Tensor term decomposition for remote photoplethysmography

Cheremy Pongajow[1], Brecht Dhuyvetters[2], Tanguy Sanglet[2], Maxime Mattelin[2], Joeri Tulkens[2]
[1]*Department of Microelectronics, Delft University of Technology, Faculty of Mechanical, Marine and Materials Engineering, Delft, The Netherlands*
[2]*IntelliProve, D'haenestraat 22, Heusden, Belgium*

***Abstract*—Remote photoplethysmography (rPPG) allows for the measurement of vital parameters by capturing subtle light changes in skin through a video camera without the need for physical contact. Heart rate (HR) is one of the essential vital signs used to indicate the physiological health of the human body. With the vast potential of this technology in the future of digital healthcare, remote monitoring of physiological signals has gained significant traction in the research community. This research paper presents a comprehensive overview of the literature and discusses its limitations. In further detail conventional methods and recent advances in deep learning-based methods of rPPG. Additionally, we analyze the implications of research findings and discuss research gaps to guide future explorations.**

***Index Terms*—remote photoplethysmography (rPPG), deep learning, Non-contact heart rate measurement**

I. Introduction

Following the COVID-19 pandemic, multiple industries were brought under a strain. Heavy measures had to be taken as restrictions were placed to further limit the spread of the health crisis. Although all industries were victims of the pandemic, the healthcare sector was hit pervasively. The healthcare sector is built on social interaction between patients and medical professionals. However, due to protocols physicians minimized in-person contact to avoid spreading the disease (Leibner, Stokes, Ahmad, & Legome, 2021). To limit the spread of disease, increasing pressure was placed on the healthcare sector to introduce and adopt innovative solutions to deliver and optimize patient care. As a result, telemedicine was widely adopted in numerous healthcare settings (Ali et al., 2020). Prior to the pandemic, fewer than 2% of medical professionals delivered care through telemedicine. Currently, at least half of medical professionals have utilized telemedicine for their appointments. The shift towards telemedicine and the availability of low-cost smartphones emphasises the need for the development of technology that can provide medical professionals with vital information about their patients.

The monitoring of changes in vital signals traditionally was an invasive procedure, involving the insertion of sensors into the body. However, advancements in technology have now made it possible to measure vital signs non-invasively (Burritt, 1998). Non-invasive methods can further be divided into two main categories: those that require contact through the skin, and those that measure remotely. The former known as the contact-based method measures physiological signals through changes in physical properties such as pressure, temperature and transmitted light (Van Egmond, Hasenbos, & Crul, 1985). While the latter, known as non-contact methods, collects information on physiological signals through using video, audio, infrared, or ultrasound, Doppler-based methods (Tohma, Nishikawa, Hashimoto, Yamazaki, & Sun, 2021; Villarroel et al., 2014). Non-contact methods have gained momentum as it provides an outcome for telemedicine because it doesn't require a constrained clinical environment. Not only is it cost-effective, but it is also suitable for continuous and long-term monitoring without being inconvenient or uncomfortable.

Human vital signs provide crucial information about the person's physiological status and emotional state. Commonly used indicators for measuring physiological state include body temperature (BT), respiratory rate (RR), blood oxygen saturation (SpO2), heart rate variability (HRV), and blood pressure (BP) (Li, Chen, Zhao, & Pietikainen, 2014). One of the most important physiological parameters which indicate a person's health is the heart rate (HR). HR indicates the number of times a person's heart beats per minute. HR fluctuation depends on a person's physical activity as well as emotional state. It is an important parameter as HR monitoring can help detect and prevent cardiovascular problems such as atherosclerosis, arrhythmias, angina, and coronary artery disease.

A. Traditional Methods

Traditional measurements of vital parameters are done by electrocardiogram (ECG), sphygmomanometer, and pulse oximeters. ECG is widely considered to be the gold standard for measuring HR, due to its high accuracy and reliability (Qiao, Ayesha, Zulkernine, Jaffar, & Masroor, 2022). However, as mentioned before, these methods have a tedious process as it causes discomfort for the patients due to the need for a gel on the chest area to attach the electrodes (Qiao et al., 2022). Mercury and sphygmomanometer are typically used for BP measurements. These devices measure the BP by gradually increasing and decreasing the pressure of the cuff around the upper arm by inflation. These methods proved an accurate measurement, however, they can cause severe discomfort and even pain for some individuals. The pressure applied to the arm during the measurement process may be uncomfortable or even painful for some patients, particularly those with sensitive skin or underlying conditions that affect the circulatory system. Therefore, ECG, Mercury, and sphygmomanometer aren't always practical for continuous monitoring of HR and

BP, as they can cause discomfort and requires the patient to be in a constrained clinical environment.

