · J. Fierrez · A. Morales
Universidad Autonoma de Madrid, Madrid, Spain
e-mail: javier.hernandezo@uam.es

R. Tolosana
e-mail: ruben.tolosana@uam.es

J. Fierrez
e-mail: julian.fierrez@uam.es

A. Morales
e-mail: aythami.morales@uam.es

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_12

12.1 Introduction

DeepFakes have become a great public concern recently [5, 8]. 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 subject with the face of another subject. This type of digital manipulation is also known in the literature as Identity Swap, and it is moving forward very fast [46].

Currently, most face manipulations are based on popular machine learning techniques such as AutoEncoders (AE) [25] and Generative Adversarial Networks (GAN) [15], achieving in general very realistic visual results, specially in the latest generation of public DeepFakes [45], and the present trends [24]. However, 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 [17], health care [31], human-computer interaction [44], and security [29], among many other tasks.

In physical face attacks, a.k.a. Presentation Attacks (PAs), real subjects are often impersonated using artefacts such as photographs, videos, makeup, and masks [13, 29, 38, 39]. Face recognition systems are known to be vulnerable against these attacks unless proper detection methods are implemented [14, 19]. Some of these detection methods are based on liveness detection by using information such as eye blinking or natural facial micro-expressions [4]. 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 [20]. When applying this technique to a video sequence with a fake face, the estimated heart rate signal is significantly different from the heart rate extracted from a real face [12].

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 chapter, 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 real faces (a.k.a. bona fide presentations [1]). 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.

This chapter analyzes the potential of DeepFakesON-Phys, which was originally analyzed in [21] for the detection of DeepFakes videos at frame level, and it is further studied in this chapter for the detection at short-term video level. DeepFakesON-Phys is a fake detector based on deep learning that uses rPPG features previously learned for the task of heart rate estimation and adapts them for the detection of DeepFakes by means of a knowledge-transfer process, thus obtaining a novel fake detector based on physiological measurement. This chapter also includes new additional experiments using DeepFakesON-Phys, comparing the accuracies of DeepFake detection based on scores from single frames and on the temporal integration of scores from consecutive frames.

In particular, the information related to the heart rate is considered to decide whether a video is real or fake. DeepFakesON-Phys 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 [3, 30] and also on the artefacts/fingerprints inserted during the synthesis process [32], which are highly dependent on a specific fake manipulation technique.

DeepFakesON-Phys is based on DeepPhys [6], 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 handcrafted approaches. In [21], we used the architecture of DeepPhys, but making changes to suit the approach 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 into rPPG-related features from the images in the face videos.

In this context, in this chapter, we:

- Perform an in-depth literature review of DeepFake detection approaches with special emphasis on physiological techniques, including the key aspects of the detection systems, the databases used, and the main results achieved.
- Describe DeepFakesON-Phys,[1] a recent approach presented in [21] based on the physiological measurement to detect DeepFake videos. Figure 12.1 graphically summarizes DeepFakesON-Phys, which is based on the original architecture DeepPhys [6], 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.
- Include a thorough experimental assessment of DeepFakesON-Phys, considering two of the latest public databases of the second DeepFake generation: Celeb-DF v2 [28] and DFDC Preview [11]. We evaluated DeepFakesON-Phys doing both analysis of fake detection at frame level and also at the short-term video level. DeepFakesON-Phys achieves high-accuracy results in both evaluations, 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. Section 12.2 summarizes previous studies focused on the detection of DeepFakes. Section 12.3 describes

