Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Rupak Katwal

Liveness Detection for 3D Face Mask Attacks

Master's thesis in Master in Information Security
Supervisor: Associate Professor Kiran Raja
June 2021

NTNU
Norwegian University of
Science and Technology

Rupak Katwal

Liveness Detection for 3D Face Mask Attacks

Master's thesis in Master in Information Security
Supervisor: Associate Professor Kiran Raja
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Liveness Detection for 3D Face Mask Attacks

Rupak Katwal

June 1, 2021

Abstract

The serviceable and convenient nature of the Face Recognition System (FRS) makes it a preferred way for access control and authentication for a wide range of application areas, from biometric passport, surveillance system, health care, law enforcement, banking services to user verification in the smartphone. Most of the current day FRS have a number of open challenges such as weaker liveness detection, makeup attacks, morphing attacks and privacy issues. As the FRS do not actively query for the liveness of the subject and verify if the person is alive. Taking the advantage of the vulnerabilities in current day FRS, intruders can fool the FRS using the presentation attacks (PA) (a.k.a spoofing attacks). An attacker can mimic being an authentic user by presenting a spoof biometric data (e.g., printed photo, face videos, 3D face mask). Such an attack can be addressed by adding a layer of security to the FRS to detect them and these approaches are generally called Presentation Attack Detection (PAD). In this work, we propose Remote Photoplethysmography (rPPG) based PAD to distinguish presentation attacks (spoofing attempts) between the real face and 3D mask face videos. Remote photoplethysmography has been used to determine the liveness of a subject in PAD by biological signals such as pulse from the face videos. In this thesis, we propose a set of complementary features for making the PAD better against 3D face masks. We evaluate the performance of the proposed approach on two publicly available 3D mask datasets - 3DMAD [1] and HKBVMarsV1+ [2] using the standard protocols. The proposed approach outperforms the performance under similar protocols as against the state-of-the-art. Further, the thesis also investigates the use of proposed approach for cross dataset evaluation by training on one kind of 3D face masks and test on unseen data 3D mask types in an effort towards generalization.

Preface

I would to express my sincere gratitude to my supervisor Assoc. Prof Kiran Raja for the continuous guidance and support throughout the thesis works. With his competent supervision guide me to work on thesis in precise manner.
Secondly, I want to acknowledge Mobai AS, for support, motivation and guidance during thesis work.

Rupak katwal

1st June 2021

Contents

Figures

Tables

Acronyms

AUC Area Under Curve. 12, 13

EER Equal Error Rate. 12, 13, 15, 16, 18

FRS Face Recognition System. iii, xi, 1–4, 7–9, 14, 65, 66

HTER Half Total Error Rate. 12, 13, 15, 17

LBP Local Binary Pattern. 14–16

LGI Local group Invariant. 47

ML Machine Learning. 6, 40, 44, 55, 58, 59, 63, 65, 83, 84

MTCNN Multi-Task Cascaded Convolution Neural Network. 29, 30, 44

NMS Non-Maximum Suppression. 29

PA Presentation Attack. xi, 2–5, 7–10, 14, 18, 65

PAD Presentation Attack Detection. iii, vii, viii, xiii, 2–9, 11, 14–16, 18–20, 27, 55, 63, 65, 67

PAI Presentation Attack Instrument. 10, 11

PPG Photoplethysmography. 2, 3, 24

ROI Region of Interest. 5, 6, 12, 21, 22, 27, 32, 43, 46, 67

rPPG Remote Photoplethysmography. iii, vii, viii, 3–7, 9, 12, 20–22, 26–28, 33, 35, 36, 43, 46, 47, 49, 63, 66, 67

SVM Support Vector Machine. 12, 15, 40, 53, 59

Glossary

3D mask 3D reconstruction of the face and to mimicry the genuine user. 7, 15

biological parameters refers to biological functionality of human body. . 5

feature set group of properties or characteristics to describe aspect of something . 4

Genuine face face biometric traits of real and authorized user. 4

Presentation Attack Spoofed biometric traits to circumvent biometric system. 2, 4

Presentation Attack Detection automatic detection of Presentation Attack. 2

Remote Photoplethysmography Contactless approach to measure estimate physiological parameter. 3

rPPG signal Heart pulse signal estimated from Remote Photoplethysmography. 4

Chapter 1

Introduction

1.1 Introduction

Every person has unique physiological and behavioural characteristics such as face, fingerprint, iris and way of walking [3]. In computer science, the measurement and statistical analysis of solitary person characteristics are referred to as biometrics [3]. Based on the biometrics data, a biometric recognition system is perceived, which refers to the identification and authentication of the user using the unique biometrics traits, e.g. retinas, irises, voices, facial characteristics, and fingerprints [4]. Among the biometrics trait for user authenticity, face biometrics is common and widely acceptable in the biometrics recognition system. The Face Recognition System (FRS) refers to identifying or verifying the user authenticity with their facial characteristics. It aims to extract distinctive details from the face and verify user identity based on the facial features such as distance between the shape of the chin, depth of eye sockets, the distance between forehead to chin, contours of lips, ears and chin or chin mapping face into three dimensional geometric. The research work on the Face Recognition System (FRS) can be traced back to the 1960s, and studied its relevancy in 1990s evolving the computer vision technology [5]. Analyzing the current biometrics recognition scenario, face biometrics traits are the ones with the highest economic and social impact since it is widely used approach after fingerprints and adopted it in unique identification documents such as International Civil Aviation Organization (ICAO)-compliant biometrics passport [6], national ID cards, border access control, surveillance, banking services, smartphone authentication and so on [5].

The upsurge increases in technological advancement have their own cost of severity. As the deployment and applicability of Face Recognition System (FRS) soar up, attacks on face biometric security are now not limited to theoretical scenarios but emerging with a severe threat. The majority of research work on face recognition is focused on improving the performance at the verification and identification task (dealing with occlusions, illumination, low resolution and so on) [5]. In the past few years, the study based on security vulnerabilities on biomet-

rics traits has become the foremost concern, since several attacks (photo attack, face video, 3D face mask etc) can evade the biometric recognition system such as Presentation Attack. The attacks in the biometrics system presented at the biometric sensor level are called Presentation Attack (PA). In these type of attacks, attackers present the biometric data which are obtained directly from a person or furtively from online sources (e.g. face printed photo or a printed iris image) and synthetic generation (e.g. face silicone mask, synthetic fingerprint), to circumvent the biometric recognition system by mimicking as a genuine user. Presentation Attack Detection provides the security on biometric systems to distinguish whether the presented biometric data is a real biometric trait or Presentation Attack (PA). Since Face Recognition System (FRS) is mainly concerned with user authenticity (difference between the real users), instead of determining the presented face biometrics traits is genuine or fake, it eases the intruders to perform the Presentation Attack (PA). Here the fake represents the PA biometrics traits while genuine represents the biometrics traits from a living subject. The security layer able to detect fake face and genuine face presented in FRS is called the face Presentation Attack Detection (PAD) [4].

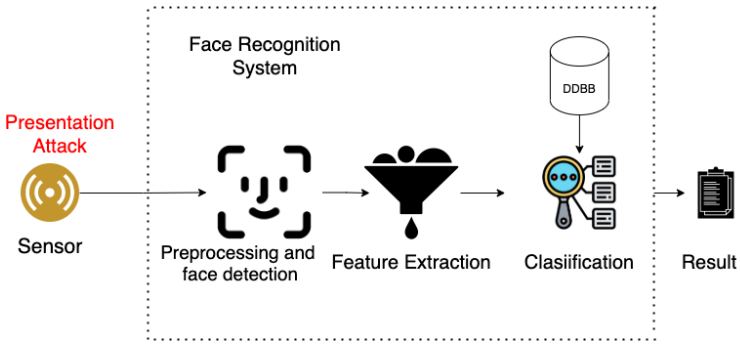


Figure 1.1: Pipeline of face recognition system with Presentation Attack (PA) scenario (inspired by figure Hernandez-Ortega *et al.* [5])

Ramachandra and Busch [7] classify the face Presentation Attack Detection (PAD) in two categories: hardware-based (characteristics of human face detected using hardware component integrated with FRS) [8] [9] and software-based (algorithm determining fake face sample and live face sample) [10] [11]. Among the software-based techniques in face Presentation Attack Detection (PAD), liveness detection of person is a functional approach, where the physiological parameter estimation such as Heart Rate [12], Respiratory Rate [12], blood oxygen saturation [13], and so on, is undertaken to verify the user liveness. The physiological parameters estimated by Photoplethysmography (PPG) [14] approach is based on the assumption that light is attenuated when illuminated on the skin surface, and the attenuated light shows variations, which depend upon the volume of blood under the observable skin surfaces [14]. The attenuation of light depend on the skin surface, skin structure, blood oxygen saturation, skin temperatures [14].

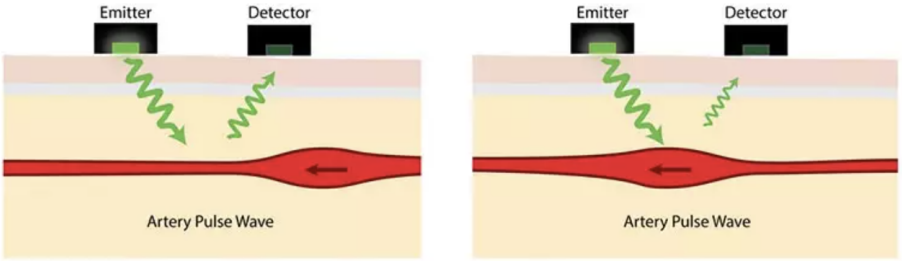


Figure 1.2: Optical heart rate sensing. Left: lower pressure preceding the pulse wave means narrower arteries and less absorption (higher reflectivity) of the green light source. Right: a higher blood pressure pulse causes wider arteries and more light absorption (lower reflectivity)(figure taken from `https://theconversation.com/how-reliable-is-your-wearable-heart-rate-monitor-98095`)

The estimation of PPG can be achieved by placing optoelectronic sensors on the skin. Alternatively, contactless acquisition of the PPG signal is also possible by estimating it from the videos/images captured from digital cameras. The contactless acquisition of pulse-based physiological parameters estimation built upon the concepts of photoplethysmography is popularly known as Remote Photoplethysmography . Other popular names based on the literature are video-based PPG, i-PPG, contactless PPG [15] based on the estimation approach employed. The Remote Photoplethysmography (rPPG) approach can be integrated into Face Recognition System (FRS) as face PAD to detect the liveness of the user; as a result, FRS can detect the presentation attacks.

1.2 Keywords

Remote photoplethysography (rPPG), Presentation Attacks (PA), Presentation Attack Detection(PAD), Face Detection, Region of Interest, signal Preprocessing, Frequency Spectrum, Machine learning

1.3 Problem Description

The Face Recognition System (FRS) is explicitly designed to capture the face variability which is caused due to illumination, occlusion, orientation and, to some extent, to detect facial makeup and face grooming rather than dealing with genuine and fake face biometrics [5]. So, the FRS framework dealing with only face variability leaves the door open to Presentation Attack (PA). The biometrics data (e.g. photos and videos) are now heavily exposed at different social media sites, revealing the faces, voice and personal behaviour [5]. Attackers are taking advantages from such biometrics traits, and operate to evade FRS by presenting a

printed photo [16], or replay face videos [17], or 3D mask [18]. The other possibilities to deceive FRS are using makeup [19] or plastic surgery. However, using photographs and videos are the most common type of attack due to its availability (social sites and video surveillance) and low cost. Among these type of attacks, the 3D masks are more likely to succeed due to the high realism of the presentation attack samples. With the advancement of 3D based face reconstruction, realistic 3D face mask can be created at affordable cost, imitating the complete structure of the face. Recently, the rapid development of 3D face printing and reconstruction technology is generating highly realistic 3D mask. This kind of technology not only the model 3D structure but also construct detailed facial textures such as hair, wrinkle, or even eye vessels, which even makes it difficult for human eyes to identify whether it is fake or genuine[20]. For Face Presentation Attack Detection (PAD) there arises a challenge against a 3D mask attack resulting in difficulty detect it whether it is a Genuine face or Presentation Attack [5]. As a result, PAD is of utmost importance for secure and reliable biometric recognition 3D mask PAD.

1.4 Justification, Motivation and Benefits

Integration of biometrics to secure digital security systems shows its applicability and effectiveness; hence, high-security systems adopting a biometrics-based security system. With the loose ends created by PA, put FRS security risk and vulnerable to access control, leading to devastating threat scenario. The PA with a face image or video are two spoofing methods that can be conducted through a digital screen or high-quality prints. Significant efforts and research works have been devoted to face PAD based on print face and video attacks[20]. Analyzing the threats caused by the 3D face attacks, the thesis is motivated to devise better approaches to detect and classify the attacks from genuine face video. Specifically, the thesis focuses on detecting the 3D mask attacks by investigating the rPPG approach in the Face Recognition System (FRS). The feature set of rPPG signal estimated from the face region has been well studied and has been demonstrated to distinguish a given face video sample genuine face or 3D face mask videos. However, we note the performance limitations of the existing methods []. Motivated by such a limitation, the thesis intends to seek for alternative features from estimated rPPG signals to make the PAD better.

1.5 Research Questions

In order to make the PAD better to detect the 3D masks, tis thesis aims formulates two critical research questions on Remote Photoplethysmography (rPPG) based 3D mask Presentation Attack Detection (PAD):

1. What are the complementary feature(s) from rPPG based pulse signal to improve classification of the given input face videos as a genuine or 3D face mask?

2. Can these complementary features help in detecting cross-dataset attacks when the different attack data is unseen during the training?

1.6 Contribution

The thesis work focuses on the use of Remote Photoplethysmography (rPPG) signal for 3D face mask Presentation Attack Detection (PAD). To cope with the spatial noise, the spatial average of all the skin pixels from the Region of Interest (ROI) is computed along with preprocessing step, which governs the series of the noise filtering process. The pulse or rPPG signal is estimated, with low effect from the spatial noise and subject motion. To cope with the subject in motion, a face tracking algorithm is employed, tracking the face (single face) across the video frames. The contribution of the thesis work was highlighted below:

- The thesis provides a better understanding and extensive analysis about the Remote Photoplethysmography (rPPG) based 3D mask Presentation Attack Detection (PAD).
- We introduce ten complimentary features defining a pulse signal generated by the proposed rPPG approach, from which we were able to distinguish between genuine face videos from 3D mask attack videos. The proposed feature set uses the biological parameters estimated from the face videos.
- The extensive analysis of the proposed approach against State-of-the-art is conducted on two 3D face mask dataset, with publicly available 3DMAD [1] and HKBVMArsV1+ [21] dataset. The proposed approach gains a performance EER of $7.9 \pm 4.3\%$ in 3DMAD [1] and EER of 18.18 ± 11.11 % in HKBVMArsV1+ [21].
- To generalize the proposed methodology, complementary feature is also evaluated under cross dataset evaluation on publicly available 3DMAD [1] and HKBVMArsV1+ [21] resulting favourable results. The proposed approach gains a performance of EER of 14.7%for cross-dataset evaluation.

1.7 Thesis Outline

This section provides an overview of every chapter that is presented in the thesis work. The thesis chapters initiated with the literature and background knowledge of 3D mask based PAD, background methodology, proposed methodology, experiment result, discussion, conclusion and future work sequentially.

1. The second chapter details the information about,concepts on PA and PAD, metrics for Presentation Attack Detection (PAD), background knowledge and related works on Presentation Attack (PA),Presentation Attack Detection (PAD) and face PAD approaches. The chapter start by presenting the conceptual knowledge on PA and PAD. The second section describes the metrics that are used to determine the performance of PAD. The third section

provides the details of five distinctive approaches on 3D mask based Presentation Attack Detection. The fourth section provides information about rPPG based pulse estimation in three categories, namely, face video processing, estimation of rPPG signal and Machine Learning (ML) approach. In the last section, details about the literature survey about feature vectors extracted from the rPPG signal can distinguish between real and fake face videos.

2. The third chapter enlightens about the background methodology of the thesis work. The chapter comprised of seven sections; the first section informs about the face detection and tracking algorithm to detect the face from the face videos in the proposed method; the second section provides the information colour channel and Region of interest selection for the best extraction of rPPG signal, third section details on the signal preprocessing of raw rPPG signal, fourth selection informed about the rPPG signal extraction method from the RGB colour space, the fifth section introduced spectral method on rPPG signal, the sixth section includes brief background about the binary classifier.
3. The fourth chapter provides information about the proposed methodology. The first section explains the implementation of existing face extraction and tracking technique to detect and track the face sequence and the video frames. The second section informs about the Region of Interest (ROI) selection from detected faces across face videos, and the third section details the signal preprocessing steps implemented in the proposed methodology. In the fourth section, introduced frequency domain analysis for feature extraction of the rPPG signal in the proposed methodology. The last section informs about learning the machine learning classifiers to distinguish genuine face videos and 3D mask face videos in the proposed methodology.
4. The fifth chapter informs about the experiment and result that were obtained from the proposed methodology. The first section introduces a brief description of the dataset included to conduct the experiment. In the second section, the experimental evaluation of each dataset is included. The last section reports the result from each dataset, produced from the proposed methodology.
5. The sixth chapter provides the analysis and discussion of the proposed methodology. The first section discuss about the rPPG approach for face PAD, second section discuss about the proposed methodology and results obtained, third section discuss about the knowledge guided in thesis work and societal consequences of proposed methodology is discussed.
6. The seventh chapter details the summary and significant finding about the research question from the proposed methodology.
7. The eighth chapter provides information about the possible future work, which can improve the proposed methodology.

Chapter 2

Related Work

This chapter provides brief information on PA and Presentation Attack Detection (PAD), background knowledge and literature survey about the metrics on PAD, Presentation Attack (PA), Presentation Attack Detection (PAD) approaches specified on the 3D mask, machine learning approaches for pulse estimation and rPPG signal based feature selection. The section The first section provides details the concepts on PA and PAD. In the second section, the evaluation metrics on Presentation Attack Detection (PAD) is described based on the literature's; the third section provides the information about related works on 3D mask based Presentation Attack Detection (PAD). Similarly, in the fourth section Remote Photoplethysmography (rPPG) signal estimation techniques based on the two key stages and literature survey on Machine learning approaches on rPPG. In fifth section, literature's on the feature group selection for Presentation Attack Detection (PAD) based on Remote Photoplethysmography (rPPG) signal is described.

2.1 Presentation Attack and Presentation Attack Detection

The biometric recognition system has one particular system vulnerabilities, called the Presentation Attack (PA), where a subject A attempts to impersonate the victim subject B using synthetic biometric data, e.g. (printed photo, videos or 3D mask, fingerprints etc.) to biometrics sensor. The biometrics traits used for the Presentation Attack (PA) is also called the Presentation Attack Instrument (PAI). Taking the scenario for face Presentation Attack Detection (PAD), Ming *et al.* [22], face PA can be classified into two categories: (a) impersonation (spoofing) attacks (b) obfuscation attacks. Imposters or intruders generally perform impersonation attacks to impersonate legitimate users; this kind of attack can be achieved with photo attacks, video replay attack, highly realistic 3D mask. On the other hand obfuscation attacks, aims to trick the Face Recognition System (FRS) to avoid being recognized, which can be performed by facial makeup, plastic surgery or face occlusion (use of scarves, glasses, masks). A taxonomy of different kind of attacks

on FRS can be seen in Figure 2.5.

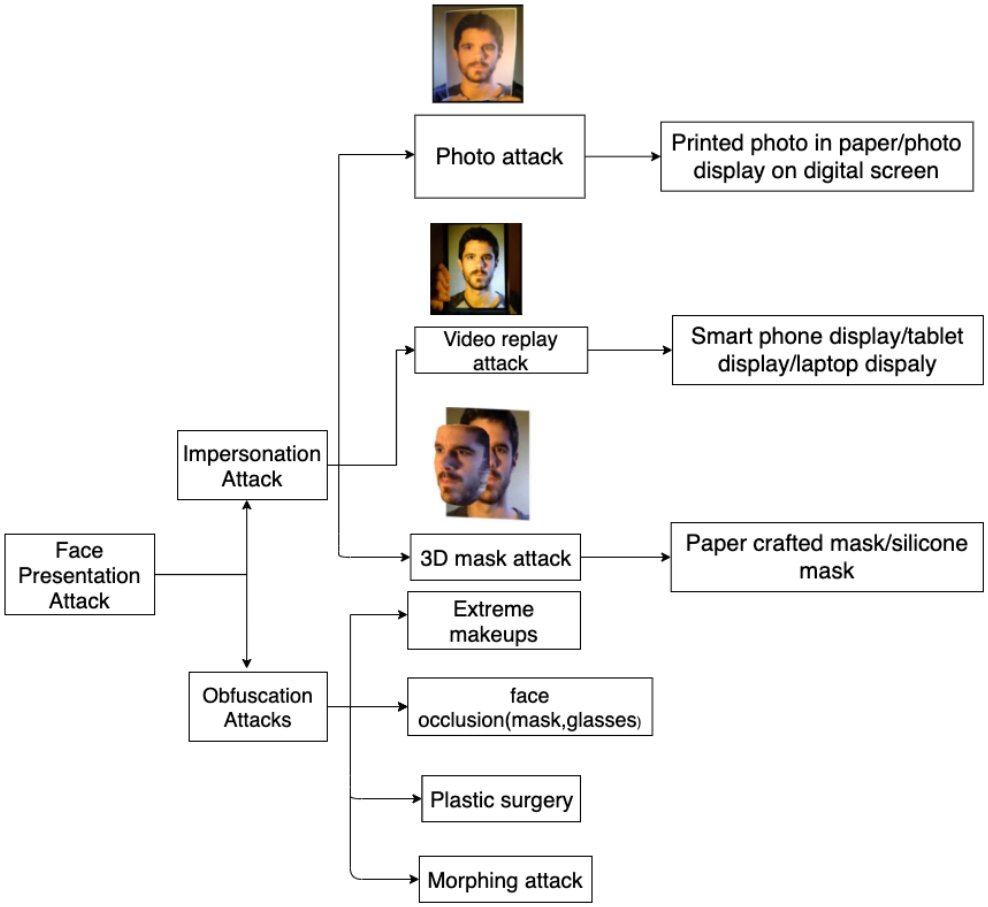


Figure 2.1: Face Presentation attack topology[22].

In the photo print attack, it is initiated by presenting the face photo, which can be hard copy printed on the paper or the digital screen to biometric sensor. The photo print attack is more common, due to readily available of biometric face traits on different social media sites or video surveillance data etc. Similarly, biometric sensor can be evade with the face video attack and can be more successful than print attacks. As high definition face video sequence consists subject motion, which can mimicry the subject liveness by subject motion in the videos (unless PAD is challenge-response). To circumvent the FRS, face video is presented on the biometric sensor with use of digital display such as smartphone display/tablet display/ laptop display etc. The 3D face mask represent the synthetic face construction with paper and more realistic by silicone. With the information about a set of face in various angles or view, the 3D based face reconstruction is performed. The results of 3D face build up highly realistic face mask, imitating complete structure of faces. The Presentation Attack (PA) initiated with 3D face mask is called 3D face

mask attack. The highly realistic face reconstruction approach makes difficult for PAD to detect 3D face mask attack.

