Remote Photoplethysmography Correspondence Feature for 3D Mask Face Presentation Attack Detection

Si-Qi Liu, Xiangyuan Lan, and Pong C. Yuen(✉)

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong
{siqiliu,lanxiangyuan,pcyuen}@comp.hkbu.edu.hk

Abstract. 3D mask face presentation attack, as a new challenge in face recognition, has been attracting increasing attention. Recently, remote Photoplethysmography (rPPG) is employed as an intrinsic liveness cue which is independent of the mask appearance. Although existing rPPG-based methods achieve promising results on both intra and cross dataset scenarios, they may not be robust enough when rPPG signals are contaminated by noise. In this paper, we propose a new liveness feature, called rPPG correspondence feature (CFrPPG) to precisely identify the heartbeat vestige from the observed noisy rPPG signals. To further overcome the global interferences, we propose a novel learning strategy which incorporates the global noise within the CFrPPG feature. Extensive experiments indicate that the proposed feature not only outperforms the state-of-the-art rPPG based methods on 3D mask attacks but also be able to handle the practical scenarios with dim light and camera motion.

Keywords: Face presentation attack detection · 3D mask attack
Remote photoplethysmography

1 Introduction

Face recognition technique has been widely deployed in a number of application domains, especially the widespread access control of mobile devices and e-commerce. Consequently, security issues of a face recognition system attract increasing attention. Despite its practicability and convenience, face recognition systems are also vulnerable to presentation attacks because one's face can be obtained and abused at very low costs with the booming of social networks. Prints and screen are the two traditional medias to conduct face presentation attacks and great effort has been devoted on detecting them in the last decades [1–16]. A wide variety of liveness cues have been studied and achieved promising results, such as texture [5,9,12,14], image quality [15], reflection patterns [13] and context of presentation attack instrument [16], and motion cues including eyes movement [8], mouth motion [17] and facial expression [9].

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11220, pp. 577–594, 2018.
https://doi.org/10.1007/978-3-030-01270-0_34

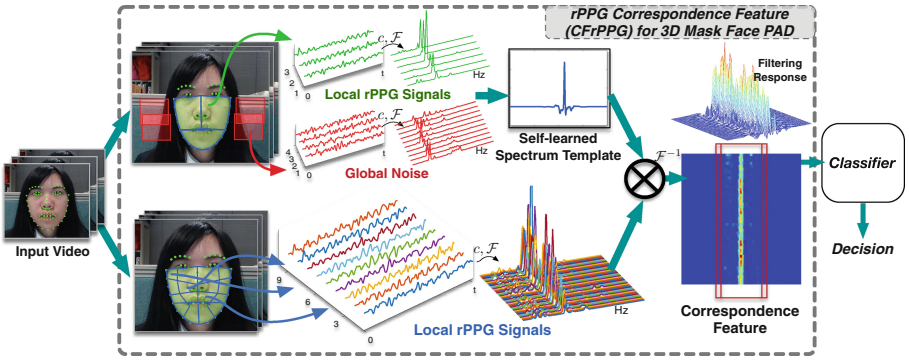


Fig. 1. Block diagram of the proposed CFrPPG feature

Recently, 3D mask attack attracts increasing attention with the rapid development of 3D reconstruction and 3D printing techniques. One can easily customize a 3D mask at an affordable price with a frontal face image[1]. Although the texture based methods can achieve promising results on detecting Thatsmyface mask [18], Liu *et al.* point out the challenges of super-real masks and poor generalization ability under practical cross dataset scenarios [19]. As such, they propose a new liveness cue based on the facial heartbeat signals — remote photoplethysmography (rPPG), which measures the blood pulse flow by modeling the skin color variations caused by the heartbeat. Due to the low transmittance of 3D mask material, such a liveness signal can only be observed on genuine faces but not on masked faces. Since rPPG is not related to the appearance, this approach can detect super-real masks well and achieve encouraging performances under both intra and cross dataset scenarios.

It is intuitive to extract liveness features by analyzing rPPG signals in the frequency domain. Li *et al.* extract the rPPG signal from the center of the face and design a spectrum feature [20]. Liu *et al.* propose the local rPPG solution to obtain spatial structure information from facial rPPG signals. Provided that the background noise is non-periodic and the subject's face does not move much, the cross-correlation operation can amplify the shared heartbeat frequency while suppressing the random interferences [19].

However, existing methods implicitly assume that the maximum value of the signal spectrum can reflect the heartbeat strength. Such an assumption is not always valid in practical scenarios where noise can dominate the observed signal. For instance, when there exists global noise such as camera motion, a mask may be misclassified as a real face since the large periodicity appears on the signal spectrum. The cross-correlation of rPPG signals from local facial regions [19] may not work as well in this case since it not only boosts the pulse signal but also amplifies the shared global noise. Moreover, the rPPG signals on a genuine face can be noisy under dim light or with small facial resolution. A genuine face

[1] www.thatsmyface.com.

may be wrongly rejected when the heartbeat strength is lower than that of the environmental noise.

Therefore, how to precisely identify the heartbeat information from the observed noisy rPPG signals is critical for rPPG-based face presentation attack detection (PAD). To achieve this, we propose a novel rPPG-based 3D mask PAD feature based on the property that the local facial regions share the same heartbeat pattern [21]. For an input video, we first learn the heartbeat as a verification template using the rPPG signal spectrums extracted from local facial regions. Then we use the correspondence between the learned spectrum template and the local rPPG signals as the verification response to construct the novel liveness feature, namely **rPPG C**orrespondence **F**eature (CFrPPG). The proposed CFrPPG can reflect the liveness evidence more precisely since the template estimation summarizes the shared heartbeat component from multiple references. Besides, the correspondence not only contains the amplitude of the signal at heartbeat frequency but also encodes the detailed spectrum information. Since the spectrum template estimation is designed to extract the commonality, the global noise is also maintained in practice. To address this issue, we further take the global interference extracted from the background into account and propose a novel learning strategy to incorporate it into the spectrum template estimation. The block diagram of CFrPPG is illustrated in Fig. 1.