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

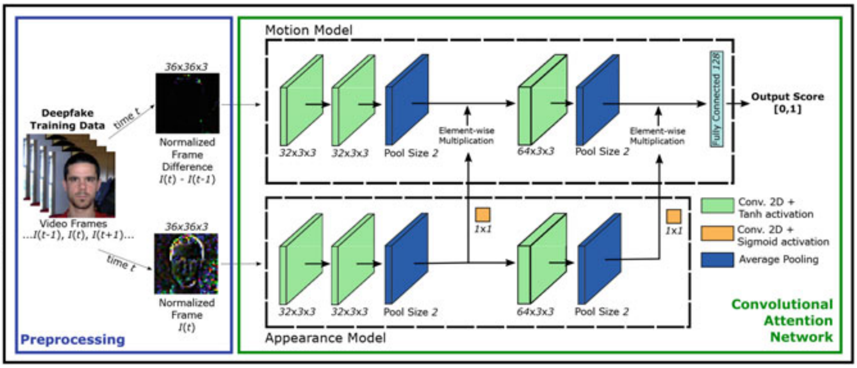

Fig. 12.1 **DeepFakesON-Phys architecture** [21]. It comprises two stages: (i) a preprocessing step to normalize the video frames, and (ii) a Convolutional Attention Network composed of Motion and Appearance Models to better detect fake videos

DeepFakesON-Phys. Section 12.4 summarizes all databases considered in the experimental framework of this study. Sections 12.5 and 12.6 describe the experimental protocol and the results achieved in comparison with the state of the art, respectively. Finally, Sect. 12.7 draws the final conclusions and points out future research lines.

12.2 Related Works

Different approaches have been proposed in the literature to detect DeepFake videos. Table 12.1 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 accuracy (Acc.), 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 [28].

The first studies in the area focused on the visual artefacts existed in the first generation of fake videos. The authors of [30] proposed fake detectors based on simple visual artefacts such as eye color, missing reflections, and missing details in the teeth areas, achieving a final 85.1% AUC.

Approaches based on the detection of the face warping artefacts have also been studied in the literature. For example, [27, 28] proposed detection systems based on CNN in order to detect the presence of such artefacts from the face and the surrounding areas, being one of the most robust detection approaches against unseen face manipulations.

Table 12.1 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 et al. [30]	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 et al. [27, 28]	Face Warping Features	CNN	AUC = 97.7	UADFV
			AUC = 93.0	*FF++ / DFD*
			AUC = 75.5	*DFDC Preview*
			AUC = 64.6	*Celeb-DF*
Rossler et al. [40]	Mesoscopic Features Steganalysis Features Deep Learning Features	CNN	Acc. ≃ 94.0	FF++ (DeepFake, LQ)
			Acc. ≃ 98.0	FF++ (DeepFake, HQ)
			Acc. ≃ 100.0	FF++ (DeepFake, RAW)
			Acc. ≃ 93.0	FF++ (FaceSwap, LQ)
			Acc. ≃ 97.0	FF++ (FaceSwap ,HQ)
			Acc. ≃ 99.0	FF++ (FaceSwap, RAW)
Nguyen et al. [33]	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. [10]	Deep Learning Features	CNN + Attention Mechanism	AUC = 99.4
			EER = 3.1
				DFFD

Dolhansky et al. [11]	Deep Learning Features	CNN	Precision = 93.0
			Recall = 8.4
				DFDC Preview

Sun et al. [43]	Deep Learning Features	CNN	AUC = 98.5	FF++
			AUC = 61.4	Celeb-DF
			AUC = 69.0	*DFDC Preview*

(continued)

Table 12.1 (continued)

Study	Method	Classifiers	Best performance (%)	Databases

Sabir et al. [41]	Image + Temporal Features	CNN + RNN	AUC = 96.9
			AUC = 96.3
				FF++ (DeepFake, LQ)
				FF++ (FaceSwap, LQ)

Trinh et al. [47]	Image + Temporal Features	CNN	AUC = 99.2	FF++
			AUC = 68.2	Celeb-DF
Tolosana et al. [45]	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. [9]	Physiological Features	–	Acc. = 100	Own
Li et al. [26]	Physiological Features	LRCN	AUC = 99.0	UADFV
Agarwal et al. [3]	Physiological Features	SVM	AUC = 96.3	Own (FaceSwap, HQ)

Ciftci et al. [7]	Physiological Features	SVM/CNN	Acc. = 94.9
			Acc. = 91.5
				FF++ (DeepFakes)
				Celeb-DF