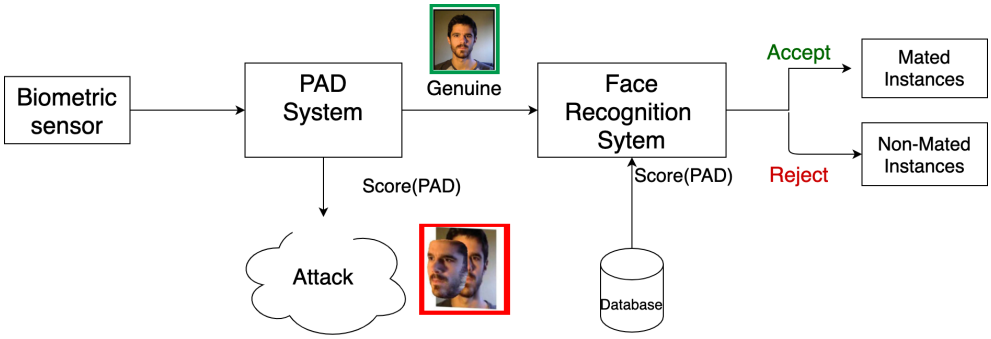


Figure 2.2: Integrated PAD with Face Recognition System (FRS) [5]

To ensure a secure biometric recognition system,it must detect and reject fake biometric traits. The PAD approach is defined as a technique that can detect and distinguish between the real biometrics traits and synthetic or forged biometrics traits presented in biometric sensor [5]. Hernandez-Ortega *et al.* [5] reported that PAD could be achieved in four different ways: (i) biometric sensor able to detect genuine biometric traits signal pattern (ii) hardware to detect evidence of genuine biometric attempts (iii) challenge-response system, where the PAD challenge the user to interact in a specific way and analyzing user response (iv) deploying recognition algorithms intrinsically robust against attacks. Detailed review related to PAD is presented in section 2.3.
Among the multiple face PA categories, 3D face mask creates a real threat on Face Recognition System (FRS) due to its appearance. This thesis focuses only on PAD approaches for detection of 3D face mask Presentation Attack (PA) on FRS using Remote Photoplethysmography (rPPG) based PAD approach.

2.2 Metrics for Presentation Attack Detection (PAD)

This section defines the evaluation of the PAD system in terms of PAD metrics, which tells how well the Presentation Attack Detection (PAD) is performing to detect and classify genuine and fake biometric data. In the domain of PAD, there are two types of biometric traits; genuine biometric samples and fake biometrics(related with Presentation Attack (PA)). As Presentation Attack Detection (PAD) aims to detect and distinguish between the given biometric sample is genuine or fake, it can be treated as a binary classification problem[1]. The two major types of error rates subjected to binary classification problem are False Positive(positive sample labelled as negative samples) and False Negative (Negative example labelled as positive samples). The corresponding error rate with False Positive is called False Positive Rate(FPR), which is the ratio between False Positive to a total

[1]classification tasks that have two labels.

number of negative samples as opposed False Negative Rate is the ratio of False Negative to a total number of positive examples. Moreover, there are other two error rates called True Positive Rate (TPR), which refers to the ratio of correctly classified positive samples and True Negative Rate (TNR), which corresponds to correctly classified negative samples, these metrics(notation) are more prevalent in terms of binary classification.
To compute the error rate, the system needs to calculate the decision threshold τ, which is the decision boundary between the genuine attempts and Presentation Attack (PA). τ is the trade-off between the FPR and FNR and often desired to choose its optimal values. Most common and popular to determine the threshold is Equal Error Rate (EER), whereτ_{EER} ensures that the difference between FPR is slight as possible. The optimal threshold is also called Operating Point (OP), which is determined using the data in the development set.

$$\tau_{EER} = argmin[FPR\tau_{dev} - FNR\tau_{dev}] \tag{2.1}$$

Once the evaluation criteria $\tau_{EER}$is determined, Half Total Error Rate (HTER) decided on the test set.

$$HTER(\tau_{test}) = argmin[(FPR\tau_{test} - FNR\tau_{test})/2] \tag{2.2}$$

Receiver Operating Curve (ROC) is a popular approach for visualizing the performance of binary classification problem, which plot the trade-off between the FPR and TPR depending on different threshold values. The corresponding terminology along with the ROC is Area Under Curve (AUC), which measures the entire two-dimensional area underneath ROC. Further, AUC also provides comprehensive performance measurement across all possible thresholds. Detection Error trade-off (DET) plot the trade-off between the FPR and FNR depending on different threshold values
In the domain of PAD, positive samples or genuine samples are named bonafide samples and negative or fake samples are named as Presentation Attack (PA). According to ISO standards[2], performance assessments of PAD are renamed into different terminology. In terms of PAD performance metrics, FPR is termed as Attack Presentation Classification Error Rate (APCER), and False Negative Rate (FNR) is termed as Bona-Fide Presentation Classification Error (BPCER) The APCER is calculated as follows:

$$APCER = \frac{1}{N_{PAIS}} \sum_{i=1}^{N_{PAIS}} (1 - RES_i) \tag{2.3}$$

WhereN_{PAIS} represent the number of attack presentation from the given Presentation Attack Instrument (PAI).The value of RES_i is 1 if the i^{th} presentation classifies as attack presentation and 0 if presentation classifies as Bona Fide presentation

[2]https://www.iso.org/standard/67381.html

[23]. Similarly, the BPCER is calculated as:

$$BPCER = \frac{1}{N_{BF}} \sum_{i=1}^{N_{BF}} RES_i \tag{2.4}$$

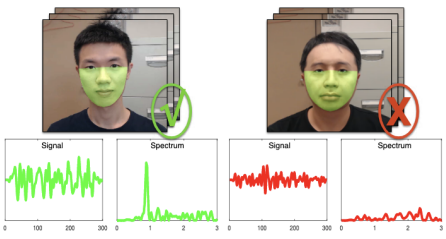
WhereN_{BF} represent the number of Bona Fide presentation from the given Presentation Attack Instrument (PAI).The value of RES_i is 1 if the i^{th} presentation classifies as attack presentation and 0 if presentation classifies as Bona Fide presentation [23]. The Average Classification Error Rate (ACER) is defined as the average of APCER and BPCER, which can be represented as:

$$ACER = \frac{APCER + BPCER}{2} \tag{2.5}$$

2.3 Approaches on face Presentation Attack Detection

This section details the information about and literature survey on the approaches of Presentation Attack Detection (PAD). Although there is not any straight forward neat topology on existing face PAD approaches [20]. Ramachandra and Busch [7] categorize the face PAD algorithms into two categories, namely, Hardware-Based (physical devices to capture or detect presentation attack) and Software-Based (program or algorithm-based PAD detection). Inspired from Ramachandra and Busch [7], Ming *et al.* [22] proposed thoroughly on a face PAD into five category; liveness cue-based,texture cue based methods, 3D geometric cue-based methods, multiple cues-based methods and methods using new trends.Jia *et al.* [24] proposed reflectance/multi-spectral properties based, texture based, shape based, deep features based, and other cues/liveness based methods based on the 3D mask PAD. Following Ming *et al.* [22] and Jia *et al.* [24], we present 3D mask PAD into five categories: liveness clue-based methods, texture clue-based methods, deep learning methods, 3D geometric clue-based methods and multiple/hybrid clue-based methods.

2.3.1 Liveness clue-based methods



(a) Genuine face (b) Masked face

Figure 2.3: rPPG signal from genuine face and mask face (figure taken from Liu *et al.* [25])

Liveness clue-based methods aim to detect a physiological sign of life such as eye blinking, head movement, mouth open/close, face expression changes, lip movement, and pulse from the capture face image(s). Based on liveliness clues,in 2016 Li *et al.* [26] proposed the first facial PAD with pulse-based estimation. In Li *et al.* [26] work, the video frames are decomposed into RGB colour space, applied temporal filtering, and used FFT to convert RGB signal into the frequency domain. Based on the assumption that when RGB signal projects in Power Spectrum Density Curve (PSD) spoofed or fake videos contain multiple random peaks with low power labels, opposing live videos contain dominant peaks [26]. Six-dimensional features from each RGB color channel were extracted and then input for training Support Vector Machine (SVM) classifier. Li *et al.* [26] reported the result of EER and HTERof 4.71% and 7.94% resp in 3DMAD [1] and EER and HTERof 4.29% and 1.58% in their private high Quality REAL-F Mask [26] Attack. A similar approach is followed by Hernandez-Ortega *et al.* [10], but compute the rPPG signal in each second across the video sequence. Hernandez-Ortega *et al.* [10] applied the feature set same as [26] and implemented SVM classifier. To report the result, Li *et al.* [26] implemented 3DMAD [1] with EER of 22.1% and 40.1% in their private HR database [10]. In 2019, Morales [27] proposed an improved version of their work [10] by integrating combination of skin detection module and CHROM [28], method to extract the rPPG signal. Results were based on 3DMAD [1]and private BidaHR [10] with the EER of 18.8% and 26.2%.

In 2016,Liu *et al.* [2] proposed a novel approach for 3D mask face anti-spoofing with a local rPPG correlation model. Liu *et al.* [2] extracted the rPPG signal from the local face region and model the local rPPG pattern by directly extracting the features of signal, such as the signal-to-noise ratio (SNR), maximum amplitude, or power spectrum density [2]. Then this feature are fed into the classifier to made the final decision about the genuine and fake face video attempts. Liu *et al.* [2] conducted the experiment on COMB (combined 3DMAD and self created Supplementary Dataset) dataset and Supplementary dataset(SUP) [2]. Liu *et al.* [2] reported the result achieving EER of 9.9%, HTER of $9.7 \pm 12.6\%$, AUC of 95.5% in COMB datasset, while EER of 16.2%, HTER of $14.7\% \pm 10.9\%$, AUC of 91.7% in SUP dataset. In 2017 Nowara *et al.* [29], proposed PPGSecure, where the rPPG extracted from five faces ROIs, two from the background region and three from the face region(left cheek, right cheek and forehead). The background ROI is subtracted from the face ROI for robustness against noise due to illumination fluctuation. For the feature vectors, the magnitude Fourier spectrum is selected from each filtered rPPG signal from ROI [29]. Then feature vectors were fed into SVM and Random Decision Forest Classifier for facial PAD [29]. On the replay attack dataset, [29] reported accuracy of 100%.

In 2020 Liu *et al.* [30] proposed the rPPG based fast 3D mask face PAD. Liu *et al.* [30] assumed that extracted the local rPPG signal from the live face in terms of their shape phase and amplitude properties. However, for a face mask, this property was different. Based on this assumption Liu *et al.* [30] introduced TSrPPG Feature Operator to measure the similarity using the distance metrics(Euclidean

distance). Besides, to boost the discriminability between genuine and fake face videos, Liu *et al.* [30] also take the background ROI, with the assumption rPPG signal for the masked faces should be identical with the rPPG signal extracted from the background region since both have random noise. In contrast, for the genuine face there is a lesser similarity. Afterwards, the TSrPPG feature from between local facial regions and background regions is concatenated. The three sets of local rPPG similarity feature fed into SVM for the classification [30]. In 3DMAD [1] dataset Liu *et al.* [30] reported HTER of $13.4 \pm 11.2\%$, EER of 13.3%, AUC of 93.8%, similarly on HKBU-MARsV1+ dataset, reported HTER of $22.3 \pm 8.8\%$, EER of 22%, AUC of 85.2%.

Reference	Contribution	Database	Performance
Nowara *et al.* [29]	face and background regions from each face and calculated the spectral density on green channel.	Replay-Attack	Acccuracy 100%
Li *et al.* [26]	green signal as a pulse signal	3DMAD,REAL-F Mask Attack	EER-3DMAD=4.73%, HTER(3DMAD)=7.94%, EER(REAL-F mask)=4.29%, HTER(REAL-F Mask) = 1.58%
Hernandez-Ortega *et al.* [10]	compute the rPPG signal in each seconds across the video sequence	3DMAD,HR	EER-3DMAD = 22.1%,EER-HR = 40.1%
Morales [27]	combination of skin detection module and CHROM a method to extract the rPPG signal	3DMAD,BidaHR	EER-3DMAD=18.8%,EER-BidaHR= 26.2%

Reference	Contribution	Database	Performance
Liu *et al.* [2]	local rPPG correlation model.	COMB,SUP	EER(COMB) = 9.9% HTER(COMB)=9.7 ± 12.6% AUC(COMB) = 95.5%% EER(SUP) = 16.2% HTER(SUP)=14.7% ± 10.9% AUC(SUP)= 91.7%
Heusch and Marcel [31]	long-term spectral statistical features of the pulse signal to discriminate the attack.	Replay-Attack,Replay-Mobile,MSU-MSFD,3DMAD	HTER(ReplayAttack)=13%, HTER(Replaymobile)=25.7% HTER(MSU-MSFD)=20.6% HTER(3DMAD)=19%
Liu *et al.* [30]	local rPPG signal from the live face in terms of their shape phase and amplitude properties	3DMAD, HKBU-MARsV2	EER(3DMAD) =13.3% HTER(3DMAD) = 13.4±11.2% AUC(3DMAD)=93.8% EER(HKBUMARsV+)=22.0% HTER(HKBUMARsV+)=22.3 ± 8.8% AUC(HKBUMARsV+)=85.2%

Table 2.1: Related work about liveness clue based 3D face mask Presentation Attack Detection (PAD)

2.3.2 Texture clue based methods

Unlike liveness clues, texture clue methods explore the micro-textual properties of the biometric face samples presented on Face Recognition System (FRS). Analyzing micro-textual properties, texture clue method performed binary classification among genuine and fake faces.The most popular and widely used texture clues-based method to overcome face Presentation Attack (PA) is Local Binary Pattern (LBP) [32]. LBP-based does not rely on any physical model (Lambertian models [33]) but captures local primitives (LBP features) due to the differences between the surface properties and light reflection between a real face and a plane photo

attack.

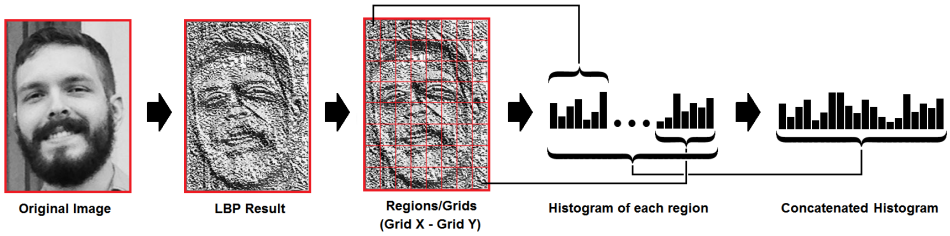


Figure 2.4: Texture based LBP with histogram calculation(figure taken from `https://towardsdatascience.com/face-recognition-how-lbph-works-90ec258c3d6b`)

Kose and Dugelay [11] first implement the static texture-based approach in PAD to detect 3D mask attacks, make used of texture or depth maps of the input image to distinguish 3D mask or live input. Kose and Dugelay [11] achieved 88.1% accuracy using the Morpho database. The Kose and Dugelay [34] proposed an improved version of their previous work by score level fusing of both texture images and depths maps. Thus, accuracy increased by 93.5% with the same Morpho database. Erdogmus and Marcel [35] also proposed their work, based on various LBP operators with different classifiers(Linear Discriminant Analysis and Support Vector Machine). The result from their proposed method showed that classification of block-based LBP features with the Linear Discriminant Analysis gives the best results for both colour and depth images [18].
In 2014,Raghavendra and Busch [36] proposed a novel approach for detecting 3D spoofed attempts; both local features, which corresponds to the eye (periocular) and nose region that is expected to provide a clue on the presence of the mask and micro-texture variation as a global feature were extracted using Binarized Statistical Image Features. Raghavendra and Busch [36] showed satisfactory performance with Half Total Error Rate (HTER) of 0.03% on a linear Support Vector Machine (SVM) in 3DAMD [1] using the weighted sum rule before making the decision about a real face or an artefact [36]. Similarly, Siddiqui *et al.* [37] combine with motion estimation using the Histogram of Oriented Optical Flow features on both 2D and 3D face spoofed attempts and achieve an Equal Error Rate (EER) of 0% on the 3DMAD [1] database. Pinto *et al.* [38] introduce new concepts aiming to detect photo, videos and 3D masks. In their work, a discriminative signature from noise and artefacts while recapturing biometrics samples is generated and characterize these artefacts by extracting time-spectral feature descriptors from the video as low-level feature descriptors. Pinto *et al.* [38] use the visual codebook concept to find mid-level feature descriptors computed from the low-level ones. Pinto *et al.* [38] result the accuracy of 96.16% on the 3DMAD [1]. Agarwal *et al.* [39] proposed block-wise Haralick texture features from redundant discrete wavelet transformed frames obtained from a video, showing the satisfactory performance on HTER of 0% on 3DMAD.

Reference	Contribution	Database	Performance
Kose and Dugelay [11]	Multi-sacle LBP texture images Multi-sacle LBP depth map images	Morpho	Accuracy = 88.1% Accuracy = 86.0%
Kose and Dugelay [34]	score level fusing of both texture images and depths maps	Morpho	Accuracy = 93.5%
Erdogmus and Marcel [18]	block-based LBP features for both color and depth image	3DMAD	HTER = 0.95%
Raghavendra and Busch [36]	local features and global feature using Binarized Statistical Image Features	3DMAD	HTER = 0.05%
Siddiqui *et al.* [37]	motion estimation using the Histogram of Oriented Optical Flow features	3DMAD	EER = 0%
Pinto *et al.* [38]	time-spectral feature descriptors and visual codebook concept to find mid-level feature descriptors	3DMAD	Accuracy= 96.16%
Agarwal *et al.* [39]	block-wise Haralick texture features	3DMAD	EER = 0%

Table 2.2: Brief information about texture based 3D face mask PAD.

2.3.3 Deep learning methods

Deep learning methods are implemented to abstract the distinguish between the discriminative appearance features for the 3D face mask and genuine face. Menotti *et al.* [40] proposed two approaches for detecting spoofed attempts: hyperparameter optimization of network architecture (AO) and learning filter weights via backpropagation (FO). Menotti *et al.* [40] conduct the AO approach on the 3DMAD dataset and achieve EER 0%, while FO approaches and combination of AO+FO scheme achieved HTER of 24% and 40%, respectively. Similarly, Lucena *et al.* [41] proposed FAS-Net, which is based on transfer learning using pre-trained VGG-16 model architecture. Menotti *et al.* [40] showed the excellent performance of 0% HTER on the 3DMAD dataset. Feng *et al.* [42] proposed the hybrid approach for both 2D and 3D spoofing detection and combined image quality ques(Shearlet) and motion cues(dense optical flow) with the use of hierarchical network architecture. The result from their network achieved 0% HTER in the 3DMAD [1]. Man-

jani *et al.* [43] introduced the first silicone mask database and proposed a novel multilevel deep dictionary which is formulated to learn by efficient greedy layer by layer training approach followed by SVM to classify the genuine and spoofing attacks. Results from their work are promising with 0% HTER on 3DMAD and 13.44% HTER on SMAD [43]. Liu and Kumar [44] introduce convolution neural networks based on 3D face masks under visible and near-infrared(multi-spectral) illumination using two separate sensors. The results from their experiment indicate that near infrared-based imaging of the 3D mask is better as compared under visible illumination.

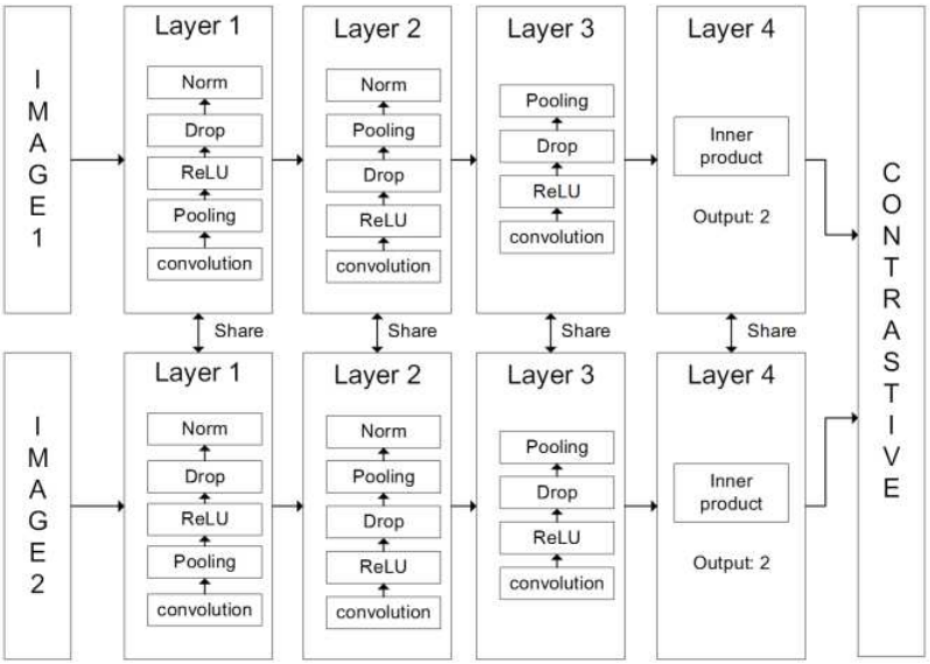


Figure 2.5: Convolution Neural Network(CNN) based 3D face masks under visible and near infrared (multi-spectral)(Figure taken from Liu and Kumar [44])

Reference	Contribution	Database	Performance
Menotti *et al.* [40]	hyper-parameter optimization of network architecture(AO) and learning filter weights via back propagation(FO)	3DMAD	HTER = 0%(AO),HTER = 0%(BO),HTER = 40%(AO + FO)
Lucena *et al.* [41]	FAS-Net transfer learning using pre-trained VGG-16 model	3DMAD	HTER = 0%
Feng *et al.* [42]	Combine image quality ques(Shearlet) and motion cues(dense optical flow) with the use of hierarchical network architecture	3DMAD	HTER= 0%
Manjani *et al.* [43]	multilevel deep dictionary learning	3DMAD SDMAD	HTER(3DMAD) = 0.95% HTER(SDMAD) = 13.1%
Liu and Kumar [44]	convolution neural networks based 3D face masks under visible and near infrared(multispectral) illumination	Private data	ACER = 3.19%

Table 2.3: Brief information about deep learning based 3D face mask Presentation Attack Detection (PAD)

2.3.4 3D geometric clue-based methods

Three-dimensional geometric cues calculate 3D geometric features from presentation images distinguishing genuine and fake face images. Basically, with a genuine face presented biometric sensor possess better 3D structure characteristics than 2D planer Presentation Attack (PA) (e.g., photo attack or video replay attack). Tang and Chen [46] applied 3D shape analysis based on principle curvatures measures that describe the meshed facial surface. The experiment was conducted on Morpho[3], and FRGcv2 dataset, with the EER of 6.91%. Hamdan and Mokhtar [47] proposed Angular Radial Transformation to extract a feature vector from the whole image and input it to a Maximum Likelihood classifier for discriminating

[3]www.morpho.com

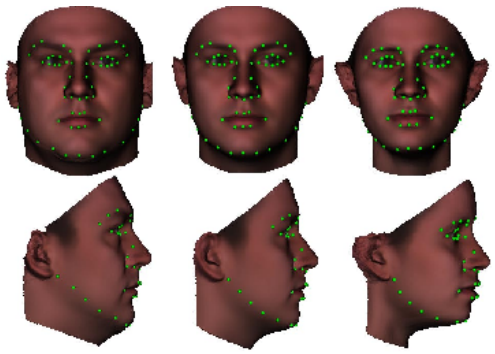


Figure 2.6: 3D Morphable shapes of face (figure taken from Zhou *et al.* [45])

between genuine and fake faces. Results were achieved using a 3DMAD with the HTER of 0.91% [47]. The same author Hamdan and Mokhtar [48] proposed another Presentation Attack Detection (PAD) against mask spoofing attacks, with a combination of Legendre Movements Invariants decomposition and the linear discriminant analysis for characteristic features extraction, and the maximum likelihood for classification on the 3DMAD dataset. The obtained spoof false acceptance rate was close to 65%, which proves that approach is vulnerable to 3D masks attack [48]. Wang *et al.* [49] proposed novel methods to detect 3D spoofed attempts, which combines texture as well as shape features. Precisely, geometry cues are reconstructed from RGB images through 3D Morphable Model. Then, hand-crafted elements and deep ones are extracted to represent texture and shape differences between real and fake faces with EER 0%.

