Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN

Bofan Lin
CMVS, University of Oulu
Oulu, Finland
Bofan.lin@student. oulu.fi

Xiaobai Li
CMVS, University of Oulu
Oulu, Finland
Xiaobai.li@.oulu.fi

Zitong Yu
CMVS, University of Oulu
Oulu, Finland
Zitong.yu@oulu.fi

Guoying Zhao*
CMVS, University of Oulu
Oulu, Finland
Guoying.zhao@oulu.fi

ABSTRACT

Face anti-spoofing plays a vital role in security systems including face payment systems and face recognition systems. Previous studies showed that live faces and presentation attacks have significant differences in both remote photoplethysmography (rPPG) and texture information, we propose a generalized method exploiting both rPPG and texture features for face anti-spoofing task. First, multi-scale long-term statistical spectral (MS-LTSS) features with variant granularities are designed for representation of rPPG information. Second, a contextual patch-based convolutional neural network (CP-CNN) is used for extracting global-local and multi-level deep texture features simultaneously. Finally, weight summation strategy is employed for decision level fusion, which helps to generalize the method for not only print attack and replay attack but also mask attack. Comprehensive experiments were conducted on five databases, namely 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD, and OULU-NPU, to show the superior results of the proposed method compared with state-of-the-art methods.

CCS Concepts

• **Security and Privacy→Authentication→Biometrics.**

Keywords

Face anti-spoofing; rPPG; mask; Contextual Patch-Based CNN.

1. INTRODUCTION

Security systems have always been a significant research area. Biometric systems are widely used in our daily lives. Fingerprint, voiceprint, iris, and face are most commonly used biometric modalities. As one of the most popular modalities, the face is widely used in designing artificial intelligence security systems, e.g., face recognition systems [1][2], access authorization systems, and payment authorization systems. At the same time, many presentation attacks have been developed for spoofing the face security system. The presentation attacks can be divided into multiple types, print attack, video replay attack, mask attack, and so on.

Print attack: an attacker presents a printed photo or image of the legitimate user on a mobile phone to the face authentication system. However, the 2D structure lacks of the depth information and pulse information, and the consistent face image limited by face life signs, such as blinking, head movement, and facial expression. Most of face authentication systems require users to present several specific facial expressions or head motions. So, the print attack is the weakest attack among all kinds of attacks.

Video replay attack: an attacker presents video sequences containing the legitimate user's face on displays. The video replay attack is more difficult than print attack since video could contain the movements that the system requires and the pulse information. However, the reflection of the screen makes the texture of the input video weird, hence it can be discriminated easily.

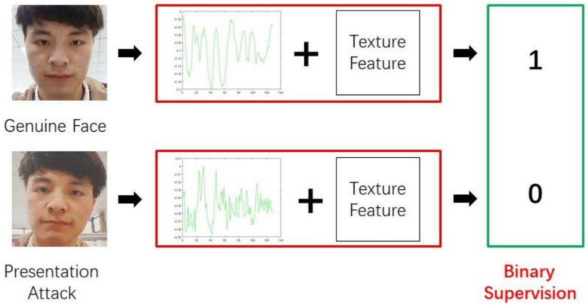

Figure 1. Our method uses both rPPG feature and texture feature for prediction.

Mask attack: an attacker wears a 3D face mask to cheat the system. Mask attack is the most difficult attack to be detected because the high-quality mask is quite similar to the real face. It also contains the depth information which is widely used in face anti-spoofing. However, it is pricey to be implemented and lack of pulse information.

During the past decade, many researchers made efforts for face anti-spoofing area [3][4]. Most of the researches were based on 2D sensors. However, the limitation of 2D sensors is that a single 2D sensor cannot record 3D structure information. Most recent research works also focused on estimating depth map from 2D images. Even though the 3D structure information contributes to face anti-spoofing, the disadvantage is also obvious. When high-quality mask attack presents, the 3D structure information could be

ICBEA 2019, May 29–31, 2019, Stockholm, Sweden

ACM ISBN 978-1-4503-6305-1/19/05…$15.00
https://doi.org/10.1145/3345336.3345345

the same as genuine faces, which makes the system based on solely depth information vulnerable.

To solve this security problem in face anti-spoofing, we propose a model that employs both remote photoplethysmography (rPPG) feature and texture feature. As shown in Figure 1, rPPG feature is discriminant for photo attack and mask attack, while texture feature is effective for detecting replay attacks. Our contributions include: 1) we design multi-scale long-term statistical spectral (MS-LTSS) features with variant granularities for representation of rPPG information; 2) we propose a contextual patch-based convolutional neural network (CP-CNN), which is able to extract global-local and multi-level texture features simultaneously; 3) we fuse this two information in decision level, which allows the method to be able to detect all three common attack types, i.e., photo attack, replay attack and mask attack.

The paper is organized as follows: Section 2 introduces the related work of face anti-spoofing. Section 3 describes the details of the proposed method. Then in section 4, we show the experiment results on five anti-spoofing databases, namely 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD, and OULU-NPU. Finally, we summarize the work and list future work.

2. RELATED WORKS

In this section, we review previous works related to face liveness detection. Previous methods can be divided into four groups as follows.