In summary, the main contributions of this paper are: (1) A rPPG correspondence feature (CFrPPG) for 3D mask PAD is proposed to precisely identify the heartbeat vestige from the observed noisy rPPG signals. (2) A novel learning strategy which incorporates the global noise with CFrPPG is proposed to further overcome the global interferences in practical scenarios. To evaluate the discriminability and robustness of the proposed CFrPPG, we conduct extensive experiments on two 3D mask attack datasets and a replay attack dataset with continuous camera motion and different lighting conditions. The results indicate that CFrPPG not only outperforms the state-of-the-art rPPG based methods on 3D mask attacks but also be able to handle the real environment with poor lighting and camera motion.

2 Related Work

Face presentation attack detection (PAD) has been studied for decades and existing methods can be mainly divided into three categories according to the liveness cues employed: appearance-based approach, motion-based approach and rPPG-based approach.

Appearance-based Approach. The appearance-based approach uses the artifacts of the attacking media to detect face presentation attack. Texture-based methods have been used for face anti-spoofing and achieve encouraging results [5,9,14,22]. Maatta *et al.* use multi-scale LBP (MS-LBP) to mine the detailed texture differences. Agarwal *et al.* analyze the input image from different scales using redundant discrete wavelet transform [22]. Although they

perform well on both traditional presentation attack and 3D mask attack detection [18], they expose limited generalization ability under different camera settings or lighting conditions [13,19]. The color texture analysis (CTA) [14] improves the discriminability and generalizability of MS-LBP by employing the characteristic of different color space (HSV and YCbCr), while it may fail on 3D mask attack as the color defects of masks can be different or small [23]. The image quality analysis [15] based approach identifies quality defects of attacking instrument, such as the reflectance pattern [13] and the moiré patterns [24], using different kinds of image quality measurement features. Although better generalizability is validated on traditional presentation attacks, this approach may not work on 3D masks since they do not contain the quality defects like videos or images. Deep features have been adopted in face PAD recently with the booming of deep learning and exhibit promising discriminability [25,26]. However, the over-fitting problem due to the intrinsic data-driven nature remains unsolved. Recently, studies indicate that the mask can be well detected with invisible light,e.g., infrared or thermal cameras [27]. However, it requires additional devices which may not be economical for existing face recognition systems using RBG camera.

Motion-based Approach. Facial motion is effective in detecting photo attack using the patterns like eye-blink [8], mouth movement [9] based on human-computer interaction (HCI), or unconscious subtle facial muscle motion [28]. However, these methods may not work on 3D mask attack since the aforementioned motion can be well preserved on masks that expose eyes and mouth [29]. In addition, the motion patterns of non-rigid 3D genuine faces and 2D planar attacking media are different and can be modeled using optical flow field [30] or the correlation of background region [31]. Similarly, these cues can hardly perform well against 3D mask attacks since 3D masks preserve both the geometric and appearance properties of genuine faces. Moreover, the soft silicone gel mask is able to preserve the subtle movement of the facial skin, which make the motion based approach less reliable.

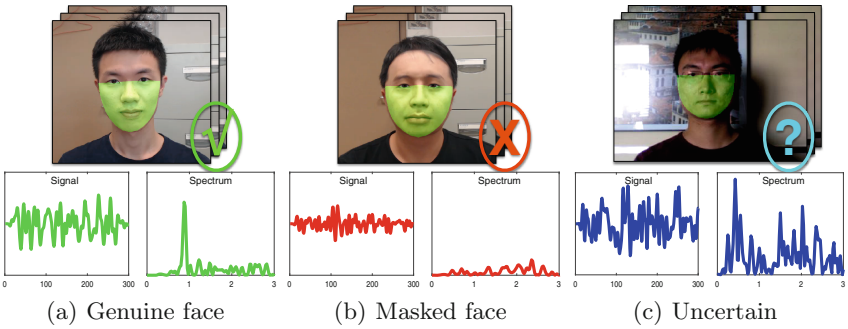
rPPG-based Approach. rPPG is a new research topic in the biomedical community and few methods have been proposed in recent years [32–35]. Because of the non-contact property, rPPG has broad application prospects in clinic, health care and emotion analysis [34]. The use of rPPG for 3D mask face PAD has been explored in previous work [19,20]. Li *et al.* extract the global rPPG signal (green channel) from the center region of a face and quantify it using the maximum value of the spectrum and the signal to noise ratio (SNR) [20]. Since the global signal lacks spatial information, Liu *et al.* propose a local solution [19] with the rPPG signals extracted from local facial regions using CHROM [33]. To suppress the random environmental noise, they apply cross-correlation of each two signals and concatenate the maximum spectrum value as the final feature. Although they achieve encouraging results on 3DMAD [18] and HKBU-MARsV1 [19], the assumption that the maximum value of the signal spectrum can represent the heartbeat may not be valid in real applications. In addition, the cross-correlation will boost the periodic global noises as they also share simi-

lar frequencies on different local facial regions. Ewa *et al.* use background rPPG to overcomes this [36]. However, the direct use of spectrum may not generalize well since the rPPG signal strength varies under different settings.

3 Analysis of rPPG Based Face PAD

This section revisits and analyzes the pros and cons of rPPG-based approach for face presentation attack detection.