Reference	Contribution	Database	Performance
Tang and Chen [46]	principle curvatures measures which describes the meshed facial surface	Morpho and FRGcv2	EER = 6.91%
Hamdan and Mokhtar [47]	Legendre Movements Invariants(LMI) decomposition and the linear discriminant analysis for characteristic features extraction,	3DMAD	SFAR = 65%
Wang *et al.* [49]	reconstruct geometry cues from RGB images through 3D Morphable Model	3DMAD	EER= 0%

Table 2.4: Brief information about 3D geometric based 3D mask detection.

2.3.5 Multiple/Hybrid clues-based methods

Multiple clues, in other words, a hybrid approach, combine multiple clues to address facial PAD. Assume a multi-modal system that is more difficult to spoof than a uni-modal system. In 2017, Pan *et al.* [50] proposed the two collaborative approaches; liveness clues (eye-blinking detection model) based on Conditional Random Field (CRF) and texture clues (check the coherence between LBP features of background region of the subject and actual background of the reference image). Another hybrid clues approach is proposed by Feng *et al.* [42], where static texture-Shearlet based image quality features [51] [52],and a scenic motion clues, face motion based on dense optical flow [53], trained into neural network and then fine-tune with PAD datasets. In 2018, Liu *et al.* [54] used CNN Recurrent Neural Network (RNN) architecture and fused Remote Photoplethysmography (rPPG) cue and pseudo-depth map clue for face Presentation Attack Detection (PAD). Similarly, Atoum *et al.* [55] fused patch-based texture clue and pseudo-map clue in two-stream CNN for facial Presentation Attack Detection (PAD).

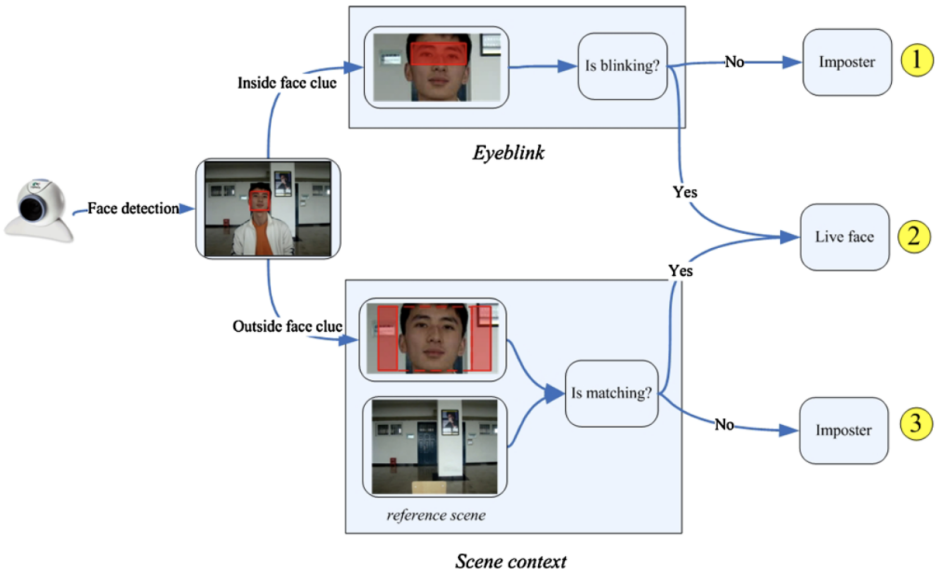


Figure 2.7: Liveness clue (Eye-blinking detection model) and texture clue based hybrid approach(Figure taken from Pan *et al.* [50]).

2.4 Remote Photoplethysmography based pulse measurement

This section describes essential steps for raw PPG extraction, which is concerned with face detection, Region of Interest selection, colour channel decomposition for

raw rPPG extraction. Remote Photoplethysmography is a contact-less approach for the physiological measurement of human signs(Pulse Rate, Pulse Rate Variability, Heart Rate, Heart Rate Variability, Respiratory Rate). The idea behind this approach is: when the skin surface is illuminated with light, then there is a subtle change in colour variations due to the blood pulse. Sikdar *et al.* [56], classify Remote Photoplethysmography estimation techniques into two classes:(1) Image-guided techniques (2) motion-guided techniques. In the image-based technique, the pulse signal is estimated from colour variation due to the change in the intensity of light from the skin surface in each cardiac cycle, in opposing motion-based technique extracted the pulse signal from the subtle head oscillations, which occurs due to blood pluming to the aorta in each cardiac cycle. We followed the image-based technique for rPPG signal extraction, presented in literature Rouast *et al.* [15], and Wang *et al.* [57], we subdivided rPPG signal extraction framework into two key stages: Face video processing, Estimation of rPPG signal.

2.4.1 Face video processing

The digital camera used to take the videos are mainly web cameras and portable device cameras. According to the Nyquist-Shannon sampling theorem, the minimum frame rate to capture the Heart Rate(HR) is eight frames per second(fps).

Face detection and Region of Interest selection

Raw rPPG signal extraction starts with Region of Interest detection; the idea is to detect the face or subregion(s) of the face in a video frame, where the rPPG signal is rich to found. The step proceed with the face detection, in most of the rPPG approach [58] [59][60] [61] [62] [63] used Viola and Jones [64] as a face detector algorithm is. Viola-Jones [64] which is used a cascade of features to classify faces and returned bounding box of the face. It is also available in OpenCV Computer Vision[4]. The ROI might be the combination sub-region(s) of the face such as cheek, face, forehead. In the preliminary phase of the rPPG study, ROI bounding boxes were selected manually from one frame to another [65]. Alternative to ROI bounding box, skin detection method is usually applied, where skin region pixel is extracted within the bounding box of the face extracted using face detection algorithm [66][67][68]. In a recent study, ROI optimization is undertaken to improve the raw signal, where the ROIs are captured in smaller patches from the forehead or cheek regions. Then quality indices(SNR) is evaluated from all patches and determine the candidate ROIs [69][70] [71] [72]. To deal with the subject in motion, accurate localization and tracking of facial landmarks in the video frames are crucial. Simple approach of ROI tracking is to re-detect ROI from every video frame, this approach is followed by[73][74][75][76][77].
In [67] [78], Kanade-Lucas-Tomasi(KLT) [79] face tracking algorithm was implemented to localize the face in every frame of the video which is more automatic

[4]https://opencv.org/

then re-detecting ROI(s). Li *et al.* [21] and Kumar *et al.* [70] fused good-feature-to-track [80] for selecting good feature points and KLT to track these features across the video. Similarly Feng *et al.* [81] implemented Speeded-up-robust-feature [82] for the facial feature point selection and KLT for tracking these feature during the subject in motion. To update or track the face skin pixel across the frame Lee *et al.* [68] used kernel [83] approach. Another new approach is applied by Wang *et al.* [84] used tracking-by-detection with kernels [85] to compensate the rigid subject motion across the video frames.

Colour channel for raw rPPG signal extraction

In the colour-based method, raw signal extraction depends on the colour pixel value captured by the camera. Based on the existing literature for raw signal extraction, the colour pixel value from the Region of Interest (ROI) across the video frame is calculated mainly on three colour space Red-Green-Blue(RGB), Hue-Saturation-Intensity(HSI), and YCbCr, where Y stands for the luminance component, Cb, Cr refer to blue-difference and Red-difference chrominance components respectively. Tsouri and Li [86] used H channel for raw pulse signal extraction. Sahindrakar *et al.* [87] investigate the pulse detection in the YCbCr channel and conclude YCbCr produces a better result than HSI. Among the three colour channels, RGB colour space is much popular raw pulse signal extraction. Based on literature survey [73][66] [65] implement RGB channels for the raw rPPG estimation, while [21] [74][70] [72] use green channel and [77][88] combine Red and green channel to estimate pulse signal. A novel approach is proposed by Rahman *et al.* [63], the RGB colour space was converted into three independent signal Lab, where L represents the lightness of the images and a (red/green) and b (yellow/blue) represent the combination of other colour channels. After the pixel(intensity) value estimation, from the ROIs in each frame, the value of each colour channel is calculated by averaging each colour pixel value from frame ROI. The method is also called spatial pooling or spatial average. Spatial average is most common in raw pulse extraction and followed in different literature's [74][67][70]

2.4.2 Estimation of rPPG signal

Upon reaching this step, raw RGB signal is estimated; it is assumed that the signal consists of illumination and motion noise which need to be removed. As mentioned above, extracted raw RGB signal is coupled with unwanted noise caused by illumination variation, subject motion, and another factor. Hence it is necessary to exclude those unwanted noises from the signal for robust and accurate pulse measurement. Most researchers use one or more filters based on the literature, which is a design based on noise frequencies and range of Heart Rate frequencies. Wang *et al.* [57] classify the noise reduction filter into two types a) temporal filters (remove the irrelevant information from the signal, thus including colour frequencies with the range of heart beat) (b) Background noise estimation(undertaking

background signal to remove the illumination noise). The temporal filtering includes bandpass filters, detrending, and moving averages. In addition to that, statistical methods such as centralization, normalization, detrending, and mean average techniques were also introduced. Both centralization and normalization are applied to remove, periodicity of the signal; centralizing refers to mean values of the signal that are calculated first and are subtracted from the individual values; the normalization technique adds the step of dividing the signal with standard deviation. The bandpass filter is applied within the range of HR frequencies based on blood pulse per minute. There is no concise Heart Rate for the measurement as Heart Rate changes per human age, health condition; due to this, HR frequency is also not concrete, so assume different frequency ranges. However, most of the frequency range is within 0.6Hz to 4Hz [10][26][89]. Similarly, the moving average filter sliding window size is defined, and the average value is calculated within that sliding window. Detrending is more applicable for signal smoothness by removing the long-running trend from the signal.Verkruysse *et al.* [65], Balakrishnan *et al.* [59] , Irani *et al.* [60], Kumar *et al.* [70] applied butter-worth bandpass filter in fourth-order butter-worth coefficient in phase neutral digital filter[5]. One or more filters were applied for noise reduction,De Haan and Van Leest [90] applied normalization and bandpass filter;Li *et al.* [21] fused three filters detrending, moving average, and bandpass filters. Similarly McDuff *et al.* [75] introduced detrending and normalization. Adaptive filters correspond with the concept of background noise estimation, which assumes that, first, ROI(s)and background share the same white light and background remain static. Based on this assumption Feng *et al.* [81], Feng *et al.* [91] applied the adaptive bandpass filters.
RPPG methods follow the noise reduction step; basically, the rPPG method refers to rPPG signal extraction from the pre-processed colour channel(common to the RGB colour channel). Based on the rPPG category presented on Wang *et al.* [28] , we categorize the rPPG method into two groups: Dimension reduction/BSS approach and Model-based methods.

- Dimension reduction/BSS approach
  A dimensional reduction algorithm applied as rPPG signal extraction methods since rPPG is concatenated with a linear combination with different sources. The classical linear algorithms for dimensionality reduction are Blind Source Separation(BSS) methods included two popular approach: Independent Component Analysis [92] and Principal Component Analysis(PCA) [93]. In ICA, linear separation of the sources is accomplished by maximizing the statistical independence of the sources. Joint Approximate Diagonalization of Eigen-matrices (JADE) [94] among the ICA algorithms implemented by Poh *et al.* [61]. The work followed on the ICA approach [73] [75] [90]. Compared with ICA Principal Component Analysis (PCA), compute to finding the direction, on the data which have maximum variance. Based on PCA [59][60] proposed for detecting pulse signal in motion-based method,

[5]filtfilt in MatLab

where the frequency spectra of PCA with the highest periodicity is selected. To handle multiset(colour channel signal from multiple facial sub-regions), Joint BSS(JBSS) methods had introduced in Guo *et al.* [95] apply Independent Vector Analysis(IVA) to analyze the colour signal from the multiple sub-regions.

- Model based methods
  As oppose to dimensional reductions, model-based methods use the information about colour vectors components to assure the demixing of the sources. Among the various approaches based on the model-based methods, start with the simplest method called the Green method. In works [21][74] [70] [78], it has been reported that the green channel provides the strongest PPG signal. It is the simplest method because it calculates the average colour intensity of the green channel value from the averaging RGB colour channel in ROI(s). In 2013 De Haan and Jeanne [66] proposed the novel method CHROM, which reduces the dimensionality of demixing by eliminating the specular component (colour or illuminate with no pulse signal) by the colour difference. With the same goal as CHROM in 2016 Wang *et al.* [28] introduce Plane-Orthogonal-to-Skin(POS), which define the plane orthogonal to skin tone in a temporarily normalized RGB plane. Similarly, in 2014 De Haan and Van Leest [90] proposed a novel Blood Volume Pulse(PBV) method, which utilized the signature of blood volume change by restricting all the colour variations to the pulsatile direction. In 2018, Pilz *et al.* [96] proposed the novel method called the Local Group Invariance (LGI) method to find a new feature space from the raw colour signal in which the rPPG method is most robust to subject movements and lightness variations. Wang *et al.* [97] Spatial Subspace Rotation" (2SR or) SSR, which is based on the assumption of 1) spatially redundant pixel-sensors of a camera and 2) a well-defined skin mask, our core idea is to estimate a spatial subspace of skin-pixels and measure its temporal rotation for pulse extraction, which does not require skin-tone or pulse-related priors in contrast to existing algorithms.

2.4.3 Machine learning approach for rPPG estimation

In a recent development, rPPG based HR estimation is applied with machine learning techniques. Song *et al.* [98] classify the existing ML-based rPPG method into two categories: feature-decoder and end-to-end methods. According to Song *et al.* [98] feature-decoder method needs to define hand-crafted feature, and overall performance depends on the quality of feature maps. Niu *et al.* [99] proposed the feature decoder approach. In their proposed network, ImageNet [100] is implemented, thus generating a large amount of synthetic rhythm spatial-temporal maps to pretrain deep heart rate regression model. Then the pre-trained Model was transferred to the real HR estimation task. Similarly, Niu *et al.* [101] also generates spatial-temporal maps from small video clips sampled from the original video using a fixed sliding window; afterwards, the data augmentation method is

applied and feed into ResNet-18 [102] architecture to estimate the HR per video clip. In Qiu *et al.* [103] applied a different approach and fused Eulerian Video Magnification (EVM) [104] and CNN; specifically, EVM is used to extract the feature image that corresponds to the heart rate information within a time interval. The extracted feature is fed into a CNN is then applied to estimate HR from the feature image, which is formulated as a regression problem. However, in end-to-end methods learns the feature from the network itself, the overall Model is interpreted as the black box referring hard to interpret every step. In 2018 Chen and McDuff [105] first, propose the end-to-end system called DeepPhys to estimate the HR from the video; authors introduced a soft attention mask to learn simultaneously, thus improving the estimation. Another approach based on the end-to-end method is introduced by Špetlıék *et al.* [106] with a two-step Neural Network(NN). The first step is called the extractor step. The sequences of images produce a sequence of scalar output called an NrPPG signal; afterwards, this signal is fed into the second step called the estimator, which outputs the HR. The input of the network is T-frame face images with RGB channels. Similarly, Yu *et al.* [107] proposed spatial-temporal networks for rPPG estimation; the network is designed with several convolutions and pooling operations and feed with T-frame face images in RGB channels [107]. Finally, the latent manifolds are projected into signal space using channel-wise convolution operation with $1 \times 1 \times 1$ kernel to generate the predicted rPPG signal length.

2.5 Remote Photoplethysmography for face PAD

This section is a review of the feature set from the rPPG signal, which can distinguish fake and genuine face videos when feeding into the binary classifier. The feature set was extracted from the genuine videos and fake videos provided with a binary label. From the estimated rPPG signal, features need to be extracted to classify whether the pulse signal is estimated with the spoofed face videos or genuine face videos. In 2016 Li *et al.* [26] proposed the first facial PAD with pulse-based estimation. Li *et al.* [26] works projects the video frames into RGB colour space, and computed Power Spectrum Density (PSD) Curve.Li *et al.* [26] assumed that spoofed or fake videos contain multiple random peaks with low power labels in opposing genuine video contain dominant peaks. Based on this assumption, two features were extracted from each colour channel projecting into PSD: maximum power value and the ratio of maximum value to the total Power. Altogether, Li *et al.* [26] created selected six-dimensional features from each RGB colour space and input for training the SVM classifier.
A similar approach is followed by Hernandez-Ortega *et al.* [10], the distinctive approach followed by them, are Hernandez-Ortega *et al.* [10], compute the rPPG signal in each second across the video sequence. Hernandez-Ortega *et al.* [10], applied the feature set same as [26] and implemented an SVM classifier. The same author Morales [27] proposed an improved version of their work in 2019 and implemented a combination of skin detection module and CHROM De Haan and

Jeanne [66], a method to extract the rPPG signal [27]. The nine different features based on time and frequency domain feature rPPG signal is introduced on their proposed method. In the time domain features, Morales [27] introduced two features zero-crossing rate (Number of times the signal crossed the zero value) and quotient between the temporal maximum and minimum. Similarly, in the frequency domain, Morales [27] introduced seven features: maximum power response from the rPPG signal in the PSD curve, the ratio of maximum Power and the total Power in (0.6–4) Hz frequency range, mean value of rPPg signal, Mean value of each frequency component multiplied by its magnitude, Sum of the N biggest values of the frequency signal divided by N, sum of the energy between 0 Hz and 4 Hz and sum of the energy between 2 Hz and 4 Hz.

In 2016, Liu *et al.* [2] proposed a novel approach for 3D mask face anti-spoofing with a local rPPG correlation model. The rPPG signal is extracted from the local face region and model the local rPPG pattern by directly extracting the features of the signal, such as the signal-to-noise ratio (SNR), maximum amplitude, PSD [2]. Then this feature is fed into the classifier to made the final decision about the live and spoofed video attempts. In 2017, Nowara *et al.* [29], proposed PPGSecure, where the rPPG signal was from extracted the face five ROI, two from the background region and three from the face region (left cheek, right cheek and forehead). With the background ROI, Nowara *et al.* [29] subtract the face ROIs for robustness against noise due to illumination fluctuation. For the feature vector, Nowara *et al.* [29] selected the magnitude Fourier spectrum of each filtered PPG signal and concatenated these spectral features from three facial regions and two background regions to obtain spectral feature vector for classification [29].

In 2018, a new approach was proposed by Heusch and Marcel [31] which utilizes long-term spectral statistical features of the pulse signal to discriminate the attack. Heusch and Marcel [31] applied DFT in each window to calculate the DFT coefficient vector. When a DFT coefficient is lower than 1, it is clipped to 1 such that the log-magnitude remains positive. Afterwards, Heusch and Marcel [31] compute the mean and variance of the (Discrete Fourier Transform) DFT coefficient of each window. Then mean and variance vectors of each window were concatenated to represent a single feature vector. The features vector is then fed into SVM to classify a given video sequence as fake or live.

In 2020 Liu *et al.* [30] proposed the rPPG based fast 3D mask face PAD. Liu *et al.* [30] assumed that extracted the local rPPG signal from the live face in terms of their shape phase and amplitude properties; however, for a masked face, this property was different. Based on this assumption, Liu *et al.* [30] introduced the TSrPPG feature operator to measure the similarity using the distance metrics (Euclidean distance). Afterwards, the TSrPPG features: amplitude, gradient and phase, from between local facial regions and background regions are concatenated.

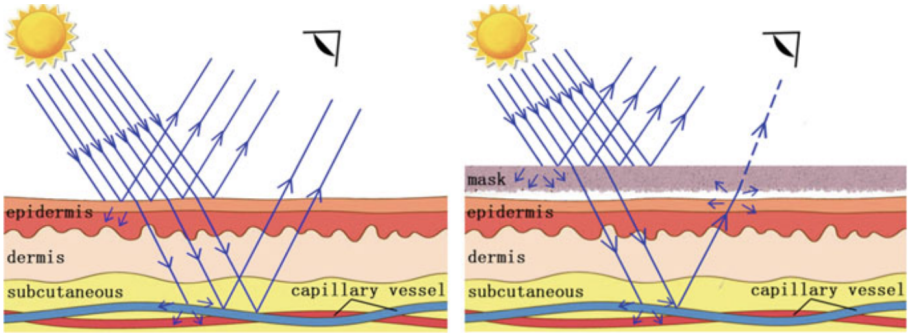
Chapter 3

Background Methodology

The chapter enlightens about the background methodology of the thesis work. The chapter comprised of seven sections; the first section informs about the face detection and tracking algorithm to detect the face from the face videos in the proposed method; the second section provides the information about colour channel and ROI selection for the best extraction of rPPG signal, third section details on the signal preprocessing of raw rPPG signal, fourth selection informed about the rPPG signal extraction method from the RGB colour space, the fifth section introduced spectral method on rPPG signal for feature extraction, the sixth section includes brief background about the binary classifier.

3.1 Principle of rPPG and applicability for 3D mask PAD

The rationale behind Remote Photoplethysmography (rPPG) approach is when the light source(s) illuminates skin, then some portion of the light penetrates through skin layers reach the capillary vessel; based on the amount of haemoglobin in the blood, that small portion of the light is absorbed, causing subtle colour change(also depend upon the volume of blood under the observable skin surface). The two distinctive approaches for rPPG signal extraction is motion-based and colour-based, where colour based focuses on extracting the rPPG signal by determining the colour variation caused by blood volume. In contrast, the motion-based approach focuses on head motion caused by the pulse and other involuntary head movements. The intensity-based approach is more popular and widely used in the research field as opposed to motion-based. We followed Rouast *et al.* [15], general classification of existing rPPG signal extraction on colour-based method. The general framework for the rPPG from the face video includes face detection, ROI selection, ROI tracking, Raw rPPG signal extraction, signal preprocessing (filtering and rPPG signal extraction methods). Since we are focusing on 3D face mask PAD, we tried to obtain only pulse signal from the face region and exclude out the heart rate estimation computation. The main principle to implement the pulse signal is describes as; from the genuine faces pulse signal is generated with high amplitudes as the light sources directly illuminate the skin surface as opposed, on



(a) Genuine Face (b) Masked Face

Figure 3.1: Comparison of rPPG from the genuine face and mask face (figure is taken from Liu *et al.* [20]).

mask attack; first, the light sources need to penetrate the masked surface before reaching to the skin and blood capillaries which result in very noisy pulse signal with low amplitude Liu *et al.* [20]. Based on this principle, we proposed the rPPG based PAD for 3D mask face videos. The mathematical approach is presented by Liu *et al.* [20] on the analysis of rPPG signal for genuine and 3D mask face videos. For the genuine face light(s) directly illuminates the skin surface and reach to capillary vessel and rPPG signal penetrates skin to be observed.The observed rPPG signal from genuine face is represented as:

$$\hat{s_l} = T_s I s + \epsilon \tag{3.1}$$

where $\hat{s_l}$ represent observed rPPG signal, s represent the raw rPPG signal from capillary vessel, T_s represent the transmittance properties of skin, I denotes, mean intensity of facial skin pixel and ϵ represent environmental noise [20].
For the mask face, the light(s) need to penetrate first face mask layer before reaching to skin surface and capillaries. Also, rPPG signal need to penetrate the face mask again to capture by camera.

$$\hat{s_m} = T_s T_m I_m s + \epsilon \tag{3.2}$$

$$I_m = T_m I \tag{3.3}$$

Combining above equation:

$$\hat{s_m} = \hat{s_m} T_m^2 + \epsilon \tag{3.4}$$

where $\hat{s_m}$ represent observed rPPG signal from face mask, T_m represent the transmittance properties of skin, $I_m = T_m I$ denotes mean intensity of skin under face mask [20].