B. Remote Photoplethysmograph

Photoplethysmography (PPG) is a common method for measuring HR, it offers an inexpensive and straightforward alternative with a 98% level of accuracy (Kim, Lee, & Sohn, 2021). PPG is a contact-based method that extracts pulse signals through the illumination of the skin with a light-emitting diode (LED) while measuring the amount of reflected light by the skin. By acquiring the optical property changes in blood vessels on the skin pulse waveform is extracted to find out the state of the HR. Beer-Lambert's Law is the principle that describes the amount of light absorbed by a single substance as it's proportional to its concentration (Taparia, Platten, Anderson, & Sniadecki, 2017). Haemoglobin is a protein in the bloodstream that has a high absorption rate of light at 532 nanometers. When the light of the LED is passed through the skin, some of it's reflected and some are transmitted. (Swinehart, 1962). By measuring the amount of the absorbed light at the specific green wavelength of 532 nm is possible to determine the PPG signal by the difference in concentration of haemoglobin (Kim et al., 2021).

During the last decade, considerable research has been published on non-contact methods that provide insight into the extraction of pulse signals by evaluating motion and colour-based methods (Li et al., 2014; Balakrishnan, Durand, & Guttag, 2013). Motion based-methods, retrieve pulse signals from the cyclical flow of blood from the heart to the head through the abdominal aorta and carotid arteries causing the head to move or change in colour pattern. Recent research proposed a novel algorithm where they tracked head oscillations caused by cardiovascular circulation and principal component analysis (PCA) to extract pulse signals (Balakrishnan et al., 2013). The motion-based method relies on subjects staying stationary and upright during video recording. Unfortunately, this HR estimation method is incorrect and error-prone in real-life situations due to the effect of temporal deviation, facial expressions, and illumination variations that result in noise (Gupta, Bhowmik, & Pal, 2018).

Recent advancements in rPPG methods have resulted in the colour-based method. The colour-based method acquires PPG waveforms by analyzing subtle colour changes of the facial skin from a digital camera (e.g. webcam, RGB, camera, near-infrared camera) (J. Chen et al., 2016). In this instance, the digital camera functions as the photodetector that captures the subtle colour changes of the skin. Also, instead of using a LED with a fixed wavelength, the ambient light functions as the luminosity source. In Figure 1 a schematic overview of the rPPG principle is displayed.

rPPG methods are very promising as they not only eliminate the discomfort of intrusive but also allow for continuous monitoring of vital parameters, without the need for physical contact. In a recent paper, they detected a region of interest (ROI) like the face area and they compute the mean pixel values of each image frame from the RGB channels for extracting

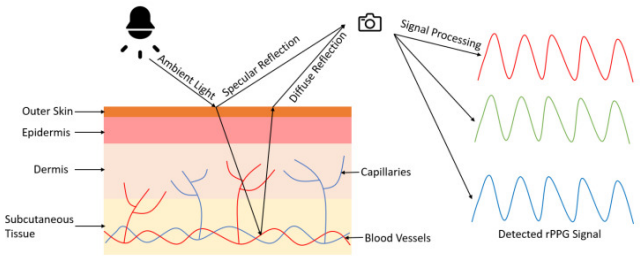

Fig. 1. Principle of remote photopethysmography (Cheng et al., 2021)

PPG signals (Poh, McDuff, & Picard, 2010a). This method was further improved by implementing several temporal filters before and after independent component analysis (ICA) (Poh et al., 2010a). With ICA, three independent signals are defined by a linear combination of all three colour channels, and non-Gaussianity is used as the criterion for independence. Other advanced ICA methods have been proposed that achieve very high accuracy for measuring HR on their data. The researchers used built-in video cameras of a smartphone for recording facial videos and utilised the raw green trace signal and ICA separation sources to extract the PPG wave-signal (Kwon, Kim, & Park, 2012). Later variations of the method were introduced that defined three independent linear combinations of the RGB channels with principal component analysis (PCA) (Lewandowska, Rumiński, Kocejko, & Nowak, 2011). The ICA and PCA methods are both regarded as blind source separation methods (BSS). However, these BSS methods retain motion artifacts which result in limited accuracy in PPG wave signal extraction. Therefore, the chrominance-based (CHROM) rPPG was introduced which improved the robustness. Other research proposed an alternative HR measurement framework which utilised near-infrared (NIR) channels. This technique suggests a more robust method as it isn't influenced by noise of changes in environment illuminations (J. Chen et al., 2016).

