FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

<Bofan Lin>

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

Master's Thesis
Degree Programme in Computer Science and Engineering
<04 2019>

Lin B. (2019) Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN. University of Oulu, Degree Programme in Computer Science and Engineering. Master's Thesis, 50 p.

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, we design multi-scale long-term statistical spectral (MS-LTSS) features with variant granularities for the 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 of the two types of features, which allow the proposed system to be generalized for detecting 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.

Keywords: Face anti-spoofing, rPPG, mask, Contextual Patch-Based CNN

Lin B. (2019) Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN. Oulun yliopisto, tietotekniikan tutkinto-ohjelma. Diplomityö, 50 s.

TIIVISTELMÄ

Kasvojen anti-spoofingilla on keskeinen rooli turvajärjestelmissä, mukaan lukien kasvojen maksujärjestelmät ja kasvojentunnistusjärjestelmät. Aiemmat tutkimukset osoittivat, että elävillä kasvoilla ja esityshyökkäyksillä on merkittäviä eroja sekä etävalopölymografiassa (rPPG) että tekstuuri-informaatiossa, ehdotamme yleistettyä menetelmää, jossa hyödynnetään sekä rPPG: tä että tekstuuriominaisuuksia kasvojen anti-spoofing -tehtävässä. Ensinnäkin rPPG-informaation esittämiseksi on suunniteltu monivaiheisia pitkän aikavälin tilastollisia spektrisiä (MS-LTSS) ominaisuuksia, joissa on muunneltavissa olevat granulariteetit. Toiseksi, kontekstuaalista patch-pohjaista konvoluutioverkkoa (CP-CNN) käytetään globaalin paikallisen ja monitasoisen syvään tekstuuriominaisuuksiin samanaikaisesti. Lopuksi, painoarvostusstrategiaa käytetään päätöksentekotason fuusioon, joka auttaa yleistämään menetelmää paitsi hyökkäys- ja toisto-iskuille, mutta myös peittää hyökkäyksen. Kattavat kokeet suoritettiin viidellä tietokannalla, nimittäin 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD ja OULU-NPU, ehdotetun menetelmän parempien tulosten osoittamiseksi verrattuna uusimpiin menetelmiin.

Avainsanat: Kasvojen anti-spoofing, rPPG, naamio, kontekstuaalinen korjaustiedostoon perustuva CNN

TABLE OF CONTENTS

FOREWORD

The topic was suggested by Professor Guoying Zhao. The purpose of this thesis is to combine the rPPG-based method from a previous work and the proposed Contextual Patch-based CNN method to implement a robust face anti-spoofing system. The research and experiments started in September 2018 and the writing of the thesis from the end of February 2019. Before the writing, experiments were almost done, when writing the thesis, I was still trying to improve models. I will introduce all the methods that I tried and finally implemented for the thesis.

Here, I would like to thank Professor Guoying Zhao, who is the supervisor of my master study. She has given great help for my study not only providing academic guidance but also with the financial support of scholarship funding. Next, I would also like to show gratitude to Dr. Xiaobai Li, she gave me instructions on how my thesis work should be implemented and checked my progress regularly. Then I would like to thank Mr. Zitong Yu, who gave me the detailed instructions of my experiments and taught me many state-of-the-art methods that are helpful for improving the experimental results.

Secondly, I would like to give special gratitude to my friend and my colleague Mr. Haoyu Chen. He helped me a lot and gave me many advices for both study and life during my two-year master study in Oulu, Finland.

Finally, I would like to express my sincere gratitude to my parents and my girlfriend. Their encouragements and supports are the best force that help me to finish the master study.

Oulu, 1.3.2019

Bofan Lin

ABBREVIATIONS

ACER	Average Classification Error Rate
AdaBoost	Adaptive Boosting
APCER	Attack Presentation Classification Error Rate
BPCER	Bona Fide Presentation Classification Error Rate
bpm	Beat-per-minute
CE	Cross Entropy
CNN	Convolutional Neural Networks
CovBlock	Convolutional Block
CP-CNN	Contextual Patch-based Convolutional Neural Network
CRFs	Conditional Random Fields
DeFA	Dense Face Alignment
DFT	Discrete Fourier Transform
DoG	Difference of Gaussian
DRLSE	Distance Regularized Level Set Evolution
DRMF	Discriminative Response Map Fitting
ECG	Electrocardiography
FC	Fully Connected
FFT	Fast Fourier Transform
FL	Focal Loss
FNR	False Negative Rate
FPR	False Positive Rate
FPS	Frames per second
FRR	False Rejection Rate

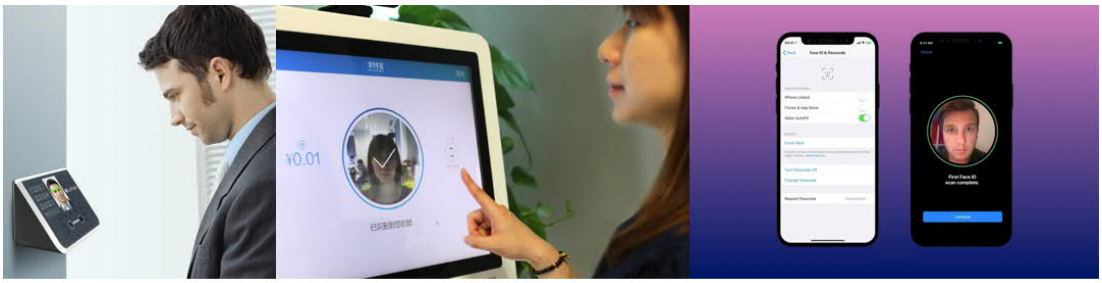
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
HoG	Histogram of Oriented Gradient
HTER	Half Total Error Rate
HRV	Heart Rate Variability
KLT	Kanade-Lucas-Tomasi
LBP	Local Binary Pattern
LBP-TOP	Local Binary Patterns from Three Orthogonal Planes
LDA	Latent Dirichlet Allocation
LOCO	Leave One Camera Out
LOOCV	Leave-one-out Cross-validation
LSTM	Long Short-Term Memory
LTSS	Long-term Spectral Statistics
MS-LTSS	Multi-scale Long-term Spectral Statistics
PA	Presentation Attack
PAD	Presentation Attack Detection
PAI	Presentation Attack Instruments
PSD	Power Spectral Density
RGB	Red-Green-Blue
RNN	Recurrent Neural Network
ROI	Region of Interest
rPPG	remote Photoplethysmography
SIFT	Scale-invariant Feature Transform
SSR	Spatial Subspace Rotation

SVM	Support Vector Machine
SURF	Speeded Up Robust Features
TPR	True Positive Rate

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 the 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 [17] [18], access authorization systems, and payment authorization systems. Several face-based security systems are shown in Figure 1. 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. Samples of the three most common types of attack, e.g., print attack, video reply attack and mask attack, are shown in Figure 3.

(1) Print attack: an attacker presents a printed photo or image of the legitimate user to the face authentication system. However, the 2D structure lacks of the depth information, pulse information, and life signs in face, such as blinking, head movement, and facial expressions. Most of face authentication systems require users to present facial expressions or head motions. So, the print attack is the weakest attack among all kinds of attacks.
(2) 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 wired, hence it can be discriminated easily.
(3) 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 get realistic masks and mask lacks of pulse information.



(a) (b) (c)

Figure 1. Examples of systems of face-based security. (a) Access authorization system; (b) Payment authorization system; (c) Face unlock system on mobile phone.

(a) (b) (c)

Figure 2. There common types of face attacks. (a) Print attack; (b) Video replay attack; (c) Mask attack.

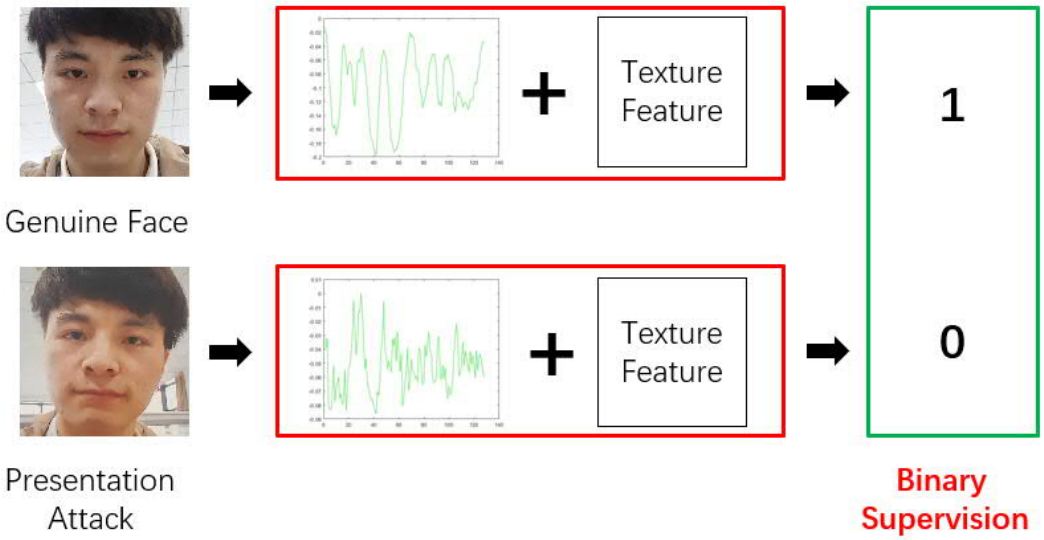


Figure 3. Our model uses the rPPG feature and the texture feature.

During the past decade, many researchers have made efforts for face anti-spoofing area [20]-[29]. 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. The 3D structure information is widely utilized in face anti-spoofing. With the help of 3D structure information, most of the face presentation attacks can be handled easily. An example is the iPhone X face unlock technique; it uses the dot projector to project more than 30,000 invisible dots onto the user's face to build the 3D facial map. Most recent research works also focus on estimating the depth map from a 2D image [13][14][30]. 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 the same as genuine faces, which makes the system based on sole depth information vulnerable.

To resolve this security problem in face anti-spoofing, we proposed a model that employs both remote photoplethysmography (rPPG) feature and texture feature. As

shown in Figure 2, rPPG feature is effective for photo attack and mask attack, while texture feature is effective for detecting video replay attacks.

When light shines on the skin of the body, for example, face, hands, and arms, hemoglobins in superficial vessels absorb part of the lights, as illustrated in Figure 4. Heartbeats periodically change the number of hemoglobins in specific areas. The rPPG can capture the subtle colour changes so that to estimate the pulse. Prior studies have reported several methods to estimate rPPG signal from facial videos recorded by RGB cameras [31][32][33].