3.2 Selection of face detection and tracking algorithm

The selection of faces from the video is a critical factor influencing the rest of the framework as selection non skin pixel result poor rPPG signal estimation. Hence, it

is essential to detect the best face region frame across face videos. Taking video as an input, it is necessary to detect the faces across the video frames; as mentioned in chapter 2, based on the existing literature's, most of the research work adopted the Viola-Jones [64] algorithm for face detection.

3.2.1 Multi-Task Cascaded Convolution Neural Network

One of the popular approaches based on the deep learning method achieving state-of-art results on the range of benchmark datasets is Multi-Task Cascaded Convolution Neural Network (MTCNN) [108]. Zhang *et al.* [108] highlighted that the network could handle pose variations in images, occlusion, illuminations, and extreme lighting to some extend. The network architecture is comprised of three networks in a cascade structure, where the outputs from the previous steps are fed as input to the next stage before feeding the image onto the networks [108]. Initially, it does some preprocessing where the input image is resized to different scales to build an image pyramid [108].

Stage 1

The First stage is called the Proposal Network (P-net), a complete convolution network candidate facial window is obtained with their bounding box regression vectors [108]. The obtained window is calibrated according to the estimated bounding box regression vector [108]. After that, highly overlapped candidates are merged with Non-Maximum Suppression (NMS) [108].

Stage 2

The output from the first step is then fed into another Convolution Neural Network (CNN) called Refine Network, also called R-Net. In this stage, false candidates generated from the P-net are rejected and further calibration with bounding box regression and Non-Maximum Suppression. This stage is called the O-Net or Output Network, where it refine the image with more detailed face regions [108].

Stage 3

Lastly, similar with respect to second stage, but more of action in this stage to describe the face in more detail manner [108]. The network in this stage output five facials landmark's positions [108].
Zhang *et al.* [108] mentioned three major task to train the their CNN detectors; face/non-face classification , bounding box regression and facial landmark localization. The facial landmark localization is formulation as binary classification problem, where each sample x_i make use of entropy loss:

$$L_i^{det} = -(y_i^{det} log(p_i) + (1 - y_i^{det})(1 - log(p_i))) \tag{3.5}$$

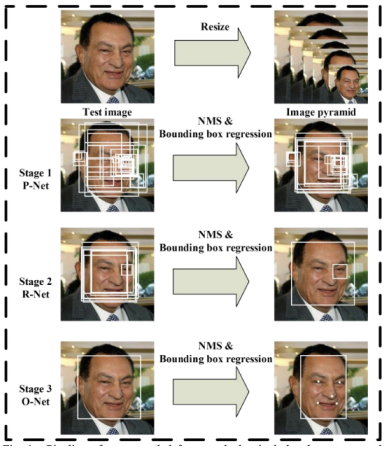


Figure 3.2: Cascaded Network architecture in MTCNN. Figure taken from Zhang *et al.* [108].

p_i indicates probability value produced by the network to a real face and$y_i^{det} \in 0, 1$ denotes the ground truth label [108].
The bounding box bounded the face in four regions left top, height and width localizing the face in each candidate window [108]. For the learning objective regression problem is formulated with the euclidean loss for each sample x_i [108]. With the similar approach of regression problem and euclidean loss function, facial landmark coordinates were obtained. The five facial landmarks obtained from the MTCNN techniques are left eye, right eye, nose, left mouth corner and right mouth corner.

3.2.2 Kalman filter for face tracking

According to Welch, Bishop *et al.* [109], Kalman filter estimates the face tracking process with a feedback control environment; first, the Kalman filter estimates the process state at some time and then obtain the feedback from that process state in the form of a noisy environment [109]. The Kalman filter equation in-cooperate two steps cycle: time update equation (predictor) and measurement update equations (corrector) [109]. The time update equation aims to project the current state into forwarding time state; concurrently, error covariance is computed to obtain a prior estimate for the next step [109]. The measurement update equation aims to provide feedback to obtain an improved predicted time equation.

Derivation of state matrix

In the kalman filter the state of an object is represent with the state matrix and process covariance matrix. The step of Kalman filter start with the state estimation of discrete-time controlled process $x \in \Re^n$ which is governed by the linear

stochastic difference equation:

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \tag{3.6}$$

with the measurement $z \in \Re^m$

$$z_k = Hx_k + v_k \tag{3.7}$$

In the above equation v_k and $w_k - 1$ represent the process covariance noise and measurement covariance noise. Similarly $n \times n$ matrix A defines state at previous time step k-1 to the current time step, and the absence of either a driving function or process noise [109]. The $n \times l$ matrix B defines optional control input $u \in \Re^l$ to the state x and $m \times n$ matrix H represent that state to the measurement z_k, might changes with each time step [109].

New predicted state

The state matrix is demonstrated by two state: first, prior state $\widehat{x_k}^-$ estimate at step k given knowledge of the process prior to step k, and second posterior state $\widehat{x_k}$ estimate at step k in a given measurement. Then a priori estimate error covariance is:

$$P_k{}^- = AP_{k-1}A^T + Q \tag{3.8}$$

Then a posteriori estimate error covariance is:

$$P_k = (I - K_k H)P_k{}^- \tag{3.9}$$

In Welch, Bishop *et al.* [109], find the equation that computes a posterior state as a linear combination of a prior state and a weighted difference between an actual measurement z_k. In more simplified version, the equation is represented by

$$\hat{x_k} = \hat{x_k}{}^- + K(z_k - H\hat{x_k}{}^-) \tag{3.10}$$

The difference $z_k - H\hat{x_k}{}^-$ also called innovation or the residual and reflect the discrepancy between the predicted measurement $H\hat{x_k}{}^-$and the actual measurementz_k.

Kalman gain

The kalman gain or blending factor minimized the a posteriori error covariance, which is obtained by taking derivative of the equation P_k with respect to K and setting that result equal to zero then solving for K.

$$K_k = P_k{}^- H^T (HP_k{}^- H^T + R)^{-1} \tag{3.11}$$

In Qian *et al.* [110] implement the Kalman filter as the face tracking algorithm. With the face detection algorithm, the location and size of the face are extracted out. Assume that face center position x_c, y_c and size of face is (w,h). Then Kalman filter smooth temporal projector in the face centre position x_c, y_c and size of the face is (w,h) [110].

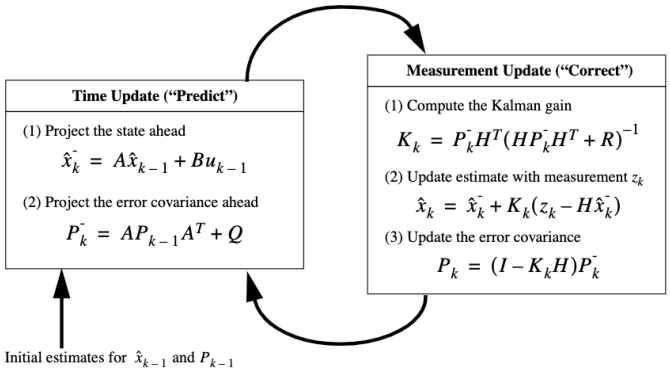


Figure 3.3: Complete Operation of Kalman filter. Figure taken from Zhang *et al.* [108]

3.3 Selection of face colour channel and ROI

As the face is detected from the frames and tracked across the videos, it is essential to select those facial regions (s) that can provide the most informative signal components to estimate the raw pulse signal. Region of Interest (ROI) selection is a critical process for rPPG estimation and prevents face segmentation errors, reduces noise, and ultimately preserves reliable pulse detection components. From related work in chapter 2, most of the predefined rectangular coordinates of face region(s) include the forehead, cheeks, nose and lower face regions for ROI. In addition to rectangular coordinates-based ROI selection, another approach called skin detection module [111] is also implemented in Hernandez-Ortega *et al.* [89] work. In the skin detection algorithm proposed by Kolkur *et al.* [111], first, the given image is converted into the two-dimensional matrix, with width and height values of images. Each entry of the matrix represents the pixel of the picture. Then ARGB value is calculated by representing each image pixel as a 32-bit value. The alpha value is calculated by shifting right by 24 bit of red, green, blue, and alpha (opacity channel) and getting alpha value. A pixel with a value of 0 % in its alpha channel represents transparency. In comparison, a value of 100 % in the alpha channel represents an opaque pixel. Similarly, to compute the red value, they shift by 16 bits; for the green matter in the pixel, they change by 8 bits. The remaining value in the pixel is the blue colour. The shifting procedure is applied to each pixel of the image. To make the design more robust, ARGB value is converted into HSV as well as YCbCr value using conversion factors and built-in functions [111]. To decide whether the pixel is a skin pixel or not, a comparative analysis of the HSV, YCbCr, and ARGB values of each pixel is done with typical values of a skin pixel. According to [111] algorithm, range of skin pixel in different colour space are represented as:

$$0.0 <= H <= 50.0 \text{ and } 0.23 <= S <= 0.68 \text{ and}$$

$$R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and } R > G \text{ and } R > B$$

$$\text{and } |R-G| > 15 \text{ and } A > 15$$

From the selection of ROI(s), most of the literature has implemented RGB colour signals. At the same time, some researchers extracted the green channel from the RGB colour channel [26]. Since RGB colour space is popular and practical to implement to extract pulse components, then YCbCr and HSV signal. RGB colour space is a widely used colour space for representing digital images consisting of three primary colours Red, Green, and Blue. Any other colour can be obtained by mixing each base colour.

3.4 Choice of signal preprocessing filters

Here, filtering represents the processing of the mean RGB signal extracted from ROI(s) to suppress noise and other artefacts, thus keeping relevant rPPG information in the signal. In addition, the filter neither adds any types of frequency components nor change signal component frequencies. However, there some chnages changes the amplitude or phase relationships. Most of the rPPG search work combination of more than one filter was used to minimize the noise from the mean RGB signal. A commonly used suppressing filtering operation is bandpass pass filters, which can cut off frequency components outside the heart rate bandwidth. The bandpass filters which are widely adopted for rPPG estimation are:

3.4.1 Butterworth IIR bandpass filter

Infinite Impulse Response (IIR) have a non-linear phase response causing frequency related signal delay but faster then FIR when designed with low cut-off frequency. The difference equation represented by the IIR filter is [112]:

$$y[n] = -\sum_{k=1}^{N} a[k]y[n-k] + \sum_{k=0}^{M} b[k]y[n-k] \tag{3.12}$$

The transfer function defined by the difference equation is[112]:

$$H(z) = \frac{b_0 + b_1 z^{-}1 + ... + b_M z^{-}M}{1 + a_1 z^{-}1 + a_2 z^{-}2 + ... + a_N z^{-}N} \tag{3.13}$$

$$H(z) = \frac{B(z)}{A(z)} \tag{3.14}$$

where

$$B(z) = \sum_{n=0}^{M} b[n] z^{-}n \tag{3.15}$$

$$A(z) = \sum_{n=0}^{N} a[n] z^{-}n \tag{3.16}$$

Combining the equation

$$H(z) = \frac{z^{-}M}{z^{-}N}.\frac{b_0 z^M + b_1 z^M - 1 + b_2 z^M - 2 + ... + b_M}{z^n + a_1 z^{-}1 + a_2 z^{-}2 + ... + a_N z^{-}N} \tag{3.17}$$

Now, the zeros represent the numerator of equation

$$b_0 z^M + b_1 z^M - 1 + b_2 z^M - 2 + ... + b_M \tag{3.18}$$

the poles represent the denominator of equation

$$z^n + a_1 z^{-}1 + a_2 z^{-}2 + ... + a_N \tag{3.19}$$

Here, the IIR filter design is based on an analog prototype transfer function. The transfer function maps the s-plane poles and zeros of the analog filter into the z-plane using bilinear transformation [112]. With bilinear transformation, there is a non-linear relationship between the analog frequency ω_a and digital frequency ω_d and also s-domain is mathematically transform from s-domain to the Z-domain

$$z = e^j \omega T$$

Preserving the frequency characteristics [112].

$$s = \frac{2}{T}\frac{1 - z^{-}1}{1 + z^{-}1} \tag{3.20}$$

$$s = j\omega_a \tag{3.21}$$

combining the equation

$$\omega_a = \frac{2}{T} tan \frac{\omega_d T}{2} \tag{3.22}$$

Before designing an IIR filter, it is necessary to tangentially wrap the cut-off frequencies of a digital filter compared with the cut-off frequencies of an analog filter citekim2018design. The position of the poles impacts the stability of the IIR filter system. To get the desired frequency response, all the poles must lie within the unit circle on the z-plane [112]. By deviating the poles from the unit circle on the z-planes, the stability of the IIR filter deteriorates.

3.4.2 Moving average bandpass filter

The sliding window is defined in moving average filter, which calculates average signal covered by the sliding window, introducing a new average signal value to each window. The small sliding window on the moving average filters results in a smoothing effect on the signal; as opposed, a wide sliding window generates the general trend of the signal [113]. The previous work on rPPG signal presents the moving average for removing high-frequency noise and intermittent motion artefact. However, it is difficult to remove a large amplitude motion artefact; also

higher sliding window degrades the quality of the waveform [113]. The equation of moving average filter is represented by [113]:

$$y[n] = \frac{1}{N}\sum_{i=0}^{N-1}(x[n-i]), n = N, N+1, ..., L \tag{3.23}$$

Where N is the window size, L denotes data length.

3.4.3 Detrending

Detrending refers to the removal of a general trend in the signal by improving fluctuation. The extraction of the rPPG signal results in unevenly sampled RR interval time series [114].

$$z = (R_2 - R_1, R_3 - R_2, ...R_N - R_{N-1}) \tag{3.24}$$

Discrete event RR series can be represented with two components [114]:

$$z = z_{stat} + z_{trend} \tag{3.25}$$

where z_{stat} denotes nearly stationary RR series of interest and z_{trend} refers to low frequency aperiodic trend component. Further the trend components can be modeled with the linear equation

$$z_{trend} = H\theta + \upsilon \tag{3.26}$$

where H is the observation matrix, θ is the regression parameters and υ is the observation error. Now the task is to fit the parameter in such a way that $z_{trend} = H\theta$, can estimate the trend [114]. The procedure strongly depend upon the column of the matrix, often called basis vectors in the fitting. The widely used solution to estimate the theta is the least square method. [114] introduced regularized least square solution for θ estimation:

$$\theta_\lambda = (H^T H + \lambda^2 H^T D_d^T D_d H)^{-}1 H^T z \tag{3.27}$$

$$z_{trend} = H\theta + \lambda \tag{3.28}$$

3.5 Choice of rPPG method

This section describes reflection model of rPPG signal or pulse signal estimation by introducing three popular rPPG methods: Blind Source Separation(BSS) [66], Model-based approach [66], and Data-driven method [66].

3.5.1 Reflection model of rPPG

Consider a light source to illuminate the skin area, and a remote colour camera captures the reflection of light from the skin area. Further, assume that the light source is composed of constant spectral composition but varies on intensity and intensity of reflected light from the skin surface captures by the camera. The reflected light to the camera depends on the distance between the camera sensor and the skin tissue. The skin colour measured by the camera sensor is the combination of the intensity of the light source, intrinsic skin colour, and sensitivities of the colour channel, which varies over time as motion-induced intensity /specular variations and pulse-induced subtle colour changes [28].

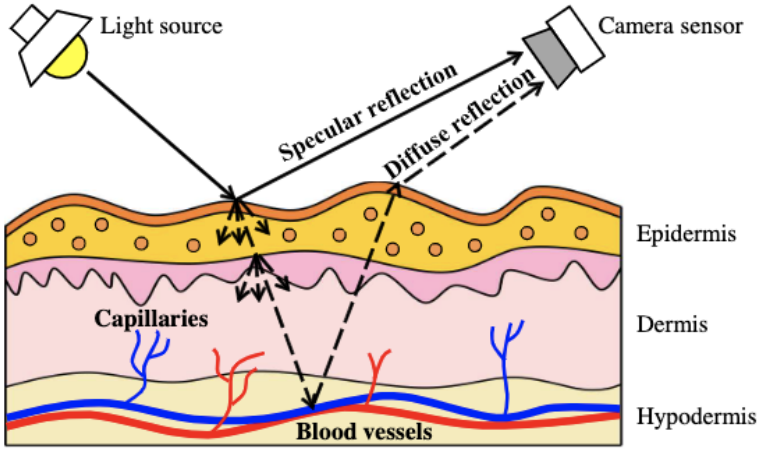


Figure 3.4: The skin reflection model illuminating with light source with specular and diffuse reflection.Figure taken from Wang *et al.* [28]

.

According to the dichromatic model, reflection from each skin pixel recorded in an image sequence can be defined as a time-varying function in the RGB colour channel:

$$C_k(t) = I(t).(V_s(t) + V_d(t)) + v_n(t) \tag{3.29}$$

where $C_k(t)$ represent RGB channels which is ordered in column of the k-th skin-pixel, $I(t)$ denotes the luminance intensity label, and it variate according to change of intensity due to lights sources, distance changes from the camera and skin pixels, subject in motion, absorption of intensity by skin tissue. $I(t)$ is modulated by two major components; specular reflection $v_s(t)$ and diffuse reflection $v_d(t)$,and the last component $v_n(t)$ is quantization noise caused by the camera sensor. The specular reflection, is mirror like reflection, in the sense that most of the light are reflected from the skin surface, containing color of illumination and do not contain any information of pulse signal. It can be represented in the general equation:

$$v_s(t) = u_s(s_0) + s(t) \tag{3.30}$$

where u_s refers unit color vector of light spectrum, s_0 and $s(t)$ refers the stationary and time varying specular reflection. In diffuse reflection on the other hand, some part of the light intensity traverse into the skin surface, due to the absorption of light by haemoglobin and melanin, resulting subtle change in color change by cardiac cycle (blood pulse). Time varying $v_d(t)$ is indicated as the pulse changes in each cardiac cycle.

$$v_d(t) = u_d(d_0) + u_p p(t) \tag{3.31}$$

where, u_d refers unit color vector of the skin pixel, d_0 refers strength of stationary reflection, u_p denotes pulsatile strength in RGB channels and $p(t)$ refers pulse signal. Substituting the value of $v_d(t)$ and $v_s(t)$

$$C_k(t) = I_0(1 + i(t))(u_c.c_0 + u_s.s(t) + u_p.p(t)) + v_n(t) \tag{3.32}$$

where u_c represent unit color vector of the skin reflection and c_o represent reflection strength. In the existing rPPG methods, with the exception of 2SR [97], spatially average RGB values of skin-pixels is calculated, resulting temporal RGB signal for pulse estimation. The spatial pixel averaging procedure reduce quantization noise caused by the camera sensor. Additionally, it is assumes that color vectors are not dependent on skin pixels position in an image thus the equation becomes:

$$C(t) = (u_c c_0 I_0 + u_c c_0 I_0.i(t) + u_s.I_0 — s(t) + u_p.I_0 p(t)) + v_n(t) \tag{3.33}$$

The task of the overall existing rPPG method is to pulse signal $p(t)$ from the decomposition equation $C(t)$

3.5.2 Local Group Invariance

Pilz *et al.* [96] introduce Local Group Invariance(LGI), implemented the prior knowledge about the invariance for estimating HR from the face videos, which is tolerant to motion noise, nuisance factors, facial noise and natural illumination factor. The invariance is model into two categories, feature space and model space, with it's property to remain constant on each equivalence class [96]. In the context of heart rate estimation from the face videos, the features are generally computed as the pixel intensity value aligned in face regions and model space over a set of suitable frequencies [96]. The feature invariant is described with the group of rigid transformations, the Special Euclidean group SE(3) and a stochastic frequency representation invariant concerning the quasi-periodic nature and non-stationary of Heart Rate [96].

Feature Space

The equation of RGB optical sensor signal, as spatial expectation over a skin operator s and function time t expressed as [96]:

$$\vec{p} \in \Re^n = R, G, B, n = 3 \tag{3.34}$$

$$\vec{x(t)} = \int_0^\infty E[\vec{p}|s(\vec{o})]dt \tag{3.35}$$

Since the local variance of blood volume changes with respect to the function of time for each input feature $\vec{x}(t)$. The input feature undergoes transformations of a differential local group of local transformation L_T [96].

$$\frac{\partial}{\partial T}|_{T=0} = f(L_T, \vec{x}(t)) = 0 \tag{3.36}$$

The covariance matrix of the observation$x_i : i = \vec{1}, ..., l$ with the above local transformation L_T.

$$C = \frac{1}{l}\sum_{i=1}^{l}(\frac{\partial}{\partial T}|_{T=0}L_T, \vec{x_i}))(\frac{\partial}{\partial T}|_{T=0}L_T, \vec{x_i}))^T \tag{3.37}$$

Which results the corresponding symmetric eigen value problem:

$$CV = V\Lambda \tag{3.38}$$

With the derivation of operator P with corank k = 1, the corresponding feature vector $\vec{x(t)}$ can be expressed as:

$$\lim_{l\to\infty} P = I - VV^T \tag{3.39}$$

$$\vec{x_v(t)} = P.\vec{v(t)} \tag{3.40}$$

where I is the identity matrix $\vec{x_v(t)}$ is a new feature space.

3.6 Spectral analysis for rPPG signal

The section present the spectral analysis for transforming time domain signal to frequency domain signal. Fast Fourier Transform [115] and Welch method [116] were detailed for spectral analysis.

3.6.1 Fast Fourier Transform

Fourier Fourier transform [115] is applied to transform the time domain into frequency domain [117]. To obtain frequency spectrum from the continuous function, first the continuous function is decompose into discrete function [117]. Fast Fourier Transform(FFT) [115] is applied to compute the Discrete Fourier Transform (DFT) [118] based on the spectrum forming the frequency component of the signal. DFT is defined by the equation of [117]:

$$Y(k) = \sum_{k=0}^{n-1} X(k)W^{nk}, k = 0, 1, ...N-1 \tag{3.41}$$

$$W = exp\frac{-j2\pi}{N} \tag{3.42}$$

3.6.2 Welch Periodogram

Welch method [116] is an averaging modified periodogram aim to calculate the power spectrum [116]. In the welch method, the time-series signal is split into overlapping segments or window, then the periodogram is calculated in each window seperatly[119]. Lastly, averaging the periodograms, welch periodogram is calculated.

$$P_l(f) = \frac{1}{M}\frac{1}{P}\sum_{n=1}^{M}[v(n)x_l(n)e^{-j2\pi f n}]^2 \tag{3.43}$$

The Welch's power spectrum is formulized from the average of these modified periodograms as [119]:

$$P_w(f) = \frac{1}{S}\sum_{l=1}^{s} P_l(f) \tag{3.44}$$

where, $P_l(f)$ is the periodogram of each window signal,$P_w(f)$ is the Welch PSD.