Texture-based methods for face anti-spoofing. Using texture information was a common method to deal with face anti-spoofing task because most face recognition systems use one single 2D sensor camera. Many research works proposed various hand-crafted features, for example, Local Binary Patterns (LBP) [7], Histogram of Gradient (HoG) [8], Difference of Gaussian (DoG) [9], Scale-invariant Feature Transform (SIFT) [10] and Speeded Up Robust Features (SURF) [11]. The traditional classifiers such as support vector machine (SVM), Random Forest, and Latent Dirichlet Allocation (LDA) were utilized in those works. They tried to transform input data into different domains, e.g., to transform input images from RGB color space into HSV or YCbCr color spaces [12], or from time domain to frequency domain, to obtain more robust results. However, traditional methods cannot perform perfectly, because the extracted features can be affected by various conditions, such as camera qualities, illumination conditions, and presentation attack instruments.

In recent years, deep learning methods have shown their power in many research areas, especially in computer vision tasks. Several works used CNN-based features for face anti-spoofing [13][14]. Li et al. [14] used CNN as the feature extractor and fine-tuned a model which was pretrained on ImageNet. Feng et al. [13] fed various drafts of the samples of face images into CNN and obtained the result of classification, i.e., live vs. spoof. In more recent research works, deep learning methods achieved perfect performance on anti-spoofing databases and in competitions. For old databases, such as NUAA [16], Replay-Attack [17], which were collected several years ago, the video resolution and quality are very poor. Deep learning methods can achieve 100% accuracy, which is far more robust than traditional methods. On the other side, newer databases such as OULU-NPU and SiW include higher-quality videos recorded from a large number of subjects in different conditions, which might be more challenging. We still need to explore new deep learning methods on those databases. Compare to other computer vision tasks, e.g., object detection, object identification, and facial expression classification, solving the task of face anti-spoofing using deep learning methods still have a long way to run. One goal of this paper is to construct a novel CNN model, which is able to extract more sophisticated and reliable texture features for face liveness detection.

Temporal-based methods for face anti-spoofing. Temporal-based methods can be further divided into two categories. The first category of methods is based on facial motion patterns, such as eye-blinking and mouth or lip movements. Prior works reported that the frequency of spontaneous blinking of normal people is about 0.25 to 0.5 blinks per second. Hence, Sun et al. [18] proposed a blinking-based face anti-spoofing method, which utilized Conditional Random Fields (CRFs) to analysis eye blinking actions (eye closed and opening state). The main idea of the proposed CRFs method is to represent actions by face images. In addition, Sun compared the performance of CRFs with Adaptive Boosting (AdaBoost) and Hidden Markov Model (HMM), and CRFs achieved outstanding results. Pan et al. [19] reported a method which combines eyeblinks and scene context for face recognition to imitate the contextual relationships of eyeblinks among eye image sequences by using undirected conditional graphical structure. The other category of methods relied on the movement between face and background. Kollreider et al. [20] proposed to use optical flow methods to track the movement of face to differentiate real vs. fake faces. Besides, Haralick features [34] and motion mag were also used for face anti-spoofing task. For deep learning, Feng et al. [13] proposed to extract features from optical flow map and Shearlet image using CNN. Xu et al. [15] also proposed an LTSM-CNN model using multiple frames of a single video to obtain fused results of classification.

However, the motion clues that these temporal based methods rely on can be easily manipulated. For example, an attacker can hold a print face photo with cropped eye holes (e.g., cropped print attack in CASIA). On the other side, there are other forms of temporal clues for face liveness, which are much harder (or impossible) to fake, that is the temporal color fluctuation of live facial skin caused by pulsation (a.k.a. the rPPG), which is invisible to human eyes. Thus, another goal of the current work is to extract and utilize the rPPG information for face liveness detection.

3D structure-based methods for face anti-spoofing. Most of existing anti-spoofing databases have 2D videos, which do not contain 3D structure information. 3D structure-based face anti-spoofing methods can be roughly divided into two categories. One category of methods extract depth information from 2D images. Liu et al. [21] proposed a CNN model which uses the depth map with auxiliary supervision instead of binary supervision. They estimated the 3D structure of the face by implementing the most recent dense face alignment (DeFA) methods.

The other category of methods analyze 3D shape information recorded with 3D sensors such as a dot projector. They compared the 3D model of the input sample with that of a genuine face to obtain the result of live vs. spoof. However, this method requires special 3D devices which might be costly and not commonly accessible.

However, methods mentioned in this session may become vulnerable when an attacker wears a high-quality 3D mask. Because the mask has the same 3D structure information as the genuine face it can easily cheat these methods. Hence, we may need multiple methods to work together in order to detect various types of attacks.

rPPG-based methods for face anti-spoofing. Remote Photoplethysmography (rPPG) is a method to extract pulse signal from facial videos without contacting any skin. Li et al. [5] proposed the first pulse-based face anti-spoofing method with a simple six-dimensional feature. In order to achieve stronger representation ability, Heusch et al. [6] designed a long-term statistical spectral (LTSS) approach for face liveness detection. Liu et al. [22] also proposed to use rPPG signals to discriminate 3D mask attack. Pulse signals can be extracted from live faces, while there is only random noise if we try the same way of extraction from 3D masks. They calculated the correlation features to classify the face video as live vs. spoof faces. In another related work, Nowara et al. [23] analyzed five different rPPG signals extracted from three face regions and two background regions to classify print and video attacks. Even though in video attacks, pulse signals still exist, the analysis of specific regions can discriminate live vs. spoof. In Liu et al. [21], they proposed to use rPPG to supervise the CNN-RNN model, and they trained the rPPG model by using the estimated rPPG signals for live faces and flattened signals for all kinds of attacks.