Jung et al. [23]	Physiological Features	Distance	Acc. = 87.5	Own
Qi et al. [35]	Physiological Features	CNN + Attention Mechanism	Acc. = 100.0	FF++ (FaceSwap)
			Acc. = 100.0	FF++
			Acc. = 64.1	*DFDC Preview*
DeepFakesON-Phys[21]	**Physiological Features**	**CAN**	**AUC = 99.9**	**Celeb-DF v2 (Frame Level)**
			AUC = 98.2	**DFDC Preview (Frame Level)**
			AUC = 100	**Celeb-DF v2 (Short-Term Video Level)**

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 [11, 40], novel ones such as Capsule Networks [33], and novel training techniques based on attention mechanisms [10]. In particular, we highlight the work presented in [43], focused on improving the generalization ability of the models to detect DeepFake videos. The authors defined a Learning-To-Weight (LTW) framework based on meta-learning that is composed of two branches: the first one performs binary detection, extracting features from the images and determining if an image is real or a fake, while the second branch aims to assign domain-adaptive weights to each sample, helping the model to extract more domain-general features.

Fake detectors based on the image and temporal discrepancies across frames have also been proposed in (DeepFake) the literature [41, 47]. In [41], the authors proposed a Recurrent Convolutional Network similar to [16], trained end-to-end instead of using a pre-trained model. Their proposed detection approach was tested using FaceForensics++ database [40], achieving AUC results above 96%.

In [47], Trinh et al. proposed a human-centered approach for detecting forgery in face images. Their approach looked for temporal artefacts within DeepFake videos, detecting them efficiently while providing explanations of DeepFake dynamics, useful for giving useful information to supervising humans.

Although most approaches are based on the detection of fake videos using the whole face, in [45], 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 first and second generations.

We also pay special attention to the fake detectors based on physiological information. The eye blinking rate was studied in [23, 26]. Li et al. [26] proposed Long-Term Recurrent Convolutional Networks (LRCN) to capture the temporal dependencies that existed in human eye blinking. Their method was evaluated on the UADFV database, achieving a final 99.0% AUC. More recently, [23] proposed a different approach named DeepVision. They fused the Fast-HyperFace [37] and EAR [42] 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 [3], 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 [9] where the authors preliminarily 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 (six 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 [7, 35] using DeepFake videos. In [7], 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 [35], 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 pre-processing 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 second generation such as the DFDC Preview (Acc. = 64.1%).

Regarding DeepFakesON-Phys originally presented in [21], in addition to the proposal of a different DeepFake detection architecture, we enhanced previous approaches, e.g. [35], 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 12.1 the results achieved with our proposed method in comparison with key related works, showing the good results on both Celeb-DF v2 and DFDC Preview databases for the frame-level analysis and on Celeb-DF v2 for the temporal integration of consecutive scores, AUC = 100%.

12.3 DeepFakesON-Phys

Figure 12.1 graphically summarizes the architecture of DeepFakesON-Phys [21], 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 [6], 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 and noise.

As can be seen in Fig. 12.1, 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 [6], 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 [6] is not publicly available, instead of training a new CAN from scratch, we decided to initialize DeepFakesON-Phys with the weights from the model pre-trained for heart rate estimation presented in [18], which is also an adaptation of DeepPhys but trained using the COHFACE database [22]. 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 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 the 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.

12.4 Databases

Two different public databases are considered in the experimental framework of this study. In particular, Celeb-DF v2 [28] and DFDC Preview [11], 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 subject 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 on rPPG features but adapted for DeepFake detection.

12.4.1 Celeb-DF v2 Database

Celeb-DF v2 is one of the most challenging DeepFake databases up to date [28]. The aim of the Celeb-DF v2 database was to generate fake videos of better visual quality compared with the previous UADFV database [26]. 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.

12.4.2 DFDC Preview

The DFDC database [11] 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, and glasses. After a given pairwise model was trained on two identities, the identities were swapped onto the other's videos.

12.5 Experimental Protocol

Celeb-DF v2 and DFDC Preview databases have been divided into non-overlapping datasets, development and evaluation. 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 [11] as the authors already considered this concern.