3.6.3 Physiological Parameter estimation from rPPG

After the successful rPPG signal extraction, the second sub-framework comes to play, named as ML classifier. Since the rPPG signal features from the 3D face mask and genuine face videos are different in various features point, which can classify whether the input videos are genuine or 3D face mask videos. Poh *et al.* [12], introduced the two physiological parameter estimation from rPPG PSD curve terms as high-frequency component and low-frequency component. Poh *et al.* [12] state low-frequency component is computed as the area under the PSD curve corresponding to (0.04 - 0.15) Hz and, the component is modulated by the baroreflex activity, which includes both sympathetic and parasympathetic influence.
Similarly, the high-frequency component is computed as the area under the PSD curve corresponding to the (0.15 - 0.4)Hz [12]. This component reflects the parasympathetic influence on the heart through efferent vagal activity [12]. It is connected to respiratory sinus arrhythmia (RSA) [12]. Respiratory Sinus Arrhythmia is a cardiorespiratory phenomenon characterized by Inter Beat Interval (IBI) fluctuations in phase with inhalation and exhalation [12]. Another physiological parameter is the ratio of the term as LF/HF, which is simply the ratio of Low-Frequency component to High-frequency component; the LF/HF ratio is considered to mirror sympatho/vagal balance reflect sympathetic modulations [12]. Li *et al.* [26] state, PSD patterns of genuine face video has a dominant peak in PSD corresponding to the pulse frequency as opposed to the 3D mask face video. A PSD usually contain multiple random noise peaks at a much lower level. The ratio of the utmost value of power to sum of total power in PSD as two feature work [26]. Li *et al.* [26] say fake video has multiple random noise peaks at a much lower level.

3.7 Selection of Machine Learning(ML) classifier

The feature of the pulse signal is extracted from the given face videos; now, it is necessary to estimate whether the input video is genuine or 3D face mask videos. Now, we can achieve this goal based on the Machine Learning (ML) classifier. More specifically, the label of the feature set is two (genuine or 3D face mask videos); we employed a binary classifier. Binary classifier refers to those classification tasks where the unknown set is classified only into two labels based on the classification rule. We have planned to use Support Vector Machine as a binary classifier to classify genuine or 3D face mask videos.

3.7.1 Support Vector Machine

Support Vector Machine (SVM) is designed for dichotomization between two class. In the case of more classes, the classification problem is divided into sub-problems. One class is discriminated against by the rest of the classes. The SVM method undertakes provided attributes of the object, even if some of them do not have much importance, so it essential beforehand to perform the attribute selection(choosing appropriate attribute). In addition, SVM methods are suitable for learning with an extensive training set, possibly with many features.
The basic principle of SVM methods is to place an optimal hyperplane in the space of attributes for classifying the labels. The optimal hyperplane is equally distant from the nearest learning examples of both labels. These closest learning examples to the hyperplane are called the support vectors. Moreover, the distance between the hyperplane and its support-vector is called the margin. The optimal class separating hyperplane is selected such that it is nearest to support vectors. Selecting this hyperplane strengthens the SVM prediction of unseen examples. This underlying concept is also called a maximum-margin hyperplane. Although the maximum-margin hyperplane makes the SVM more robust and classy the unknown samples to their correct label, the case is not always the same. In contrast, the learning examples contain an error. The SVM misclassify the erroneous data to incorrect label or wrong side of the hyperplane. To handle a case like this, the SVM algorithm must be tuned with a soft margin, which allows erroneous data to push their way through the separating hyperplane without affecting the final result [120]. Introducing a wide range of soft margin allow SVM to misclassify the data, so it has to be tuned according to erroneous learning examples in the dataset. Another critical point in SVM is the kernel function. This mathematical computation projects the data from low-dimensional space to a higher dimension. The kernel function not always projects the data to a higher-dimensional space but also reduces the soft margin. We should also consider the very high dimensional space problem called the "curse of dimensionality."

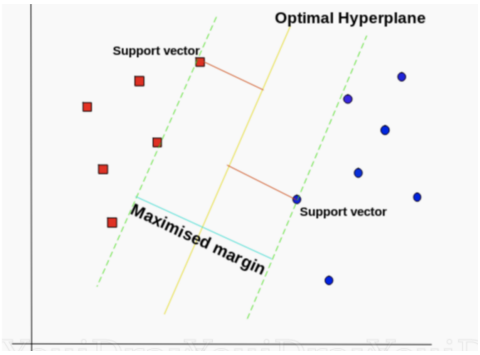


Figure 3.5: Support vector machine for binary classification. (figure undertaken from `https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989`)

Chapter 4

Proposed Methodology

The chapter provides information about the proposed methodology. The proposed methodology is described according to its pipeline first it explains the implementation of face extraction and tracking technique to detect and track the face sequence across video frames. Second, it informs about the Region of Interest (ROI) selection from detected faces across video frames. Third it details about signal preprocessing steps, fourth, frequency domain analysis for feature extraction of the rPPG signal and last machine learning classifiers to distinguish genuine face videos and 3D mask face videos in the proposed methodology.

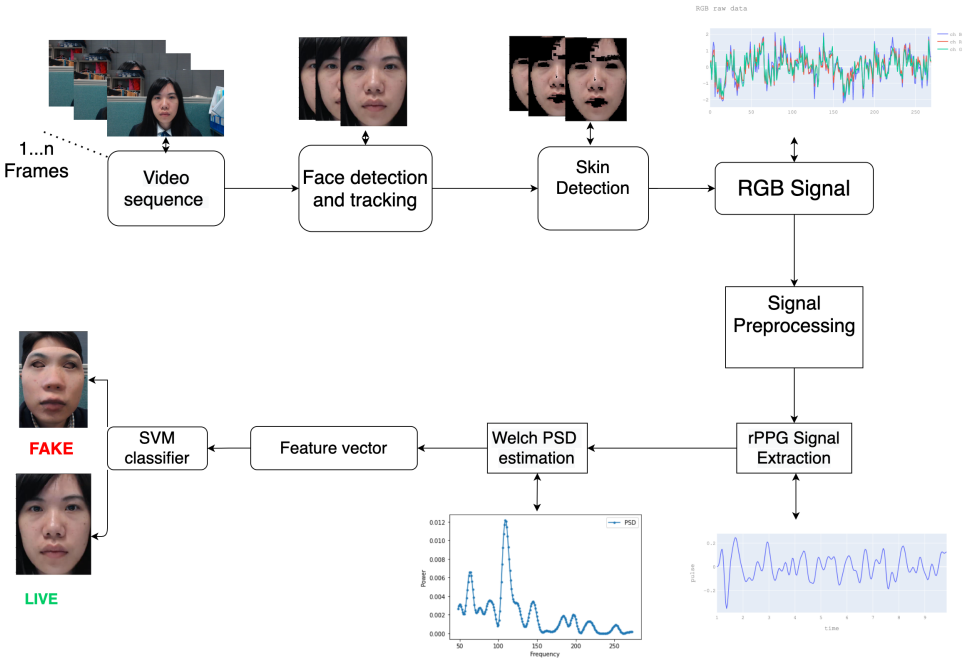


Figure 4.1: Framework of the proposed methodology.

4.1 Proposed Approach

This section describes the proposed functional framework that is carried out for the experiment. To better understand its granularity, we divided the proposed framework into two available units: rPPG feature signal extraction and ML classifier. The first part of the proposed method contributes to rPPG signal extraction from the video input. The proposed method is inspired by the pyVHR[1] framework developed by Boccignone *et al.* [121]. In the second part ML classifier, corresponding features vectors extracted from the rPPG signal; we take the classification as a binary classification as video inputs consist of genuine and 3D face mask video attempts. The feature vector for the genuine face videos was labelled as 1, and the 3D face mask video was labelled as 0. To know its granularity, we divided the rPPG feature signal extraction into six sub-functional-units:

4.1.1 Face extraction and tracking

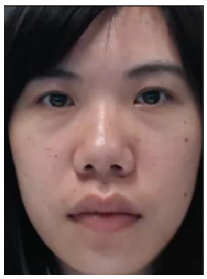
The process starts by taking in face video as an input; for simplicity, assume that the input video sequence is v(t), where t is video length in seconds. Each input video is processed frame by frame. To accomplished this task, we integrated the scikit-video[2] a python package to read the video file and load it frame-by-frame. We also extracted the video metadata such as video frame rate, total frames in a video, total video duration, and video coding from this module which provides general information about the input video. After the video is split up in the frame, the face portion of the image corresponding to each frame is extracted with the face detector approach. We had undertaken the assumption; face video contains a single face across the face video sequence. In the framework, we integrated MTCNN [3] [108] from the python library as a face detector.
The MTCNN [108] detects face from the corresponding frames across the video sequence. The face detector MTCNN [108] has proven its effectiveness when faces present in spatial or appearance distortions [121]. MTCNN localized face within the bounding box, and the face region is cropped following the size of the bounding box.
We coherent Kalman filter to track the face in the video frames to handle the subject motion across the video frames. As the face is detected from the first frame(the first frame refers to the first face detected from the video frame), the Kalman filter exploited the coordinates from the face bounding box and then updated the face bounding box coordinates in subsequent frames. Here we administer, Kalman Filter from the OpenCV [4] python package . From the face extraction and face tracking, the signal c(t)(which is the cropped face images from each frame t = 1,2,3,....., T) is computed with the dimension w × h × r × d, w and h are width

[1] https://github.com/phuselab/pyVHR

[2] `http://www.scikit-video.org/stable/modules/generated/skvideo.io.vreader.html#skvideo.io.vreader`

[3] `https://pypi.org/project/mtcnn/`

[4] `https://pypi.org/project/opencv-python/`



(a) Face detection from the genuine face video frame.

(b) Face detection from the 3D face mask video frame.

Figure 4.2: Face detection from the face videos using MTCCN. A random frame across the video in undertaken to demonstrate the genuine and 3D face mask detection approach.

and height of the bounding box containing face, D represent the depth, according to the time window and r represent 3 channels being coded in the RGB-color space.

4.1.2 Region of Interest Processing

The ROI processing starts by taking the cropped faces that are generated from the face detection and tracking. This step aims to detect the skin pixels that are rich with pulse information. It is the most crucial process in the framework because selecting inaccurate skin pixels from the face regions leads to poor rPPG signal computation. Prior work of rPPG signal extraction is mainly based on the rectangular ROIs and skin detection module. Among them, rectangular ROIs is much expected in the context of rPPG signal extraction. The rectangular region indicates the extraction of skin pixels from predefined rectangular patches, for instance, forehead, nose, or cheeks. Skin detection module instead of rectangular ROI; skin detection module, cut-off non-skin regions such as (beards, hair, background) and tries to extract only skin pixels from the face. Among these two techniques, we performed skin detection modules for each cropped face image; precisely, from the face-cropped image in the ith frame is transferred into the HSV colour space, at the same time lower and upper threshold of the HSV value is defined. The rationale to set the threshold value is that most pixel values will belong to the skin; thus, the threshold value should cut off all non-skin pixel [121]. From each frame t=1,2,3....., T, where T is the end frame of the video. The detected face across the video frame, the skin region is distinguished from the non-skin area; from the selected skin region, we average over all the selected skin pixels to compute corresponding RGB values (frame by frame); this process continues the end of video frames. We denote the summed RGB signal from the skin pixel can be represented q(t).

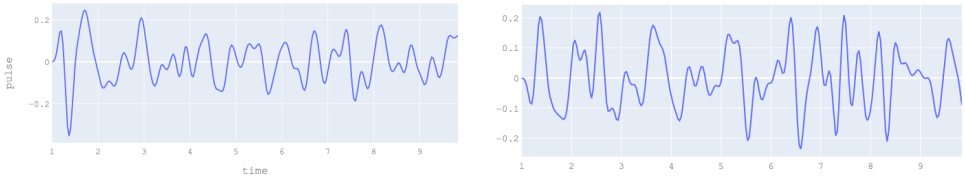
(a) Skin detection from genuine video frame.

(b) Skin detection from 3D face mask video frame.

Figure 4.3: Skin detection module distinguish the skin and non skin pixel from face region in HSV color space from video frame.

4.1.3 Signal Preprocessing and rPPG Estimation

As the process reaches the third step, it started with the prepossessing of RGB signal extracted from ROI. The preprocessing integrates with three-step, firstly, removing noise and other redundant artefacts by moving the average filter. Second, removing the trend from the signal, third suppressing the signal within HR bandwidth (ranging from 40 to 220 BPM) from the signal q(t).In the first step, the signal q(t) is averaged by moving window fashion where we set the moving average window of size 3, then detrending[5] filter from the sciPy[6] package is applied. The aim of detrending in RGB signal is to minimize the deficient frequency trend components present in RGB signal, which can distort the low-frequency components in spectrum [121]. To exclude the frequency outside the blood pulse, we integrated Butterworth IIR bandpass filter, Butter[7] module is implemented from the sciPY[8] package, the filter is designed with the order of six, and the frequency range is set to (0.60 to 4.0)Hz. According to Wang *et al.* [28], the main principle



(a) Remote Photoplethysmography (rPPG) signal estimated from the genuine face video frames.

(b) rPPG signal estimated from the 3D face mask video.

Figure 4.4: The rPPG signal is extracted from the LGI method across the video frames.

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.detrend.html

[6] https://www.scipy.org/

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html#scipy.signal.butter

[8] https://scipy.org/

of existing rPPG methods is to extract the rPPG signal based on the skin reflection model. Among the existing rPPG arithmetic principles, the LGI [96] method is integrated into the framework to compute the rPPG signal. The algorithm is explained in section 3.4(LGI method). From the preprocessed c(t) signal, the new rPPG signal is estimated is calculated by :

$$P = I - VV^T \tag{4.1}$$

$$y(t) = P.c(t) \tag{4.2}$$

4.1.4 Frequency Domain Analysis For Feature Extraction

In this point, clean rPPG signal is extracted from the face videos, now corresponding feature vectors computed from the extracted rPPG signal. Since our aim deviates from calculating accurate HR measurement to extract the distinct features vectors to distinguish between genuine face video and 3D mask attack. Keeping that in mind, we have selected distinct features from the rPPG signal spectrum and also decided to use the features from Li *et al.* [26]. Based on Li *et al.* [26], we transformed preprocessed RGB signal into the frequency domain. The Power Spectrum Density is calculated from each R, G and B colour signal.We integrated Welch[9] python package provided by SciPY. Welch's method estimated the PSD by dividing the data into a segment and computing a periodogram from each segment. Lastly, averaging the periodogram value, the hamming window is implemented with the each segment size, and FFT value is set to 2048. The welch PSD returns power as a function of the frequency
According to the assumption proposed by Li *et al.* [26], in the face video, there will be the dominant peak in PSD patterns to the pulse frequency as opposed to fake videos; the PSD patterns usually contains just multiple random noise peaks with much low power level. Under this assumption,Li *et al.* [26] also constructed the two features set from each RGB colour channel. The first features are the maximum value of power P at the frequency range of [0.6,4.0] and the second feature, R, represents the ratio of maximum power P to the sum of power in the frequency range [0.6,4.0] [26].

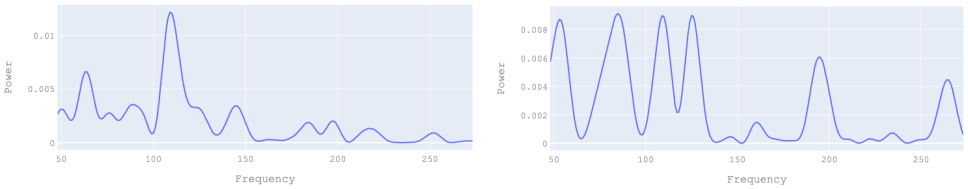
$$R = \frac{P}{\sum_{\forall f \in [0.6,4]} p(f)} \tag{4.3}$$

The process is executed in each on the color channel producing six dimensional feature vector$[P_r, P_g, P_b, R_r, R_g, R_b]$. With the same spirit, Li *et al.* [26] P and R are evaluated on the RGB signal with the frequency range of [0.6,4]. We introduce these first six feature set in our proposed methodology.

[9]urlhttps://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.welch.html#scipy.signal.welch

Feature vector	Description
P_r	Maximum power value on the frequency range [0.6,4] in Red channel.
P_g	Maximum power value on the frequency range [0.6,4] in Blue channel.
P_b	Maximum power value on the frequency range [0.6,4] in Green channel.
R_r	Ratio of maximum power P to the sum of power in the frequency range [0.6,4.0] in red channel.
R_g	Ratio of maximum power P to the sum of power in the frequency range [0.6,4.0] in green channel.
R_b	Ratio of maximum power P to the sum of power in the frequency range [0.6,4.0] in blue channel.

Table 4.1: Brief information about feature vector of computed in Li *et al.* [26]



(a) PSD curve of genuine face video rPPG signal with the frequency range of 0.6-4.0 Hz using Welch method

(b) PSD curve of 3D face mask rPPG signal with the frequency range of 0.6-4.0 Hz using Welch method

Figure 4.5: PSD curve from the genuine face video show a dominant peak as apposed 3D mask face videos curve show a random low level noise like rPPG signal

4.1.5 Rationale behind the complementary feature vector

The complementary feature computed from the proposed methodology mainly describe the physiological parameter estimation. Poh *et al.* [12] proposed the quantification of physiological parameter estimation based rPPG principle. Basic assumption about the complementary feature is; rPPG signal is simulated with the amount of blood flow under the skin, under this scenario physiological parameters estimation becomes more informative, and possess high energy value. As apposed to 3D face mask, light(s) need to passed, first into masked then reaching to skin surface generating rPPG signal with less energy. Furthermore, 3D mask does not coherent any types of biological feature, although its realistic face reconstruction technique. Hence, We consider the three feature as the physiological parameter estimated by Poh *et al.* [12], High Frequency Component informs the breathing

activity and Low Frequency Component informs about baroreflex activity generating significant peak in PSD spectrum. On third, we consider LF/HF feature vector which is the ratio of Low Frequency Component to High Frequency Component represent reflect the sympathetic modulation. Similarly the dominant peak of HF and LF component in the PSD curve from the genuine face regions,contains high energy value. So, we consider the sum of LF and HF component. According to Liu *et al.* [20], rPPG signal produced from the genuine face have high dominant peak as apposed 3D face mask videos generated random small peak. Hence area under PSD curve is computed based on the assumption that genuine face generate significant area in PSD than 3D face mask. We followed Liu *et al.* [20] compute the ratio of maximum power to total power from PSD of BVP signal. The standard deviation and mean from the estimated pulse signal to improve the generalization features of pulse signal.we decided to add the physiological parameter specified in subsection 3.5.4 as the complementary feature vector to improve the classification of genuine and 3D mask face rPPG signal. In total, we proposed ten complementary feature vector; the first three feature vector were undertaken as physiological parameter explains in Poh *et al.* [12] and the rest seven vectors were proposed to more generalize the rPPG signal.

- The AUC_LF feature vector corresponds to the low-frequency component of the rPPG signal with in [0.04,0.15] frequency range, and PSD is estimated on the given frequency range.

$$AUC\_LF = \int_0^x y(t)dt \tag{4.4}$$

  y(t) is the rPPG in frequency range of [0.04,0.15] where x is the maximum frequency in y(t) signal.
- The AUC_HF feature vector corresponds to high-frequency component of rPPG signal with in range of [0.15,4] Hz and area under PSD are calculated.

$$AUC\_HF = \int_0^x y(t)dt \tag{4.5}$$

  y(t) is the rPPG in the frequency range of [0.15,4.0] where x is the maximum frequency in y(t) signal.
- The LF/HF feature vector is the ration of AUC_LF and AUC_HF.

$$LF/HF = \frac{AUC\_LF}{AUC\_HF} \tag{4.6}$$

- Sum of low frequency power component from the PSD in frequency range [0.04,0.15]

$$Sum_{LF} = \sum P_{lf}(f) \tag{4.7}$$

  P_lf (f) is the power in PSD within frequency range [0.04,0.15]

- Sum of high frequency power component from the PSD in frequency range [0.15,4.0]

$$Sum_{HF} = \sum P_{hf}(f) \tag{4.8}$$

  P_hf (f) is the power in PSD within frequency range [0.15,4.0]
- The feature vector M represent Mean of the obtained rPPG signal from face video in the frequency range of [0.6,4].

$$M = \frac{y(1).....y(t)}{T} \tag{4.9}$$

  where T is the total length of y(t)
- The feature vector σ_{rPPG} represent Standard deviation of the obtained rPPG signal from face video in the frequency range of [0.6,4].

$$\sigma_{rPPG} = \sqrt{\frac{\sum(y(t)-M)}{T}} \tag{4.10}$$

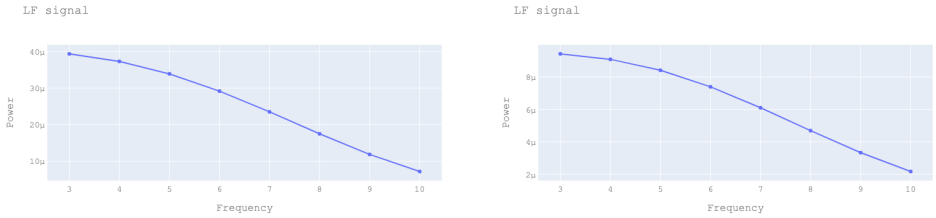
- P_rPPG feature vector represent maximum power value in the PSD curve on frequency range of [0.6,4] in the rPPG signal.

- R_rPPG is the ratio of maximum power value to sum of power value in the PSD curve on the frequency range of [0.6,4] in the rPPG signal.

$$R_{rPPG} = \frac{P}{\sum_{\forall f \in [0.6,4]} p(f)} \tag{4.11}$$

- AUC_rPPG is the Area under the PSD curve in the frequency range of [0.6,4] in the rPPG signal

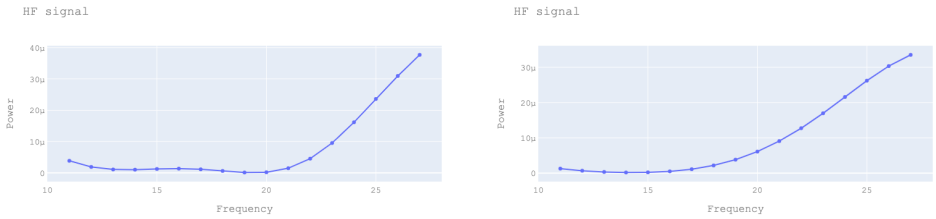
$$AUC\_rPPG = \int_0^x y(t)dt \tag{4.12}$$



(a) PSD curve of low frequency component in the 3D face mask video rPPG signal with the frequency range of [0.04,0.15]

(b) PSD curve of low frequency component in the fre-genuine face video rPPG signal with the frequency range of [0.04,0.15]

Figure 4.6: PSD curve of Low Frequency component from the rPPG signal from 3d mask and genuine face video, within the range of 0.015 to 4.0 Hz.



