Time Analysis of Pulse-based Face Anti-Spoofing in Visible and NIR

Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, Pedro Tome
Biometrics and Data Pattern Analytics - BiDA Lab
Universidad Autonoma de Madrid, Madrid, Spain
javier.hernandezo@uam.es, julian.fierrez@uam.es, aythami.morales@uam.es, pedro.tome@inv.uam.es

Abstract

In this paper we study Presentation Attack Detection (PAD) in face recognition systems against realistic artifacts such as 3D masks or good quality of photo attacks. In recent works, pulse detection based on remote photoplethysmography (rPPG) has shown to be a effective countermeasure in concrete setups, but still there is a need for a deeper understanding of when and how this kind of PAD works in various practical conditions. Related works analyze full video sequences (usually over 60 seconds) to distinguish between attacks and legitimate accesses. However, existing approaches may not be as effective as it has been claimed in the literature in time variable scenarios. In this paper we evaluate the performance of an existent state-of-the-art PAD scheme based on rPPG when analyzing short-time video sequences extracted from a longer video.

Results are reported using the 3D Mask Attack Database (3DMAD), and a self-collected dataset called Heart Rate Database (HR), including different video durations, spectrum bands, resolutions and frame rates.

Several conclusions can be drawn from this work: a) PAD performance based on rPPG varies significantly with the length of the analyzed video, b) rPPG information extracted from short-time sequences (over 5 seconds) can be discriminant enough for performing the PAD task, c) in general, videos using the NIR band perform better than those using the RGB band, and d) the temporal resolution is more valuable for rPPG signal extraction than the spatial resolution.

1. Introduction

Face recognition is one of the most extended biometric traits together with fingerprint and iris. Biometric systems based on human faces have interesting properties that differentiate them from those based on fingerprint and iris, e.g. the possibility to acquire the facial information at a distance, and its non-intrusiveness [28]. At present, face is one of the biometric traits with the highest economical and social impact, being included in high impact products like Face ID technology from Apple[1]. Due to this high level of deployment, attacks against face recognition systems have become a real threat. Regarding that, it is worth noting that the factors that make face an interesting trait for person recognition, i.e. images can be taken at distance in a non-intrusive way, also make it specially vulnerable to attackers who may easily get and use facial biometric information in an illicit manner.

In presentation attacks, an assailant presents to the sensor an artifact for trying to impersonate a genuine user [15]. There are different spoofs and artifacts that a face recognition system may confront, and the same anti-spoofing technique may not be useful against all of them, each situation normally needing a specific countermeasure. Techniques for countermeasuring those attacks are also known as Presentation Attack Detection (PAD) methods [19].

Additionally to the ease of getting information of the real users, face recognition systems are known to respond weakly to presentation attacks for a long time [15, 17], and are easily spoofed, for example using one of these three categories of attacks [11]: i) using a photograph of the user to impersonate [24]; ii) using a video of the user to impersonate (aka replay attack) [7]; and iii) building and using a 3D model of the enrolled persons face, for example an hyperrealistic mask [9].

There are a high number of PAD measures in the literature for trying to deal with those three (and others) types of presentation attacks [19, 11, 23]. For example, texture-based techniques perform the analysis of the facial texture to discover unnatural characteristics that may be related to presentation attacks [7, 12, 1]. This type of approaches may be useful to detect photo-attacks, video-attacks, and also low quality mask-attacks. The major drawback of texture-based presentation attack detection is that high resolution input images are required in order to extract fine details from the faces. These countermeasures will not work properly with bad illumination conditions that make the captured images to have bad quality in general.

[1] https://support.apple.com/en-my/HT208108

Another type of PAD techniques use the depth information for detecting the spoofs. In photo and replay attacks the artifacts are 2D surfaces that can be detected using depth information. This depth information can be captured by specialized sensors, like Microsoft Kinect[2] or Intel RealSense[3]. Recent works have shown that it is even possible to extract depth information from a single RGB image [16]. Nevertheless, this type of countermeasures becomes inefficient when dealing with 3D presentation attack artifacts like realistic masks.

Specifically when dealing with 3D mask attacks, in which the attacker manages somehow to build a highly realistic 3D reconstruction of a genuine face and presents it to the sensor-camera [9, 13], it becomes difficult to find effective countermeasures due to the high realism of the spoofs. As has been said before, the use of depth information becomes inefficient as the artifacts present the same volume of a real face, and the texture information is also useless when dealing with hyper-realistic masks that imitate the real texture of the human skin. Social media and video streaming web sites (e.g. Facebook and YouTube) contain a huge amount of facial recordings of people, making easy to access to the information required to manufacture 3D masks or another face spoofs. Additionally, over the Internet there exist a variety of online services (e.g. "ThatsMyFace"[4]) for ordering a highly realistic 3D mask at an affordable cost (200-300 USD).