All the works mentioned above show the effectiveness of rPPG-based methods for face anti-spoofing detection. In this work, we propose a novel multi-temporal-resolution rPPG feature as one part of the whole model for solving the face anti-spoofing task.

3. THE PROPOSED APPROACH

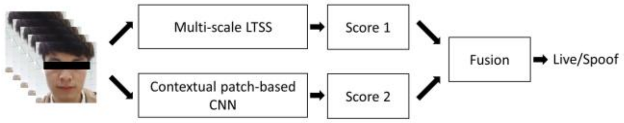

Figure 2. The overview of the proposed method.

In this section, firstly, we propose Multi-scale long-term spectral statistics method for rPPG feature representation, which improves the original LTSS [6] method in multi-scale level. Secondly, we will describe the Contextual patch-based CNN. The overview of

the proposed method is shown in

.

3.1 Multi-scale LTSS

First of all, we use Li's method to extract rPPG signal from input face sequence as in paper [5]. Compared to the original LTSS extracting the spectral features only on constant temporal dimension, the Multi-scale LTSS (MS-LTSS) combines the spectral statistics of sliding windows with different length and different overlapping size. As a result, we expect our proposed MS-LTSS can extract more elaborate rPPG information due to the pyramid-like multi-scale segmentation. The Framework of the proposed MS-LTSS is shown in

Figure 3.

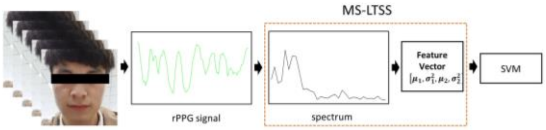

Figure 3. The Framework of Multi-scale LTSS.

For each sliding window w, we convert the extracted rPPG signal from time domain into frequency domain by using an N-point discrete Fourier transform (DFT). Then we receive a sequence X_w of dimension $k = 0 \dots N/2 - 1$ which contains DFT coefficients. We consider log-magnitude of the frequency bins of the spectrum for statistics. The DFT coefficient $|X_w(k)|$ is set to 1 if it is lower than 1, so that the log-magnitude is always positive. The mean and variance statistics of the coefficient vectors $(X_1, X_2, \dots, X_w)$ are computed as following:

$$\mu(k) = \frac{1}{W}\sum_{i=1}^{W} log|X_i(k)|\,. \tag{1}$$

$$\sigma^2(k) = \frac{1}{W}\sum_{i=1}^{W} log|X_i(k) - \mu(k)|\,. \tag{2}$$

The first and second order for vectors (for $k = 0 \dots N/2 - 1$) are concatenated as the feature of the signal. Then we introduced the MS-LTSS, we concatenated the LTSS features that calculated by different settings of length of sliding windows w and the overlapping size of the sliding window o.

$$F = [\mu_1, \sigma_1^2, \dots, \mu_n, \sigma_n^2]. \tag{3}$$

Where, F is the final MS-LTSS feature of the given signal, μ_1, σ_1^2 is calculated by w_1, o_1, similarly, μ_n, σ_n^2 is calculated by w_n, o_n. Then we used the SVM as the classifier to classify, whether it is a genuine presentation or an attack.

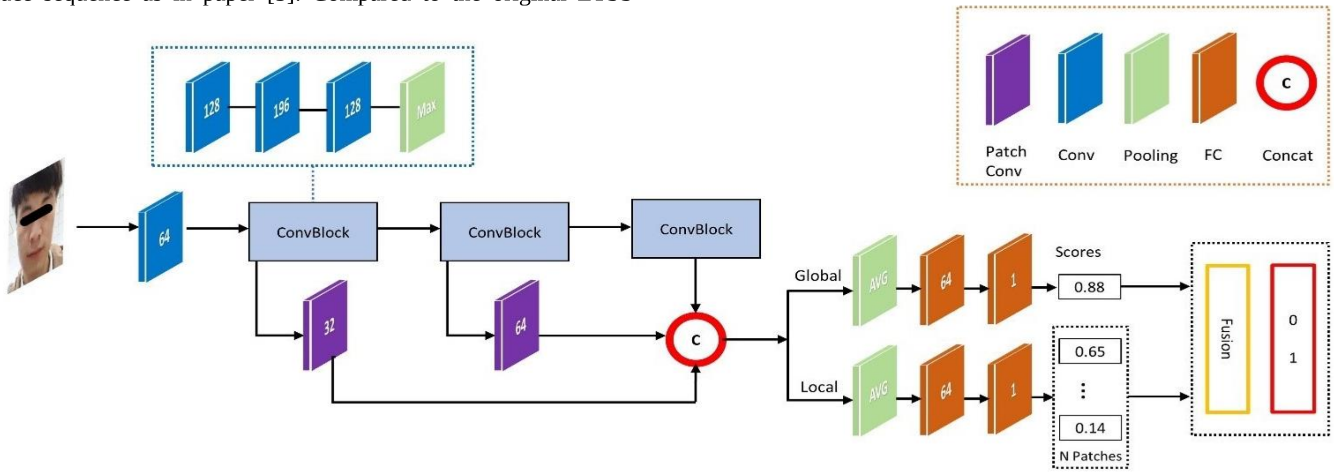

Figure 4. The pipeline of the proposed architecture. The number of filters is shown in the middle of each layer. *Color code* used*: purple=Patch convolution layer, blue=convolution layer, green=pooling layer, orange=fully connect lay*

3.2 Contextual Patch-based CNN

Previous patch-based CNN methods [24] divide the original RGB face image into several patches and extract the local texture feature from the patches directly, which ignores the mutual information interaction between global and local features. In order to better represent both global and local texture features, a contextual patch-based CNN (CP-CNN) is proposed and its architecture is shown in Figure 4.