In this chapter, we followed two different strategies for DeepFake detection. First, for Celeb-DF v2 and DFDC Preview, we perform detection based on single scores obtained by DeepFakesON-Phys where the evaluation is carried out at a frame level as in most previous studies [46], not video level, using the popular AUC and accuracy metrics. Second, we also perform for Celeb-DF v2 videos temporal integration of DeepFake detection scores combining the single scores from non-overlapped temporal windows of T seconds to form a final fused DeepFake detection score. We decided to combine the individual scores following three different strategies:

- **Mean Score**: The DeepFake detection scores of individual frames from each temporal window (T seconds) are averaged to obtain the integrated score.
- **Median Score**: We computed the median of the individual DeepFake detection scores into each temporal window (T seconds).
- **Quickest Change Detection (QCD)**: This is a statistical method that first estimates match and non-match distributions of the scores, i.e., real face and DeepFakes. Then it tries to detect the specific moment in which the incoming detection scores change from one type of distribution to the other. This approach needs prior data in order to build the match and non-match distributions. Some variants of QCD also require to know the probability of a DeepFake in advance, so we decided to implement the MiniMax QCD (MQCD) algorithm from [34], which only needs the score distributions that we obtained in advance using a development data subset.

12.6 Fake Detection Results: DeepFakesON-Phys

This section evaluates the ability of DeepFakesON-Phys to detect some of the most challenging DeepFake videos of the second generation from Celeb-DF v2 [28] and DFDC Preview [11] databases.

12.6.1 DeepFakes Detection at Frame Level

Table 12.2 shows the fake detection results for the case in which we perform an analysis at frame level, following the traditional procedure in the literature [45, 46]. 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 to the ones 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 [18]), and we only train the last fully connected layers, we can conclude that the proposed detection approach based on physiological measurement is successful using pulse-related features for distinguishing between

Table 12.2 Comparison of different state-of-the-art DeepFake detectors with the **frame-level detection** based on 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 [28]	DFDC [11]
Yang et al. [48]	Head pose features	SVM	*54.6*	*55.9*
Li et al. [28]	Face warping features	CNN	*64.6*	*75.5*
Afchar et al. [2]	Mesoscopic features	CNN	*54.8*	*75.3*
Dang et al. [10]	Deep learning features	CNN + Attention mechanism	*71.2*	–
Tolosana et al. [45]	Deep learning features	CNN	83.6	91.1
Qi et al. [35]	Physiological features	CNN + Attention mechanism	–	*Acc. = 64.1*
Ciftci et al. [7]	Physiological features	SVM/CNN	Acc. = 91.5	–
Sun et al. [43]	Deep learning features	CNN	*61.4*	*69.0*
Trinh et al. [47]	Image + Temporal features	CNN	*68.20*	–

DeepFakesON-Phys [21]	Physiological Features	CNN + Attention Mechanism	AUC = 99.9
			Acc. = 98.7
				AUC = 98.2
				Acc. = 94.4

real and fake faces. These results prove that the current face manipulation techniques do not pay attention to the heart-rate-related physiological information of the human being when synthesizing fake videos.

In Table 12.2, we also compare the results achieved with the single score DeepFake detection approach against other state-of-the-art DeepFake detection methods: head pose variations [48], face warping artefacts [28], mesoscopic features [2], pure deep learning features [10, 45], and physiological features [7, 35]. Results in *italics* indicate that the evaluated database was not used for training. Some of these results are extracted from [28]. Note that the comparison is not always made under the same datasets and protocols; therefore, it must be interpreted with care. Despite of that, it is patent that DeepFakesON-Phys has achieved state-of-the-art results. In particular, it has further outperformed popular fake detectors based on pure deep learning approaches such as Xception and Capsule Networks [45] and also other recent physiological approaches based on SVM/CNN [7].