The rPPG originates from PPG, a biomedical technique that uses a pulse oximeter to illuminate the skin and measure the changes in light absorption caused by the pumping of blood to the dermis and subcutaneous tissue during cardiac cycles [37]. Different from contact PPG, rPPG measures the heartbeat caused skin color variations remotely through a conventional RGB camera under an environmental light. When applying rPPG on face PAD, 3D masks that cover the live faces block the heartbeat signal so that attacks can be detected by identifying whether the signals can be observed or not (Fig. 2). Following this principle, a rPPG-based solution not only can be effective in 3D mask detection but also works on traditional presentation attacks such as the prints and screen attacks, because these materials block the heartbeat signals in the same way [20].



(a) Genuine face (b) Masked face (c) Uncertain

Fig. 2. Three typical rPPG signal patterns in 3D mask presentation attack detection (PAD). Ideally, the difference of rPPG signals from genuine face and masked face is significant. However, the rPPG signal is fragile to interference in practical scenario

Ideally, the rPPG based solution can achieve high performance under intra and cross dataset scenarios since the observed heartbeat signal is independent of the appearance of the attacking media. Most of existing methods measure the heartbeat strength by directly using the maximum amplitude of the rPPG signal spectrum in frequency domain [19,20]. Although these methods achieve promising results on existing 3D mask attack datasets, we found two critical drawbacks: (1) The assumption that the maximum amplitude can reflect the heartbeat strength may not be valid in real applications. Due to the principle of rPPG is measuring the subtle color variation caused by heartbeat, the rPPG

signal is fragile in practical scenarios. For instance, the heartbeat amplitude can be hardly be observed under poor lighting conditions since the signal strength relies on the amount of light that reaches the blood vessels [19]. When there exist global noise such camera motion, the observed rPPG signals is easy to get contaminated [34] As such, there may be more than one dominant peaks in the rPPG signal spectrum and the one with maximum amplitude may not reflect the heartbeat in some cases (see Fig. 2(c)). In addition, strong peaks caused by noise may also appear on rPPG signals extracted from masked faces and lead to false acceptance error. Although Liu *et al.* use cross-correlation of local rPPG signals to suppress random noise [19], it may still fail when there exists global noise such as handhold caused camera motion since the cross-correlation will not only enhance the heartbeat component but also amplify noises that share similar frequencies. (2) Even when the assumption is valid, the detailed information contained in the distribution of signal spectrum is missing. For instance, on a genuine face, the harmonic peaks of the heartbeat frequency hiding among the noise can be used to boost the discriminability.

4 rPPG Correspondence Feature for 3D Mask PAD

To overcome the limitations of existing rPPG-based 3D mask PAD methods, this paper proposes a novel rPPG correspondence feature (CFrPPG) that can precisely identify the liveness evidence from the observed noisy rPPG signals.

4.1 CFrPPG

Before the identification of liveness information, we first need to figure out what is the real heartbeat component in the observed rPPG signals. Based on the property that the local facial skin shares same heartbeat frequency, we propose to extract the heartbeat by summarizing the commonality of the local rPPG signals. Instead of directly extracting its signal form from the observed rPPG, we propose to model the heartbeat as a template using the correlation filter framework and use it as a detector to identify the liveness component of the local rPPG signals. Specifically, the proposed CFrPPG is constructed by taking the correspondence between the local rPPG signal spectrum and the template learned on themselves.

Learning Spectrum Template. Intuitively, we want to train a template that summarizes the commonality of local rPPG signals which reflects the heartbeat information. As shown in Fig. 1, for an input face video, local rPPG signals are extracted from the local region of interests defined based on facial landmarks. To reduce random noise, we perform cross-correlation of local rPPG signals as preprocessing and obtain their frequency spectra $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N$ (details can be found in Sect. 4.3). Then the spectrum template is learned by solving the following ridge regression problem:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{N} \|\boldsymbol{S}_i \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \tag{1}$$

Note that the learned spectrum template is denoted by the vector $\boldsymbol{w}$. The square matrix $\boldsymbol{S}_i \in \mathbb{R}^{n \times n}$ contains all circulant shifts of the local rPPG signal spectrum $\boldsymbol{s}_i$ and the regression target $\boldsymbol{y}$ is the vector of 1D Gaussian with variance σ.

The objective function in Eq. 1 is strictly convex and has a unique global minimum. By taking its derivative and setting it equal to zero, we can obtain the close form solution for the learned spectrum template.

$$\boldsymbol{w} = (\sum_{i=1}^{N} \boldsymbol{S}_i^{\mathsf{T}} \boldsymbol{S}_i + \lambda \boldsymbol{I})^{-1} \sum_{i=1}^{N} \boldsymbol{S}_i^{\mathsf{T}} \boldsymbol{y} \tag{2}$$

Since $\boldsymbol{S}_i$ is circulant, we have $\boldsymbol{S}_i = \boldsymbol{F} diag(\hat{\boldsymbol{s}}_i) \boldsymbol{F}^{\mathsf{H}}$ and $\boldsymbol{S}_i^{\mathsf{T}} = \boldsymbol{F} diag(\hat{\boldsymbol{s}}_i^{*}) \boldsymbol{F}^{\mathsf{H}}$, where $\boldsymbol{s}^{*}$ is conjugate, $\boldsymbol{F}$ is the DFT matrix, $\hat{\boldsymbol{s}}$ is Discrete Fourier Transform (DFT) $\sqrt{n}\boldsymbol{F}\boldsymbol{s}$ and H is Hermitian transposition. The matrix inversion of Eq. 2 can be solved efficiently in the Fourier domain [38]. The DFT of the spectrum template $\boldsymbol{w}$ in Eq. 2 can be obtained efficiently by the element-wise operation $\odot$ in frequency domain as shown in Eq. 3, and then by taking inverse Fast Fourier Transformation (FFT), the spectrum template w can be obtained.