3.2.1 Network Architecture

As illustrated in Figure 4, the network backbone contains one convolution layer with a 5x5 kernel and three ConvBlocks, which intends to obtain global texture features. The ConvBlock consists of three convolutional layers with a 3x3 kernel and one maximum pooling layer. Every convolutional layer is followed by a batch normalization layer and ReLU layer.

Patch-based Convolution Module. The key component in CP-CNN is the patch-based convolution module, which extracted semantic patch feature in deep feature level instead of RGB level. As shown in Figure 5, the deep features are divided into spatial uniform N patches, then each patch features are convoluted with an independent 3x3 filter. It is mentioned that the parameters of the 3x3 filter for each patch are not shared, which helps to learn the discriminant features for each local patch position. There are two patch-based convolution modules with convolutional two stride and four stride embedded after the first and second ConvBlock respectively, which outputs the patch features with the same spatial dimension for further fusion. In our experiment, local patch number is set as$\mathrm{N} = 4, 9,$ and 16.

Contextual Fusion. After obtaining the patch features and global feature, fusion is needed for feature integration. The patch features are merged in spatial dimension and reshape to the same spatial size as the global feature. Then all global and patch features are channel-wise concatenated, as illustrated in Figure 4. Hence, the fused multi-level features are with rich contextual global-local information and strong representation ability.

Global and Local Classifier. With the deep contextual features, a global classifier and N local classifiers are designed for confidence prediction of face liveness. As for the global classifier, global average pooling is used and then cascaded with two fully connected layers and a sigmoid function. Similarly, local classifiers use the same operations for corresponding patch positions.

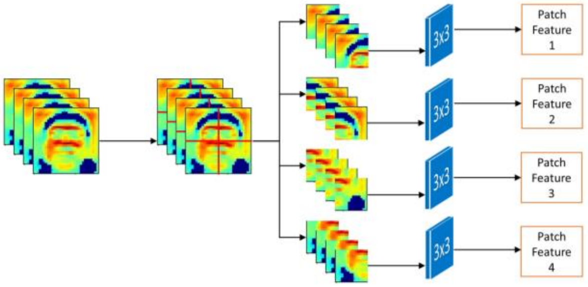

Figure 5. Illustration of Patch-based Convolution Module.

Loss Function and Network Inference. In the training stage, it can be regarded as a binary classification task. The network input is the RGB face image $I^{128\times128\times3}$, and the output is the predicted global score S_g and patch based local scores $S_i(i = 1 \dots N)$ and N is the number of patches. If the input face image is a genuine face, we set the binary label 1 while the label is set to 0 when there are attacks. We adopt binary cross entropy as the loss function, so the overall loss can be formulated as

$$Loss = wL_g + \frac{(1-w)\sum_{i=1}^{N} L_i}{N}, \tag{4}$$

where w is a hyper-parameter to tradeoff the global loss L_g and all the patches losses $L_i(i = 1 \dots N)$.

In the inference stage, we use the weighted scores among all the global score S_G and local scores S_L with the same hyper-parameter w. It can be formulated as

$$Score = wS_G + \frac{(1-w)\sum_{i=1}^{N} S_{Li}}{N}. \tag{5}$$

Then we can obtain the fused score $Score \in [0,1]$ from the single frame. The final result of the input video is the average predicted score across all video frames, as

$$S_F = \frac{\sum_{i=1}^{N_f} Score_i}{N_f}, \tag{6}$$

where N_f is the number of input video frames. $S_F \in [0,1]$ is the final result of the input video.

3.2.2 Multi-modality Fusion

In order to fuse these two modalities features, i.e., rPPG features and texture features, we employ the weight summation strategy in the decision level. to fusion the scores output from the Multi-scale LTSS model and Contextual Patch-based CNN model. So, the fusion method can be formulated as

$$S = w_f S_{MS-LTSS} + (1-w_f) S_{CP-CNN}, \tag{7}$$

where w_f is the tradeoff weight, and $S_{MS-LTSS}$ is the predicted score of MS-LTSS while S_{CP-CNN} is the predicted score from CP-CNN.

4. EXPERIMENT RESULTS

4.1 Experimental Setup

We evaluated our method on five databases, 3DMAD [25], HKBU-Mars V1 [22], MSU-MFSD [26], CASIA-FASD [27], OULU-NPU [28], to evaluate its performance under three different types of attacks, e.g., print, video replay, and 3D mask attacks. The performance on each database is compared with the results of state-of-the-art methods.

4.1.1 Databases

3D Mask Attack Database (3DMAD): It contains 255 video clips recorded from 17 subjects, including 3D mask attack. The FPS of videos is 30 and the resolution is 640×480. The length of each video is 10 seconds.