Figure 12.2 shows some examples of successful and failed detections when evaluating the fake detection at the frame level. In particular, all the failures correspond

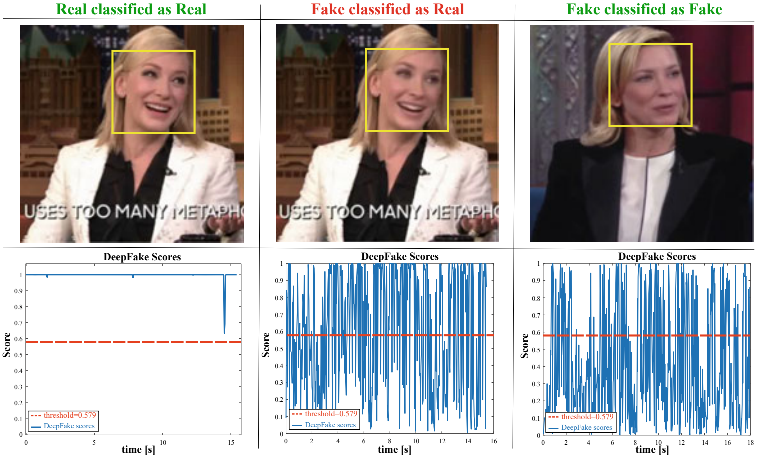

Fig. 12.2 Examples of successful and failed DeepFake detections. Top: sample frames of evaluated videos. Bottom: detection scores for each evaluated video (frame level). For the fake video misclassified as containing a real face, the DeepFake detection scores present a higher mean compared to the case of the fake video correctly classified as a fake

to fake faces generated from a particular video, misclassifying them as real faces. Figure 12.2 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).

Looking at the score distributions along time of the three examples (Fig. 12.2, 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 of each frame changes constantly, making the score of some fake frames to cross the detection threshold and consequently misclassifying them as real.

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 it difficult to distinguish those illumination changes from the color changes caused by blood perfusion. Anyway, DeepFakesON-Phys is more robust to this kind of illumination perturbations than handcrafted methods, thanks to the fact that the training process is data-driven, making it possible to identify those interferences by using their presence in the training data.

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 [46]). In case we consider the temporal information available in short-time segments of the video, e.g., in a similar way as described in [20] for continuous face anti-spoofing, DeepFakesON-Phys could achieve better detection results. This analysis at the short-term video level (not frame level) is described in the next section.

12.6.2 DeepFakes Detection at Short-Term Video Level

With the objective of detecting the type of errors illustrated in Fig. 12.2, in this section, we perform combination of the frame-level scores inside a temporal window of variable length (T) using three different combination strategies, i.e., mean score, median score, and QCD score [34]. The output for each one of these combination methods will be an individual DeepFake detection score for each temporal window. Therefore, the analysis carried out in this section is at the short-term video level.

We evaluate these methods on Celeb-DF v2 considering values of T going from 5 to 15 seconds in order to have a relevant number of scores to combine inside each time window. In this case, a DeepFake detection decision will be generated with a Delay of T seconds (video segments are not overlapped in time in our experiments). Additionally, the QCD algorithm also needs prior data in order to build the match and non-match distributions. To compute those distributions, we use all the single scores of 50 different time windows (25 real, 25 fake) from the evaluation dataset, leaving them out of the final testing process and results included in this section.

Table 12.3 shows the results for the evaluation of the DeepFake detector when varying the duration of the temporal window T. QCD has shown to be the most accurate integration method, obtaining the highest levels of AUC and accuracy even with slightly shorter values of T than the other combination strategies.

It can be seen that, in general, the highest AUC (i.e., the best DeepFake detection performance) is not obtained when using the largest T value, but lower ones (T = 6-7 seconds). For example, for the QCD scores, we have achieved an AUC and an accuracy of 100.0% using temporal windows of 6 seconds, while using higher values of T makes performance to get slightly worse. With shorter values of T (less of 5 s.), the small amount of available frame-level scores within each decision time window may diminish the reliability of each combined score. On the other hand, the combined scores obtained with large values of T may be less reliable as they are more prone to errors due to variations inside each window.

Finally, we decided to test the evolution of the different strategies for temporal integration of scores in cases like the one shown in Fig. 12.2 (right), where the single frame-level scores vary constantly. With temporal integration of scores, we expect to avoid that changeful behavior, obtaining more stable DeepFake detection results.