$$\hat{\boldsymbol{w}} = \frac{\sum_{i=1}^{N} \hat{\boldsymbol{s}}_i^{\,*} \odot \hat{\boldsymbol{y}}}{\sum_{i=1}^{N} \hat{\boldsymbol{s}}_i^{\,*} \odot \hat{\boldsymbol{s}}_i + \lambda} \tag{3}$$

Constructing Correspondence Feature. Given the self-learned spectrum template $\boldsymbol{w}$, the correspondence between local rPPG signals and learned spectrum template can be obtained by convolving $\boldsymbol{w}$ with local rPPG signal $\boldsymbol{s}_i$, i.e.:

$$\hat{\boldsymbol{r}}_i = \hat{\boldsymbol{s}}_i \odot \hat{\boldsymbol{w}} \tag{4}$$

Given the convolution output array, the correspondence can be reflected by the peak value. Since correlation filters are designed to detect the target with the sharp peaks, we use the peak sharpness to measure the correspondence to achieve better discrimination properties. One of the most commonly used peak sharpness metrics is the peak-to-sidelobe ratio (PSR) defined as $\hat{r}_i = (peak_{\hat{\boldsymbol{r}}_i} - \mu_{\hat{\boldsymbol{r}}_i})\sigma_{\hat{\boldsymbol{r}}_i}$ where $peak_{\hat{\boldsymbol{r}}_i}$, $\mu_{\hat{\boldsymbol{r}}_i}$ and $\sigma_{\hat{\boldsymbol{r}}_i}$ is the center value, average and standard deviation of the response, respectively. Finally, we construct the liveness feature as the concatenation of local responses: $\boldsymbol{x} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_N]$.

Comparing with the maximum amplitude of the frequency spectra, the proposed CFrPPG can reflect the liveness sign more accurately since the learned spectrum template summarizes the heartbeat component from local rPPG signals. By taking the correspondence between the learned template and local rPPG themselves (Eq. 4), both the response of heartbeat frequency and the detailed spectrum information are employed in CFrPPG. Besides, our CFrPPG is robust to random noise since the template estimation of the input local rPPG spectrums (Eq. 1) explicitly suppress the diversity that reflects the random noise. Consequently, rPPG signals on a genuine face share the commonality from the heartbeat so that these signals and learned spectrum template could yield strong correspondence. For a masked face, observed signals are less consistent and the

response shall be faint correspondingly. The computation of CFrPPG is fast since the main cost lies on DFT and IDFT. The computational complexity is is $\mathcal{O}(NDlogD)$, where N is the number of local rPPG signals and D is the dimension of each signal spectrum $\boldsymbol{s}_i$.

4.2 Noise-Aware Robust CFrPPG

As mentioned in Sect. 3, global interferences have a big impact on rPPG-based face PAD. For instance, facial expression or motion may contaminate the heartbeat signal of a genuine face and leads to false rejection. Also, the periodic noise such as handhold caused camera motion may be regarded as heartbeat and introduces false acceptance error. Therefore, we take the global noise extracted from background regions into account and incorporate it into the spectrum template learning (see Fig. 1). In addition, since rPPG signal quality varies with different facial regions [21], we use signals extracted from larger reliable regions to learn robust global spectrum template. To maintain sufficient spatial information in the final CFrPPG feature, the rPPG signals for the calculation of correspondence are extracted from finer regions (see Fig. 1).

For an input face video, we extract M and N local rPPG signals and use their spectrum $\boldsymbol{s}_i^t \in \mathbb{R}^n$ and $\boldsymbol{s}_j^l \in \mathbb{R}^n$ to train the global spectrum template and obtain the final liveness feature respectively. K rPPG signal spectrum $\boldsymbol{s}_k^n \in \mathbb{R}^n$ are acquired from the background within similar region size to model the global noise. Detailed region selection strategy can be found in Sect. 4.3. Their corresponding circulant matrix are $\boldsymbol{S}_i^t \in \mathbb{R}^{n\times n}$, $\boldsymbol{S}_j^l \in \mathbb{R}^{n\times n}$ and $\boldsymbol{S}_k^n \in \mathbb{R}^{n\times n}$, respectively. The background noise spectrum can be regarded as the hard negative samples during the template learning. Our objective is to learn a filter $\boldsymbol{w} \in \mathbb{R}^n$ that yields high response for heartbeat signals while nearly zero response for global noise. To achieve this, we formulate the global noise suppression as a regularizer controlled by the parameter γ into Eq. 1:

$$\min_{\boldsymbol{w}} \sum_{i=1}^{M} \|\boldsymbol{S}_i^t \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 + \gamma \sum_{k=1}^{K} \|\boldsymbol{S}_k^n \boldsymbol{w}\|_2^2 \tag{5}$$

It is noted that the summary of K noise signals implicitly picks up the shared global noise and reduce the others so that the learned template will not be suppressed by random noise.

Similarly, since the objective function Eq. 5 is also strictly convex, the closed-form solution can be obtained by setting the gradient to zero:

$$\boldsymbol{w} = (\sum_{i=1}^{M} \boldsymbol{S}_i^{t\intercal} \boldsymbol{S}_i^t + \lambda \boldsymbol{I} + \gamma \sum_{k=1}^{K} \boldsymbol{S}_k^{n\intercal} \boldsymbol{S}_k^n)^{-1} \sum_{i=1}^{M} \boldsymbol{S}_i^{t\intercal} \boldsymbol{y} \tag{6}$$

Then, $\boldsymbol{w}$ can be calculated efficiently in frequency domain through FFT due to the circulant property of $\boldsymbol{S}_i^t$ and $\boldsymbol{S}_k^n$:

$$\hat{\boldsymbol{w}} = \frac{\sum_{i=1}^{M} \hat{\boldsymbol{s}}_i^{t\,*} \odot \hat{\boldsymbol{y}}}{\sum_{i=1}^{M} \hat{\boldsymbol{s}}_i^{t\,*} \odot \hat{\boldsymbol{s}}_i^t + \lambda + \gamma \sum_{k=1}^{K} \hat{\boldsymbol{s}}_k^{n\,*} \odot \hat{\boldsymbol{s}}_k^n} \tag{7}$$

Provided the learned template $\boldsymbol{w}$, we calculate correspondence between local rPPG signals spectrum by $\hat{\boldsymbol{r}}_j = \hat{\boldsymbol{s}}_j^l \odot \hat{\boldsymbol{w}}, j = 1, \ldots, N$. Then we concatenate the PSR as the final liveness feature: $\boldsymbol{x} = [r_1, \hat{r}_2, \ldots, \hat{\boldsymbol{r}}_N]$.