HKBU 3D Mask Attack with Real World Variations V1 Database (HKBU-MARs V1): This database contains 120 videos, 80 genuine videos recorded from 8 subjects, and the other 40 videos are two different kinds of masks. The FPS of each video is 30 and the resolution is 1280×720. The length of each video is 10 seconds.

The MSU Mobile Face Spoofing Database (MSU-MFSD): This database contains 280 video clips recorded from 35 subjects, including print attack, and video replay attack. The FPS of videos is 30 and the resolution is 640×480 or 720×480 because of different recording devices. The length of each video is 10 seconds.

The CASIA Face Anti-spoofing Database (CASIA-FASD): It contains 600 video clips recorded from 50 subjects, including warped photo attack, cut photo attack, and video replay attack. The FPS of videos is 25. This database has high-quality and low-quality videos, and the resolution is 1280×720 or 640×480. The length of each video is approximately 5 seconds.

OULU-NPU: It contains 4950 video clips recorded from 55 subjects, including print attack and video replay attack. The FPS of videos is 30 and the resolution is 1080×1920. The length of each video is 5 seconds. This database provides four protocols for testing. The protocol I is designed to evaluate the performance of methods under unseen environmental conditions, specifically illumination and background scene. The protocol II is designed to evaluate the generalization of the methods under different types of printers or displays. The protocol III is a Leave One Camera Out (LOCO) protocol to evaluate the performance of sensor interoperability. The protocol IV is the most challenging protocol, it contains all above three conditions together to evaluate the performance of methods.

4.1.2 Hyperparameter setting

The proposed Multi-scale LTSS method is implemented in MATLAB. The two pairs of the settings of the length of sliding windows w and the overlapping size of the sliding window o are [64, 16] and [128, 64] respectively. For the video which is shorter than 256 frames, we added several beginning frames to the end of the frame sequences to supplement the video to 256 frames.

The proposed Contextual Patch-Based CNN method is implemented in PyTorch v1.0.1 with the learning rate of 1e-4, and the training phase is 30 epochs. The batch size of the Contextual patch-based CNN stream is 16. We set the tradeoff parameter w and w_f to 0.5 and 0.4 respectively. We trained and tested the model on Nvidia K80.

4.1.3 Evaluation metrics

In order to fairly compare the performance, we follow previous studies on each of the databases and use the same evaluation metrics. The metrics are from the standardized ISO/IEC 30107-3 metrics.

In OULU-NPU, we utilized 1) Attack Presentation Classification Error Rate (APCER), which evaluates the highest error rate from all presentation attack instruments (PAI), e.g., print or display, 2) Bona Fide Presentation Classification Error Rate (BPCER), which calculates the error rate of real accesses, and 3) ACER, the average of APCER and BPCER:

$$APCER = \frac{\sum_{i=1}^{N_{PAI}}(1 - Res_i)}{N_{PAI}}, \qquad (8)$$

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}, \qquad (9)$$

$$ACER = \frac{APCER + BPCER}{2}, \qquad (10)$$

where N_{PAI} is the total number of attack presentations for the given PAI, N_{BF} is the number of bona fide presentations. Res_i sets as 1 when the ith presentation is classified as an attack presentation. However, set Res_i as 0 if classified as bona fide presentation.

For evaluations on 3DMAD and HKBU-Mars V1, we adopted HTER (Half total error rate) and EER (equal error rates). TPR and EER were adopted when evaluated MSU-MFSD and CASIA-FASD. HTER is the mean of False Negative Rate (FNR) and False Positive Rate (FPR). FNR and FPR are commonly used in presentation attack detection (PAD).

$$TPR = \frac{\sum_{i=1}^{N_G}(1 - Res_i)}{N_G}. \qquad (11)$$

$$FNR = \frac{\sum_{i=1}^{N_A}(1 - Res_i)}{N_A}. \qquad (12)$$

$$FPR = \frac{\sum_{i=1}^{N_G} Res_i}{N_G}. \qquad (13)$$

$$HTER = \frac{FNR(\tau^*) + FPR(\tau^*)}{2}. \qquad (14)$$

N_A is the total number of attack presentations s, and N_G is the number of genuine presentations. Same as the APCER and BPCER, Res_i is 1 when the ith presentation is classified as an attack presentation, and Res_i is 0 if it is classified as a genuine presentation. The threshold τ^* corresponds to the EER when testing on the development set.

4.2 Experimental Comparison

4.2.1 Comparison with State-of-the-art Methods

We measured performance on 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD, and OULU-NPU databases. All the CP-CNN is the proposed 16 patch-based CNN. For 3DMAD and HKBU-Mars V1 databases, we report their EER and HTER in Table 1 and Table 2. The EER, and TPR (when FNR = 0.1) of MSU-MFSD and CASIA-FASD are reported in Table 3 and Table 4 respectively. For OULU-NPU, we report ACER, APCER, and BPCER of four protocols in Table 5.