Figure 12.3 shows the evolution of the different detection scores for a former fail case video, both for single frame-level scores and for mean and QCD integrated scores. The results in the figure show that the temporal integration of scores can reduce the shakiness of the single scores (both for mean and QCD combinations), what is translated into an improved AUC and accuracy rates like the ones seen in Table 12.3. Even though QCD scores have achieved the highest improvement in performance,

Table 12.3 **DeepFakes Detection at Short-Term Video Level**. The study has been performed on Celeb-DF v2, changing the length of the time window T of the video sequences analyzed. Values are in %. The highest values of AUC for each type of combination of score are highlighted in bold

Mean score

Window Size T[s]	5	6	7	8	9	10	11	12	13	14	15
AUC [%]	99.97	99.98	**99.99**	99.97	99.98	99.96	99.97	99.98	99.97	99.97	99.93
Acc. [%]	99.24	99.47	99.47	99.24	99.46	99.15	99.32	99.63	99.14	99.06	99.37

Median score

Window Size T[s]	5	6	7	8	9	10	11	12	13	14	15
AUC [%]	99.97	99.98	**99.99**	99.97	99.98	99.96	99.97	99.98	99.97	99.97	99.93
Acc. [%]	99.24	99.47	99.47	99.24	99.46	99.15	99.32	99.63	99.14	99.06	99.37

QCD score

Window Size T[s]	5	6	7	8	9	10	11	12	13	14	15
AUC [%]	99.97	**100.0**	99.98	99.96	99.98	99.96	99.97	99.98	99.97	99.97	99.93
Acc. [%]	99.49	100.0	99.73	99.24	99.46	99.15	99.32	99.63	99.14	99.06	99.37

the mean scores also obtain the same stability benefits with the additional advantage of not needing any previous knowledge of the real and fake scores distributions.

12.7 Conclusions

This chapter has evaluated the potential of physiological measurement to detect DeepFake videos. In particular, we have described the recent DeepFake detector named DeepFakesON-Phys, originally presented in [21]. DeepFakesON-Phys is based on a Convolutional Attention Network (CAN) initially trained for heart rate estimation using remote photoplethysmography (rPPG). The proposed CAN approach consists of two parallel Convolutional Neural Networks (CNN) that extract and share temporal and spatial information from video frames.

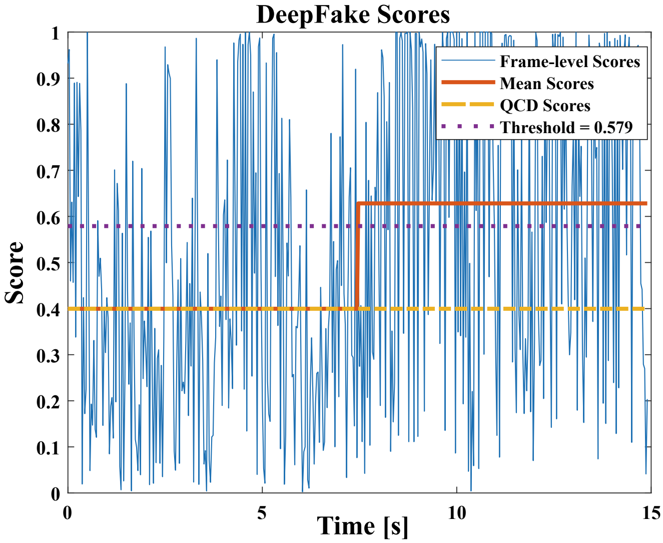

Fig. 12.3 Examples of successful temporal integration of frame-level scores. The figure shows the single scores, the mean scores, and QCD integrated scores ($T = 7$ sec.) for a DeepFake video of Celeb-DF v2. For the single frame-level score detection, the scores go over and under the threshold causing numerous false acceptances. For the temporal integration strategies (short-term video analysis), the mean detection score is under the threshold for the first temporal window (successful DeepFake detection), but for the second window, the score crosses the threshold causing a false acceptance. On the contrary, the QCD score is under the threshold for both temporal windows thanks to its statistical nature

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.

Two different evaluations have been performed using DeepFakesON-Phys, the first one consisted in detecting DeepFakes using frame-level scores, proving the soundness and competitiveness of the detection model with Area Under the Curve (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.