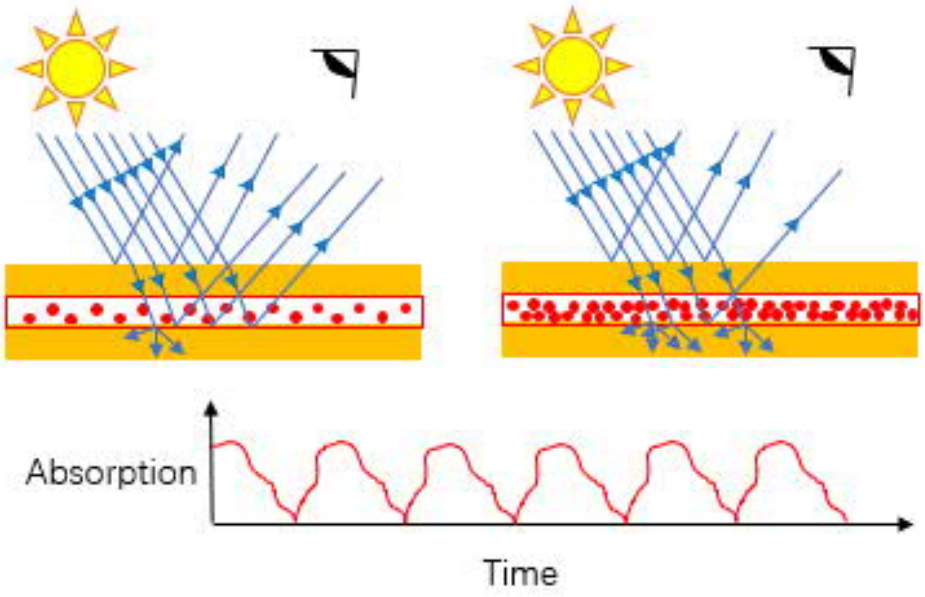


Figure 4. Illustration of how an rPPG works.

Inspired by the pulse-based face anti-spoofing method [34], we explore pulse signal features extracted by the rPPG algorithm. Our pulse signal feature extraction method is inspired by a prior work, where the long-term statistical spectral (LTSS) [7] feature is first used for face anti-spoofing research. The original LTSS was utilized in speaker presentation attack detection. We propose to use the Multi-scale LTSS (MS-LTSS) features in our study. This method combines different length of sliding windows and different overlapping sizes of the sliding window of LTSS features to represent both local and global features of frequency spectrum of the rPPG signal respectively to improve classification results.

The Convolutional Neural Network (CNN) showed its excellent performance in anti-spoofing tasks. In order to strengthen our anti-spoofing model, we propose a Contextual Patch-based CNN (CP-CNN) to analyse the texture features. The proposed CNN model extracts the local and global features from face images, where the local features are extracted from fixed patches of the deep feature image, and the global features are extracted from the whole deep feature image. The face anti-spoofing databases contain various video qualities. Hence, the combination of local and global features assists the model to learn independent patterns with the integral patterns of spatial face areas. “Contextual” means that the connection of outputs from Convolutional blocks, which allows the network to utilize different texture features extracted from multiple layers. Contributions of the thesis include: 1) we design multi-

scale long-term statistical spectral (MS-LTSS) features with various 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 these two parts of 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 structure of this thesis is organized as follows: in chapter 2 review face anti-spoofing methods including pulse-based approaches and other approaches. Next, in chapter 3 we describe details of the proposed methods for face anti-spoofing, including the Multi-scale LTSS with the contextual Patch-based CNN. Then in chapter 4, we illustrate the process of our experiments, the setups, and the comparison of the experimental results on five anti-spoofing databases, namely 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD, and OULU-NPU. Discussion of the experimental results are also reported in this chapter. Finally, in chapter 5 we summarize the thesis work and list future work.

2. PRIOR WORKS

Most important previous studies about the face anti-spoofing problem are reviewed and summarized in this section, as the background for the thesis work. Based on the key clue information that these methods reply on, we divide previous face anti-spoofing methods into five groups: texture-based methods, temporal-based methods, 3D structure-based methods, hardware-based methods, and remote photoplethysmography-based methods, each will be discussed in details in the following subsections.

2.1. Texture-based methods for face anti-spoofing

Using texture information to deal with the face anti-spoofing task is one major and common approach, because most face recognition systems use one single 2D sensor camera. Many research works proposed various hand-crafted features for this purpose, such as Local Binary Patterns (LBP) [35][36][37], Histogram of Gradient (HoG) [38][39], Difference of Gaussian (DoG) [42][43], Scale-invariant Feature Transform (SIFT) [40] and Speeded Up Robust Features (SURF) [41]. Traditional classifiers such as support vector machine (SVM), Random Forest, and Latent Dirichlet Allocation (LDA) were utilized in those works. In [34], they utilized the LBP-*ms-color* features as a part of their proposed cascaded model. Where *ms* indicates that multi-scale LBP features combining with the $LBP_{8,1}$, $LBP_{8,2}$, $LBP_{8,3}$, $LBP_{8,4}$ and $LBP_{16,2}$ features. And *color* indicates that the LBP features extracted from RGB channels of the face image separately. In [35], they proposed to use the LBP-TOP (local binary patterns from three orthogonal planes) features which analyses three-dimensional texture features to discriminate a face video as live vs. spoof. Prior works also tried to transform input data into different domains, e.g., to transform input images from RGB colour space into HSV or YCbCr colour spaces [44][45], or from time domain to frequency domain, to obtain more robust results. However, these texture-based methods share some common limitations, 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 showed their power in many research areas, especially in computer vision tasks. Several works used CNN-based features for face anti-spoofing [47][48][49][50]. Li et al. [48] used CNN as the feature extractor and fine-tuned a model which is pretrained on ImageNet for face anti-spoofing. Feng et al. [47] fed the samples of face images into CNN and obtained the result of classification, i.e., live vs. spoof. Atoum et al. [13] proposed to use patch-based CNN as a part of their two-stream CNN model, which segments the input face image into small patches to analyze the texture features for face liveness detection. This method solved the problem of insufficient training samples for neural network models. In more recent research works, deep learning methods achieved superior performance on anti-spoofing databases and in anti-spoofing competitions. For some old databases, such as NUAA [52], and Replay-Attack [53], which were collected several years ago, video resolution and quality are very poor. Deep learning methods can achieve 100% accuracy, which is more robust than traditional texture methods. On the other side, newer databases such as OULU-NPU [4] and SiW [14] include higher-quality videos

recorded from a large number of subjects in different conditions, which might be more challenging. New deep learning methods need to be developed on those databases.

Compared to other computer vision tasks, e.g., object detection, object identification, and facial expression classification, the task of face anti-spoofing using deep learning methods still have a long way to run. One goal of this thesis is to construct a novel CNN model, which is able to extract more reliable texture features for face liveness detection.

2.2. 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. [54] proposed a blinking-based face anti-spoofing method, which utilized Conditional Random Fields (CRFs) to analyze 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. [55] 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. Phan et al. [74] proposed to utilize the Local Derivative Pattern from Three Orthogonal Planes (LDP-TOP) [73] features to describe motions of the facial. Tirunagari et al. [75] proposed to utilize the dynamic mode decomposition (DMD) to describe the feature of movements. Li et al. [76] proposed to utilize the Difference of Gaussian filter to analyze the variation of motion patterns which are aroused by different dimensions of objects.

The other category of methods relied on the movement between the face and the background. Anjos et al. [77] considered to measure the motion correlation coefficient of the face and the background, and set a threshold to classify the print attack or real access. Kollreider et al. [56] proposed to use optical flow methods to track the movement of face to differentiate real vs. fake faces. Besides, Haralick features [70] and motion mag [69] were also used in face anti-spoofing. For deep learning, Feng et al. [47] proposed to extract features from optical flow map and Shearlet image using CNN. Xu et al. [51] also proposed the LSTM-CNN model using multiple frames of the single video to obtain the fused result 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 colour fluctuation of live facial skin caused by pulsation (a.k.a. the rPPG) invisible to human eyes. Thus, another goal of the thesis work is to extract and utilize the rPPG information for face liveness detection.

2.3. 3D structure-based methods for face anti-spoofing

Most of existing anti-spoofing databases have 2D videos, which do not contain 3D structure information. A popular method of 3D structure-based face anti-spoofing methods is to extract depth information from 2D images. Liu et al. [14] 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 [57][58]. Then they trained a CNN model to extract the depth map and the feature map together to classify the input video. In [13], the authors proposed a depth-based CNN model as a part of their two-stream CNN model. They utilized the same 3D face model fitting algorithm as in Liu's work to estimate the shape parameters and projection matrix. Then they utilized the 3DMM model [71] for computing the dense 3D face shape. Then they extracted the feature from the estimated depth mask and utilized the SVM to classify the input video as live vs. spoof.

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, and it can easily cheat these 3D structure-based methods.

2.4. Hardware-based methods for face anti-spoofing

With the help of various advanced hardware, such as depth cameras, light-field cameras and infrared cameras, we can analyse additional information to detect face attacks. IPhone X can analyse the 3D shape information recorded by a dot projector which could project more than 30,000 invisible dots onto an user's face to build the 3D facial map. Then the 3D model of the input sample is compared with that of a genuine face to obtain the result of live vs. spoof. Erdogmus et al. [78] proposed to record depth information using a Kinect camera, then compared the depth information to classify the video as live vs. spoof. Pavlidis et al. [79] considered the variation of light reflection, and they proposed to calculate the upper-band of near-infrared (NIR) spectrum recorded by a multi-spectral camera. Furthermore, Zhang et al. [80] utilized two photodiodes to capture the light reflectance, and used this feature to detect the presentation attack.

The light-field camera is becoming more and more popular recently. It can record the disparity information with the depth information from a single capture. Kim et al. [81], Ji et al. [83], and Moghaddam et al. [82] utilized this kind of cameras to tackle the face presentation attack detection task.

Although these methods introduced above achieved good performance on video replay attack and print attack, they could be influenced by the variations in a realistic environment. For instance, the illumination condition could cause serious interferences for infrared cameras and depth cameras. As well these methods could be vulnerable when an attacker wears a high-quality 3D mask, since the 3D mask contains the same depth information as the real face. Furthermore, the hardware-based method requires special devices which might be costly and not commonly accessible.

2.5. Remote Photoplethysmography-based methods for face anti-spoofing

Remote Photoplethysmography (rPPG) is a method to extract pulse signal from facial videos without any skin contact [32][59][60][61][62]. The analysis of extracted rPPG signal can be used for face anti-spoofing. In the first part of this session, we review three benchmark algorithms for extracting pulse signals from face videos. In the later part of this session, we review several rPPG-based methods for face anti-spoofing.

The first method named as Li CVPR was proposed in [31], the second method named as CHROM was introduced in [32], and the third method named as Spatial Subspace Rotation (SSR) was proposed in [66]. All three methods made a great impact in the field of remote pulse signal measurement from facial videos and inspired later studies on this topic. Sample figures of extracted pulse signals from the same facial video using these three rPPG algorithms are shown in Figure 5, and the details of each method are introduced below.