Table 1. Results on 3DMAD

Method	**3DMAD**	
	EER	HTER
LBP [35]	1.40%	-
LPQ [35]	4.70%	-
Li et al. [5]	4.71 %	7.94 %
MS -LTSS	3.52 %	6.86 %
CP-CNN	0.00 %	0.00 %
MS -LTSS + CP-CNN	**0.00 %**	**0.00 %**

Results on 3DMAD database. We compare with rPPG and texture-based method and achieved the perfect result on 3DMAD. With only the CP-CNN method, we could classify all the videos correctly. The proposed Multi-scale LTSS method also outperformed previous results. Our CP-CNN model could extract the more useful texture features of 3D masks, and the 16 patch-based methods solved the problem of insufficient samples for deep learning.

Table 2. Results on HKBU-Mars V1

Method	**HKBU-Mars V1**	
	EER	HTER
Liu et al. [22]	14.7 %	22.6%
MS -LTSS	0.95 %	3.11 %
CP-CNN	0.00 %	0.00 %
MS -LTSS + CP-CNN	**0.00 %**	**0.00 %**

Results on HKBU-Mars V1 database. We compare the proposed MS-LTSS and CP-CNN methods with two previous results, and the results showed superior performance of our methods on HKBU-

Mars V1. With only the CP-CNN method we could achieve perfect classification with zero error. The proposed Multi-scale LTSS method made a few errors but still outperformed the baseline results with a significant range. The HKBU-Mars V1 data contains both low quality and high-quality 3D mask attacks which makes it more challenging than 3DMAD data. Our results indicate that the proposed methods (both CP-CNN and MS-LTSS) have good generalization ability over mask types. We used LOOCV protocol, which means that even if an attacker uses a high-quality mask which is not seen in the training, our proposed methods are still able to reliably detect the attack.

Table 3. Results on MSU-MFSD

Method	**MSU-MFSD**	
	EER	TPR (FNR=0.1)
LBP Baseline	14.7 %	69.9 %
DoG-LBP Baseline	23.1 %	62.8 %
IDA [26]	8.6 %	92.8 %
MS-LTSS	3.4 %	96.5 %
CP-CNN	1.1 %	99.1 %
MS -LTSS + CP-CNN	**0.00 %**	**100.00 %**

Results on MSU-MFSD database. We compared with two baselines and a texture-based method. With only MS-LTSS or CP-CNN, we are able to achieve better performance than previous results, and when these two are fused, we achieved the perfect result. The MSU-MFSD contains video replay attack and print attack which evaluates the robustness of method over different kinds of attacks. The result indicates that our proposed method has good generalization ability over different kinds of attacks.

Table 4. The results on CASIA-FASD

Method	**CASIA-FASD**	
	EER	TPR (FNR=0.1)
LBP+LDA [29]	21.0 %	75.7 %
IDA [26]	12.9 %	86.7 %
CDD [30]	11.8 %	88.8 %
Dynamic [31]	10.0 %	89.1 %
Patch-based CNN [24]	4.44%	-
SPMT + SSD [32]	**0.04 %**	**100.0 %**
MS -LTSS	8.3 %	92.6 %
CP-CNN	3.2 %	96.6 %
MS -LTSS + CP-CNN	1.8 %	98.7 %

Results on CASIA-FASD database. We compared with five state-of-the-art methods and our method achieved the second-best performance which is quite close to the best one. Our proposed CP-CNN achieved better performance than the prior patch-based CNN [24].

Results on OULU-NPU database. We reported the results of four protocols of OULU-NPU and compared with several state-of-the-art methods. In Protocol one, we achieved the best result of APCER. However, the protocol 4 is the most challenging one, we achieved poor result.

Table 5. The results on OULU-NPU

Prot.	**Method**	**OULU-NPU**		
		APCER (%)	BPCER (%)	ACER (%)
1	CPqD	2.9	10.8	6.9
	GRADIANT	1.3	12.5	6.9
	FAS-BAS [21]	1.6	1.6	1.6
	Wang et al. [33]	2.5	**0.0**	**1.3**
	MS -LTSS	3.0	18.6	10.8
	CP-CNN	2.1	8.3	5.2
	MS -LTSS + CP-CNN	**1.2**	7.6	4.4
2	MixedFASNet	9.7	2.5	6.1
	FAS-BAS	2.7	2.7	2.7
	GRADIANT	3.1	**1.9**	2.5
	Wang et al.	**1.7**	2.0	**1.9**
	MS -LTSS	1.8	15.3	9.5
	CP-CNN	6.5	2.2	4.3
	MS -LTSS + CP-CNN	4.7	2.5	3.6
3	MixedFASNet	5.3 ± 6.7	7.8 ± 5.5	6.5 ± 4.6
	Wang et al.	5.9 ± 1.0	5.9 ± 1.0	5.9 ± 1.0
	GRADIANT	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	MS -LTSS	5.9	16.5	11.2
	CP-CNN	2.5 ± 1.7	5.0 ± 3.3	3.7 ± 2.5
	MS -LTSS + CP-CNN	**1.9 ± 1.0**	4.4 ± 2.6	3.1 ± 1.8
	FAS-BAS	2.7 ± 1.3	**3.1 ± 1.7**	**2.9 ± 1.5**
4	Massy HNU	35.8±35.3	8.3±4.1	22.1±17.6
	GRADIANT	**5.0±4.5**	15.0±7.1	10.0±5.0
	FAS-BAS	9.3±5.6	10.4±6.0	9.5±6.0
	Wang et al.	14.0±3.4	**4.1±3.4**	**9.2±3.4**
	MS -LTSS	19.8±10.5	24.6±17.8	21.6±6.8
	CP-CNN	18.7±16.2	23.5±23.5	20.5±12.5
	MS -LTSS + CP-CNN	16.3±11.2	18.1±17.5	15.1±7.5