4.3 Implementation Details

rPPG Signals Extraction. Given an input video, we first extract and track 68 points facial landmarks using CLNF proposed in [39] to ensure that each local region can be precisely located. The rPPG signals used for template training and the construction of correspondence feature are different. As shown in the above image in Fig. 1, we extract rPPG signals from larger facial regions to learn a robust rPPG signal spectrum template. As shown in the bottom image in Fig. 1, rPPG signals used in the correspondence feature are extracted from finer overlapped regions to obtain sufficient spatial structural information. Since the proposed feature relies on rPPG signals extracted from small facial regions, we select the CHROM [33] that allows the varying size of the input region as the rPPG sensor. To ease the effect of random noise, we perform the cross-correlation operation used in [19] on the raw rPPG signals for preprocessing.

Global Noise Extraction. Since it has been demonstrated that the global noise from the background and the facial region share similar patterns [36], we model the global noise by extracting rPPG signals using CHROM [33] from background regions. To obtain stable locations under camera motion, facial landmarks are used as the reference to locate the rectangular background regions around the check (see Fig. 1). Empirically, the number and size of these regions are set to be similar to the facial regions used for template estimation as shown in Fig. 1.

5 Experiments

We conduct experiments on the 3D Mask Attack Dataset (3DMAD) [29] and the HKBU Mask Attack with Real World Variations Dataset Version 2 (HKBU-MARsV2) [23], and their combination to evaluate the effectiveness of our proposed CFrPPG feature. To further validate the robustness to global noise, we select the Replay Attack Dataset (RAD) [40] which includes more challenging and practical cases, such as the continuous handhold camera motion and different lighting conditions. The experiment is conducted under intra-dataset and cross-dataset testing protocols. Three appearances-based methods and two rPPG-based methods are selected as the baseline methods.

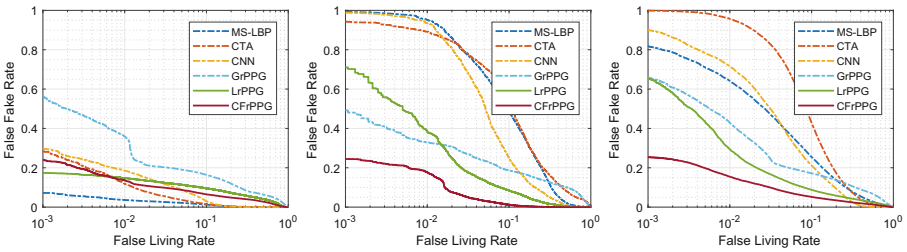
5.1 Baseline Methods and Implementation

Baseline Methods. The MS-LBP is selected as a baseline due to the promising performance reported on 3DMAD [18]. We extracted a set of LBP from a normalized face image to form an 833-dimensional feature vector following settings in [18]. The color texture analysis (CTA) that uses LBP in HSV and

YCbCr color spaces is also compared, following the setting in [41]. Inspired by the success of deep learning, we also add a deep feature extractor (CNN for short), which uses a pre-trained VGGNet [42] to obtain a 4096-dimensional feature vector. For the state-of-the-art rPPG-based methods, the LrPPG [19] and GrPPG [20] are selected for comparison. Since the face PAD can be regarded as a two-class classification problem, SVM with their original kernel settings is used as the classifier for all the baseline methods.

Parameter Settings. As shown in Fig. 1, for all evaluation, we select 3 facial regions and 4 background regions for the spectrum template learning. The correspondence feature is obtained from rPPG signals extracted from 9 overlapped regions in smaller sizes. Each of these regions is the combination of 4 unit regions and they are half overlapped. Details are described in supplementary material. We set the parameter $\{\sigma, \lambda, \gamma\}$ as $\{0.1, 0.5, 0.4\}$,$\{1, 0.5, 0.4\}$ and $\{0.1, 20, 0.1\}$ on 3DMAD, HKBU-MARsV2 and RAD, respectively. SVM with linear kernel is used for classification.

Evaluation Criteria. AUC, EER, Half Total Error Rate (HTER) [18], and False Fake Rate (FFR) when False Liveness Rate (FLR) equals 0.1 and 0.01 are used as the evaluation criteria. For the intra-dataset evaluation, HTER on the development set (HTER_dev) and testing set (HTER_test) is measured, respectively. ROC curves with FFR and FLR are plotted for qualitative comparisons.



(a) 3DMAD (b) HKBU-MARsV2 (c) Combined

Fig. 3. Average ROC curves of three datasets under intra-dataset protocol

5.2 Intra-Dataset Evaluation

The intra-dataset experiments are conducted on 3DMAD, HKBU-MARsV2, and Combined dataset.

3DMAD. The 3DMAD dataset contains 17 subjects with the Custom Wearable Masks made from Thatsmyface.com, which has been proven to be able to spoof popular face recognition system [29]. The dataset is recorded at 640×480, 30fps using Kinect under controlled lighting condition. We follow the leave-one-out-cross-validation (LOOCV) protocol settings in [19] with random subject index on

3DMAD. Specifically, after leaving one subject out as the testing set, 8 subjects are selected as the training set and the rest 8 are used as the development set. Due to the random subject index, we conduct 20 rounds of LOOCV (each contains 17 iterations) and results are summarized in Table 1 and Fig. 3(a).