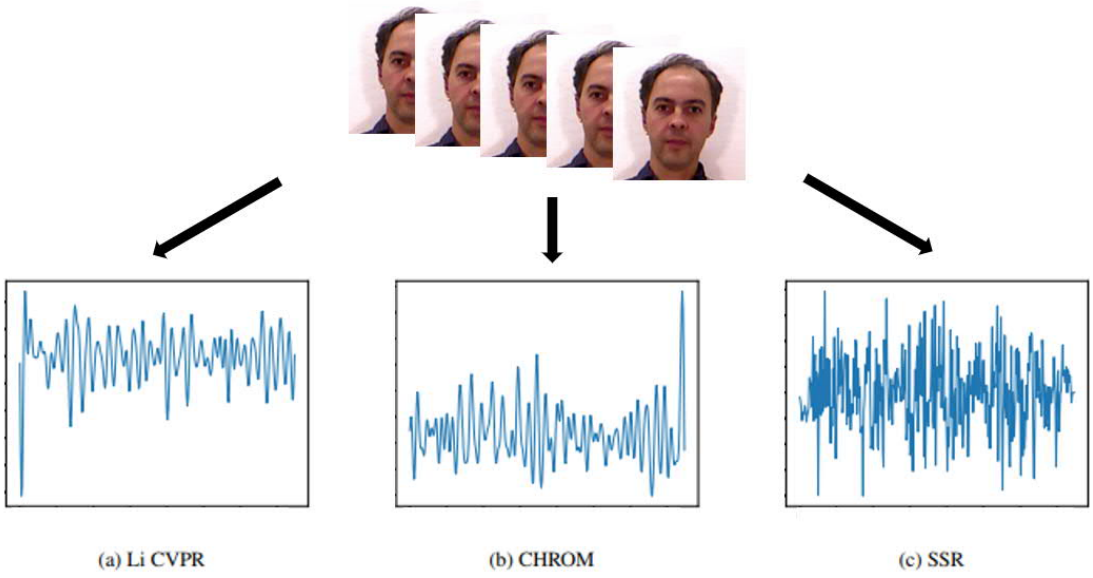


Figure 5. Examples of pulse signals extracted from a real face video by using different rPPG algorithms.

Li CVPR. In [31], Li et al. proposed a practical rPPG algorithm which can measure pulse signal from facial videos recorded under realistic human-computer interaction scenes. A simple version of this approach was first implemented in [34] for face liveness detections. First, they used the Discriminative Response Map Fitting (DRMF) approach to detect facial landmarks and determine the face region of interest (ROI) in the first frame of the input video. Then they utilized the Kanade-Lucas-Tomasi (KLT) method to track the region of interest (ROI) in the following image sequences. They calculated the mean value of the green channel of the ROI for extracting the raw pulse signal.

Considering illumination variation, they utilized the Distance Regularized Level Set Evolution (DRLSE) algorithm to segment the background and regarded the mean of the background region green channel as a reference to correct the raw pulse signal.

After that, Non-rigid Motion Elimination was implemented by segmenting the pulse signal into clips and getting rid of the clips containing the non-rigid movements. Finally, three filters were implemented to obtain the final signal, including (1) a detrending filter to make the trend of the signal flattened; (2) a Moving-average filter, to reduce the noise and smooth the signal; and (3) a bandpass filter with Hamming window, to cut the signal frequency into [0.7, 4] Hz which covers the ordinary heart rate range from 42 to 240 beat-per-minute (bpm).

CHROM. In [32], Hann et al. proposed the CHROM algorithm which is powerful but comparatively simple. The first step of this algorithm is to find the skin region of the input image by analysing the skin colour and compute the average colour of that region. Then, the authors projected the value of the average skin colour onto a specific colour space in order to capture the subtle changes of the skin colour caused by heartbeats. Finally, a 5-point averaging filter used for reducing the random noise and smoothing the signal, and a band-pass filter used for cutting the signal frequency into [0.7, 4] Hz which covers the heart rate between 42 and 240, were implemented to obtain the final pulse signal. This is a common algorithm that is widely used in tracking rPPG signals. This algorithm shares some similarity to the method proposed in Li *et al.* [34]. One key contribution of CHROM is that, it used the whole skin region instead of the specific face region of interest, which might work better as more facial pixels were analysed. However, in the step of skin region detection, the threshold of the skin colour is hard to determine due to the different races, e.g. Caucasian, Asian or African, and the illumination condition can also make impact on the recorded skin colours.

Spatial Subspace Rotation. Wang et al. [66] proposed the spatial subspace rotation (SSR) method to extract rPPG signal from videos. Considering the RGB spatial distribution of skin region pixels, the authors proposed to utilize spatial RGB correlation. Rotation and scaling were implemented simultaneously when modelling the two subspaces. They proposed to use the eigenvectors for computing the skin colour region correlation matrix. The direction change of the eigenvectors and energy changes of the eigenvalues were analysed in temporal domain. This method is able to retrieve the pulse signal directly without filtering progress.

These previous rPPG algorithms focused on improving the accuracy of averaged heart rate. On the one side, they are complex to be implemented because many preprocess steps are required for input face sequences, such as ROI selection and skin-pixels detection; on the other side, they also involved signal processing steps aimingfor improving HR accuracy which is not necessary for the purpose of face liveness detection. There is a novel deep learning rPPG algorithm proposed in our research group, the PhysNet [8], which is an end-to-end system using spatio-temporal convolutional neural network to extract rPPG signals from raw facial videos. We consider the PhysNet to be a good tool for exploring rPPG-based features for the face anti-spoofing problem and it will be explored in this thesis work.

For the task of face liveness detection, the main target is not computing the actual reading of the heart rate, but differentiating whether there is reasonable pulsation or not in order to justify whether it is a live or spoof face. So, we need to design and propose features for face liveness detection based on these works.

Several studies have explored rPPG-based methods for the face anti-spoofing problem. Li et al. [34] proposed the first pulse-based face anti-spoofing method. They extracted three pulse signals one from each of RGB channel, then utilized several filters to process the raw pulse signals. The filtered pulse signals are transformed from time domain to frequency domain and the maximum values of the power in frequency range of [0.7, 4] Hz, and the ratio of max power with the total power from each RGB signals are calculated. Finally, they got a simple six-dimensional feature, and utilized the SVM classifier to discriminate the input video as live vs. spoof.

In order to achieve stronger representation ability, Heusch et al. [7] designed the long-term statistical spectral (LTSS) approach for face liveness detection. The original LTSS approach was designed for speaker presentation attack detection [11]. This approach utilized the discrete Fourier transform (DFT) to transform the pulse signal from time domain into frequency domain and considered the log-magnitude of the frequency bins of the spectrum for statistics. Then they used the mean and variance statistics of the DFT coefficient as the feature of the pulse signal and classified by SVM to discriminate the input video as live vs. spoof.

Liu et al. [63] 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, and achieved supportive results on 3DMAD and HKBU-MARs V1 database. However, the cross-correlation can increase periodic global noises, and frequencies of different facial regions are similar to each other. In another related work, Nowara et al. [64] overcame this problem by analysing 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 deep learning area, Liu et al. [14] proposed to use rPPG supervision instead of analysing rPPG signal directly. They utilized rPPG signal to supervise their CNN-RNN model, and trained the rPPG model by using the estimated rPPG signal as the label of live face video, and flattened signal as the label of all kinds of attack videos. Then they utilized the RNN model to classify the input video as live vs. spoof.

All the works mentioned above show the effectiveness of rPPG-based methods for face liveness detection. However, rPPG based methods are vulnerable for video replay attack, since the re-captured video could record the pulse signal from the original face as well. Hence, we may need multiple methods to work together in order to detect various types of attacks. In this work, we propose an innovative multi-temporal-resolution rPPG feature as one part of the whole model for solving the face presentation detection task.

3. THE PROPOSED APPROACH

In this section, firstly, we will introduce an innovative approach, the PhysNet to extract rPPG signals from input facial videos. Secondly, we will present the proposed Multi-scale long-term spectral statistics method for rPPG feature representation, which improves the original LTSS [7] method on multi-scale level. Thirdly, we will describe the Contextual Patch-based CNN as texture-based feature analysis for face anti-spoofing. Finally, we will introduce our fusion method to merge the contribution of the two features on the decision level. The overview of the proposed method is shown in Figure 6.

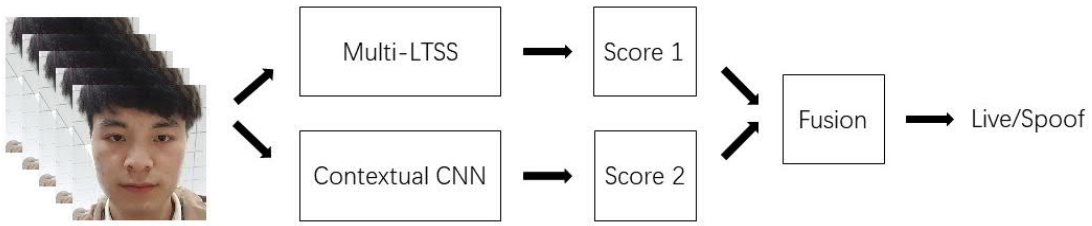


Figure 6. The overview of the proposed method.

3.1. rPPG-based Features for Anti-spoofing

The approach takes videos as inputs. For a video with *n* frames, we use PhysNet to extract rPPG signals. Then we utilize Multi-scale LTSS (MS-LTSS) to extract features and feed the features to SVM as a classifier for the face liveness detection task. The framework of the proposed rPPG-based method is shown in Figure 7.

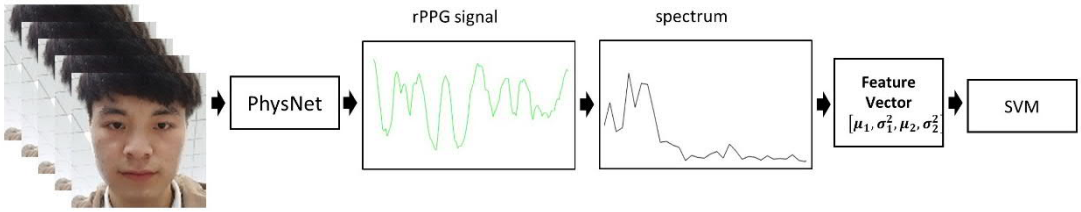


Figure 7. The framework of the rPPG-based method.

3.1.1. Face cropping

We utilize the Dlib [9] *'68 facial landmark'* detector to locate the coordinates of the nose and chin of face in each frame in a video. Then we use those coordinates to estimate the bounding box of facial frames by the following equations and use OpenCV to crop all the frame to 128×128 pixels. Figure 8 illustrates the facial landmarks we use and how the image is cropped. We set the nose be the centre of the cropped face and utilize the nose and chin coordinates to determine the bounding box of the face region by following equations:

$$y = y_n - 4 \times \frac{y_c - y_n}{3} = \frac{7y_n - 4y_c}{3}, \quad (1)$$

$$x = x_n - 4 \times \frac{y_c - y_n}{3}. \quad (2)$$

Where, $(x_n,\ y_n)$ and $(x_c,\ y_c)$ is the coordinates of the nose and chin respectively. $(x,\ y)$ is the top left coordinates for the cropped image.