However, in some specific cases, the detection of DeepFakes using frame-level scores has shown some instability that leads to misclassified DeepFakes and real videos. To solve these issues, we have included a second evaluation on Celeb-DF v2, in which we have performed temporal integration of the scores inside a temporal window of T seconds (analysis at short-term video level). We have calculated three different integrated scores: mean, median, and Quickest Change Detection (QCD) scores. The results of this second evaluation have improved those obtained with the single scores (analysis at frame level), achieving both an AUC and an accuracy of 100% when using the QCD score with a temporal window of T=6 seconds. We can conclude that 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 will be oriented to the analysis of the robustness of the proposed fake detection approach against face manipulations unseen during the training process [46], and the application of the proposed physiological approach to other face manipulation techniques such as face morphing [36].

Acknowledgements This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (MINECO/FEDER RTI2018-101248-B-I00), and COST CA16101 (MULTI-FORESEE). J. H.-O. is supported by a PhD fellowship from UAM.

References

1. Information Technology-Biometric Presentation Attack Detection-Part 3: Testing and Reporting. Tech. rep., ISO/IEC JTC1 SC37 Biometrics (2017)
2. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) MesoNet: a compact facial video forgery detection network. In: Proceedings of IEEE international workshop on information forensics and security (2018)
3. Agarwal S, Farid H (2019) Protecting world leaders against deep fakes. In: Proceedings of IEEE/CVF conference on computer vision and pattern recognition workshops (2019)
4. Bharadwaj S, Dhamecha TI, Vatsa M, Singh R (2013) Computationally efficient face spoofing detection with motion magnification. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition workshops (2013)
5. Cellan-Jones R (2019) Deepfake videos double in nine months (2019). https://www.bbc.com/news/technology-49961089
6. Chen W, McDuff D (2018) DeepPhys: video-based physiological measurement using convolutional attention networks. In: Proceedings of European Conference on Computer Vision, pp 349–365
7. Ciftci UA, Demir I, Yin L (2020) FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell
8. Citron D (2019) How deepfake undermine truth and threaten democracy. https://www.ted.com
9. Conotter V, Bodnari E, Boato G, Farid H (2014) Physiologically-based detection of computer generated faces in video. In: Proceedings IEEE international conference on image processing
10. Dang H, Liu F, Stehouwer J, Liu X, Jain A (2020) On the detection of digital face manipulation. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition
11. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv:1910.08854
12. Erdogmus N, Marcel S (2014) Spoofing face recognition with 3D masks. IEEE Trans Inf Forensics Secur 9(7):1084–1097
13. Galbally J, Fierrez J, Ortega-Garcia J (2007) Vulnerabilities in biometric systems: attacks and recent advances in liveness detection. In: Proceedings Spanish workshop on biometrics, SWB
14. Galbally J, Marcel S, Fierrez J (2014) Biometric anti-spoofing methods: a survey in face recognition. IEEE Access 2:1530–1552
15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings advances in neural information processing systems