Due to the easiness to perform this type of sophisticated attack, efficient countermeasures against 3D mask attacks are nowadays highly relevant. In this scenario, detecting pulse from face videos using remote photoplethysmography (rPPG) has proved to be an effective countermeasure against 3D mask attacks [18]. Even though rPPG-based PAD is quite promising, all the current approaches still have their limitations. They usually consist in taking a complete video sequence for extracting its rPPG signal. In the databases employed in published studies, the length of the video sequences uses to be long enough to give the rPPG system sufficient data for making a robust estimation without analyzing the time interval size. Nevertheless, not in all situations will be possible to have a long video sequence (e.g. 1 minute) for making the analysis. In order to perform a low latency/short-time study of the rPPG signal, where the selected length of the videos is as short as possible, it becomes necessary to study the performance when varying the video length.

On the other hand, even when working with favorable conditions: long enough videos, good illumination, high resolution, perfect face detection and tracking, etc, the pulse detection algorithms must deal with variable scenarios, e.g. an attacker that puts on a mask in the middle of the video, in which case existing approaches may not be able to give a consistent estimation of pulse and/or presentation attack probability. In such cases, a short-time approach is more adequate. Additional problems arise when these algorithms are applied to real scenarios where they may also have to deal with other factors like bad illumination conditions or failures in the face detection module, making their performance to drop significantly. With a short-time analysis of the rPPG signal, the frames without a properly detected face could be discarded without affecting the global performance.

In this paper we: i) present a new dataset of photo attacks with long video sequences, in visible (RGB) and Near InfraRed (NIR); ii) study the performance of video-based pulse detection depending on the length of the video sequences, both in an existing benchmark (3DMAD) and our new dataset; and iii) simulate and test pulse-based PAD in a scenario in which the attacking conditions vary over time.

The rest of this paper is organized as follows: Section 2 introduces Remote Photoplethysmography and summarizes works that use it for pulse detection and/or face PAD. Section 3 describes the proposed system. Section 4 describes the employed databases and the experimental protocol. Section 5 shows the results obtained. Finally, concluding remarks are drawn in Section 6.

2. Remote Photoplethysmography

Plethysmography refers to techniques for measuring the changes in the volume of blood through human vessels. This information can be used to estimate parameters such as heart rate, arterial pressure, blood glucose level, or oxygen saturation levels. The variant called Photoplethysmography (PPG) includes low-cost and noninvasive techniques associated with imagery and the optical properties of the human body [2]. For example, human heart rate can be measured detecting the periodic changes between oxygenated and deoxygenated blood through the veins.

Recently, related studies have proven that it is possible to measure the changes in the amount of oxygenated blood through facial video sequences [22]. These techniques are called Remote Photoplethysmography (rPPG) and their operating principle consists in looking for slight changes in the skin color at video recordings using signal processing methods. When applying this technique to a 3D mask attack or a photo print attack, the estimated pulse signal is highly different from a genuine pulse signal [18].

Table 1 summarizes related works in rPPG, from where we can see that most research in this area use self-collected datasets not publicly available. One of the few public datasets available for 3D mask PAD is 3DMAD, which contains RGB videos of genuine users and of 3D mask attacks.

[2] https://developer.microsoft.com/en-us/windows/kinect

[3] https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

[4] http://thatsmyface.com/

Method	Type of Images	Database used	Video Length	Parameter Estimated	Performance
Garbey et al. 2007 [14]	Thermal	self-collected	120 secs	Heart Rate	Accuracy = 99%
Poh et al. 2011 [22]	RGB	self-collected	60 secs.	Heart Rate	RMSE = 5.63%
Tasli et al. 2014 [26]	RGB	self-collected	90 secs.	Heart Rate	MAE = 4.2%
Chen et al. 2014 [6]	Hyperspectral	self-collected	30-60 secs	Stress Level	Qualitative
McDuff et al. 2014 [20]	Multiband (RGBCO)	self-collected	120 secs	Heart Rate	Correlation = 1.0
Chen et al. 2016 [5]	RGB + NIR	self-collected	90 secs	Heart Rate	RMSE = 1.65%
Li et al. 2016 [18]	RGB	3DMAD and self-collected	10 secs	Face PAD	EER = 4.71%
Present Work	RGB & NIR	3DMAD and self-collected	10 & 60 secs	Face PAD	EER = 25% & 0%

Table 1: Related works that use different types of images to implement rPPG for pulse extraction and related tasks like face Presentation Attack Detection or stress detection. For our system, performances obtained with RGB and NIR videos are shown separately.