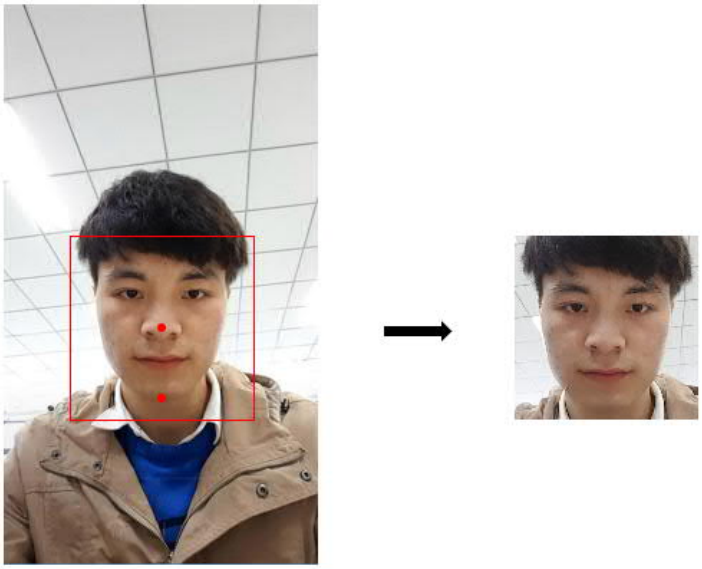


Figure 8. Example of using facial landmarks of nose and chin to crop the facial image.

3.1.2. PhysNet

PhysNet is an innovative method to extract rPPG signal from raw face sequences directly by using deep spatio-temporal convolutional networks. This is an end-to-end model with no additional filter requires, which is very convenient. Hence, we utilize this approach in this thesis. The architecture of PhysNet is shown in Figure 9.

The input of the network is T-frame face images. After forwarding several operations of spatial and temporal convolution with average pooling, multi-channel manifolds are formed to represent the spatio-temporal features. Finally, the latent manifolds are projected into signal space using simple pseudo 3D convolution operation with 1 × 1 × 1 kernel to generate the predicted rPPG signal.

We separate the cropped facial images into RGB channels, and feed the green channel images into PhysNet pretrained on OBF [10] database to extract rPPG signals.

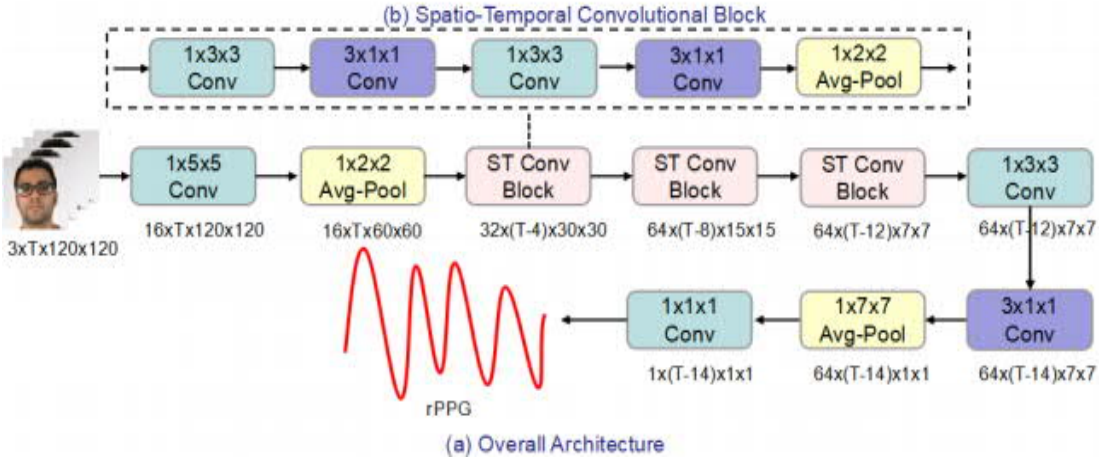


Figure 9. The architecture of PhysNet. The four-dimensional tensors showed under each block are the corresponding output dimensions *(channel × depth × height × width).*

3.1.3. Multi-scale LTSS

It is not suitable to directly use the extracted rPPG signal as the feature vector for face liveness detection because it is time defined and can be high dimensional. Thus, we need to design proper features for the task and here we utilize Multi-scale LTSS for this purpose. 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 elaborated rPPG information due to the pyramid-like multi-scale segmentation.

For each sliding window w, we convert the 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. As in the original LTSS [11] used in speaker presentation attack detection, we set the DFT coefficient $|X_w(k)|$ 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:

$$\mu(k) = \frac{1}{W}\sum_{i=1}^{W} log|X_i(k)|\,. \tag{3}$$

$$\sigma^2(k) = \frac{1}{W}\sum_{i=1}^{W} log|X_i(k) - \mu(k)|\,. \tag{4}$$

The first and second order statistics for vectors (for $k = 0 \dots N/2 - 1$) is concatenated as the part of the feature of the signal. Then we introduced the Multi-

scale LTSS, we concatenated the different LTSS features that calculated by different settings of the length of sliding windows w and the overlapping size of the sliding window o.

$$F = [\mu_1, \sigma_1^2, \dots, \mu_n, \sigma_n^2], \tag{5}$$

where, the 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 to acquire the score of the input video, whether it is a genuine presentation or an attack.

3.2. Contextual Patch-based CNN for Anti-spoofing

Previous patch-based CNN methods [13] divide the original RGB face image into several patches by randomly cropping the original face image and extracting 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 10.

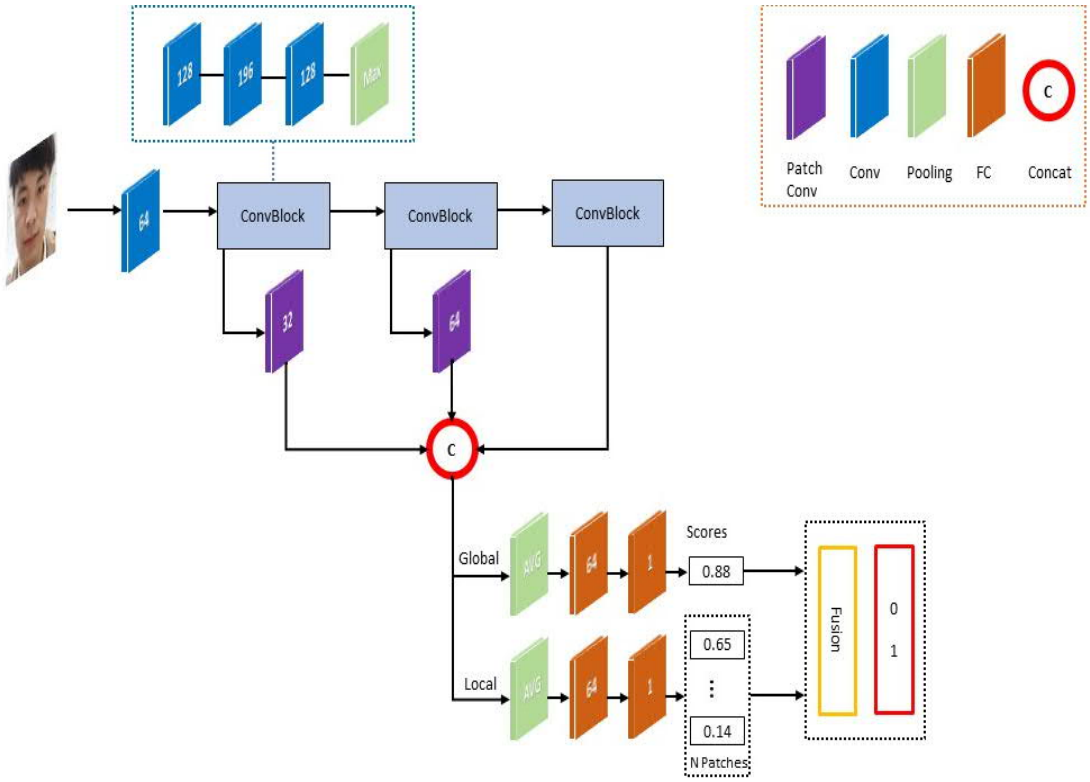


Figure 10. The pipeline of the proposed CP-CNN architecture. The number of filters are shown in the middle of each layer. *Colour code* used: *purple=Patch convolution layer, blue=convolution layer, green=pooling layer, orange=fully connect layer.*

3.2.1. Network Architecture of CP-CNN

As illustrated in Figure 10, 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 3x3 kernel and one maximum pooling layer. Every convolutional layer is followed by a batch normalization layer and ReLU layer.

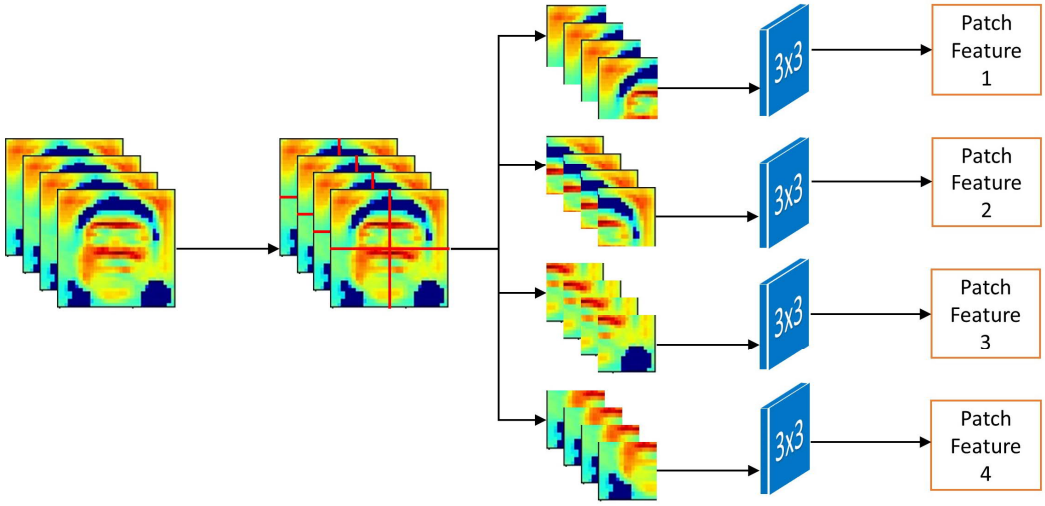


Figure 11. Illustration of Patch-based Convolution Module when the local patch number N = 4.