4.2.2 Ablation Study

Table 6. The results of CP-CNN with the different number of patches, on OULU-NPU Protocol 2

Method	**OULU P2**		
	APCER (%)	BPCER (%)	ACER (%)
Global	15.50	3.67	9.58
Global + 4P	13.76	2.63	8.19
Global + 9P	13.19	3.33	8.26
Global + 16P	**6.53**	**2.22**	**4.38**

The performance with different patches. We compare four architectures to illustrate the advantages of the proposed contextual patch-based CNN. We train models on OULU-NPU based on Protocol 2, and the result of each model is shown in Table 6. The first model without patches shows poor performance due to lack of local features. In comparison, by using the local features together with the global features, we achieve better performance in the second model (Global+4P). Moreover, with an increased number of patches, the 16 patches model (Global+16P) achieves even better results. Hence, we can see the advantage of combining global and local features using patch-based CNN.

Table 7. The results of different settings of LTSS and Multi-scale LTSS on OULU-NUP Protocol 1

Method	OULU P1		
	APCER (%)	BPCER (%)	ACER (%)
[64, 16]	7.2	17.4	12.3
[128, 64]	5.0	18.2	11.6
[64, 16] + [128, 64]	**1.8**	**15.3**	**9.5**

The effectiveness of Multi-scale LTSS. In order to show the advantage of Multi-scale LTSS for extracting the rPPG features, we test two different settings of LTSS ([64,16] and [128,64]) and compare with the combined MS-LTSS (([64,16] + [128,64])) on OULU-NPU based on Protocol 1, and the results are shown in Table 7. It can be seen that the combined MS-LTSS achieved significantly better performance than the other two, especially for metrics of APCER and ACER.

The performance of sequences length. In Table 8, we report results of different length of sequence for the proposed method on OULU-NPU based on Protocol 1. Results show that by increasing the length of input sequences, we can reduce the APCER and ACER. A possible reason could be that with the longer sequence we can achieve more accurate estimated rPPG signal and more samples for the proposed CP-CNN model.

Table 8. The results of the Multi-scale LTSS with different length of sequences on OULU-NUP Protocol 1

Method	OULU P1		
	APCER (%)	BPCER (%)	ACER (%)
64 Frames	13.1	18.1	15.6
128 Frames	12.5	15.3	13.4
256 Frames	**1.8**	**15.3**	**9.5**

4.2.3 Visualization and Analysis

In this proposed method, the Multi-scale LTSS is analyzed by the rPPG signal which is important for the whole model, Figure 6 shows examples of succeeded and failed samples in rPPG signals. The estimated rPPG signal of the failed sample contains periodic peaks like the real ECG signal. However, it also has too much noise like small and random peaks. This kind of signals is hard to be discriminated as live vs. spoof.

.

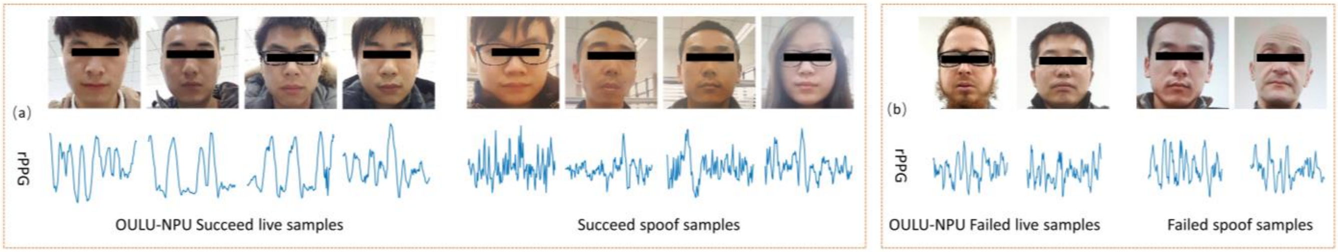

Figure 6. (a) 8 succeeded examples and their estimated rPPG signals. The first four are live and the other four are spoof. (b) 4 failed examples. The first two are live and the other two are spoof.

5. CONCLUSION

In this work, we propose a method of combining rPPG based Multi-scale LTSS features and a contextual patch-based convolutional neural network (CP-CNN) for face anti-spoofing task. The method was evaluated and achieved robust performance on five databases including various evaluations. According to the analysis of results, the rPPG based face anti-spoofing method is extremely effective for mask and print attack detection. The Contextual Patch-based CNN achieved great performance for all kinds of attacks. This combined model also improves efficiency and accuracy. In the future, we will continue focusing on exploring the hidden information of the rPPG signal which can be used for face anti-spoofing detection.

6. ACKNOWLEDGMENT

The support of Business Finland project (Grant No. 3116/31/2017), Tekes Fidipro program (Grant No. 1849/31/2015), Academy of Finland funded ICT 2023 project (313600) and Infotech Oulu is gratefully acknowledged.

7. REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition*.

[2] L. Tran, X. Yin, and X. Liu. 2017. Disentangled representation learning GAN for pose-invariant face recognition. In *Conference on Computer Vision and Pattern Recognition*.