Table 1. Comparison results under intra dataset protocol on 3DMAD

	HTER_dev(%)	HTER_test(%)	EER(%)	AUC	FFR@ FLR=0.1	FFR@ FLR=0.01
MS-LBP [18]	**1.25 ± 1.9**	**4.22 ± 10.3**	**2.71**	**99.7**	**1.28**	**3.62**
CTA [41]	2.78 ± 3.6	4.40 ± 9.7	4.24	99.3	1.60	11.8
CNN	4.28 ± 3.5	6.07 ± 11.3	6.63	98.9	2.98	18.5
GrPPG [20]	13.5 ± 4.3	13.3 ± 13.3	14.4	92.2	16.4	36.0
LrPPG [19]	9.06 ± 4.4	8.57 ± 13.3	9.64	95.5	9.51	14.8
CFrPPG	5.95 ± 3.3	6.82 ± 12.1	7.44	96.8	6.51	13.6

HKBU-MARsV2. To evaluate the performance under more realistic scenarios, we also carry out the experiment on HKBU-MARsV2 dataset, a subset of the HKBU-MARs [23] dataset that contains 12 subjects with two types of masks: 6 Thatsmyface masks and 6 high-quality masks from REAL-f[2]. This dataset is recorded under room light using a web-camera Logtech C920 at 1280×720, 25fps. We conduct 20 rounds of LOOCV where each iteration contains 5 subjects for training and the rest 6 subjects for developing after leaving 1 testing subject out. The experimental results are summarized in Table 2 and Fig. 3(b).

Table 2. Comparison results under intra dataset protocol on HKBU-MARsV2

	HTER_dev(%)	HTER_test(%)	EER(%)	AUC	FFR@ FLR=0.1	FFR@ FLR=0.01
MS-LBP [18]	20.5 ± 8.9	24.0 ± 25.6	22.5	85.8	48.6	95.1
CTA [41]	22.4 ± 10.4	23.4 ± 20.5	23.0	82.3	53.7	89.2
CNN	13.7 ± 10.8	14.8 ± 22.2	15.2	91.4	25.1	93.5
GrPPG [20]	15.4 ± 6.7	16.1 ± 20.5	16.4	89.4	18.6	32.9
LrPPG [19]	8.43 ± 2.9	8.67 ± 8.8	9.07	97.0	8.51	38.9
CFrPPG	**3.24 ± 1.9**	**4.42 ± 5.1**	**4.04**	**99.3**	**1.24**	**17.8**

Combined Dataset. To further evaluate the performance under various application scenarios, we enlarge the diversity of existing 3D mask attacks dataset by merging the 3DMAD and HKBU-MARsV2 as the Combined dataset. The combined dataset contains 29 subjects, 2 types of masks, 2 camera settings, and 2 lighting conditions. We conduct 20 rounds LOOCV with random subject index on the combined dataset. In each iteration, we randomly select 8 subjects for training and the rest 20 for developing after leaving 1 testing subject out. The experimental results are summarized in Table 3 and Fig. 3(c).

[2] http://real-f.jp.

Table 3. Comparison results under intra dataset protocol on the Combined dataset

	HTER_dev(%)	HTER_test(%)	EER(%)	AUC	FFR@ FLR=0.1	FFR@ FLR=0.01
MS-LBP [18]	15.7 ± 4.2	16.2 ± 22.6	16.6	91.0	25.4	64.2
CTA [41]	18.4 ± 5.8	19.5 ± 21.5	18.9	87.7	42.9	95.7
CNN	13.5 ± 5.9	14.6 ± 20.6	14.5	93.5	21.2	71.5
GrPPG [20]	15.3 ± 2.9	15.5 ± 18.5	15.2	91.1	17.2	42.8
LrPPG [19]	8.69 ± 1.5	9.16 ± 11.9	9.21	95.7	8.79	29.4
CFrPPG	**6.22 ± 1.4**	**6.62 ± 11.0**	**6.54**	**97.6**	**5.18**	**15.5**

It is noted that the proposed CFrPPG feature outperforms the state-of-the-art rPPG based methods on the three mask attack datasets and achieves the best on HKBU-MARsV2 and the Combined. In particular, the CFrPPG outperforms the LrPPG in a larger gap on HKBU-MARsV2 and the Combined dataset. This is because the HKBU-MARsV2 is recorded under uncontrolled room lights (compared with 3DMAD) which leads to noisy rPPG signals. The proposed CFrPPG can extract the heartbeat information more precisely under severe environment so that it can exhibit better robustness compared with existing methods.

On the other hand, the appearance based methods reach the best performances on 3DMAD since the distinguishable quality defects of texture of Thatsmyface masks. However, they can hardly detect the hyper real RAEL-f masks on HKBU-MARsV2 and fail on adapting to the variation of mask types and lightings on the Combined dataset. It is noted that the CNN exceeds MS-LBP on generalizability due to the property of deep features. But it also exposes the weakness of appearance-based approach on HKBU-MARsV2 and Combined dataset that contain more diversity. In contrast, the rPPG signal is independent of the mask appearances so the rPPG-based methods can generalize better in practical scenarios.