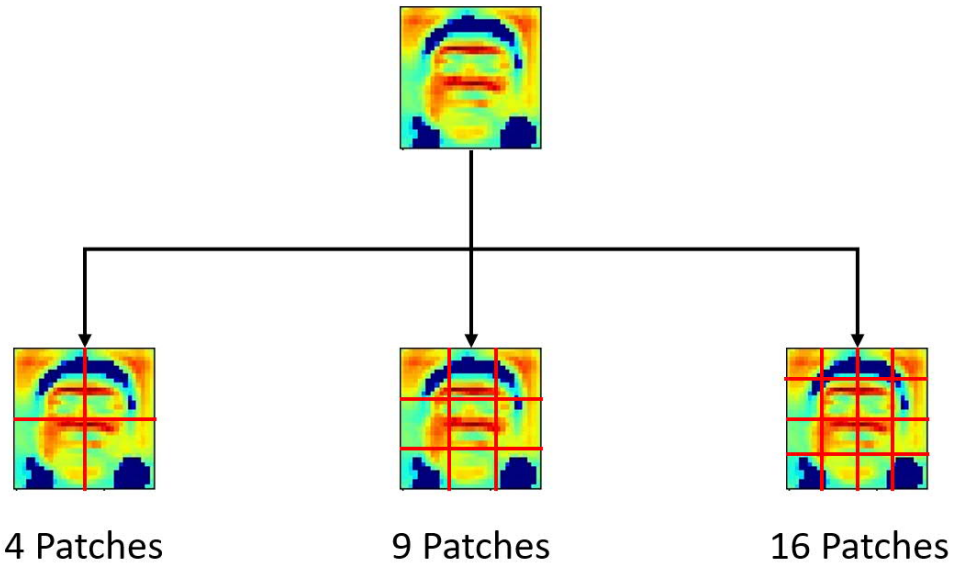


Figure 12. Illustration of different numbers of patches.

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 11, the deep features are divided into 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, the local patch number is set as $N = 4, 9, \text{and } 16$. The illustration of different numbers of patches is shown in Figure 12.

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 the global and patch features are channel-wise concatenated, which is illustrated in Figure 11. 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 real face confidence prediction. As for the global classifier, global average pooling is used and then cascaded with two fully connected layers and a sigmoid function. Similarly, the local classifiers use the same operations on the corresponding patch positions.

3.2.2. Loss Function and Network Inference

In this work, we utilize Focal loss as our based loss function for the CNN model, which was proposed by Lin *et al.* [16]. This loss function was designed for dense object detection to solve the problem of class imbalance. The class imbalance could lead to two consequences. (1) Training could be inefficient because most negative samples are easily classified and not effectively contribute to model learning. (2) The easy negative samples are overwhelmed during training and make the model degenerated. Authors proposed to reduce the weight of easy negative samples, then lead the model to focus more on the difficult samples. The face liveness detection databases are class imbalanced as well. Hence, we utilize Focal loss in this thesis.

This loss function is modified based on cross entropy (CE) loss for binary classification as follow:

$$CE(p, y) = \begin{cases} -\log(p) & if\ y = 1 \\ -\log(1-p) & otherwise. \end{cases} \tag{6}$$

In equation (7), $y \in \{\pm 1\}$ represents the ground-truth class and $p \in [0, 1]$ is the estimated probability for the class by the model with label equals to 1. They defined p_t instead of p for the national convenience,

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise. \end{cases} \tag{7}$$

And they rewrote $CE(p, y) = CE(p_t) = -\log(p_t)$.

Here, they designed a weighting factor $\alpha \in [0,1]$ for class 1 and $1 - \alpha$ for class 0. For notational convenience, they defined α_t for p_t, and rewrote the α-balanced CE loss as follow:

$$CE(p_t) = -\alpha_t \log(p_t). \tag{8}$$

As the equation above, α controls the balance of positive and negative samples, but cannot discriminate the easy and hard samples. In other to allow the model to focus on hard samples by reducing the weight of easy samples. They introduced a modulating factor $(1 - p_t)^{\gamma}$ to the cross entropy loss with the tunable focusing parameter$\gamma \geq 0$, and the focal loss is written as:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t). \tag{9}$$

Finally, the focal loss which combines the weighting factor α and the modulating factor is written as:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t). \tag{10}$$

In the training stage, it can be regarded as a binary classification task. The network input is an RGB face image $I^{128\times128\times3}$, 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 face image is a genuine face, we set the binary label as 1, while the label is set to 0 for attacks. We adopt Focal loss as the loss function, so the overall loss can be formulated as

$$Loss = wL_g + \frac{(1 - w)\sum_{i=1}^{N} L_i}{N}, \tag{11}$$

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

$$Score = wS_G + \frac{(1 - w)\sum_{i=1}^{N} S_{Li}}{N}. \tag{12}$$

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

$$S_F = \frac{\sum_{i=1}^{N_f} Score_i}{N_f}, \tag{13}$$

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

3.3. 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

$$S = w_f S_{MS-LTSS} + (1 - w_f) S_{CP-CNN}, \quad (14)$$

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

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

We evaluate our method on five databases, 3DMAD [1], HKBU-Mars V1 [63], MSU-MFSD [2], CASIA-FASD [3], and OULU-NPU [4], to demonstrate its generalizability under three different types of attacks, e.g., print, video replay, and 3D mask attacks. We perform intra testing among all five databases and cross testing between 3DMAD and HKBU-Mars V1, and compare our results with several state-of-the-art methods.

4.1.1. Databases

3D Mask Attack Database (3DMAD): It contains 255 video clips recorded from 17 subjects. It has 170 real accesses and 85 3D mask attack videos. Each subject recorded 10 real videos, and 5 attack videos while wearing a 3D mask. All masks recorded in the 3DMAD database were bought from ThatsMyFace.com. The frame rate of videos is 30 frame per second (fps) and the resolution is 640×480. The length of each video is 10 seconds. Two samples of 3DMAD video clips are shown in Figure 13.

Figure 13. Examples of 3DMAD video clips. The left image shows a real face, and the right one shows a mask attack.

HKBU 3D Mask Attack with Real World Variations V1 Database (HKBU-MARs V1): This database contains 120 videos recorded from eight subjects. It has 80 real access videos and 40 mask attack videos. Mask attack videos are recorded with two different kinds of 3D masks, of which six masks were bought from the ThatsMyFace.com with the same quality as those in 3DMAD, and the rest two masks with higher quality were bought from REAL-F[1]. All videos were captured by a Logitech C920 web camera. The frame rate of each video is 30 fps and the resolution is 1280×720. The length of each video is 10 seconds. Four samples of HKBU-Mars V1 video clips are shown in Figure 14.

[1] Detail for REAL-F: http://real-f.jp.

Figure 14. Examples of HKBU-Mars V1 video clips. The top two images present the real face and the mask bought from ThatsMyFace.com and the bottom two images show the real face and the mask bought from real-f.jp.

Figure 15. Examples of MSU-MFSD videos. The first row presents images recorded from an Android Phone, while the second row shows images recorded with a Laptop. From left to right: real faces, replay video attack presented by iPad, replay video attack presented by iPhone, and print attack.

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. It has 70 real accesses and 210 attack videos. Each subject recorded two genuine videos, four print attack videos, and two replay attack videos. Each type of videos was recorded by two different kinds of cameras, MacBook Air 13'' built-in camera and Google Nexus 5 Android Phone. For the replay videos, they used two methods. The one used Canon 550D for recording and iPad Air presenting. The other one used iPhone 5S for both recording and presenting. The frame rate of videos is 30

fps and the resolution is 640×480 and 720×480. The length of each video is 10 seconds. Samples of MSU-MFSD video clips are shown in Figure 15.

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. It has 150 real accesses and 450 attack videos. Each subject recorded three genuine videos, three print videos, three warped videos, and three replay attack videos. This database has high-quality, normal-quality, and low-quality videos, and the resolution is 1280×720, 480×640, 640×480 respectively. The low-quality videos were recorded using a USB webcam which is used for a long time and degrade the video quality automatically. The normal-quality videos were recorded by a new USB webcam which could keep the normal video quality. For high-quality videos, they used a Sony NEX-5 camera which has the Full HD resolution. The replay videos were all displayed using an iPad. The frame rate of videos is 25 fps and the length of each video is approximately 5 seconds. Sample images of CASIA-FASD video clips are shown in Figure 16.

Figure 16. Complete videos of one subject in CASIA-FASD. Top four images show the low-quality videos, mid four are the normal-quality videos, and the bottom are the high-quality videos. For the columns, from left to right: genuine, warped photo attack, cut photo attack and video attack.

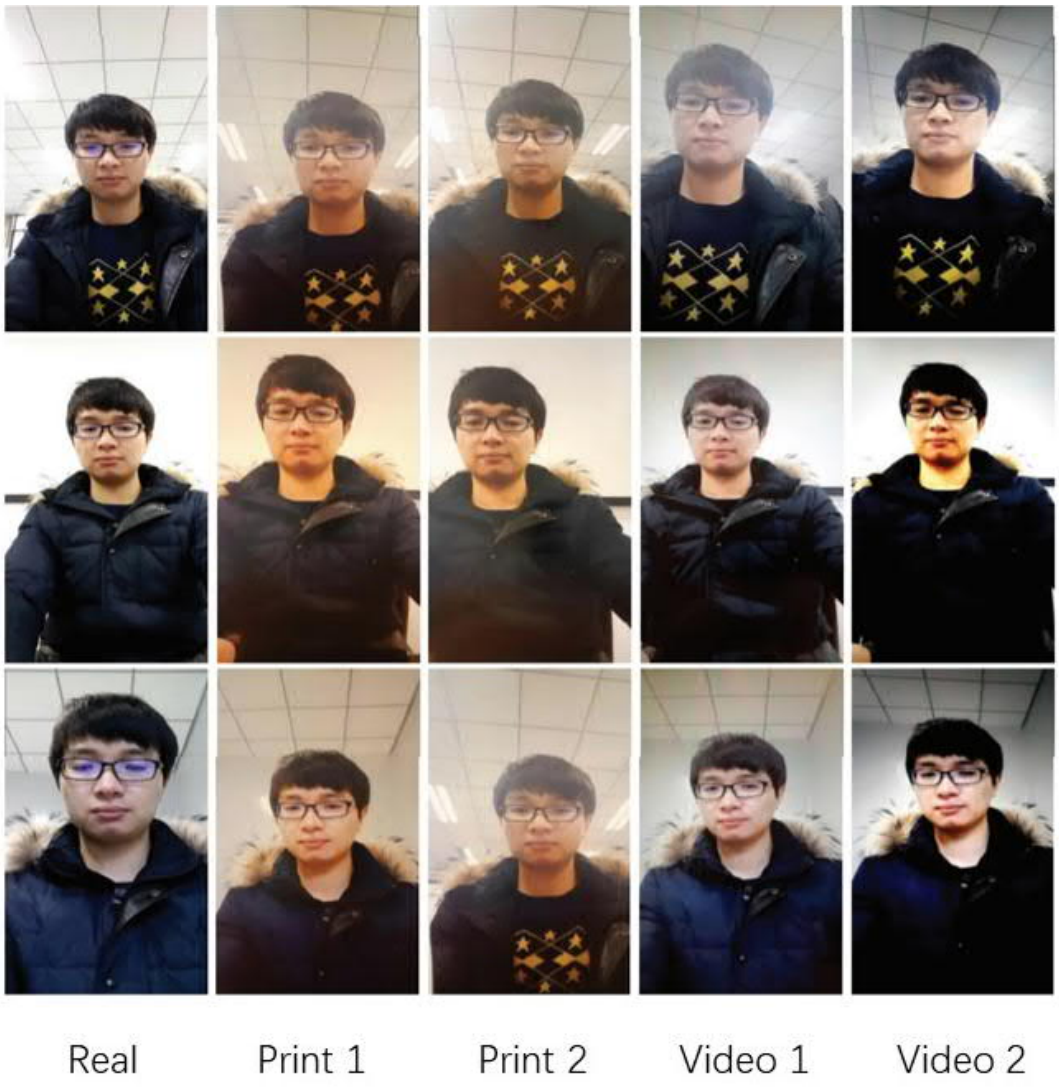


Figure 17. The examples of OULU-NPU videos. The first row presents video clips taken from the first scenario, the second row shows video clips from the second scenario, and the third row presents video clips from the third scenario. From left column to right column: real faces and printed photos 1, printed photos 2, replay videos 1 and replay video 2.