(a) PSD curve of high frequency component in the 3D face mask video rPPG signal with the frequency range of [0.15,4.0].

(b) PSD curve of high frequency component in the fre-genuine face video rPPG signal with the frequency range of [0.15,4.0].

Figure 4.7: PSD curve of high frequency component from the rPPG signal from 3d mask and genuine face video, within the range of 0.15 to 4.0 Hz.

Feature vector	Description
AUC_LF	Area under the PSD curve within frequency range of [0.04,0.15]
AUC_HF	Area under the PSD curve within the frequency range of [0.15,4]
Sum_LF	Sum of power in the PSD within frequency range [0.04,0.15].
Sum_HF	Sum of power in the PSD within frequency range [0.15,4.0].
LF/HF	Ratio of AUC_LF and AUC_HF
M	Mean of rPPG signal within frequency range of [0.6,4].
σ_{rPPG}	Standard deviation of rPPG signal in the frequency range of [0.6,4].
P_rPPG	Maximum power value within frequency range [0.6,4].
R_rPPG	Ratio of maximum power P to the sum of power in the frequency range [0.6,4.0].
AUC_rPPG	Area under the PSD curve within frequency range of [0.6,4.0].

Table 4.2: Brief information about ten complementary feature vector of rPPG signal in the proposed methodology.

4.1.6 Learning and classification

From the first functional proposed method results in sixteen different feature vector. We implemented the binary classifier to distinguish the genuine and 3D mask face videos based on the computed feature set. We allocate the label '0' for the genuine face video feature set and label '1' for the 3D mask face video feature set. The feature set is standardized with Sklearn python module Standard scaler[10],

[10] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

before feeding into the classifier. We integrated SVM [11] classifier from the sklearn python package. The Radial Basis Function(RBF) kernel with the fixed cost parameter 1000 is undertaken as an SVM parameter.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Chapter 5

Experimental Result

This chapter includes information about the result from two different experimental evaluation protocol. The first section describes the two 3D mask dataset, the second section informs about two experimental evaluation protocol; intra-dataset testing and cross dataset testing. From the intra-dataset testing and cross dataset testing, results were reported following the respective dataset protocol.

5.1 Dataset

This section gathered the information about two 3D mask dataset; 3DMAD [1] and HKBU-MARsv1+ [2]. From each dataset samples, rPPG estimation and corresponding feature vectors were extracted from the proposed methodology. The computed feature set is trained with SVM classifier to distinguish the 3D mask attack and genuine face.

5.1.1 3D Mask Attack Database(3DMAD)

Nesli and Marcel [1] introduced a public database called 3D Mask Attack Database (3DMAD), which is composed of genuine face video and 3D mask face videos of 17 different subjects recorded by Microsoft Kinect sensor. In the 3DMAD database, Nesli and Marcel [1] utilized ThatsMyFace.com for facial reconstruction and transformed 2D portraiture into 3D sculptures. The 3D face is constructed only after seconds of uploading frontal and profile face images of the person. Nesli and Marcel [1] uploaded one frontal and two profile images of 17 different subjects on ThatsMyFace.com and ordered a life-size wearable mask and a paper cut mask for each. The 17 wearable masks are made out of a hard resin which is composed of full 24-bit colour with holes at the eyes, and the nostrils [1].
The recording of all the dataset is performed using Microsoft Kinect for the Xbox 360 sensor, which generated both RGB(8-bit) and depth data(11-bit) of size 640×480 at 30 frames per second [1]. The videos collected in three different sessions; two real access sessions composed two weeks apart and a third session which is

Figure 5.1: Face masked used by the subjects in 3DMAD dataset.Figure taken from Nesli and Marcel [1].

mask attacks performed by a single person (attacker) [1]. The recording environment in all three sessions was well-controlled; the background scene is uniform, and light is set up to minimize the shadows cast on the face [1]. In each session, subjects correspond to frontal-view and neutral expression, 17 subject records 5 videos of length 10s. The first two sessions are composed to real access of 170 videos altogether and the third sessions organized to fake videos of 85 videos overall.

5.1.2 HKBU-MARsv1+

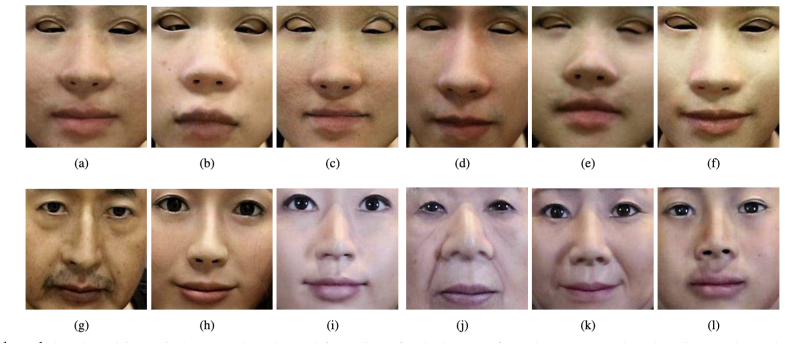


Figure 5.2: Sample mask images in the database HKBU-MARsv1+. (a)-(f) are ThatsMyFace masks and (g)-(l) are Real-F masks [2]. Figure taken from Liu *et al.* [2].

Li *et al.* [21] introduced 3D mask attack public database,a subset of HongKong Baptist University 3D Mask Attack with Real World Variations(HKBU-MARs) [2]. The dataset is composed of 12 subjects, with two different types of mask attacks,ThatMYFace and REAL-F, which increases the diversity of mask types in database[21]. Li *et al.* [21] introduced a web camera Logitech C92 is used to record the face videos with resolution of 1280× 720) and frame rate of 25fps. Li *et al.* [21] recorded the data under room light.

5.2 Experimental Evaluation Protocol

This section covers the two experiment evaluation protocol: Intra dataset evaluation and cross dataset. We had followed the evaluation protocol based on the respective dataset 3DMAD [1] and HKBU-MARsv1+ [2]. In addition, to report the result, we had included the three Presentation Attack Detection (PAD) performance metrics: APCER, BPCER ACER, EER and AUC.

5.2.1 Intradataset testing

In the intradataset testing, the train, development and test samples were generated within the same dataset. At, first datasets samples were spilt into train, development and test samples. The train samples were used to train the Machine Learning (ML) model; the development set is used to tune parameters and hyperparameters of the Machine Learning (ML) model to reproduce the best results while testing the Machine Learning (ML) model with test samples were undertaken. Test samples remain unknown to the Machine Learning (ML) model, which generalises the Machine Learning (ML) model as a real-world scenario. In our experiment, we followed the evaluation protocol based on the 3DMAD dataset and HKBU-MARsv1+ dataset.

Protocol for 3DMAD

In the 3DMAD experiment, leave one out cross-validation(LOOCV) protocol is implemented, which selects one testing subject from each iteration and divides the rest subjects into training and development sets [26]. Similarly, 17 folds of cross-validations are implemented in the 3DMAD dataset where one subject samples wer left for testing. In contrast, the remaining 16 samples are divided into two subject-disjoint halves as training and development sets, respectively [26].

Result for 3DMAD

From every 17 folds, APCER, BPCER ACER, EER and AUC. are computed in the development and test set. The results from each fold are average with a confidence interval of 95%. Based on the development set of 3DMAD, we succeeded in obtaining the EER of 7.1% with a confidence interval of 1.4% , APCER of 8.2

% with a confidence interval of 0.8% , BPCER of 8.85 % with a confidence interval of 1% , ACER of 13.2% with a confidence interval of 10.3% and AUC of 95.01% with a confidence interval of 0.4%. Similarly, on the testing set, we obtained the favourable outcome with an EER of 7.9% with a confidence interval of 4.3%, APCER of 7.6% with a confidence interval of 4.08%, BPCER of 10.8% with a confidence interval of 6.5%, ACER of 9.3% with a confidence interval of 4.9% and AUC of 95% with a confidence interval of 0.06%.

Method	EER-dev (%)	APCER-dev (%)	BPCER-dev (%)	ACER-dev (%)	AUC-dev (%)
Li *et al.* [26]	2.31	-	-	-	-
proposed	7.1 ± 1.4	8.02 ± 0.8	8.85 ± 1	13.2 ± 10.3	95 ± 0.01

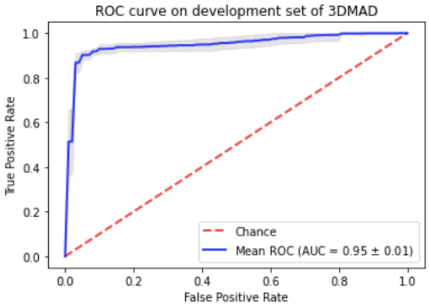
Table 5.1: Result for intra-dataset protocol on the 3DMAD dataset for development set and comparing the result with existing approach

Method	EER-test (%)	APCER-test (%)	BPCER-test (%)	ACER-test (%)	AUC-test (%)
Morales [27]	18.18	-	-	-	-
Hernandez-Ortega *et al.* [10]	22.1		-	-	-
Liu *et al.* [30]	13.3		-	-	93.8
Liu *et al.* [25]	6.54		-	-	97.6
Li *et al.* [26]	4.71	-	-	-	-
method	7.9 ± 4.3	7.6 ± 4.08	10.8 ± 6.5	9.3 ± 4.9	95 ± 0.06

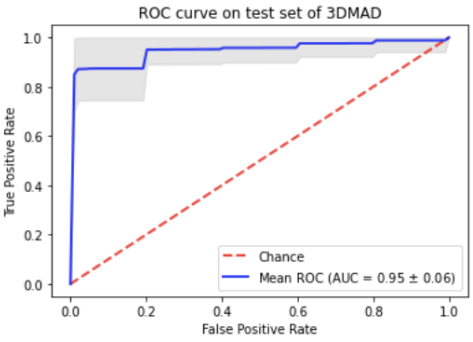
Table 5.2: Result for intra-dataset protocol on 3DMAD dataset for testing set and comparing the result with existing approach

Protocol for HKBU-MARsv1+

In HKBU-MARsv1+, we conducted leave one out cross-validation (LOOCV) protocol since subject 8 taken out due to privacy issues. We only undertook the 11 subjects; we selected one testing subject from each of 11 fold iterations and divided the rest subjects into training and development sets. We randomly selected

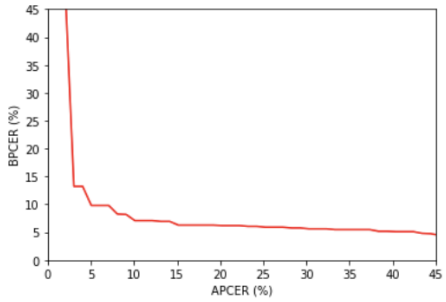


(a) Average ROC curve of 3DMAD development set in 17 folds.

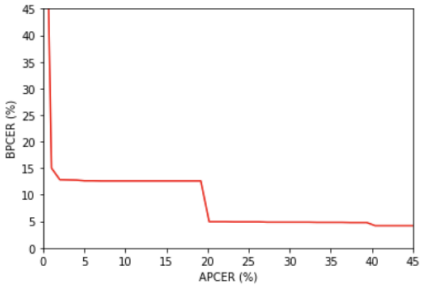


(b) ROC curve of 3DMAD test set

Figure 5.3: Average ROC curve for training and testing set in 17 3DMAD fold.



(a) DET curve of 3DMAD development set in 17 3DMAD fold..



(b) DET curve of 3DMAD test set 17 3DMAD fold.

Figure 5.4: Average DET curve for training and testing set in 17 fold 3DMAD.

5 subjects in each fold for development and 5 subjects in each fold for the testing set.

Result for HKBU-MARsv1+

The HKBU-MARsv1+ data samples were split into test, train and development into 11 folds. From every 11 folds APCER, BPCER, EER, ACER and AUC were computed in the development and test set. The results from each fold are average with a confidence interval of 95%. Based on the development set of HKBU-MARsv1+, we succeeded in obtaining the EER of 28.48% with a confidence interval of 2.9%, APCER of 27.27% with a confidence interval of 2.7%, BPCER of 25.43% with a confidence interval of 4.2 % and AUC of 78% with a confidence interval of 0.03%. Similarly, on the testing set, we obtained the favourable outcome with EER of 18.18% with a confidence interval of 11.11%, APCER of 19.1% with a confidence interval of 11.22%, BPCER of 37.2% with a confidence interval of 21.2% , ACER of 25.2% with a confidence interval of 12.4% and AUC of 81% with a confidence interval of 17%.

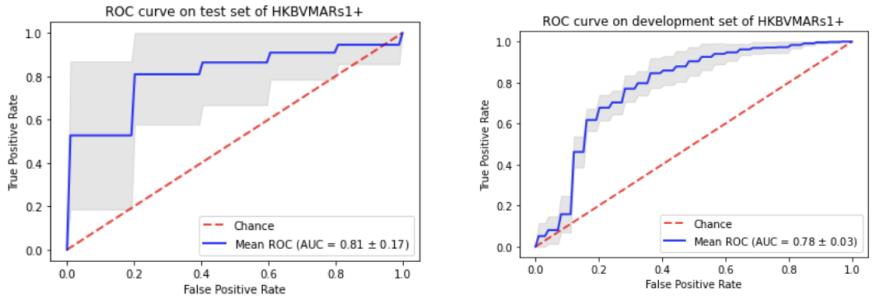
Method	EER-dev (%)	APCER-dev (%)	BPCER-dev (%)	ACER-dev (%)	AUC-dev (%)
proposed	28.48 ± 2.9	27 ± 2.7	25.43 ± 4.2	26.7 ± 2	78 ± 0.03

1

Table 5.3: Result for intra-dataset protocol on HKBU-MARsv1+ dataset for development set

Method	EER-test (%)	APCER-test (%)	BPCER-test (%)	ACER-test (%)	AUC-test (%)
Liu *et al.* [25]	4.4	-	-	-	99.3
Liu *et al.* [30]	22	-	-	-	85.2
proposed	18.18 ± 11.11	19.1 ± 11.22	37.2 ± 21.2	25.2 ± 12.4	81 ± 0.17

Table 5.4: Result for intra-dataset protocol on the HKBU-MARsv1+ dataset for testing set and comparing the result with existing approach



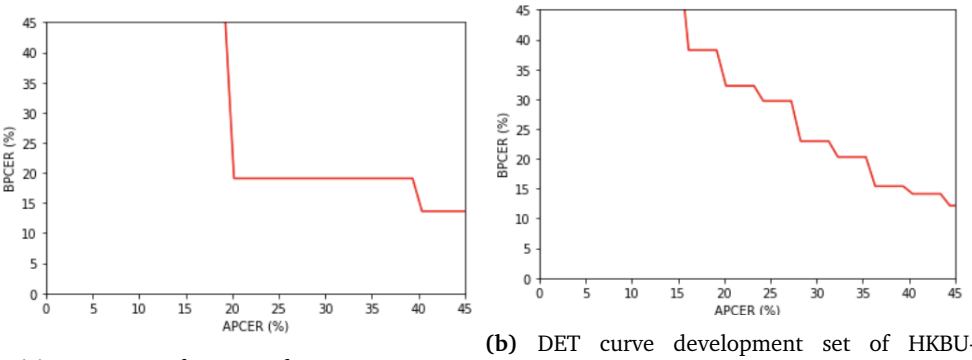
(a) ROC curve of test set of HKBU-MARsv1+ in 11 folds.

(b) ROC curve dev set of HKBU-MARsv1+ in 11 folds.

Figure 5.5: Average ROC curve for development set and test set in 11 fold HKBU-MARsv1+ dataset.

5.2.2 Cross Dataset Testing

In the cross dataset training protocol, we use a different set of datasets for training and testing the ML model, which can simulate generalised scenarios. In our cross dataset experiment, we undertake HKBU-MARsv1+ and 3DMAD; we select data



(a) DET curve of test set of HKBU-MARsv1+

(b) DET curve development set of HKBU-MARsv1+

Figure 5.6: Average DET curve for development set and test set in 11 fold HKBU-MARsv1+ dataset.

samples from one dataset, computes the feature vectors from the proposed methodology and trained ML model with those generated feature vectors. For testing, another dataset is taken to produce the feature vectors and with these feature vectors ML model was tested upon.

Protocol for HKBU-MARsv1+ and 3DMAD

For cross data setting, we undertook HKBU-MARsv1+ and 3DMAD as the combined dataset.The observed generability result from the proposed method, the ML classifier is train and test within two datasets; we used 17 subjects from 3DMAD as training and 11 subjects from the HKBU-MARsv1+ dataset samples as a testing set. Since the subjects, experimental setup, video acquisition device were completely different among these two dataset, which, from which we proposed method is tested upon with generability environment.

Result for HKBU-MARsv1+ and 3DMAD

The results were produced on training the SVM model with 17 subjects from 3DMAD and tested upon 11 subjects from the HKBU-MARsv1+. With this experimental protocol, we succeeded obtaining the EER of 14.7 %, APCER of 14.7 % , BPCER of 10.6 %, ACER of 12.6 % and AUC of 89.62% with a confidence interval of 2%.

Method	EER (%)	APCER (%)	BPCER (%)	ACER (%)	AUC (%)
Liu *et al.* [25]	5.884	-	-	-	98
proposed	14.7	14.7	10.6	12.6	89.62

Table 5.5: Result for cross-dataset protocol HKBU-MARsv1+, where HKBU-MARsv1+ is taken as training and 3DMAD as testingand comparing the result with existing approach.

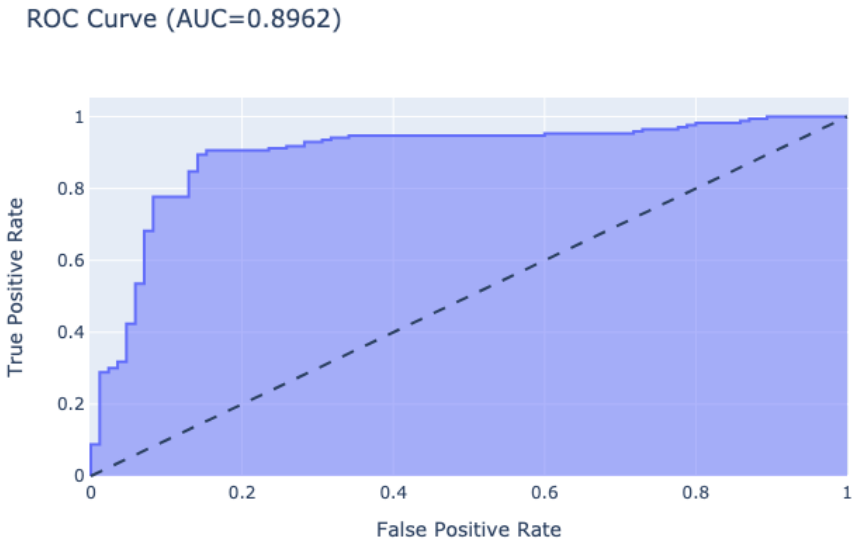


Figure 5.7: ROC curve for cross dataset 3DMAD and HKBU-MARsv1+, where HKBU-MARsv1+ is taken as training and 3DMAD as testing.

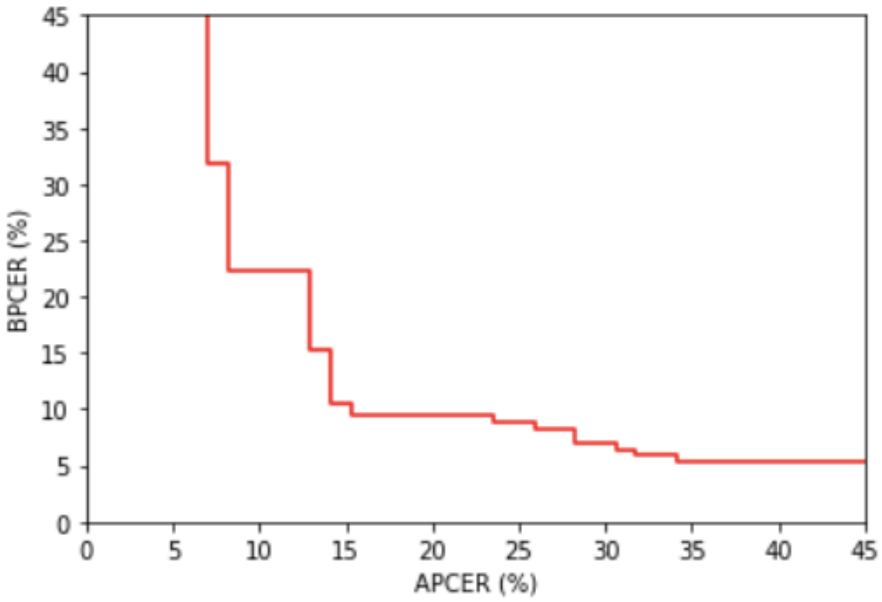


Figure 5.8: DET curve for cross dataset 3DMAD and HKBU-MARsv1+, where HKBU-MARsv1+ is taken as training and 3DMAD as testing.

Chapter 6

Discussion

6.1 Discussion about rPPG approach for face PAD

The severity of the face recognition system is due to overexposure of the personal biometrics data, specifically photos and videos with clear front face view. Taking advantages of this kind of biometric traits, attackers can circumvent the face recognition system, also called presentation attacks. Nowadays, presentation attacks are limited by photographs and videos but highly realistic 3D masks. With this persuasive approach resulted from the 3D mask, the face recognition system may fail to detect the spoofed face biometrics traits, resulting in unauthorized users accessing the system. For countermeasure of the 3D mask, we proposed a liveness based presentation attack detection with the rPPG method. The rationale behind Remote Photoplethysmography (rPPG) approach is when the light source(s) illuminates skin, then some portion of the light penetrates through skin layers reach the capillary vessel; based on the amount of haemoglobin in the blood, that small portion of the light is absorbed, causing subtle colour change(also depend upon the volume of blood under the observable skin surface). Since we are focusing on face PAD, we tried to obtain only pulse signal from the face region. Based on this approach, pulse signal from the genuine faces presented to generated with high amplitudes as the light sources directly illuminate the skin surface as opposed, on mask attack; first, the light sources need to penetrate the masked surface before reaching the skin and blood capillaries which result in very noisy pulse signal with low amplitude. Based on this principle, we proposed the rPPG based PAD for 3D mask spoofing.