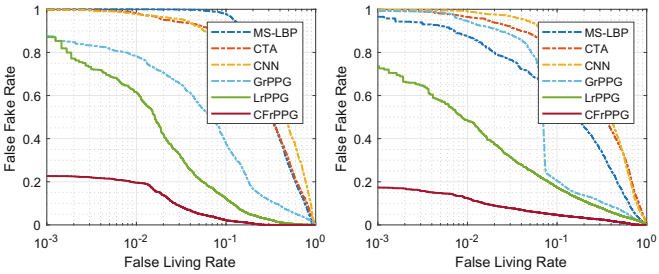
5.3 Cross-Dataset Evaluation

To evaluate the generalization ability across different datasets, we conduct the cross-dataset experiments by training and testing with different datasets. When training on 3DMAD and testing on HKBU-MARsV2, 3DMAD→HKBUMARsV2 for short, we randomly select 8 subjects from 3DMAD for training, use the remaining 9 subjects from 3DMAD for development, and use the entire of HKBU-MARsV2 for testing. For HKBUMARsV2→3DMAD, training on HKBU-MARsV2 and testing on 3DMAD, we randomly select 6 subjects from HKBU-MARsV2 for training, use the remaining 6 subjects from HKBU-MARsV2 for development, and use the entire of 3DMAD for testing. Due to the randomness in subject selection, we also conduct 20 rounds of experiments.

As shown in Table 4 and Fig. 4, the proposed CFrPPG achieves the best among the baseline methods, which demonstrates the better generalizability. Noted that the CFrPPG achieves similar performance and outperforms the GrPPG and LrPPG in a larger gap compared with the results in intra-dataset

Table 4. Cross-dataset evaluation results between 3DMAD and HKBU-MARsV2

	3DMAD→HKBUMARsV2					HKBUMARsV2→3DMAD				
	HTER (%)	EER (%)	AUC (%)	FFR@ FLR=0.1	FFR@ FLR=0.01	HTER (%)	EER (%)	AUC (%)	FFR@ FLR=0.1	FFR@ FLR=0.01
MS-LBP [18]	53.0 ± 3.6	39.8	60.4	97.8	100.0	32.8 ± 11.5	32.5	75.3	58.5	87.8
CTA [41]	40.1 ± 7.8	40.2	62.1	87.1	98.3	47.7 ± 5.4	42.5	60.5	81.2	96.5
CNN	50.0 ± 0.0	47.8	54.6	82.6	97.9	50.0 ± 0.0	44.3	58.6	87.3	99.3
GrPPG [20]	24.3 ± 7.1	18.5	86.7	37.8	78.5	15.7 ± 6.8	15.4	87.2	20.6	94.5
LrPPG [19]	16.8 ± 5.0	10.9	95.6	12.4	61.7	17.4 ± 4.4	14.0	92.3	17.4	48.7
CFrPPG	**2.51 ± 0.1**	**5.08**	**99.0**	**2.19**	**19.6**	**2.55 ± 0.1**	**5.88**	**98.0**	**4.66**	**12.4**



(a) 3DMAD→HKBUMARsV2 (b) HKBUMARsV2→3DMAD

Fig. 4. Average ROC curves under cross-dataset protocol

3D mask detection experiments. This is because CFrPPG can extract heartbeat information more precisely so that the feature distribution from the two datasets align better in the feature spaces than existing methods. It is also noted that the performance of the appearance-based methods drops compared with the intra-dataset testing, which exposes the over-fitting problem due to their data-driven property.

5.4 Evaluation of Robustness to Global Noise in More Practical Scenarios

Existing 3D mask attack datasets are recorded under controlled settings without varying lighting conditions or camera motion. To further validate the robustness of CFrPPG to global noise under more challenging and practical scenarios, we compare rPPG-based methods on the Replay Attack Dataset (RAD) that contains different lighting conditions and continuous camera motion [40]. The RAD contains photo and video attacks from 50 subjects with lower camera resolution (320×240). Although the presentation media is different from 3D mask, the rPPG-based approach works based on the same physical principle [20]. We also do self-comparison by excluding the noise-aware robustness (CFrPPG^{-NAR}), i.e., setting the $\gamma = 0$ (Eq. 5), to validate the effectiveness of the robust noise-aware learning strategy.

Table 5. Comparison of rPPG-based methods under intra-dataset protocol on RAD

	HTER_dev(%)	HTER_test(%)	EER(%)	AUC	FFR@ FLR=0.1	FFR@ FLR=0.01
GrPPG [20]	30.5 ± 3.1	30.3 ± 13.4	31.0	73.9	66.5	97.2
LrPPG [19]	19.3 ± 1.5	19.3 ± 11.2	19.4	88.2	35.4	73.5
CFrPPG^{-NAR}	10.0 ± 1.5	10.2 ± 8.2	10.2	95.4	10.3	29.9
CFrPPG	**6.00 ± 1.4**	**6.11 ± 6.9**	**6.17**	**97.9**	**4.48**	**17.8**

We conduct 20 rounds (each contains 50 iterations) LOOCV on RAD instead of using the fixed testing set partition mentioned in [40]. In each iteration, after

leaving 1 testing subject out, we randomly select 15 subjects for training and the rest 34 for developing. From the experimental results in Table 5 and Fig. 5, it is obvious that the CFrPPG outperforms the others in a larger gap than the results in 3D mask attack datasets. This is because the rPPG signals are more noisy under poor light or with camera motion due to the principle of rPPG. Consequently, the maximum amplitude of the signal spectrum may not reflect the heartbeat information. The proposed CFrPPG solves this limitation with the correspondence between the self-learned template and the local rPPG signals so that CFrPPG^{-NAR} outperforms GrPPG and LrPPG in a large margin (see Fig. 5). CFrPPG achieves better performances than CFrPPG^{-NAR}, which validates the effectiveness of the noise-aware learning strategy.