We decided to use 3DMAD to compare our results with [18]. We also decided to acquire a supplementary dataset to have larger RGB videos (1 minute) compared to the ones from 3DMAD (only 10 seconds), what will help us with our target of performing a time analysis of the PAD performance. Our dataset also contains information from the NIR spectrum band in order to allow us to compare performances between both bands.

3. Proposed System

In this section we describe the general scheme of the proposed framework. The purpose of the system is the following: given a facial video sequence of a person, our system decides if the video comes from a real face or if, on the contrary, it is really a presentation attack.

As shown in Fig. 1, our system consists in three main stages. The first stage processes a video, extracting temporal windows and measuring raw rPPG signals. The second stage, extracts discriminant pulse-related features from the rPPG signals. The third and last stage performs the matching task comparing models of presentation attacks and real faces with the extracted features. Our approach is largely based in [18]. We chose it as reference because its excellent performance for PAD shown in 3DMAD. There are slight differences between the system from [18] and ours, and between the process followed with the 3DMAD database and with the HR database (our own self-collected database). The differences are described below.

3.1. rPPG Signal Generation

The input of this stage is a temporal window extracted from the original video. The window length T is configurable in order to perform a time dependent analysis. Each video can be processed as a whole or extracting smaller video sequences. The rPPG signal extraction stage is divided into three steps, namely: face detection, ROI selection and tracking, and rPPG signal generation.

3.1.1 Face detection

For each RGB windowed video (or NIR if available) we perform face detection at the first frame using the Matlab implementation of the Viola-Jones algorithm [29]. This algorithm is known to perform reasonably well and in real time when dealing with frontal faces, as in our case. There are three possible outputs from the face detector: 1) One face is detected and the detector returns a bounding box that contains the face location. 2) Multiple faces are detected and the bounding boxes of all them are returned. Sometimes, there are false positive face detections and bounding boxes without a real face inside them are returned. In these cases, we decided to keep only the larger bounding box because in the majority of cases it contains the real face. 3) No face is detected so there is no output. This case is called Failed To Acquire (FTA). The FTA rate of this specific work is 0% for both databases as they contain good quality faces.

3.1.2 Face region selection and tracking

After the recognition stage, we selected a facial Region of Interest. An example of this region (nose and cheeks) can be seen in Fig. 1. Our ROI is different to [18], since they selected a bigger ROI that includes also the mouth and chin. We decided to use a smaller region because it is less affected by objects like hats, glasses, beards or mustaches. The next step consists of detecting corners inside that region for tracking them over time using the Kanade–Lucas–Tomasi algorithm [27], also implemented in Matlab.

3.1.3 rPPG signal extraction

For the ROI of each frame from the video segment, we calculate its raw rPPG value as the average intensity of the pixels inside the region. This sum is made separately for the three available channels of the RGB images (Red, Green and Blue) giving us three different rPPG values for each RGB video segment (one value for NIR recordings). The rPPG signal generation is executed at every frame of the