OULU-NPU: It contains 4950 video clips recorded from 55 subjects, including print attack and video replay attack. It has 1980 genuine videos and 3960 attack videos. Each subject recorded 18 genuine videos, 36 print attack videos, and 36 replay attack videos. Each type of video was recorded under three different illumination conditions and by 6 different smartphones namely Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual, and OPPO N3. For print attacks, they utilized two different printers, Canon PIXMA iX6550, and Canon imagePRESS C6011. The frame rate of videos is 30 fps and the resolution is 1080×1920. The length of each video is 5 seconds. This database provides four protocols for testing. Protocol I is designed to evaluate the performance of methods under unseen environmental conditions, specifically illumination and background scene. Protocol II is designed to evaluate the generalization of the methods under

different types of printers or displays. Protocol III is a Leave One Camera Out (LOCO) protocol to evaluate the performance of sensor interoperability. Protocol IV is the most challenging protocol, it contains all above three conditions together to evaluate the performance of methods. Samples of OULU-NPU video clips are shown in Figure 17.

4.1.2. Hyperparameter setting

The proposed Multi-scale LTSS method is implemented in MATLAB. We used the first 256 frames of the input video and fed them to the PhysNet. 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. We used two pairs of hyperparameters for the Multi-scale LTSS. 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.

The proposed Contextual Patch-Based CNN method is implemented in PyTorch v1.0.1 [5] with the learning rate of e-4, and the training phase is 30 epochs. The batch size of the Contextual patch-based CNN stream is 16. We train and test the model on a Linux server with graphic card of NVIDIA K80, and a desktop PC with graphic card of NVIDIA RTX 2080Ti.

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 [6]. For OULU-NPU database, we utilize 1) Attack Presentation Classification Error Rate (APCER), which evaluates the highest error rate from all presentation attack instruments (PAI), for example, 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:

$$APCER = \frac{\sum_{i=1}^{N_{PAI}}(1 - Res_i)}{N_{PAI}}, \tag{15}$$

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}, \tag{16}$$

$$ACER = \frac{APCER + BPCER}{2}, \tag{17}$$

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 is set as 1 when the ith presentation is classified as an attack presentation, and set Res_i as 0 if classified as bona fide presentation.

For evaluations on 3DMAD and HKBU-Mars V1, we adopt HTER (Half total error rate) and EER (equal error rates) metrics. TPR and EER are adopted for evaluations on MSU-MFSD and CASIA-FASD databases. 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).

$$TPR = \frac{\sum_{i=1}^{N_G}(1 - Res_i)}{N_G}, \tag{18}$$

$$FNR = \frac{\sum_{i=1}^{N_A}(1 - Res_i)}{N_A}, \tag{19}$$

$$FPR = \frac{\sum_{i=1}^{N_G} Res_i}{N_G}, \tag{20}$$

$$HTER = \frac{FNR(\tau^*) + FPR(\tau^*)}{2}, \tag{21}$$

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

4.2. Experimental results

4.2.1. Intra Testing

We perform intra testing on 3DMAD, HKBU-Mars V1, MSU-MFSD, CASIA-FASD, and OULU-NPU databases. For all the tests, we use 16 patch-based CP-CNN as we found in our prior tests that 16 patches gave the best performance than other partition patches. For the validation protocols, we use different testing protocols for each database as described in below:

Testing protocol of 3DMAD: We follow the leave-one-subject-out cross-validation (LOOCV) protocol which used in the original paper [1]. For all 17 subjects, we use eight subjects' videos as the development set, another eight subject's videos as the training set, and rest one subject's videos as the testing set, and rotate for 17 times so that each of all subjects' videos were tested once.

Testing protocol of HKBU-Mars V1: We follow the same testing protocol as S. Liu *et al*. [63] which is leave-one-out cross-validation (LOOCV). For all eight subjects, we randomly chose three subjects' videos for training, another three subjects' videos for developing, and the rest one subject's videos for testing, and rotated for eight times.

Testing protocol of MSU-MFSD: When experimenting on the MSU-MFSD database, we follow the protocol proposed in the original paper [2], e.g., 15 subjects' videos were used for training and the rest 20 subjects' videos for testing.

Testing protocol of CASIA-FASD: There were three testing protocols in the original paper [3]. We utilize the overall protocol which used 20 subjects' videos for training and 30 subjects' videos for testing.

Testing protocol of OULU-NPU: OULU-NPU has four different testing protocols, and we tested with all four protocols and reported the results. Protocol 1 is designed for testing the performance under different illumination conditions, for example, videos captured under high light condition and the low light condition. The detailed data assignment is shown in Table 1. Protocol 2 tests on different displays and printers, to evaluate the presentation attack detection method when facing the presentation attack instruments variation with two different printers and two different displays. Detailed information is shown in Table 2. Protocol 3 in Table 3 tests on 6 different kinds of phone cameras to evaluate the robustness of the face presentation attack detection method. Protocol 4 is the most challenging one, which combines the key points of all above three protocols and has the smallest number of training samples among all testing protocols. The detailed information is shown in Table 4.

Table 1. The detailed information about OULU-NPU testing protocol 1.

Protocol	Set	No. Real Videos	No. Attack Videos	No. All Videos
1	Train	240	960	1200
	Dev	180	720	900
	Test	120	480	600

Table 2. The detailed information about OULU-NPU testing protocol 2.

Protocol	Set	No. Real Videos	No. Attack Videos	No. All Videos
2	Train	360	720	1080
	Dev	270	540	810
	Test	360	720	1080

Table 3. The detailed information about OULU-NPU testing protocol 3.

Protocol	Set	No. Real Videos	No. Attack Videos	No. All Videos
3	Train	300	1200	1500
	Dev	225	300	1125
	Test	60	240	300

Table 4. The detailed information about OULU-NPU testing protocol 4.

Protocol	Set	No. Real Videos	No. Attack Videos	No. All Videos
4	Train	200	400	600
	Dev	150	300	450
	Test	20	40	60

For evaluations on 3DMAD and HKBU-Mars V1databases, we report achieved EER and HTER in Table 5 and Table 6. The EER and TPR (when FNR = 0.1) achieved on MSU-MFSD and CASIA-FASD are reported in Table 7 and Table 8 respectively. For OULU-NPU, we report ACER, APCER, and BPCER of the four protocols in Table 9. We also compare our results with results from previous works.

Table 5. The intra-testing results on 3DMAD.

Method	3DMAD	
	EER	HTER
Li *et al.*[34]	4.71 %	7.94 %
MS-LTSS	3.52 %	6.86 %
CP-CNN	0.00 %	0.00 %
MS-LTSS + CP-CNN	**0.00 %**	**0.00 %**

Intra-testing 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 6. The intra-testing results on HKBU-Mars V1.

Method	HKBU-Mars V1	
	EER	HTER
MS-LBP [19]	23.0%	22.6%
Liu et al. [63]	14.7 %	16.2 %
MS -LTSS	0.95 %	3.11 %
CP-CNN	0.00 %	0.00 %
MS -LTSS + CP-CNN	**0.00 %**	**0.00 %**

Intra-testing results on HKBU-Mars V1 database: We compare the proposed MS-LTSS and CP-CNN methods with two previous results, and the results showed the 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 7. The intra-testing 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 [2]	8.6 %	92.8 %
MS-LTSS	3.4 %	96.5 %
CP-CNN	1.1 %	99.1 %
MS-LTSS + CP-CNN	**0.00 %**	**100.00 %**

Intra-testing results on MSU-MFSD database: We compared with two baselines and a texture-based method, and 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 with zero error. 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 8. The intra-testing results on CASIA-FASD.

Method	CASIA-FASD	
	EER	TPR (FNR=0.1)
LBP+LDA [36]	21.0 %	75.7 %
IDA [2]	12.9 %	86.7 %
CDD [39]	11.8 %	88.8 %
Dynamic [67]	10.0 %	89.1 %
Patch-based CNN [13]	4.44%	-
MS-LTSS	8.3 %	92.6 %
CP-CNN	3.2 %	96.6 %
MS -LTSS + CP-CNN	1.8 %	98.7 %
SPMT + SSD [68]	**0.04 %**	**100.0 %**

Intra-testing results on CASIA-FASD database: We compared with five state-of-the-art methods with our proposed MS-LTSS and CP-CNN methods. Although we didn't achieve state-of-the-art performance, the performance is very close to it. Our proposed CP-CNN achieved better performance compared with the prior patch-based CNN. The CASIA-FASD data contains three kinds of quality of video replay attacks and print attacks, which make it more challenging than MSU-MFSD data. Our results indicate that our proposed methods are robust when facing different quality of video replay attacks and print attacks.

Table 9. The intra-testing results on OULU-NPU.

Prot.	Method	OULU-NPU		
		APCER (%)	BPCER (%)	ACER (%)
1	**MS-LTSS**	3.0	18.6	10.8
	CPqD	2.9	10.8	6.9
	GRADIANT	1.3	12.5	6.9
	CP-CNN	2.1	8.3	5.2
	MS -LTSS + CP-CNN	**1.2**	7.6	4.4
	FAS-BAS [14]	1.6	1.6	1.6
	Wang *et al.* [30]	2.5	**0.0**	**1.3**
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

Intra-testing results on OULU-NPU database: We reported the results of four protocols of OULU-NPU and compared with several state-of-the-art methods. The OULU-NPU data recorded under four evaluation method is more challenging than all previous data. In Protocol one, we achieved the best result of APCER, which indicates that the proposed method has good generalization ability over different illumination conditions. And in protocol 2 and protocol 3, our fused MS-LTSS and CP-CNN method performed well and is very close to the state-of-the-art result, which demonstrate that our method is robust over various displays and cameras. However, protocol 4 is the most challenging one which contains all the evaluation key points, and with much less training videos than those of the three previous protocols, our methods did not work well probably because of insufficient training data.