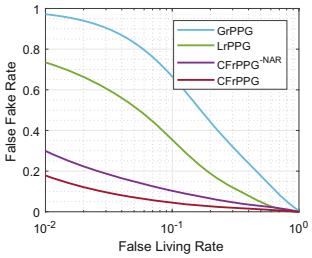


Fig. 5. Average ROC curves of RAD datasets under intra-dataset protocol

6 Conclusion

To precisely identify the heartbeat vestige from the observed noisy rPPG signals, this paper proposes a novel CFrPPG feature which takes the correspondence between the learned spectrum template and the local rPPG signals as the liveness feature. To further overcome the global interferences, a novel learning strategy which incorporates the global noise in the template estimation is proposed. We show that the proposed feature not only outperforms the state-of-the-art rPPG based methods but also be able to handle more practical and challenging scenarios with poor lighting and continues camera motion. In addition, the results of CFrPPG on RAD indicate its potential on handling general face PAD.

Acknowledgement. This project is partially supported by Hong Kong RGC General Research Fund HKBU 12201215.

References

1. Rattani, A., Poh, N., Ross, A.: Analysis of user-specific score characteristics for spoof biometric attacks. In: CVPRW (2012)
2. Evans, N.W., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification. In: Interspeech, pp. 925–929 (2013)
3. Pavlidis, I., Symosek, P.: The imaging issue in an automatic face/disguise detection system. In: Computer Vision Beyond the Visible Spectrum: Methods and Applications (2000)
4. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 504–517. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_37
5. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: IJCB (2011)
6. Anjos, A., Marcel, S.: Counter-measures to photo attacks in face recognition: a public database and a baseline. In: IJCB (2011)
7. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: ICB (2012)
8. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: ICCV (2007)
9. de Freitas Pereira, T., Komulainen, J., Anjos, A., De Martino, J.M., Hadid, A., Pietikäinen, M., Marcel, S.: Face liveness detection using dynamic texture. EURASIP J. Image Video Process. **2014**(1), 1–15 (2014)
10. Kose, N., Dugelay, J.L.: Mask spoofing in face recognition and countermeasures. Image Vis. Comput. **32**(10), 779–789 (2014)
11. Yi, D., Lei, Z., Zhang, Z., Li, S.Z.: Face anti-spoofing: multi-spectral approach. In: Marcel, S., Nixon, M.S., Li, S.Z. (eds.) Handbook of Biometric Anti-Spoofing. ACVPR, pp. 83–102. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6524-8_5
12. Kose, N., Dugelay, J.L.: Shape and texture based countermeasure to protect face recognition systems against mask attacks. In: CVPRW (2013)
13. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Trans. Inf. Forensics Secur. **10**(4), 746–761 (2015)
14. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Trans. Inf. Forensics Secur. **11**(8), 1818–1830 (2016)
15. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: application to Iris, fingerprint, and face recognition. IEEE Trans. Image Process. **23**(2), 710–724 (2014)
16. Komulainen, J., Hadid, A., Pietikainen, M.: Context based face anti-spoofing. In: BTAS (2013)
17. Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-time face detection and motion analysis with application in liveness assessment. IEEE Trans. Inf. Forensics Secur. **2**(3), 548–558 (2007)
18. Erdogmus, N., Marcel, S.: Spoofing face recognition with 3D masks. IEEE Trans. Inf. Forensics Secur. **9**(7), 1084–1097 (2014)
19. Liu, S., Yuen, P.C., Zhang, S., Zhao, G.: 3D mask face anti-spoofing with remote photoplethysmography. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 85–100. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_6

20. Li, X., Komulainen, J., Zhao, G., Yuen, P.C., Pietikäinen, M.: Generalized face anti-spoofing by detecting pulse from face videos. In: ICPR (2016)
21. Lempe, G., Zaunseder, S., Wirthgen, T., Zipser, S., Malberg, H.: Roi selection for remote photoplethysmography. In: Bildverarbeitung für die Medizin 2013. Springer, pp. 99–103 (2013)
22. Agarwal, A., Singh, R., Vatsa, M.: Face anti-spoofing using haralick features. In: BTAS (2016)
23. Liu, S., Yang, B., Yuen, P.C., Zhao, G.: A 3D mask face anti-spoofing database with real world variations. In: CVPRW (2016)
24. Patel, K., Han, H., Jain, A.K., Ott, G.: Live face video vs. spoof face video: use of moiré patterns to detect replay video attacks. In: ICB (2015)
25. Menotti, D., Chiachia, G., Pinto, A., Schwartz, W.R., Pedrini, H., Falcão, A.X., Rocha, A.: Deep representations for Iris, face, and fingerprint spoofing detection. IEEE Trans. Inf. Forensics Secur. **10**(4), 864–879 (2015)
26. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
27. Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M., Noore, A.: Face presentation attack with latex masks in multispectral videos. In: CVPRW (2017)
28. Bharadwaj, S., Dhamecha, T.I., Vatsa, M., Singh, R.: Computationally efficient face spoofing detection with motion magnification. In: CVPR (2013)
29. Erdogmus, N., Marcel, S.: Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect. In: BTAS (2013)
30. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: IASP (2009)
31. Yan, J., Zhang, Z., Lei, Z., Yi, D., Li, S.Z.: Face liveness detection by exploring multiple scenic clues. In: ICARCV (2012)
32. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express **18**(10), 10762–10774 (2010)
33. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
34. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: CVPR (2014)
35. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: CVPR (2016)
36. Nowara, E.M., Sabharwal, A., Veeraraghavan, A.: Ppgsecure: biometric presentation attack detection using photopletysmograms. In: FG (2017)
37. Shelley, K., Shelley, S.: Pulse oximeter waveform: photoelectric plethysmography. In: Lake, C., Hines, R., Blitt, C. (eds.) Clinical Monitoring, pp. 420–428. WB Saunders Company (2001)
38. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with Kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_50
39. Baltrusaitis, T., Robinson, P., Morency, L.P.: Constrained local neural fields for robust facial landmark detection in the wild. In: ICCVW (2013)
40. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: BIOSIG (2012)

41. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: ICIP (2015)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)