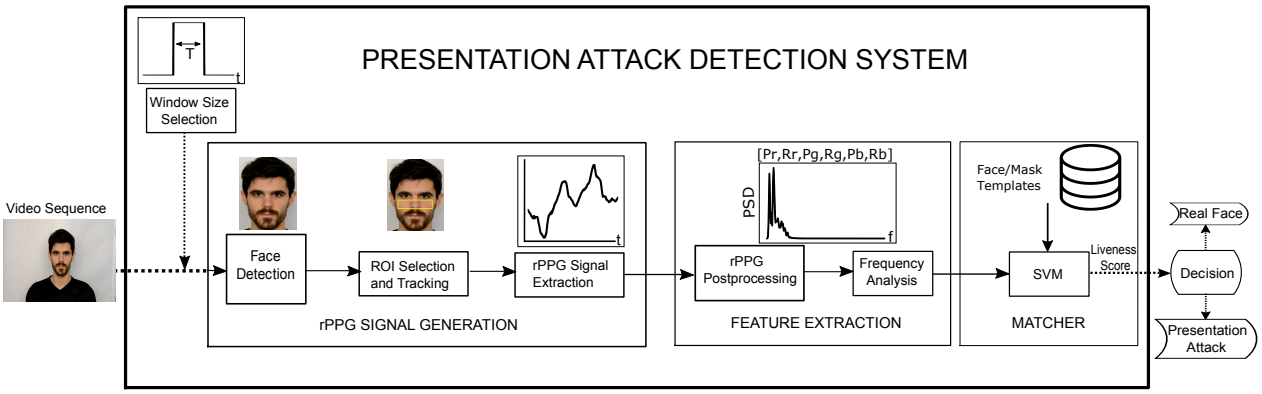


Figure 1: **Architecture of the proposed pulse-based face presentation attack detection**. Given a facial video, the face is detected and rPPG related features are extracted from the ROI in order to obtain an individual score of each video segment. Then, the video segment is classified as an attack or a legitimate access based on its score considering a database of real faces and mask attacks.

video segment, being its final output a temporal signal with the raw rPPG values for the windowed recording.

3.2. Feature Extraction

3.2.1 rPPG signal postprocessing

The raw rPPG signal contains not only the light variations due to the human pulse but also the mean environmental illumination, changes in that environmental levels and noise from other sources. Due to all those factors, a postprocessing stage is necessary.

The postprocessing method consists of three filters:

- **Detrending filter [25]:** this temporal filter is employed for reducing the stationary part of the rPPG signal, i.e. eliminating the contribution from environmental light and reducing the slow changes in the rPPG level that are not part of the expected pulse signal.
- **Moving-average filter:** this filter is designed to eliminate the random noise on the rPPG signal. That noise may be caused by imperfections on the sensor and inaccuracies in the capturing process. This filter consists in a moving average of the rPPG values (size 3).
- **Band-pass filter:** a regular human heart rate uses to be between 40-240 beats per minute (bpm), which corresponds to signals with frequencies between 0.6 and 4 Hz approximately. All the rPPG frequency components outside that range are unlikely to correspond to the real pulse signal so they are discarded.

3.2.2 Frequency Domain Analysis

The input to this stage is a clean rPPG signal that should be a robust estimation of the changes in skin tone due to the blood level evolution. At this point, if we want to extract the corresponding heart rate value for a video sequence, the highest frequency peak inside the normal frequency range should be selected. In our case, we did not want to extract the exact value of the heart rate since our task is to distinguish between real faces and mask attacks. With that target in mind, we have selected discriminant features from the final rPPG signal's spectrum.

The features we have decided to use are also the ones from [18]. We transformed the signal from the spatial domain to the frequency domain using FFT and estimated its power spectral density (PSD) distribution. An example of a PSD can be seen in the feature extraction stage in Fig. 1.

From the PSD we selected the maximum power response as the first feature P, and the ratio of P to the total power in the 0.6 - 4 Hz frequency range as a second feature R. This pair of features $[P, R]$ have been extracted for the three color channels in the case of the RGB videos, and for the unique channel in the NIR case, resulting in feature vectors of size 6 and 2, respectively.

3.3. Match Score Generation

The last block of the presentation attack detection system is the comparator. Like in [18] we use Support Vector Machines (SVM) as our classifier. A Support Vector Machine is a supervised algorithm based in a representation of the examples as points in a multidimensional space. The training process consists in choosing a hyperplane that maximizes the distance from it to the nearest data point of each class. With this type of classifiers, it is interesting to have the input data represented in a high-dimensional space of uncorrelated features, what gives more freedom to find a hyperplane with a minimum distance large enough to obtain high classification rates. We have used the 6-dimensional

(or 2-dimensional for NIR videos) feature vectors for training and testing the SVMs.

4. Databases and Experimental Protocol

4.1. Databases

In our experiments we have used two different databases: the first dataset, the 3D Mask Attack Database (3DMAD) has been used in related works of 3D mask PAD. We decided to use 3DMAD to enable direct comparison with related studies. The second dataset is a self-collected database, called from now on HR (Heart Rate database). The purpose of this dataset is to complement the results obtained with 3DMAD with images of higher resolution, extra spectrum bands and longer duration.

- **The 3D Mask Attack Database (3DMAD)** [8] is a dataset collected and distributed by the Idiap Research Institute. It contains frontal-view controlled recordings of 17 different users acquired using Microsoft Kinect. For each user there are 3 different sessions, with 5 videos for session. Two of the sessions include legitimate user videos with a time delay of 2 weeks between recordings, while the remaining session consists of a 3D mask attack using the mask of each corresponding user.

  The masks used for the attacks were obtained from ThatsMyFace.com[5], an online service that allows their clients to create wearable hard-plastic realistic masks with holes at the eyes and nostrils. To produce each mask only 3 images of each user are needed (1 frontal and 2 lateral). The pictures that were employed to create each mask presented in the database are also included in the 3DMAD corpus.