4.2.2. Cross Testing

In order to illustrate the generalization of the proposed method on mask attack, we reported the results of cross-testing between 3DMAD and HKBU-Mars V1 databases by following the baseline cross-testing protocol.

Cross-testing protocol of 3DMAD to HKBU-Mars V1. We randomly chose 8 subjects from 3DMAD database for training, and test on the whole HKBU-Mars V1 database.

Cross-testing protocol of HKBU-Mars V1 to 3DMAD. We randomly chose 5 subjects from HKBU-Mars V1 database for training, and test on the whole 3DMAD database.

We report the HTER and EER of the cross-testing results and compare with the MS-LBP baseline and S. Liu et al. [63] algorithms. The results are shown in Table 10.

Table 10. The cross-testing results between 3DMAD and HKBU-Mars V1.

Method	3DMAD to HKBU-Mars V1		HKBU-Mars V1 to 3DMAD	
	EER	HTER	EER	HTER
MS-LBP [19]	49.2 %	46.5 %	51.6 %	64.2 %
S. Liu et al. [63]	12.3 %	11.9 %	17.7 %	17.4 %
MS -LTSS + CP-CNN	**0.7 %**	**1.2 %**	**1.6 %**	**2.1 %**

It can be seen that our method (fused) overperformed the previous methods by a large margin, which demonstrated the robustness of our method. In both cross-tests, our method was able to achieve EER of less than 2%, which strongly supported that our proposed method has a superior adaptability when facing novel mask attacks. Training on 3DMAD performed slightly better than training on HKBU-Mars V1. This may be caused by the camera quality of these two databases. The HKBU-Mars V1 videos are of higher resolution which contain more detailed texture features than 3DMAD.

4.2.3. Ablation Study

The performance with different colour channels for estimating pulse signals. The input videos can be divided into R, G, B channels, so we evaluate the PhysNet trained with each of the three channels separately (as PhysNet R, PhysNet G and PhysNet B), and then of all three channels (as PhysNet RGB) together. We feed the single-color-channel video sequences and the original video sequences into the PhysNet to estimate four sets of pulse signals, and evaluate them on OULU-NPU data with protocol 1. As reported in prior works, the blue channel contains the noisiest signal, and the green channel performs the best. Our results are consistent with prior findings. As shown in Table 11, PhysNet G performed better than the other three conditions.

Table 11. The results of utilizing different channels of estimated pulse signals for extracting MT-LTSS features.

Method	OULU-NPU P1		
	APCER (%)	BPCER (%)	ACER (%)
PhysNet RGB	40.0	14.1	27.1
PhysNet R	12.3	**10.3**	17.5
PhysNet B	50.0	15.5	32.7
PhysNet G	**3.0**	18.6	**10.8**

The performance with different patches. We compare four architectures with different patches to illustrate the advantages of the proposed patch-based CNN. We train these four models on OULU-NPU based on Protocol 1, and the result of each model is shown in Table 12. The first model without patches (Global) 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. In the current setting Global + 16P achieved the best results but it doesn't mean that more patches will always lead to better results, we expect the performance will drop if the patches number is too large although not tested here.

Table 12. The results of CP-CNN with different number of patches, on OULU-NPU Protocol 1.

Method	OULU-NPU P1		
	APCER (%)	BPCER (%)	ACER (%)
Global	8.50	13.67	11.08
Global + 4P	5.75	12.63	7.69
Global + 9P	3.19	10.35	6.77
Global + 16P	**2.1**	**8.3**	**5.2**

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 scales of LTSS ([64, 16] and [128, 64]) and compare them with the combined MS-LTSS (([64, 16] + [128, 64])) on OULU-NPU based on Protocol 1, and the results are shown in Table 13. It can be seen that the combined MS-LTSS achieved significantly better performance that the other two. We choose the combination of ([64, 16] + [128, 64]), where the first set of parameters designed to extract the local rPPG feature with small window size and overlapping. Furthermore, the second set of parameters is utilized for extracting the global rPPG feature with larger window size and overlapping.

Table 13. The results of different settings of LTSS and Multi-scale LTSS on OULU-NUP Protocol 1.

Method	OULU-NPU 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 effect of sequence length. In Table 14, we report results of different sequence lengths for the proposed method (MS-LTSS + CP-CNN) 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 longer sequences we can achieve more accurate estimated rPPG signals. For extracting Multi-scale LTSS features, longer signals contain more information, so the extracted MS-LTSS features are more helpful for discriminating the input video as live vs. spoof. Another possible reason is that longer videos provide more samples for the proposed CP-CNN model.

Table 14. The results of MS-LTSS + CP-CNN model 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	**3.7**	**15.3**	**9.5**

4.2.4. Visualization and Error Analysis

We investigate some sample rPPG signals reconstructed by the proposed MS-LTSS model in order to examine the method in more details. Figure 18 shows some examples of successful and failed samples of rPPG signals and their power spectral density (PSD) curves. It can be seen from the succeeded samples that, typical live faces have periodical rPPGs and their PSD curves with a dominant peak, while typical spoofs have noisy pattern rPPGs and their PSD curves with multiple random peaks at a much lower amplitude. On the other side for the failure cases, some live faces and spoofs have similar noisy pattern of rPPGs and PSD curves for some reasons (e.g., motions), which make them difficult to be discriminated.

We also conduct statistical analysis on failed samples (53 samples, ACET=4.4%) on the OULU-NPU protocol 1 on which our proposed model did not perform so well. For each sample, we check the output scores of the Multi-scale LTSS and the Contextual patch-based CNN independently. There are $\frac{36}{53}$, $\frac{28}{53}$, and $\frac{18}{53}$ samples failed from the MS-LTSS feature, the CP-CNN feature and the overlap of both features. One possible reason of the failure is that the sampling rate of the original video is less than 30 fps or videos are shorter than 10 seconds so that the inputs are shorter than 256 frames, which caused that the estimated pulse signals are not completed. The length of

inputs influences the accuracy of the estimated pulse signals and thus the Multi-scale LTSS features. Hence, future research will focus on working with shorter samples.

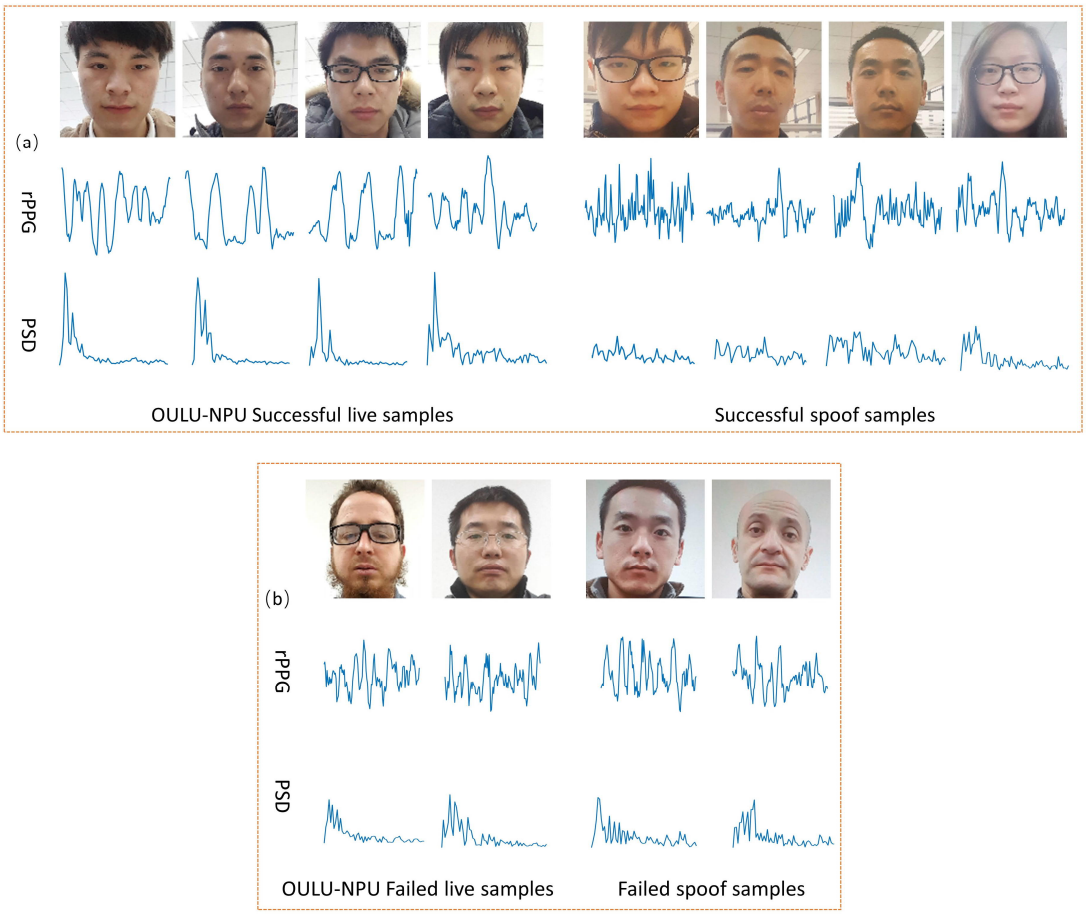


Figure 18. (a) 8 successful face examples, their estimated rPPG signals, and PSD figures. The left-side four are live and the right-side four are spoofs. (b) 4 failed examples. The left-side two are live and the right-side two are spoofs.

5. CONCLUSION

Nowadays, biometric-based security systems are threatened by various kinds of presentation attacks. Although the face presentation attack detection methods are becoming more and more robust and efficient. New face presentation attack instruments are also developing at the same time. For example, the camera sensor and the display resolution are increasing rapidly, so that the video replay attacks and print attacks are more difficult to be discriminated. With the development of 3D print technology, high-quality 3D masks are becoming more and more realistic with much detailed textures. Hence, the traditional face liveness detection methods are vulnerable with these situations. More innovation methods are still required to explore in this research area.

At the beginning of this work, we reviewed previous face liveness detection methods in recent years, from traditional methods to deep learning methods. rPPG based face liveness detection methods as a branch of the face liveness detection research orientation showed its unique advantages when facing print attacks and mask attacks. We also reviewed benchmark rPPG extraction methods of prior works. Then we proposed to use an advanced rPPG signal extraction method PhysNet to estimate rPPG signals from videos, and then utilize rPPG based Multi-scale LTSS features for the face anti-spoofing task.

During the study and the analysis of prior works of face liveness detection methods, we found that one single method could not achieve excellent performance for all kinds of spoofing, and they are also easily affected by different situations. So, most of the recent research works are all based fusing two or more methods, for example, combining temporal and spatial information, or patched and depth information. Hence, we propose to combine two deep learning models, one based on rPPG features and the other based on patch based texture features, to improve the performance of face liveness detection.