[3] J. Maatta, A. Hadid, M. Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *IET Biometrics 1* (1) (2012) 3–10.

[4] J. Li, Y. Wang, T. Tan, A.K. Jain. 2004. Live face detection based on the analysis of fourier spectra. In *Defense and Security, International Society for Optics and Photonics*, 2004, pp. 296–303.

[5] X. Li, J. Komulainen, G. Zhao. Generalized face anti-spoofing by detecting pulse from face videos. *23rd Int. Conf. Pattern Recognition (ICPR)*, 2016.

[6] G. Heusch, S. Marcel. 2018. Pulse-based Features for Face Presentation Attack Detection. *Biometrics: Theory, Applications and Systems*.

[7] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. 2012. LBP-TOP based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pp. 121–132.

[8] J. Yang, Z. Lei, S. Liao, and S. Z. Li. 2013. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics,* 2013.

[9] B. Peixoto, C. Michelassi, A. Rocha. 2011. Face liveness detection under bad illumination conditions. In *2011 18th IEEE International Conference on Image Processing (ICIP), IEEE,* 2011, pp. 3557–3560.

[10] K. Patel, H. Han, and A. K. Jain. 2016. Secure face unlock: Spoof detection on smartphones. IEEE Trans. *Inf. Forens. Security,* 11(10):2268–2283.

[11] Z. Boulkenafet, J. Komulainen, and A. Hadid. 2017. Face anti-spoofing using speeded-up robust features and Fisher vector encoding. *IEEE Signal Process. Letters,* 24(2):141–145.

[12] Z. Boulkenafet, J. Komulainen, and A. Hadid. 2016. Face spoofing detection using colour texture analysis. *IEEE Trans. Inf. Forens. Security,* 11(8):1818–1830.

[13] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. 2016. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *J. Visual Communication and Image Representation,* 38:451– 460.

[14] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. 2016. An original face anti-spoofing approach using partial convolutional neural network. In *International Pediatric Transplant Association*.

[15] Z. Xu, S. Li, and W. Deng. 2015. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pp. 141–145. IEEE.

[16] X. Tan, Y. Li, J. Liu, and L. Jiang. 2010. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pp. 504–517.

[17] I. Chingovska, A. Anjos, and S. Marcel. 2012. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference of Biometrics Special Interest Group*.

[18] L. Sun, G. Pan, Z. Wu, S. Lao. 2007. Blinking-based live face detection using conditional random fields. In *Advances in Biometrics, Springer* , pp. 252–260.

[19] G. Pan, L. Sun, Z. Wu, Y. Wang. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommun. Syst.* 47 (3–4) (2011) 215–225.

[20] K. Kollreider, H. Fronthaler, J. Bigun. 2005. Evaluating liveness by face images and the structure tensor. In *2005. Fourth IEEE Workshop on Automatic Identification Advanced Technologies, IEEE* , pp. 75–80.

[21] Y. Liu, A. Jourabloo, and X. Liu. 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Conference on Computer Vision and Pattern Recognition,* pp. 389–398.

[22] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao. 2016. 3D mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision,* pp. 85–100.

[23] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan. 2017. Ppgsecure: Biometric presentation attack detection using photopletysmograms. In *International Conference on Automatic Face & Gesture Recognition*, pp. 56–62.

[24] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. 2017. Face anti-spoofing using patch and depth-based cnns. *International Journal of Central Banking*, pp. 319–328. IEEE.

[25] N. Erdogmus, S. Marcel. Spoofing face recognition with 3d masks. 2014. *Information Forensics and Security,* IEEE Transactions on, vol. 9, no. 7, pp. 1084–1097.

[26] D. Wen, H. Han, and A. Jain. 2015. Face spoof detection with image distortion analysis. *Information Forensics and Security*, IEEE Transactions on, vol. 10, no. 4, pp. 746–761, April.

[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. 2012. A face antispoofing database with diverse attacks. In *International Conference on Biometrics*, pp. 26–31.

[28] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. 2017. Hadid.Oulu-npu: A mobile face presentation attack database with real-world variations. In *International Conference on Automatic Face & Gesture Recognition*, pp. 612–618.

[29] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. 2013. Can face anti-spoofing countermeasures work in a real world scenario?. In *International Conference on Biometrics*.

[30] J. Yang, Z. Lei, S. Liao, and S. Z. Li. 2013. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics*.

[31] T. de Freitas Pereira, J. Komulainen, A. Anjos, J.M. De Martino, A. Hadid, M. Pietikäinen, S. Marcel. 2014. Face liveness detection using dynamic texture. *EURASIP J. Image Video Process.* 2014 (1) (2014) 2.

[32] X. Song, X. Zhao, L. Fang, T. Lin. 2017. Discriminative representation combinations for accurate face spoofing detection. *IEEE International Conference on Image Processing (ICIP)*.

[33] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, Z. Lei. 2018. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv preprint,* arXiv: 1811.05118.

[34] A. Agarwal, R. Singh, and M. Vatsa. 2016Face anti-spoofing using Haralick features. In *Biometrics: Theory, Applications and Systems*.

[35] Gragnaniello, Diego, et al. An investigation of local descriptors for biometric spoofing detection. *IEEE transactions on information forensics and security* 10.4 (2015): 849-863.