  The duration of each recording is 10 seconds, captured at 30 frames per second, resulting in 300 frames per video. The Kinect sensor makes possible to capture RGB and Depth images at the same time, so the database contains both information for each recorded frame. It also contains annotated eye positions for each frame of the RGB videos.

  Summarizing, 3DMAD is formed by 17 users $\times$ 3 sessions/user $\times$ 5 videos/session $\times$ 300 frames/video = 76,500 RGB and 76,500 Depth frames, all of them with a resolution of 640$\times$480 pixels. One-third of these frames (25,500) correspond to mask attacks and two-thirds (51,000) to legitimate access attempts.

  In our experiments we have only used the RGB information of the database in order to compare our results with the ones in [18]. The Depth frames are not employed in this study, but they could be processed to perform more robust face detection and tracking.

[5] http://www.thatsmyface.com/

Figure 2: **HR Database**. The figure shows some samples from the HR database. The dataset contains information from both RGB (left) and NIR sensors (right). The samples are splited into legitimate users (up) and photo attacks (down).

- **The Heart Rate Database (HR)** is a dataset self-collected by our research group. It has not been released yet as we are now enlarging it for subsequent research in this area. We will make it public at a later stage when we finish the collection effort. It is a database collected for complementing existing databases like 3DMAD.

  3DMAD is a large dataset but it presents some limitations for performing a study like the one presented in the present paper: i) it only has one type of face spoofing artifacts; ii) the length of the videos is short, only 10 seconds; iii) it does not contain information of extra spectrum bands, only depth information that is useless for rPPG and that has been recorded using only one type of sensor. To overcome these limitations, we captured HR.

  For the preliminary experiments repeated in the present paper, we use frontal-view controlled recordings of 10 different users. We have captured facial RGB videos with a reflex digital camera NIKON D5200 with 1920$\times$1080 resolution. The database also contains facial NIR videos, captured simultaneously to the RGB video using a NIR camera with 1032$\times$770 resolution. See Fig. 2 for an example of the database images.

The RGB and NIR videos are synchronized in order to develop experiments that can take advantages of the multiband information. In this work we compare both types of videos.

The videos have a duration of 60 seconds, being captured at 25 frames per second for the RGB camera, and 30 frames per second for the NIR camera, resulting in 1,800 and 1,500 frames per video, respectively.

HR also contains face presentation attacks, in this case not using 3D realistic faces, but using HQ color printings of the attacked faces (see Fig. 2). This way are able to measure the performance of face PAD based on pulse detection with other type of easy-to-create spoofing artifacts different to 3D masks. For each user we have recorded two sessions: one legitimate access and one photo print face attack.

Summarizing, HR contains recordings of: 10 users $\times$ 2 sessions/user $\times$ 1 videos/session $\times$ 60 seconds/video = 30,000 RGB and 36,000 NIR frames, from which one-half (15,000 and 18,000 frames) correspond to attacks and the other half to legitimate access attempts.

4.2. Experimental Protocol

4.2.1 3DMAD database

From the rPPG signal we extracted the six-dimensional feature vector $(P_r, R_r, P_g, R_g, P_b, R_b)$ for each one of the following temporal window sizes: from 1 to 10 seconds (1 second steps). For the Support Vector Machines we used linear kernels with fixed cost parameter $C = 1000$ similarly to [18].

The whole dataset is divided into legitimate samples as the positive class and attack samples as the negative class. For training and testing the classifier, we use a Leave-One-Out Cross-Validation (LOOCV) protocol: for each one of the 17 users in the 3DMAD database, we use all his feature vectors for testing against a SVM model that has been trained with all the samples from the remaining 16 users.

The metric used to report results is the Equal Error Rate (EER in %). EER refers to the value where the Impostor Attack Presentation Match Rate (IAPMR, percentage of presentation attacks classified as real) and the False Non-Match Rate (FNMR, percentage of real faces classified as fake) are equal[6]. For each window size, the EER has been calculated independently for the 17 subjects (each one of the LOOCV iterations). The 17 individual results are then averaged to produce a single performance (mean and standard deviation).

[6]As error measures we have mentioned IAPMR and FNMR as defined and discussed by Galbally et al. [12]. Modifying the Decision Threshold on the right of Fig. 1 until those error rates are equal we obtain the Presentation Attack Equal Error Rate, PAEER, defined and discussed in [12]. Here we follow [12] using PAEER to evaluate the presentation attacks, but calling it as EER for simplicity.