16. Güera D, Delp E (2018) Deepfake video detection using recurrent neural networks. In: Proceedings international conference on advanced video and signal based surveillance
17. Hernandez-Ortega J, Daza R, Morales A, Fierrez J, Tolosana R (2020) Heart rate estimation from face videos for student assessment: experiments on edBB. In: Proceedings IEEE computer software and applications conference
18. Hernandez-Ortega J, Fierrez J, Morales A, Diaz D (2020) A comparative evaluation of heart rate estimation methods using face videos. In: Proceedings IEEE international workshop on medical computing
19. Hernandez-Ortega J, Fierrez J, Morales A, Galbally J (2019) Introduction to face presentation attack detection. In: Handbook of biometric anti-spoofing. Springer, pp 187–206
20. Hernandez-Ortega J, Fierrez J, Morales A, Tome P (2018) Time analysis of pulse-based face anti-spoofing in visible and NIR. In: Proceedings IEEE conference on computer vision and pattern recognition workshops
21. Hernandez-Ortega J, Tolosana R, Fierrez J, Morales A (2021) DeepFakesON-Phys: deepfakes detection based on heart rate estimation. AAAI's workshop on artificial intelligence safety (SafeAI) (2021)
22. Heusch G, Anjos A, Marcel S (2017) A reproducible study on remote heart rate measurement. arXiv:1709.00962
23. Jung T, Kim S, Kim K (2020) DeepVision: deepfakes detection using human eye blinking pattern. IEEE Access 8:83144–83154
24. Karras T et al (2020) Analyzing and improving the image quality of StyleGAN. In: Proceedings IEEE/CVF conference on computer vision and patter recognition
25. Kingma DP, Welling M (2013) Auto-encoding Variational Bayes. In: Proceedings international conference on learning represent
26. Li Y, Chang M, Lyu S (2018) In Ictu Oculi: exposing AI generated fake face videos by detecting eye blinking. In: Proceedings IEEE international workshop information forensics and security
27. Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition workshops
28. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition
29. Marcel S, Nixon M, Fierrez J, Evans N (2019) Handbook of biometric anti-spoofing, 2nd edn
30. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: Proceedings IEEE winter applications of computer vision workshops
31. McDuff DJ, Estepp JR, Piasecki AM, Blackford EB (2015) A survey of remote optical photoplethysmographic imaging methods. In: Proceedings annual international conference of the IEEE engineering in medicine and biology society
32. Neves JC, Tolosana R, Vera-Rodriguez R, Lopes V, Proença H, Fierrez J (2020) GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection. IEEE J Select Top Signal Process 14(5):1038–1048
33. Nguyen HH, Yamagishi J, Echizen I (2019) Use of a capsule network to detect fake images and videos. arXiv:1910.12467
34. Perera P, Fierrez J, Patel V (2020) Quickest intruder detection for multiple user active authentication. In: IEEE international conference on image processing (ICIP)
35. Qi H, Guo Q, Juefei-Xu F, Xie X, Ma L, Feng W, Liu Y, Zhao J (2020) DeepRhythm: exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings ACM multimedia conference
36. Raja K, Ferrara M, Franco A, Spreeuwers L, Batskos I, de Wit F, Gomez-Barrero M, Scherhag U, Fischer D, Venkatesh S, Singh JM, Li G, Bergeron L, Isadskiy S, Ramachandra R, Rathgeb C, Frings D, Seidel U, Knopjes F, Veldhuis R, Maltoni D, Busch C (2020) Morphing attack detection-database. Evaluation platform and benchmarking. IEEE Trans Inf Forensics Secur
37. Ranjan R, Patel VM, Chellappa R (2017) Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell 41(1):121–135

38. Rathgeb C, Drozdowski P, Busch C (2020) Makeup presentation attacks: review and detection performance benchmark. IEEE Access 8:224958–224973
39. Rathgeb C, Drozdowski P, Busch C (2021) Detection of makeup presentation attacks based on deep face representations. In: Proceedings international conference on pattern recognition
40. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: learning to detect manipulated facial images. In: Proceedings IEEE/CVF international conference on computer vision
41. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. In: Proceedings IEEE/CVF conference on computer vision and pattern recognition workshops
42. Soukupova T, Cech J (2016) Real-time eye blink detection using facial landmarks. In: Proceedings computer vision winter workshop
43. Sun K, Liu H, Ye Q, Liu J, Gao Y, Shao L, Ji R (2021) Domain general face forgery detection by learning to weight. In: Proceedings AAAI conference on artificial intelligence
44. Tan D, Nijholt A (2010) Brain-computer interfaces and human-computer interaction. In: Brain-computer interfaces. Springer, pp 3–19
45. Tolosana R, Romero-Tapiador S, Fierrez J, Vera-Rodriguez R (2020) DeepFakes evolution: analysis of facial regions and fake detection performance. In: Proceedings international conference on pattern recognition workshops
46. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) DeepFakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
47. Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. In: Proceedings IEEE/CVF winter conference on applications of computer vision (WACV), pp 1973–1983
48. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: Proceedings IEEE international conference on acoustics, speech and signal processing