6.2 Discussion about proposed methodology

In our proposed methodology, we developed an end-to-end pipeline and subdivided the proposed method into two functional units: rPPG signal estimation and Machine Learning (ML) classifier to distinguish the genuine or fake face

videos. We initiate by taking face video as an input and then decompose the video into respective frames. From the face videos, it is necessary to extract face regions where we can locate rich rPPG signal. To find the face regions from the video, we implemented MTCNN [108] as a face detector. First, the input video is decomposed into the frame by frame; from the frame, we extracted the face using MTCNN [108]. With the n video frames, rather than re-detecting the face on each frame, we implemented the face tracking algorithm to track the face across the videos. We implemented a Kalman filter [109] for face tracking along the video sequence. Another reason to implement the face tracking algorithm is to reduce the subject motion's effect while calculating the rPPG signal. The face detection is followed by a selection of skin regions enriched with the rPPG signal. Most of the literature introduced selective skin patches(forehead, cheeks, nose and lower lips), expecting these regions a good source for rPPG signal. We introduced the skin detection module rather than selecting the skin patches (distinguishing the skin and non-skin pixel). From all the skin pixels in each frame, the RGB value from each skin pixels is average and producing a raw RGB signal. Since the raw RGB signal is concatenated with noise artefacts, we passed it into series of filters as a preprocessing step. First, the moving average filters are introduced with the sliding window of size 3, which aims to remove high-frequency noise and intermittent motion artefact. After signal detrending is performed to remove the general trend in the signal improving the signal fluctuation; lastly, the signal is passed through bandpass filters which can cut off the frequency components outside the pulse range bandwidth. We designed the 6 order Butterworth IIR bandpass filter and allocated the frequency range of 0.6 Hz to 4.0 Hz. After the preprocessing step, a clean RGB signal is produced and input the obtained RGB signal into LGI [96] rPPG method. The LGI [96] method compute the pulse signal, in which the energy of the blood volume signal is re-arranged in vector space with a more concentrated distribution. The rPPG signal is estimated from LGI rPPG methods; we intended to determine the differences between the rPPG signal based on the various feature vectors. With the assumption 3D mask face videos produced have low rPPG signal energy-generating random and low-value noise in the power spectrum, while real videos aims has high energy generating high values spikes in the power spectrum, we performed frequency analysis in the rPPG signal projecting time domain into the frequency domain. Then PSD is calculated with the welch's method [116].

The first function unit completes with rPPG signal estimation; now, the feature group defining the rPPG signal is computed and trained the ML model to distinguish genuine face videos and 3D mask face videos. We extracted sixteen feature vector from the rPPG signal to determine the genuine and 3D mask face rPPG signal properties. The feature sets include maximum power and ratio of maximum power to sum of the total power of each RGB values, Area under the PSD curve of high frequency and low-frequency rPPG signal component, the ratio of high to low-frequency components, mean and standard deviation of rPPG signal, maximum power of rPPG signal in the power spectrum, the ratio of maximum power

to sum of total power and Area under PSD for rPPG signal frequency(0.6 to 4) Hz. Similarly, we in-cooperate the 3DMAD and HKBV-MARsv1+ dataset to compute the result—both of these data composed of real videos and fake videos(3D masks). We implemented an SVM classifier with RBF kernel, and the Cost parameter is fixed to 1000 in all experiments. The results are calculated based on PAD metrics, and the best work from the proposed method is obtained in the 3DMAD dataset EER of 7.9% with a confidence interval of 4.3%.

6.3 Discussion about knowledge guided on thesis work

The overall goal of conducting this thesis work is to acquire meaningful concepts and State of the art on 3D mask face PAD. Keeping in that direction, we had to review scholarly research articles, research works, conference papers, book, website, and other related academic documents to draw close attention to the PAD approach. In the background section, we talked about the current approaches on the PAD, which provides a robust foundation on the existing PAD approach. The thesis enlightens the concepts about the security vulnerabilities in Face Recognition System (FRS) and the State of the art method. Although the current approach, improvised to detected the photo attacks, video attacks, most of them is complex highly accurate 3D masks attacks since most of the PAD approaches is based on textures clues and challenges responses to detect the liveliness. Based on the recent advancements in the research work on 3D mask face PAD, we choose to adopt rPPG based PAD, aiming to provide attack detection on 3D face masks. Moreover, rPPG based approach is embraced to compute the heart rate from the face videos; we generalize the concepts to tackle 3D mask face PA by generating rPPG or pulse signal. It is essential to perceive the ideas on the principle of rPPG methodology and how rPPG signals detect the masks attacks. Undergoing rPPG based approach or developing the pipeline, thesis procure tangible notion on face detection and training approaches—a comprehensible face detection technique MTCNN [108] which can also be implemented in various face detection application. Secondly, thesis work implemented a face tracking algorithm through Kalman filter [109]. The State of art approaches for the pulse or rPPG signal extraction methodologies becomes more apparent. As mentioned above, the thesis focused on computing the implementation of rPPG methods on 3D masks; these approaches may suffer from illumination noise and motion noise. To handle noise artefacts, thesis work enlightens about signal processing by introducing moving average, detrending and bandpass filter concepts. Furthermore, ideas about the transformation of the time domain to the frequency domain and the calculation of Power Spectrum Density are highlighted. Lastly, the thesis work guided the concepts of Machine Learning, metrics about binary classifier and implementation of Machine Learning (ML) classifier (SVM)

6.4 Discussion about pros and cons about thesis work

Any proposed methods or the framework comes along with the pros and cons within it. The pros, our proposed plans, we were succeeded in producing the ERR of 6.15% with the confidence interval of 4.77% on the 3DMAD dataset; an ML model is also cross-validated with two different datasets to see the generalizability of the proposed framework, which computes the EER of 14.1% and ACER of 12.4%. As opposed to pros, all the dataset we had implemented in the framework were produced in the adjusted experimental environment. The whole environment setup is changed with lighting, background noise, and other factors. Hence, the thesis was not able to examine subjects from real-world scenarios. Another con in this approach is the time complexity of face detection algorithm detection; MTCNN consumes a little longer time to perform face detection than other face detection approaches. So, if we had an opportunity to redo this thesis again, we will examine the results on the more wild and realistic dataset. And to improve the applicability of the pipeline, we will try to reach out to some other face detection techniques to reduce the time complexity.

6.5 Discussion about societal consequences

The proposed method is designated to detect and distinguish genuine and 3D face mask, result in a robust and secure Face Recognition System (FRS). With the Remote Photoplethysmography (rPPG) even the super realistic 3D face mask can be detected by the FRS adding security towards 3D face mask threats.

Chapter 7

Conclusion

The main proposal of the master thesis is to achieve Presentation Attack Detection (PAD) on the 3D mask attack. Among the approaches based on PAD, we proposed rPPG based face PAD and analysed its effectiveness on 3D face mask detection. To reduce the effect on spatial noise, the proposed method integrates the spatial average of skin pixel from Region of Interest (ROI) and introduced three preprocessing filters moving average, detrending and bandpass filter. From each filter, the raw RGB signal is prepossessed to produce a much clear RGB signal. The moving average for removing high-frequency noise and intermittent motion artefact. Detrending refers to the removal of a general trend in the signal by improving fluctuation. And bandpass filter suppresses noise and other artefacts, keeping relevant pulse information in the signal.
To handle subject motion, Kalman filters are introduced to track face across the video. From the first video frame, the face is detected from the MTCNN face detection generating the facial boundary and landmark localization; based on the facial coordinates produces from the MTCCN, Kalman filters track face coordinates across video frame. The proposed design seems much robust towards the spatial and subject motion, creating a clean rPPG signal. At the end of the thesis work, we succeeded in answering the research question:

1. What are the complimentary feature(s) of the pulse signal, estimated by rPPG approach to classifying the given input videos as a genuine or 3D face mask?
   The estimated pulse signal from the proposed methodology is essential to distinguish between the pulse generated by genuine face videos and 3D face mask videos. Following the principle that pulse signal generated from 3D face mask videos produced low energy level in the PSD compared to genuine face videos. Ten complimentary features were introduced to better generalise the estimated pulse signal to distinguish between genuine and 3D face videos. The SVM classifier was undertaken with cost parameter 1000 and RBF kernel to classify the features computed feature vector produced from the proposed methodology. The experiment is conducted on the two databases 3DMAD and HKBV MARsv1+, following the experimental pro-

tocol. The results are calculated based on PAD metrics, and the best work from the proposed method is obtained with EER of $7.9 \pm 4.3\%$ in 3DMAD [1] and EER of 18.18 ± 11.11 % in HKBVMArsV1+ [21].

2. Can these complimentary features help in detecting cross-dataset attacks? From the proposed method, sixteen features sets were introduced, among them ten features were proposed complimentary features. The features set are performed in the cross dataset set testing, taking 17 subjects from 3DMAD and tested upon 11 subjects from the HKBU-MARsv1+ .To generalize the proposed methodology, complementary feature is also evaluated under cross dataset evaluation on publicly available 3DMAD [1] and HKBVMArsV1+ [21] resulting favourable results. The proposed approach gains a performance of EER of 14.7%for cross-dataset evaluation. The result show the features set show genearability towards the cross dataset analysis.

Chapter 8

Future Work

The resilience of any framework or methodology is not possible in all perspective. Similarly, in our proposal, there were some aspects which can be further improved. The dataset we had employed in this work is recorded in the experimental setting rather than a real-world scenario; hence the proposed methodology is deprived of the vital information on how well it adapts in real scenarios. As the combination of the two datasets resulted in a total of 29 subjects and they do not necessarily represent all different ethnic variations of the face, the bias factor needs to be studied. And more importantly, we are focused only on the 3D mask attack scenarios. At the same time, some other face presentation attacks remain untouched by our methodology, such as face occlusions, highly accurate silicone mask attacks, extreme makeups, and morphing attacks. All these attacks needs to be studied in a joint manner in future works.
In addition, most of the real face videos we have employed are stable (with no or less facial movement) and a clear frontal view. Based on these kind of facial input videos, we determined the rPPG signal. But it is not always the case in the real-time application or real-world scenario; the subject possesses inherent motion and face occlusion may be encountered by the face detection approach. These aspects needs to be studied in the future works in the-wild setting.
Another aspect is the continuous detection of attacks in a cohesive manner as our framework needs a set of frames before making a decision. The need for a set of frames may hinder the use in real-time scenario. A trust factor based continuous authentication can be integrated to improve the applicability of the face PAD in real-time scenarios.

Bibliography

[1] E. Nesli and S. Marcel, 'Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect,' in *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS'13)*, 2013, pp. 1–8.

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

[3] A. K. Jain, P. Flynn and A. A. Ross, *Handbook of biometrics*. Springer Science & Business Media, 2007.

[4] S. Marcel, M. S. Nixon, J. Fierrez and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection*. Springer, 2019.

[5] J. Hernandez-Ortega, J. Fierrez, A. Morales and J. Galbally, 'Introduction to face presentation attack detection,' in *Handbook of Biometric Anti-Spoofing*, Springer, 2019, pp. 187–206.

[6] B. Gipp, J. Beel and I. Rössling, 'Epassport: The world's new electronic passport: A report about the epassport's benefits, risks and its security,' 2007.

[7] R. Ramachandra and C. Busch, 'Presentation attack detection methods for face recognition systems: A comprehensive survey,' *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[8] R. Raghavendra, K. B. Raja and C. Busch, 'Presentation attack detection for face recognition using light field camera,' *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 1060–1075, 2015. DOI: 10.1109/TIP.2015.2395951.

[9] D. Yi, Z. Lei, Z. Zhang and S. Z. Li, 'Face anti-spoofing: Multi-spectral approach,' in *Handbook of Biometric Anti-Spoofing*, Springer, 2014, pp. 83–102.

[10] J. Hernandez-Ortega, J. Fierrez, A. Morales and P. Tome, 'Time analysis of pulse-based face anti-spoofing in visible and nir,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 544–552.

[11] N. Kose and J.-L. Dugelay, 'Countermeasure for the protection of face recognition systems against mask attacks,' in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, 2013, pp. 1–6.

[12] M.-Z. Poh, D. J. McDuff and R. W. Picard, 'Advancements in noncontact, multiparameter physiological measurements using a webcam,' *IEEE transactions on biomedical engineering*, vol. 58, no. 1, pp. 7–11, 2010.

[13] A. Bhattacharjee and M. S. U. Yusuf, 'A facial video based framework to estimate physiological parameters using remote photoplethysmography,' in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, 2021, pp. 1–7.

[14] A. Kamal, J. Harness, G. Irving and A. Mearns, 'Skin photoplethysmography—a review,' *Computer methods and programs in biomedicine*, vol. 28, no. 4, pp. 257–269, 1989.

[15] P. V. Rouast, M. T. Adam, R. Chiong, D. Cornforth and E. Lux, 'Remote heart rate measurement using low-cost rgb face video: A technical literature review,' *Frontiers of Computer Science*, vol. 12, no. 5, pp. 858–872, 2018.

[16] 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*, Springer, 2010, pp. 504–517.

[17] I. Chingovska, A. Anjos and S. Marcel, 'On the effectiveness of local binary patterns in face anti-spoofing,' in *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, IEEE, 2012, pp. 1–7.

[18] N. Erdogmus and S. Marcel, 'Spoofing face recognition with 3d masks,' *IEEE transactions on information forensics and security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[19] A. Dantcheva, C. Chen and A. Ross, 'Can facial cosmetics affect the matching accuracy of face recognition systems?' In *2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS)*, IEEE, 2012, pp. 391–398.

[20] S.-Q. Liu, P. C. Yuen, X. Li and G. Zhao, 'Recent progress on face presentation attack detection of 3d mask attacks,' *Handbook of Biometric Anti-Spoofing*, pp. 229–246, 2019.

[21] X. Li, J. Chen, G. Zhao and M. Pietikainen, 'Remote heart rate measurement from face videos under realistic situations,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.

[22] Z. Ming, M. Visani, M. M. Luqman and J.-C. Burie, 'A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices,' *Journal of Imaging*, vol. 6, no. 12, p. 139, 2020.

[23] C. Busch, 'Standards for biometric presentation attack detection,' in *Handbook of Biometric Anti-Spoofing*, Springer, 2019, pp. 503–514.

[24] S. Jia, G. Guo and Z. Xu, 'A survey on 3d mask presentation attack detection and countermeasures,' *Pattern Recognition*, vol. 98, p. 107 032, 2020.

[25] S.-Q. Liu, X. Lan and P. C. Yuen, 'Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection,' in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 558–573.

[26] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen and M. Pietikäinen, 'Generalized face anti-spoofing by detecting pulse from face videos,' in *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, 2016, pp. 4244–4249.

[27] A. Morales, 'Continuous presentation attack detection in face biometrics based on heart rate,' in *Video Analytics. Face and Facial Expression Recognition: Third International Workshop, FFER 2018, and Second International Workshop, DLPR 2018, Beijing, China, August 20, 2018, Revised Selected Papers*, Springer, vol. 11264, 2019, p. 72.

[28] W. Wang, A. C. den Brinker, S. Stuijk and G. De Haan, 'Algorithmic principles of remote ppg,' *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[29] E. M. Nowara, A. Sabharwal and A. Veeraraghavan, 'Ppgsecure: Biometric presentation attack detection using photopletysmograms,' in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 56–62.

[30] S. Liu, X. Lan and P. Yuen, 'Temporal similarity analysis of remote photoplethysmography for fast 3d mask face presentation attack detection,' in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2608–2616.

[31] G. Heusch and S. Marcel, 'Pulse-based features for face presentation attack detection,' in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, pp. 1–8.

[32] J. Määttä, A. Hadid and M. Pietikäinen, 'Face spoofing detection from single images using micro-texture analysis,' in *2011 international joint conference on Biometrics (IJCB)*, IEEE, 2011, pp. 1–7.

[33] M. Oren and S. K. Nayar, 'Generalization of the lambertian model and implications for machine vision,' *International Journal of Computer Vision*, vol. 14, no. 3, pp. 227–251, 1995.

[34] N. Kose and J.-L. Dugelay, 'Shape and texture based countermeasure to protect face recognition systems against mask attacks,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 111–116.

[35] N. Erdogmus and S. Marcel, 'Spoofing 2d face recognition systems with 3d masks,' in *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*, IEEE, 2013, pp. 1–8.

[36] R. Raghavendra and C. Busch, 'Novel presentation attack detection algorithm for face recognition system: Application to 3d face mask attack,' in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 323–327.

[37] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh and N. Ratha, 'Face anti-spoofing with multifeature videolet aggregation,' in *2016 23rd International Conference on Pattern Recognition (ICPR)*, IEEE, 2016, pp. 1035–1040.

[38] A. Pinto, H. Pedrini, W. R. Schwartz and A. Rocha, 'Face spoofing detection through visual codebooks of spectral temporal cubes,' *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4726–4740, 2015.

[39] A. Agarwal, R. Singh and M. Vatsa, 'Face anti-spoofing using haralick features,' in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2016, pp. 1–6.

[40] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao and A. Rocha, 'Deep representations for iris, face, and fingerprint spoofing detection,' *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[41] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle and R. Lotufo, 'Transfer learning using convolutional neural networks for face anti-spoofing,' in *International conference image analysis and recognition*, Springer, 2017, pp. 27–34.

[42] 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,' *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.

[43] I. Manjani, S. Tariyal, M. Vatsa, R. Singh and A. Majumdar, 'Detecting silicone mask-based presentation attack via deep dictionary learning,' *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, 2017.

[44] J. Liu and A. Kumar, 'Detecting presentation attacks from 3d face masks under multispectral imaging,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 47–52.

[45] D. Zhou, D. Petrovska-Delacrétaz and B. Dorizzi, '3d active shape model for automatic facial landmark location trained with automatically generated landmark points,' in *2010 20th International Conference on Pattern Recognition*, IEEE, 2010, pp. 3801–3805.

[46] Y. Tang and L. Chen, 'Shape analysis based anti-spoofing 3d face recognition with mask attacks,' in *International Workshop on Representations, Analysis and Recognition of Shape and Motion FroM Imaging Data*, Springer, 2016, pp. 41–55.

[47] B. Hamdan and K. Mokhtar, 'The detection of spoofing by 3d mask in a 2d identity recognition system,' *Egyptian Informatics Journal*, vol. 19, no. 2, pp. 75–82, 2018.

[48] B. Hamdan and K. Mokhtar, 'A self-immune to 3d masks attacks face recognition system,' *Signal, Image and Video Processing*, vol. 12, no. 6, pp. 1053–1060, 2018.

[49] Y. Wang, S. Chen, W. Li, D. Huang and Y. Wang, 'Face anti-spoofing to 3d masks by combining texture and geometry features,' in *Chinese Conference on Biometric Recognition*, Springer, 2018, pp. 399–408.

[50] G. Pan, L. Sun, Z. Wu and Y. Wang, 'Monocular camera-based face liveness detection by combining eyeblink and scene context,' *Telecommunication Systems*, vol. 47, no. 3, pp. 215–225, 2011.

[51] G. Easley, D. Labate and W.-Q. Lim, 'Sparse directional image representations using the discrete shearlet transform,' *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 25–46, 2008.

[52] Y. Li, L.-M. Po, X. Xu and L. Feng, 'No-reference image quality assessment using statistical characterization in the shearlet domain,' *Signal Processing: Image Communication*, vol. 29, no. 7, pp. 748–759, 2014.

[53] C. Liu *et al.*, 'Beyond pixels: Exploring new representations and applications for motion analysis,' Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[54] Y. Liu, A. Jourabloo and X. Liu, 'Learning deep models for face anti-spoofing: Binary or auxiliary supervision,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 389–398.

[55] Y. Atoum, Y. Liu, A. Jourabloo and X. Liu, 'Face anti-spoofing using patch and depth-based cnns,' in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2017, pp. 319–328.

[56] A. Sikdar, S. K. Behera and D. P. Dogra, 'Computer-vision-guided human pulse rate estimation: A review,' *IEEE reviews in biomedical engineering*, vol. 9, pp. 91–105, 2016.

[57] C. Wang, T. Pun and G. Chanel, 'A comparative survey of methods for remote heart rate detection from frontal face videos,' *Frontiers in bioengineering and biotechnology*, vol. 6, p. 33, 2018.

[58] P. Gupta, B. Bhowmick and A. Pal, 'Mombat: Heart rate monitoring from face video using pulse modeling and bayesian tracking,' *Computers in biology and medicine*, vol. 121, p. 103 813, 2020.

[59] G. Balakrishnan, F. Durand and J. Guttag, 'Detecting pulse from head motions in video,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3430–3437.

[60] R. Irani, K. Nasrollahi and T. B. Moeslund, 'Improved pulse detection from head motions using dct,' in *2014 international conference on computer vision theory and applications (VISAPP)*, IEEE, vol. 3, 2014, pp. 118–124.

[61] M.-Z. Poh, D. J. McDuff and R. W. Picard, 'Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.,' *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[62] A. M. Rodrıéguez and J. Ramos-Castro, 'Video pulse rate variability analysis in stationary and motion conditions,' *Biomedical engineering online*, vol. 17, no. 1, pp. 1–26, 2018.

[63] H. Rahman, M. U. Ahmed and S. Begum, 'Non-contact physiological parameters extraction using facial video considering illumination, motion, movement and vibration,' *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 88–98, 2019.

[64] P. Viola and M. Jones, 'Rapid object detection using a boosted cascade of simple features,' in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, IEEE, vol. 1, 2001, pp. I–I.

[65] W. Verkruysse, L. O. Svaasand and J. S. Nelson, 'Remote plethysmographic imaging using ambient light.,' *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[66] G. De Haan and V. Jeanne, 'Robust pulse rate from chrominance-based rppg,' *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[67] D. N. Tran, H. Lee and C. Kim, 'A robust real time system for remote heart rate measurement via camera,' in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2015, pp. 1–6.

[68] K.-Z. Lee, P.-C. Hung and L.-W. Tsai, 'Contact-free heart rate measurement using a camera,' in *2012 Ninth Conference on Computer and Robot Vision*, IEEE, 2012, pp. 147–152.

[69] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu and S.-H. Chang, 'Siamese-rppg network: Remote photoplethysmography signal estimation from face videos,' in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 2066–2073.

[70] M. Kumar, A. Veeraraghavan and A. Sabharwal, 'Distanceppg: Robust non-contact vital signs monitoring using a camera,' *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[71] L. Feng, L.-M. Po, X. Xu, Y. Li, C.-H. Cheung, K.-W. Cheung and F. Yuan, 'Dynamic roi based on k-means for remote photoplethysmography,' in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 1310–1314.

[72] R. Song, J. Li, M. Wang, J. Cheng, C. Li and X. Chen, 'Remote photoplethysmography with an eemd-mcca method robust against spatially uneven illuminations,' *IEEE Sensors Journal*, 2021.

[73] H. Monkaresi, R. A. Calvo and H. Yan, 'A machine learning approach to improve contactless heart rate monitoring using a webcam,' *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1153–1160, 2013.

[74] H. E. Tasli, A. Gudi and M. den Uyl, 'Remote ppg based vital sign measurement using adaptive facial regions,' in *2014 IEEE international conference on image processing (ICIP)*, IEEE, 2014, pp. 1410–1414.

[75] D. McDuff, S. Gontarek and R. W. Picard, 'Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera,' *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 12, pp. 2948–2954, 2014.

[76] Y. Hsu, Y.-L. Lin and W. Hsu, 'Learning-based heart rate detection from remote photoplethysmography features,' in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 4433–4437.

[77] B. Chwyl, A. G. Chung, J. Deglint, A. Wong and D. Clausi, 'Remote heart rate measurement through broadband video via stochastic bayesian estimation,' *Journal of Computational Vision and Imaging Systems*, vol. 1, no. 1, 2015.

[78] M.-C. Li and Y.-H. Lin, 'A real-time non-contact pulse rate detector based on smartphone,' in *2015 International Symposium on Next-Generation Electronics (ISNE)*, IEEE, 2015, pp. 1–3.

[79] T. Carlo and T. Kanade, 'Detection and tracking of point features,' *Int'l Journal of Computer*, 1991.

[80] J. Shi *et al.*, 'Good features to track,' in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, IEEE, 1994, pp. 593–600.