4.2.2 HR database

For the experiments with HR we used a slightly different experimental protocol in order to compare the information from RGB and NIR videos. We distinguish between: i) using only RGB videos, and ii) using only NIR videos.

For the scenario (i) we followed exactly the same experimental protocol than the one explained for 3DMAD, with the obvious changes in the total number of users and videos and the possible sizes of the temporal windows, since the higher duration of the recordings allowed us to have also into account windows sizes of 20, 40 and 60 seconds.

For the scenario (ii) we also followed the same experimental protocol but this time the feature vector only consists of two features (P, R) as the NIR images only have one color channel.

5. Results

The results obtained on 3DMAD and HR RGB are summarized in Table 2. The EER has been computed for different window sizes. The performance of the presentation attack detection becomes higher when increasing the length of the processed video sequence. For short video sequences (from 1 second to 5 seconds) the system shows almost random behavior, close to 50% EER, improving for longer videos. When dealing with those short recordings, the rPPG sequence does not have enough complete pulse cycles for extracting robustly the features from the signal spectrum. As can be seen in Table 2, Li et al. [18] obtained much lower EER results than ours when working with the 3DMAD database. Compared to their work our approach is much simpler (as can be seen in Section 3.1.2). The facial ROI extracted in our experiments is smaller and our extraction method is less robust to movement and illumination changes, affecting the final results. Despite that fact, our target, which was analyzing to what extent the length of the video affects to the final performance, is achieved.

Comparing the EER results obtained from 3DMAD data with those obtained with HR, there is a gap between performances, specially when using longer videos, achieving always better performance when working with the 3DMAD database. Though our self-collected dataset has been captured with higher resolution than 3DMAD (1900$\times$1080 pixels vs 640$\times$480 pixels), the frame rate is lower in HR (25 fps vs 30 fps). This means that even though the detected ROIs will contain a higher number of pixels, i.e. more spacial information, for each temporal window, the videos will contain less frames, i.e. less temporal information. This result indicates that for extracting a robust rPPG signal from facial videos, the temporal resolution is more relevant than the spatial resolution.

In Table 3 a comparison between the EERs obtained with RGB and NIR videos on the HR database is shown. In this

	Video Length T [s]	1	2	3	4	5	6	7	8	9	10	10 (Li et al. [18])
3DMAD	Mean EER [%]	42.8	45.0	37.8	40.7	33.1	29.7	25	26.1	24.1	**22.1**	**4.71**
	Std EER [%]	5.0	5.9	8.6	9.8	10.8	18.1	14.5	15.2	11.9	10.3	-
HR	Mean EER [%]	46.9	45.7	46.5	42.1	42.1	40.1	**34.1**	36.4	37.3	40.1	-
	Std EER [%]	3.9	5.1	3.9	8.1	9.5	10.2	12.7	11.8	11.7	9.6	-

Table 2: **EER of our implemented face PAD on 3DMAD and HR databases.** The study has been performed changing the length of the video sequences analyzed. The table also compares our results with [18] on 3DMAD. Values in %. Highlighted in bold are the best EER results for each database.

	Video Length T [s]	1	2	5	10	20	30	40	50	60
RGB videos	Mean EER [%]	46.9	45.7	42.1	40.1	40.0	40.0	36.6	30.0	**25.0**
	Std EER [%]	3.9	5.1	9.5	9.6	14.0	21.1	20.5	25.8	26.3
NIR videos	Mean EER [%]	42.4	41.7	38.4	30.9	30.0	16.6	5.0	**0.0**	**0.0**
	Std EER [%]	5.9	6.4	10.8	13.5	18.8	17.5	15.8	0.0	0.0

Table 3: **EER of our implemented face PAD on HR database.** The table shows EER values for video sequences of increasing length T and also results using NIR videos. Values in %. Highlighted in bold are the best EER for each band.

case, longer video sequences (up to 60 seconds) have been analyzed thanks to the higher duration of the recordings. The overall performance obtained with NIR videos is much better than in the RGB case, achieving a 0% EER with 50 seconds of window size.

In the process of extracting the rPPG information from the facial videos, there are several factors that can affect the final performance: head motion, ROI location, light changes, frame rate, resolution or noise sources. The NIR camera employed for acquiring the video sequences have a higher frame rate (30 fps), and it adds less noise to the final images thanks to the higher quality of its sensor. The NIR spectrum band is also more robust to environmental light variations than the RGB bands.