For the study of deep learning based face liveness detection methods, we reviewed the structures of the convolutional neural networks reported in prior works. In order to combine the advantages of different CNN structures, we designed the Contextual Patch-based CNN which combines global-local and multi-level deep texture features simultaneously. This architecture of CNN has not been implemented in previous studies yet.

As reported in prior works, the rPPG features are powerful for discriminating print attacks and mask attacks, while texture features are effective for detecting replay attacks. We evaluated these two approaches separately until we obtained well results on each model. Then we fused these two models as our final model. Our proposed method fusing MS-LTSS and CP-CNN models has shown robust performance on five databases including various evaluations of the proposed method. According to the analysis of results, 1) the rPPG based MS-LTSS model is extremely effective for mask and print attack detection, 2) the Contextual Patch-based CNN achieved great performance among all kinds of attacks, and 3) the fused model further improved accuracy.

This is the most recent research which covers all kinds of attacks and evaluated on most popular benchmark databases, and the results may inspire new research key points to future studies in face anti-spoofing area. This work also gives a clear conclusion of different methods for face anti-spoofing and did a lot of experiments on different databases to analyze the advantages and disadvantages of the proposed

method. In the future, we will continue focusing on exploring the hidden information of the rPPG signals which can be used for face liveness detection, and to improve the structure of our CP-CNN to achieve more robust results when dealing with complex conditions, for example, OULU-NPU protocol 4. We also plan to experiment on the latest face anti-spoofing database SiW and conduct cross-testing between CASIA and Replay Attacks by following the latest research orientation.

6. REFERENCES

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

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

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

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

[5] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch," 2017.

[6] ISO/IEC JTC 1/SC 37 Biometrics. Information technology – Biometric presentation attack detection – Part 1: Framework. International Organization for Standardization, 2016.

[7] G. Heusch, and S. Marcel. "pulse-based features for face presentation attack detection," In Biometrics: Theory, Applications and Systems, 2018.

[8] Z. Yu, X. Li, G. Zhao. "Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks," arXiv preprint arXiv:1905.02419, 2018.

[9] D. E. King, "Dlib-ml: A machine learning toolkit," J. Mach. Learning Res., vol. 10, pp. 1755–1758, 2009.

[10] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K, Majamaa-Voltti, M. Tulppo, G. Zhao. "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection." Proc. IEEE International Conference on Automatic Face and Gesture Recognition. pp. 242–249. IEEE (2018)

[11] H. Muckenhirn, P. Korshunov, M. Magimai. Doss, and S. Marcel. "Long-term Spectral Statistics for Voice Presentation Attack Detection," IEEE/ACM Transactions on Audio, Speech and Language Processing, 25(11):2098–2111, Nov. 2017.

[12] W. Wang, S. Stuijk, and G. de Haan. “Living-skin classification via remote-PPG,” IEEE Transactions on Biomedical Engineering, vol. PP, no. 99, pp. 1–1, Mar. 2017 (DOI: 10.1109/TBME.2017.2676160)

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

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

[15] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition,” In Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[16] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. “Focal loss for dense object detection,” International Conference on Computer Vision, 2017.

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

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

[19] J. Määttä, A. Hadid, M. Pietikainen. “Face spoofing detection from single images using micro-texture analysis,” In: 2011 International Joint Conference on Biometrics (IJCB), IEEE, 2011, pp. 1–7

[20] 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

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

[22] G. Kim, S. Eum, J.K. Suhr, D.I. Kim, K.R. Park, J. Kim. “Face liveness detection based on texture and frequency analyses,” In: 2012 5th IAPR International Conference on Biometrics (ICB), IEEE, 2012, pp. 67–72

[23] T. de Freitas Pereira, A. Anjos, J.M. De Martino, S. Marcel. “LBP-TOP based countermeasure against face spoofing attacks,” In: Computer Vision-ACCV 2012 Workshops, Springer, 2012, pp. 121–132.

[24] J. Yang, Z. Lei, S.Z. Li. “Learn convolutional neural network for face anti-spoofing,” 2014. Available from: <1408.5601>.

[25] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes, S. Sridharan. “Liveness detection based on 3d face shape analysis,” In: 2013 International Workshop on Biometrics and Forensics (IWBF), IEEE, 2013, pp. 1–4

[26] 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

[27] K. Kollreider, H. Fronthaler, M.I. Faraj, J. Bigun. “Real-time face detection and motion analysis with application in liveness assessment,” IEEE Trans. Info. For. Secure. 2 (3) (2007) 548–558.

[28] J. Yan, Z. Zhang, Z. Lei, D. Yi, S.Z. Li. “Face liveness detection by exploring multiple scenic clues,” In: 2012 12th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, 2012, pp. 188–193.

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

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

[31] X. Li, J. Chen, G. Zhao, and M. Pietikainen. “Remote Heart Rate Measurement From Face Videos Under Realistic Situations,” In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.

[32] G. de Haan and V. Jeanne. “Robust Pulse Rate From Chrominance Based rPPG,” IEEE Trans. On Biomedical Engineering, 60(10):2878–2886, 2013.

[33] W. Wang, S. Stuijk, and G. de Haan. “A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation,” IEEE Transactions on Biomedical Engineering, 2015.

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

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

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

[37] J. Maatta, A. Hadid, and M. Pietikainen. “Face spoofing detection from single images using micro-texture analysis,” In International Journal of Central Banking, 2011

[38] J. Komulainen, A. Hadid, and M. Pietikainen. “Context based face anti-spoofing,” In Biometrics: Theory, Applications and Systems, 2013

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

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

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

[42] [14] X. Tan, Y. Li, J. Liu, L. Jiang. “Face liveness detection from a single image with sparse low rank bilinear discriminative model,” In: European Conference on Computer Vision 2010, Springer, 2010, pp. 504–517.

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

[44] Z. Boulkenafet, J. Komulainen, and A. Hadid. “Face antispoofing based on color texture analysis,” In International Conference on Image Processing, pages 2636–2640, 2015.

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

[46] J. Li, Y. Wang, T. Tan, and A. K. Jain. “Live face detection based on the analysis of fourier spectra,” In SPIE (BTHI), volume 5404, pages 296–304, 2004.

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

[48] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. “An original face anti-spoofing approach using partial convolutional neural network,” In IPTA, 2016. 1

[49] K. Patel, H. Han, and A. K. Jain. “Cross-database face antispoofing with robust feature representation,” In Chinese Conference on Biometric Recognition, pages 611–619, 2016

[50] J. Yang, Z. Lei, and S. Z. Li. “Learn convolutional neural network for face anti-spoofing,” arXiv preprint arXiv: 1408 5601, 2014.

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

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

[53] I. Chingovska, A. Anjos, and S. Marcel. “On the effectiveness of local binary patterns in face anti-spoofing,” In BIOSIG, 2012.

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

[55] 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.

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

[57] A. Jourabloo and X. Liu. “Pose-invariant face alignment via CNN-based dense 3D model fitting,” Int. J. Comput. Vision, 124(2):187–203, 2017.

[58] Y. Liu, A. Jourabloo, W. Ren, and X. Liu. “Dense face alignment.” In International Conference on Computer Vision workshop, pages 1619–1628, 2017.

[59] S. Bobbia, Y. Benezeth, and J. Dubois. “Remote photoplethysmography based on implicit living skin tissue segmentation,” In International Conference on Pattern, pages 361–365, 2016.

[60] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, and K.-W. Cheung. “Block-based adaptive ROI for remote photoplethysmography,” J. Multimedia Tools and Applications, pages 1– 27, 2017.

[61] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe. “Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions,” In Conference on Computer Vision and Pattern Recognition, pages 2396–2404, 2016.

[62] B.-F. Wu, Y.-W. Chu, P.-W. Huang, M.-L. “Chung, and T.-M. Lin. A motion robust remote-PPG approach to driver’s health state monitoring,” In Asian Conference on Computer Vision, pages 463–476, 2016.

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

[64] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan. “Ppgsecure: Biometric presentation attack detection using photopletysmograms,” In International Conference on Automatic Face and Gesture Recognition, pages 56–62, 2017.

[65] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. “Robust discriminative response map fitting with constrained local models,” In Conference on Computer Vision and Pattern Recognition, 2013

[66] W. Wang, S. Stuijk, and G. de Haan. “A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation,” IEEE Transactions on Biomedical Engineering, 2015.

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

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

[69] S. Bharadwaj, T. Dhamecha, M. Vatsa, and R. Singh. “Face anti-spoofing via motion magnification and multifeature videolet aggregation,” 2014.

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

[71] V. Blanz and T. Vetter. “Face recognition based on fitting a 3d morphable model”. In IEEE Transactions on pattern analysis and machine intelligence, 25(9):1063–1074, 2003.

[72] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang. “Optical flow guided feature: A fast and robust motion representation for video action recognition.” In Conference on Computer Vision and Pattern Recognition, pages 1390– 1399, 2018.

[73] B. Zhang, Y. Gao, S. Zhao, and J. Liu, “Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor,” IEEE Transactions on Image Processing, vol. 19, no. 2, pp. 533–544, Feb 2010.

[74] Q. T. Phan, D. T. Dang-Nguyen, G. Boato, and F. G. B. D. Natale, “Face spoofing detection using ldp-top,” in IEEE International Conference on Image Processing, Sept 2016, Conference Proceedings, pp. 404–408.

[75] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, “Detection of face spoofing using visual dynamics,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 762–777, April 2015.

[76] Y. Li and X. Tan, “An anti-photo spoof method in face recognition based on the analysis of fourier spectra with sparse logistic regression,” In Chinese

Conference on Pattern Recognition, Nov 2009, Conference Proceedings, pp. 1–5.

[77] A. Anjos and S. Marcel, “Counter-measures to photo attacks in face recognition: A public database and a baseline,” in International Joint Conference on Biometrics, Oct 2011, Conference Proceedings, pp. 1–7.

[78] N. Erdogmus and S. Marcel, “Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect,” in IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, Sept 2013, Conference Proceedings, pp. 1–6.

[79] I. Pavlidis and P. Symosek, “The imaging issue in an automatic face/disguise detection system,” in IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, June 2000, Conference Proceedings, pp. 15–24.

[80] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, “Face liveness detection by learning multispectral reflectance distributions,” in IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, May 2011, Conference Proceedings, pp. 436–441

[81] S. Kim, Y. Ban, and S. Lee, “Face liveness detection using a light field camera,” Sensors, vol. 14, no. 12, pp. 22 471–22 499, Nov 2014.

[82] F. P. A. Sepas-Moghaddam, P. Correia, “Light field local binary patterns description for face recognition,” in IEEE International Conference on Image Processing, Sept 2017, Conference Proceedings, pp. 3815–3819.

[83] Z. Ji, H. Zhu, and Q. Wang, “Lfhog: A discriminative descriptor for live face detection from light field image,” in IEEE International Conference on Image Processing, Sept 2016, Conference Proceedings, pp. 1474– 1478.