[81] L. Feng, L.-M. Po, X. Xu and Y. Li, 'Motion artifacts suppression for remote imaging photoplethysmography,' in *2014 19th International Conference on Digital Signal Processing*, IEEE, 2014, pp. 18–23.

[82] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, 'Speeded-up robust features (surf),' *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[83] D. Comaniciu, V. Ramesh and P. Meer, 'Kernel-based object tracking,' *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[84] W. Wang, S. Stuijk and G. De Haan, 'Exploiting spatial redundancy of image sensor for motion robust rppg,' *IEEE transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, 2014.

[85] J. F. Henriques, R. Caseiro, P. Martins and J. Batista, 'Exploiting the circulant structure of tracking-by-detection with kernels,' in *European conference on computer vision*, Springer, 2012, pp. 702–715.

[86] G. R. Tsouri and Z. Li, 'On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras,' *Journal of biomedical optics*, vol. 20, no. 4, p. 048 002, 2015.

[87] P. Sahindrakar, G. de Haan and I. Kirenko, 'Improving motion robustness of contact-less monitoring of heart rate using video analysis,' *Technische Universiteit Eindhoven, Department of Mathematics and Computer Science*, 2011.

[88] S. Xu, L. Sun and G. K. Rohde, 'Robust efficient estimation of heart rate pulse from video,' *Biomedical optics express*, vol. 5, no. 4, pp. 1124–1135, 2014.

[89] J. Hernandez-Ortega, J. Fierrez, E. Gonzalez-Sosa and A. Morales, 'Continuous presentation attack detection in face biometrics based on heart rate,' in *Video Analytics. Face and Facial Expression Recognition*, Springer, 2018, pp. 72–86.

[90] G. De Haan and A. Van Leest, 'Improved motion robustness of remote-ppg by using the blood volume pulse signature,' *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[91] L. Feng, L.-M. Po, X. Xu, Y. Li and R. Ma, 'Motion-resistant remote imaging photoplethysmography based on the optical properties of skin,' *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, 2014.

[92] A. Hyvärinen and E. Oja, 'Independent component analysis: Algorithms and applications,' *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[93] S. Wold, K. Esbensen and P. Geladi, 'Principal component analysis,' *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[94] J.-F. Cardoso, 'High-order contrasts for independent component analysis,' *Neural computation*, vol. 11, no. 1, pp. 157–192, 1999.

[95] Z. Guo, Z. J. Wang and Z. Shen, 'Physiological parameter monitoring of drivers based on video data and independent vector analysis,' in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, pp. 4374–4378.

[96] C. S. Pilz, S. Zaunseder, J. Krajewski and V. Blazek, 'Local group invariance for heart rate estimation from face videos in the wild,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1254–1262.

[97] W. Wang, S. Stuijk and G. De Haan, 'A novel algorithm for remote photoplethysmography: Spatial subspace rotation,' *IEEE transactions on biomedical engineering*, vol. 63, no. 9, pp. 1974–1984, 2015.

[98] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng and X. Chen, 'Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks,' *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7411–7421, 2020.

[99] X. Niu, H. Han, S. Shan and X. Chen, 'Synrhythm: Learning a deep heart rate estimator from general to specific,' in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 3580–3585.

[100] O. Russakovsky, J. Deng, H. Su, J. Krorause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, 'Imagenet large scale visual recognition challenge,' *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[101] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan and X. Chen, 'Robust remote heart rate estimation from face utilizing spatial-temporal attention,' in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, IEEE, 2019, pp. 1–8.

[102] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[103] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong and A. El Saddik, 'Evm-cnn: Real-time contactless heart rate estimation from facial video,' *IEEE transactions on multimedia*, vol. 21, no. 7, pp. 1778–1787, 2018.

[104] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand and W. Freeman, 'Eulerian video magnification for revealing subtle changes in the world,' *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[105] W. Chen and D. McDuff, 'Deepphys: Video-based physiological measurement using convolutional attention networks,' in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

[106] R. Špetlıék, V. Franc and J. Matas, 'Visual heart rate estimation with convolutional neural network,' in *Proceedings of the British Machine Vision Conference, Newcastle, UK*, 2018, pp. 3–6.

[107] Z. Yu, X. Li and G. Zhao, 'Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks,' *arXiv preprint arXiv:1905.02419*, 2019.

[108] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, 'Joint face detection and alignment using multitask cascaded convolutional networks,' *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[109] G. Welch, G. Bishop *et al.*, 'An introduction to the kalman filter,' 1995.

[110] R. J. Qian, M. I. Sezan and K. E. Matthews, 'A robust real-time face tracking algorithm,' in *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, IEEE, vol. 1, 1998, pp. 131–135.

[111] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat and J. Jatakia, 'Human skin detection using rgb, hsv and ycbcr color models,' *arXiv preprint arXiv:1708.02694*, 2017.

[112] J. K. Kim and J. M. Ahn, 'Design of an optimal digital iir filter for heart rate variability by photoplethysmogram,' *International Journal of Engineering Research and Technology*, vol. 11, no. 12, pp. 2009–2021, 2018.

[113] Y. Chen, D. Li, Y. Li, X. Ma and J. Wei, 'Use moving average filter to reduce noises in wearable ppg during continuous monitoring,' in *eHealth 360°*, Springer, 2017, pp. 193–203.

[114] M. P. Tarvainen, P. O. Ranta-Aho and P. A. Karjalainen, 'An advanced detrending method with application to hrv analysis,' *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[115] H. J. Nussbaumer, 'The fast fourier transform,' in *Fast Fourier Transform and Convolution Algorithms*, Springer, 1981, pp. 80–111.

[116] P. Welch, 'The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,' *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[117] R. Aisuwarya, H. Hendrick and M. Meitiza, 'Analysis of cardiac frequency on photoplethysmograph (ppg) synthesis for detecting heart rate using fast fourier transform (fft),' in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, IEEE, 2019, pp. 391–395.

[118] S. Winograd, 'On computing the discrete fourier transform,' *Mathematics of computation*, vol. 32, no. 141, pp. 175–199, 1978.

[119] S. A. Akar, S. Kara, F. Latifoğlu and V. Bilgic, 'Spectral analysis of photoplethysmographic signals: The importance of preprocessing,' *Biomedical Signal Processing and Control*, vol. 8, no. 1, pp. 16–22, 2013.

[120] W. S. Noble, 'What is a support vector machine?' *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[121] G. Boccignone, D. Conte, V. Cuculo, A. D'Amelio, G. Grossi and R. Lanzarotti, 'An open framework for remote-ppg methods and their assessment,' *IEEE Access*, vol. 8, pp. 216 083–216 103, 2020.

[122] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa and A. Noore, 'Face presentation attack with latex masks in multispectral videos,' in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 275–283. DOI: 10.1109/CVPRW.2017.40.

Appendix A

Additional Material

A.1 Multispectral Latex Mask based Video Face Presentation Attack Database(MLFP)

Agarwal *et al.* [122] introduced first public face presentation attack database, where all the videos were recorded in three different spectrums: visible(VIS), near-infrared (NIR) and thermal. The database is collected from the 10 subjects among them 4 were females and 6 were males, the age of the subject falls in the range of 23-38 years[122]. In total there are 1350 videos from 10 subjects, among them 1200 videos are attacks videos(mask videos) and remaining 150 videos are without mask(real videos)[122]. Altogether Agarwal *et al.* [122] captured 200,000 frames. Agarwal *et al.* [122] collected the dataset in two indoor and outdoor, over a three months period and the environmental temperature ranging from (-15 to 15) degree celsius. In MLFP database, two types of masks were utilized: 3D latex masks and 2D paper masks.The properties of 3D Latex Masks were soft and they conform to the subject's face shape and allow life-like movements [122].While, 2D paper masks were created using high resolution images on high quality card paper [122]. In ten session, each suject weared seven 3D latex Masks(six masks over entire face and seventh mask cover the face region below the eyes) and three 2D paper masks[122].
The videos were collected in two different acquisition environments: indoor and outdoor with two different backgrounds: fixed and random in visible and thermal spectrum [122]. Agarwal *et al.* [122] utilized three devices were utilized under three different spectrum's. For visible spectrum videos collection Agarwal *et al.* [122] used Android smartphones 8 megapixels camera at frame resolution of,280×720 pixels. Agarwal *et al.* [122] used LIR ONE thermal imager for Android for thermal data collection,on the operating temperature range of 32°F to 95°F with 640×480video resolution. Lastly, Agarwal *et al.* [122] collected the videos in NIR spectrum using Microsoft Kinect for Windows V23with output video resolution of 424×512.

A.1.1 Experimental protocol for MLFP

We followed the experimental protocol represent in[1]. The MLFP database is divided into a sbuject and unseen mask for training and testing protocol. The datbaase in divided into three random subjects folds were selected where each subject masks has 10 mask vidoes. Out of these 9 mask,since tenth mask is half mask and utilized in the testing set, one paper mask and two latex mask were randomly chosen in the training fold for three subject, and for the testing set consists of remaining subjects.

Training				Testing			
Subject Folds	Ids	Mask Folds	Ids	Subject Folds	Ids	Mask Folds	Ids
1	1, 2, and 3	1	2, 3, and 8	1	4 to 10	1	1,4,5,6,7,9,10
	1, 2, and 3	2	4,5, and 9		4 to 10	2	1,2,3,6,7,8,10
	1, 2, and 3	3	6, 7, and 10		4 to 10	3	1,2,3,4,5,8,9
2	4,5, and 6	1	2, 3, and 8	2	1 to 3 and 7 to 10	1	1,4,5,6,7,9,10
	4,5, and 6	2	4,5, and 9		1 to 3 and 7 to 10	2	1,2,3,6,7,8,10
	4,5, and 6	3	6, 7, and 10		1 to 3 and 7 to 10	3	1,2,3,4,5,8,9
3	7, 8, 9 and 10	1	2, 3, and 8	3	1 to 6	1	1,4,5,6,7,9,10
	7, 8, 9 and 10	2	4,5, and 9		1 to 6	2	1,2,3,6,7,8,10
	7, 8, 9 and 10	3	6, 7, and 10		1 to 6	3	1,2,3,4,5,8,9

Figure A.1: Experiment protocol on MLFP Agarwal *et al.* [122]

A.1.2 Results for MLFP Dataset

For the evaluation protocol, we followed the experimental protocol presented on MLFP [122]. The subject is divided into three folds, for training and testing. For each training fold three mask(one paper mask and two latex mask) were chosen and for testing remaining seven subjects were undertaken.

Method	APCER %	BPCER%	ACER	EER%	AUC%
proposed	46.16 ± 4.418	42.03 ± 31.35	46.36 ± 10.21	46.16 ± 4.418	55 ± 0.01

Table A.1: Result for MLFP dataset

[1] http://iab-rubric.org/resources/mlfp.html

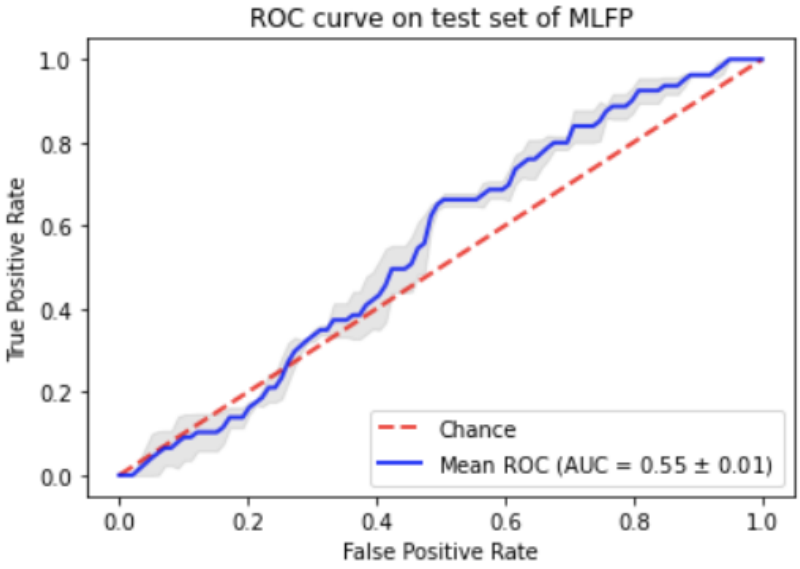


Figure A.2: ROC curve following the experiment protocol on MLFP Agarwal *et al.* [122])

A.2 Cross dataset testing

A.2.1 Results for 3DMAD and MLFP

For the evaluation protocol we undertaken,firs we train Machine Learning (ML) model with 3DMAD dataset and test with MLFP dataset. In second, we train Machine Learning (ML) model with MLFP dataset and test with 3DMAD dataset.

Method	APCER %	BPCER%	ACER	EER%	AUC %
proposed	49.1	52.7	50.0	49.1	50.0

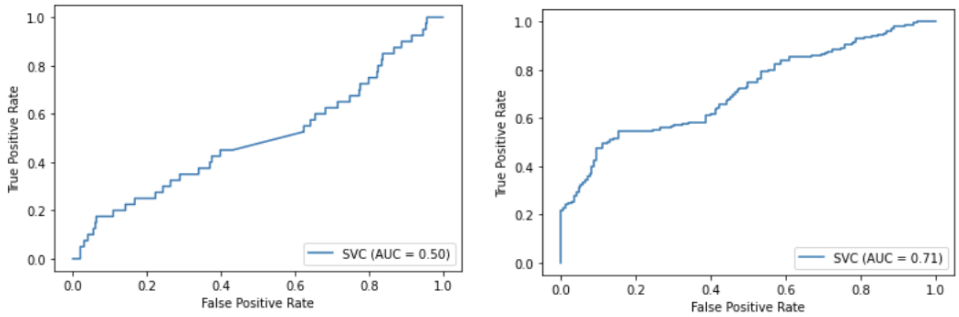
Table A.2: Result for 3DMAD as training and testing as MLFP dataset

Method	APCER %	BPCER%	ACER	EER%	AUC%
proposed	39.7	39.4	39.6	39.7	69

Table A.3: Result for MLFP as training set and testing set as 3DMAD dataset

A.2.2 Results for 3DMAD and HKBVMarsv1+

The result is presented by training the SVM classifier with 3DMAD dataset and test upon with HKBVMarsv1+ dataset.



(a) ROC curve for 3DMAD as training set and testing set as MLFP dataset

(b) ROC curve for MLFP as training and testing as 3DMAD dataset

Figure A.3: ROC curve for cross data testing in 3DMAD and MLFP

Method	APCER %	BPCER%	ACER	EER%	AUC%
proposed	50.8	60	55.4	50.8	62

Table A.4: Result for 3DMAD as training and testing as HKBVMarsv1+ dataset

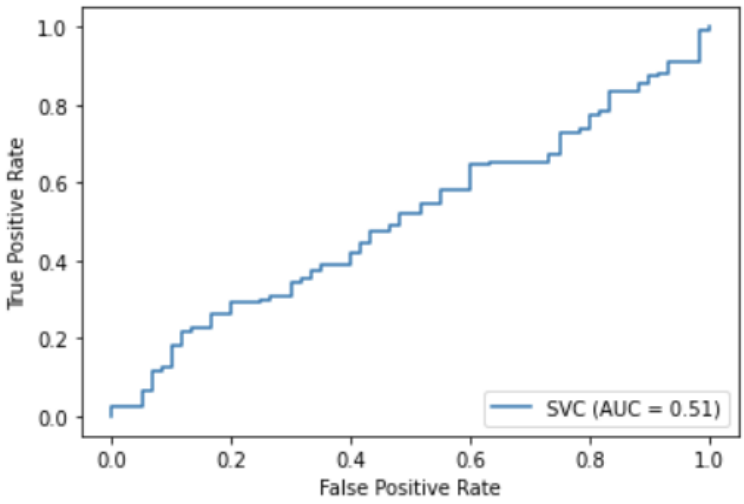


Figure A.4: ROC curve for 3DMAD as training dataset and testing as HKBVMarsv1+ dataset

A.2.3 Results for MLFP and HKBVMarsv1+

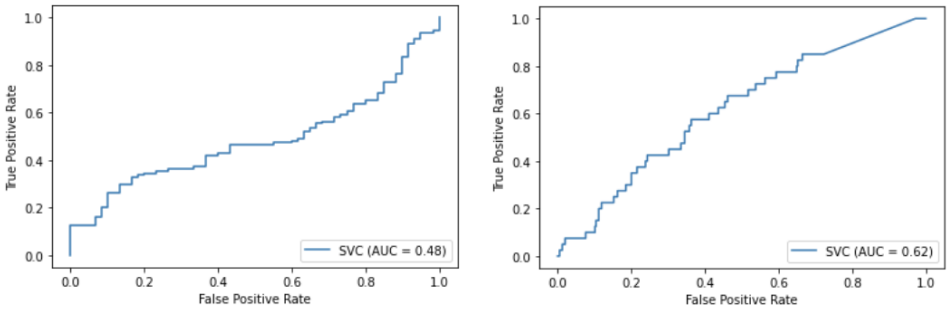
The result is presented by training the SVM classifier with MLFP dataset and test upon with HKBVMarsv1+ dataset. In second, we train Machine Learning (ML) model with HKBVMarsv1+ dataset and test with we dataset

Method	APCER %	BPCER%	ACER	EER%	AUC%
proposed	54.2	53.6	53.9	54.3	48

Table A.5: Result for MLFP as training and testing as HKBVMarsv1+ dataset

Method	APCER %	BPCER%	ACER	EER%	AUC%
proposed	40.6	42.5	41.5	40.6	62

Table A.6: Result for HKBVMarsv1+ as training and testing as MLFP dataset



(a) ROC curve for MLFP as training and HKBV-Marsv1+ testing as dataset

(b) ROC curve for HKBVMarsv1 as training and testing as MLFP dataset

Figure A.5: ROC curve for cross data testing in HKBVMarsv1+ and MLFP

A.2.4 Results for development set on HKBVMarsv1+

The result is reported on the development set on HKBVMarsv1+. The evaluation protocol is highest in section 5.1.2. This section

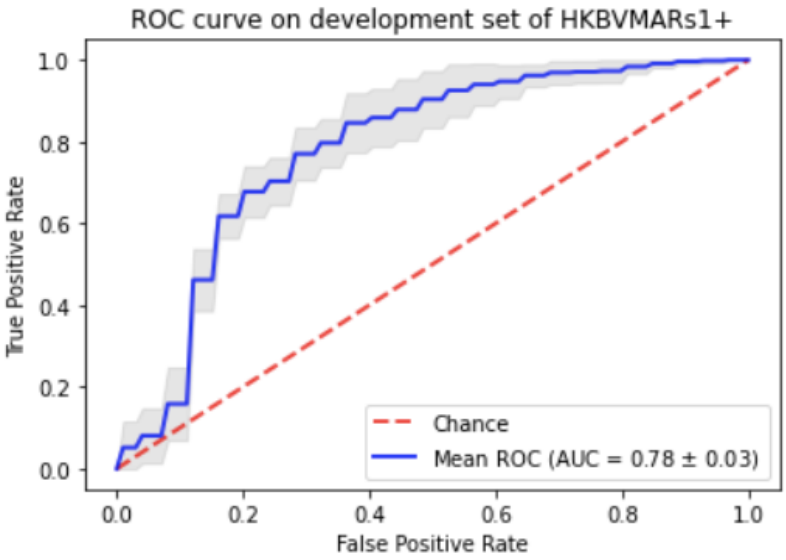
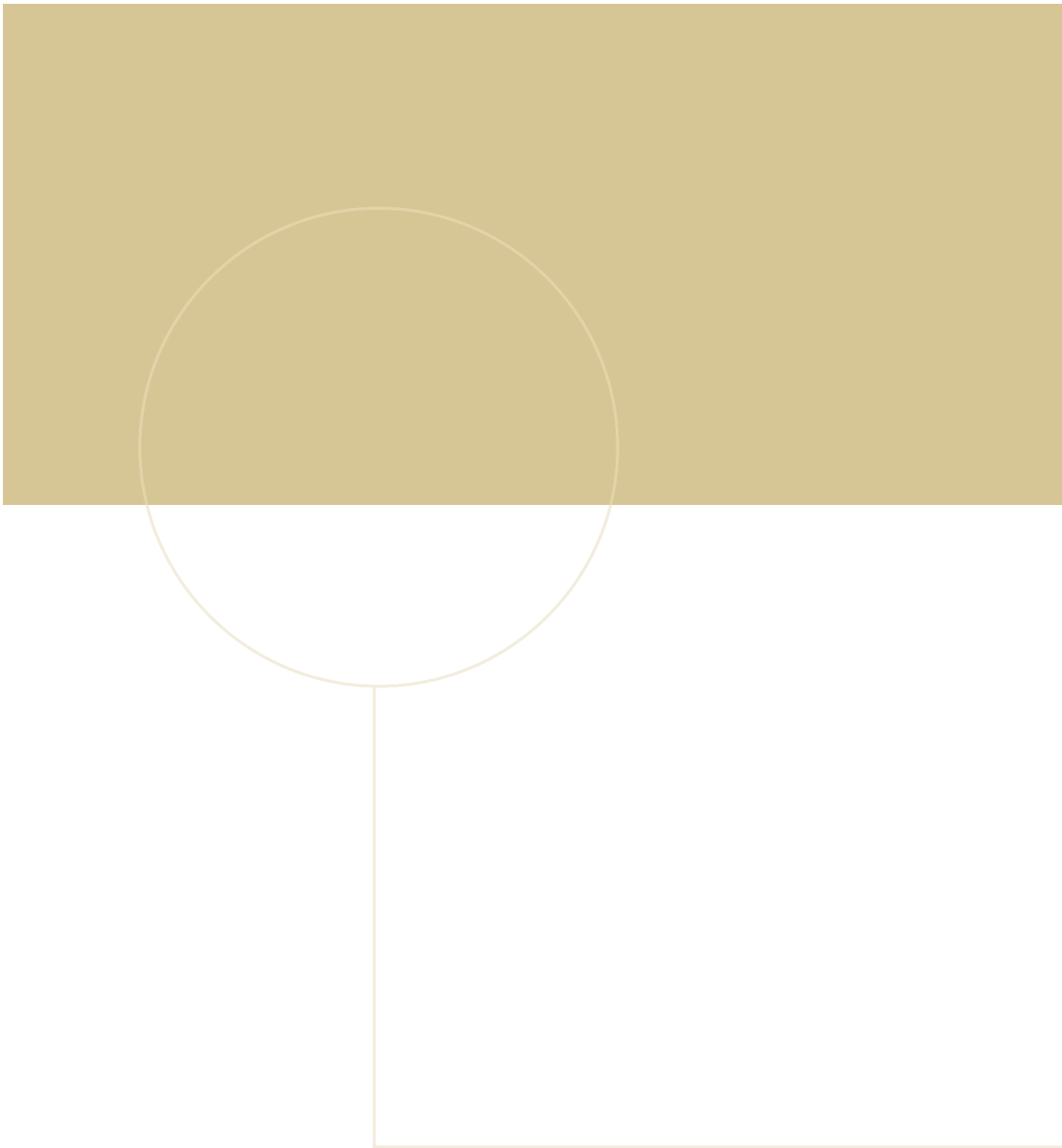


Figure A.6: ROC curve for HKBVMarsv1+ on development set



NTNU

Norwegian University of
Science and Technology