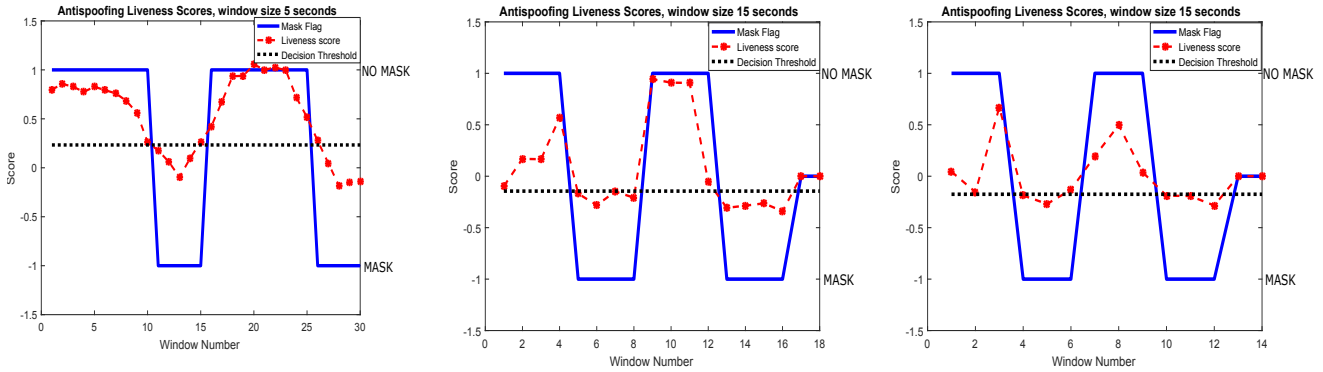
Once again, as in the comparison between the performance obtained with HR and 3DMAD, the NIR videos have lower spatial resolution than the RGB, but higher temporal resolution, reinforcing the hypothesis that the spatial resolution is not as critical as the temporal resolution in order to achieve high performance when working with rPPG. It is also remarkable the fact that the feature vector extracted from NIR videos is only 2-dimensional versus the higher dimensionality (6 features) from the RGB recordings.

Finally, Fig. 3 shows the temporal evolution of the anti-spoofing liveness scores in a variable attack scenario (the higher the scores the lower the estimated probability of presentation attack). In this scenario we wanted to simulate a situation in which an imaginary attacker puts on and removes the mask (or the printed HQ photograph) several times during the same recording. Extracting the features using the full video may not be enough for discerning between an attack and a real access in this situation due to the intra-video variability. Using a short-time/low latency approach, the PAD output will be able to evolve over the video, which justifies the usefulness of a short-time rPPG analysis. It can be seen how the liveness scores decrease when the attacker puts on the mask and viceversa. In Fig. 3a the video windows analyzed are shorter ($T = 5$ seconds) than the length of those in Fig. 3b and Fig. 3c, due to the lower performance obtained with HR compared to 3DMAD (see Table 2). Our future research will be oriented to investigating practical trade-offs towards short-time pulse detection and related liveness scores, and their integration in this kind of time-variant attacking scenarios.

6. Conclusions

We analyzed time effects of Presentation Attack Detection (PAD) against face biometrics based on video pulse detection (remote PhotoPlethysmoGraphy, rPPG). We analyzed the performance of pulse-based face PAD using two different databases, one public (3DMAD) and one self-collected (HR), with various spectrum bands (RGB and NIR), frame rates and resolutions. We also discussed a possible time-variant attack scenario in which the advantages of a short-time rPPG analysis can be exploited.

While time holistic methods may fail to adapt their decisions in variable situations, we advocate for short-time PAD able to deal with quick changes in the attacking scenario. For that, the video sequences must have a minimum length in order to obtain a robust PAD score, and we have provided some evidence on the performance of pulse-based face PAD for small video durations.



(a) Scores from 3DMAD RGB videos. (b) Scores from HR RGB videos. (c) Scores from HR NIR videos.

Figure 3: **Temporal evolution of the scores in a variable attack scenario.** The attacker puts on and removes the mask several times inside the video. Results using data from 3DMAD and HR are shown. In (a) the video sequences analyzed are shorter (5 seconds) than the length in (b) and (c), due to the lower performance obtained on HR compared to 3DMAD.

Our implemented rPPG PAD method works better with higher frame rate recordings in the NIR spectrum band. This is due to the temporal changing nature of the rPPG signal, that favors the temporal resolution against the spatial resolution, and also due to the higher robustness against illumination variations of the NIR spectrum band.

This is the first in-depth research of the temporal dependence of pulse detection for PAD. The proposed short-time analysis has potential to be generalized into many real-world use case scenarios in which a low latency analysis of the video sequence is necessary.

Future work includes: 1) Improving the baseline system for getting lower EER with short videos (e.g. using video magnification techniques [3]). 2) Temporal integration of individual scores for performing continuous PAD [21]. 3) Capturing a larger database with a higher number of users, more variate spoofing artifacts and maybe employing sensors that combine RGB and NIR information (like Intel Real-Sense) for improved PAD based on multiple evidences [4, 10]. And 4) accomplishing a more in depth study of the performance when changing spatial and temporal resolution.

7. Acknowledgments

This work was supported in part by Accenture, project CogniMetrics from MINECO/FEDER under Grant TEC2015-70627-R, project Neurometrics (CEAL-AL/2017-13) from UAM-Banco Santander, and the COST Action CA16101 (Multi-Foresee). The work of J. Hernandez-Ortega was supported by a Ph.D. Scholarship from Universidad Autonoma de Madrid.

References

[1] A. Agarwal, R. Singh, and V. M. Face anti-spoofing using Haralick features. In *Proc. IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016. 1

[2] J. Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007. 2

[3] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh. Computationally Efficient Face Spoofing Detection with Motion Magnification. In *Procs. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVRW*, pages 105–110, 2013. 8

[4] B. Biggio, G. Fumera, G. L. Marcialis, and F. Roli. Statistical Meta-Analysis of Presentation Attacks for Secure Multibiometric Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):561–575, 2017. 8

[5] J. Chen, Z. Chang, Q. Qiu, X. Li, G. Sapiro, A. Bronstein, and M. Pietikäinen. Realsense = real heart rate: Illumination invariant heart rate estimation from videos. In *Proc. Image Processing Theory Tools and Applications (IPTA)*. IEEE Press, 2016. 3

[6] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She. Detection of psychological stress using a hyperspectral imaging technique. *IEEE Transactions on Affective Computing*, 5(4):391–405, 2014. 3

[7] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Proc. Biometrics Special Interest Group (BIOSIG)*. IEEE, 2012. 1

[8] N. Erdogmus and S. Marcel. Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. In *Proc. IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems*, 2013. 5

[9] N. Erdogmus and S. Marcel. Spoofing face recognition with 3D masks. *IEEE Transactions on Information Forensics and Security*, 9(7):1084–1097, 2014. 1, 2

[10] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho. Multiple classifiers in biometrics. part 2: Trends and challenges. *Information Fusion*, 44:103–112, November 2018. 8

[11] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014. 1

[12] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, 2014. 1, 6

[13] J. Galbally and R. Satta. Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models. *IET Biometrics*, 5:83–91, 2016. 2

[14] M. Garbey, N. Sun, A. Merla, and I. Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54(8):1418–1426, 2007. 3

[15] A. Hadid, N. Evans, S. Marcel, and J. Fierrez. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal Processing Magazine*, 32(5):20–30, 2015. 1

[16] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1031–1039, 2017. 2

[17] L. Li, P. L. Correia, and A. Hadid. Face recognition under spoofing attacks: countermeasures and research directions. *IET Biometrics*, 7:3–14(11), January 2018. 1

[18] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *International Conference on Pattern Recognition (ICPR)*, pages 4244–4249. IEEE, 2016. 2, 3, 4, 5, 6, 7

[19] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans. *Handbook of Biometric Anti-Spoofing*. Springer, 2018. 1

[20] D. McDuff, S. Gontarek, and R. W. Picard. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering*, 61(10):2593–2601, 2014. 3

[21] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. Continuous User Authentication on Mobile Devices: Recent progress and remaining challenges. *IEEE Signal Processing Magazine*, 33(4):49–61, 2016. 8

[22] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011. 2, 3

[23] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.*, 50(1), 2017. 1

[24] X. Tan, Y. Li, J. Liu, and L. Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Proc. ECCV*, pages 504–517. Springer, 2010. 1

[25] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen. An advanced detrending method with application to HRV analysis. *IEEE Transactions on Biomedical Engineering*, 2002. 4

[26] H. E. Tasli, A. Gudi, and M. den Uyl. Remote PPG based vital sign measurement using adaptive facial regions. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1410–1414. IEEE, 2014. 3

[27] C. Tomasi and T. Kanade. Detection and tracking of point features. *International Journal of Computer Vision*, 1991. 3

[28] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. Nixon. Soft Biometrics and their Application in Person Recognition at a Distance. *IEEE Transactions on Information Forensics and Security*, 9(3):464–475, 2014. 1

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 3