Advancements in artificial intelligence and computer vision have led to remarkable breakthroughs. As a result of the advancements in many non-contact-based rPPG processing methods have begun to appear that leverage AI and deep neural networks. Recently research proposed a novel rPPG denoising algorithm to effectively mitigate noise from facial expressions and illuminations (Lokendra & Puneet, 2022). Another novel approach provided a deep learning algorithm that utilises 3D convolutional neural network based on an attention mechanism to predict the HR (Liu, Wei, Kuang, & Ma, 2022).

Above discussed methods extract pixels from a facial video taken by RGB cameras. Various methods for extracting PPG signals have been proposed by analyzing facial videos in which various information algorithms are utilized. The differences between estimated HR and actual physiological signals can be due to multiple sources of noise (De Haan & Jeanne, 2013). The noise of the extracted rPPG signal can vary depending on factors such as the type of camera used, measurement conditions, human error, and demographic biases (Bent, Goldstein,

Kibbe, & Dunn, 2020). Since commonly used datasets for rPPG processing, lack demographic diversity. The majority of the subjects in the dataset are of Euro-American descent (49%) and Asian descent (27%) (Gross, Matthews, Cohn, Kanade, & Baker, 2010), which limits the robustness of the algorithms. The field of rPPG has gathered significant momentum in recent years, with many studies exploring the potential of rPPG presenting a technique without the need for physical contact for measuring vital parameters such as HR (Sun & Thakor, 2015). Despite these advancements, the literature on rPPG is still limited by a number of factors, including noise, demographic variations, and sensor noise variations. These limitations have prevented the widespread adoption of rPPG in real-world settings and have hindered its ability to provide accurate and reliable vital parameter measurements. As such, it is of utmost importance to thoroughly examine the current state of rPPG and identify these limitations in order to improve the efficacy of rPPG and facilitate its implementation in practical applications.

By exploring the current state of remote photoplethysmography vital signal measurements, this research paper aims to provide a comprehensive overview of the literature and its limitations. Additionally, this paper will help to identify the challenges associated with implementing rPPG technology in real-world settings and propose a heart rate (HR) estimation method that mitigates these issues to improve the efficacy of rPPG.

II. Research methodology

In the literature, there isn't a clear consensus of terms describing the research topic. As a result, various search terms were used for examining the literature. The following search terms were used: "remote", "non-contact", "camera-based", "video-based", "contactless", "contact-free", "imaging", "photoplethysmography", "heart rate measurement", "heart rate estimation", heart rate variability, vital signs measurement, vital parameter measurement, oxygen saturation, blood pressure, blind source separation, deep learning, machine learning, convolutional neural networks, transformers, and attention mechanism.

In the process of identifying a wide range of relevant published papers, the previously mentioned terms were used to conduct searches in Google Scholar and PubMed. The research paper includes only papers that use facial video for extracting PPG signals. It's also important to note that searches for estimation of other vital parameters like RR, SpO2, and BT haven't been explicitly used. Research papers that utilized specialised equipment weren't also explicitly used. Research papers that implemented specialised equipment are excluded from the literature review.

III. Related work

The phenomenon of rPPG signal extraction is mainly based on two theoretic frameworks. First, the theory based on conventional PPG assumes penetration and reflection of light through the dermis and arteries (Poh et al., 2010a). Second, a theory that assumes that visible light won't pass down to pulsating arteries (Moco, Stuijk, & De Haan, 2015). Thus the theories differ in the assumption of the depth of penetration of light in the skin, whereas the second theory expects no interaction with deeper blood vessels.

Motion-based methods also known as ballistocardiography focus on capturing the differences in mechanical movement due to the impact of blood flow from carotid arteries (Balakrishnan et al., 2013). The working principle behind ballistocardiography is utilizing Newtonian mechanics to detect cyclical movement of the human body, caused by blood flow. This theoretical framework assumes that the human body is a stacked inverted pendulum. It hypothesizes that the circulation of blood results in an opposite reactive force that causes displacement of the head (Balakrishnan et al., 2013). Although ballistocardiogram can yield waveform signals for HR measurement this approach is left outside of the scope of this research paper.

Back in 1937, Hertzman and Spealman first described that transmission or reflectance of light on the finger could be detected by a photoelectric cell (Hertzman, 1938). The subsequent research that followed expanded upon this knowledge and led to the development of rPPG method. The research paper discovered that facial video recordings of subjects contain sufficient information for PPG wave extraction and HR estimation (Verkruysse, Svaasand, & Nelson, 2008). The researchers instructed the subject the sit motionless while recording facial videos. This paper presented the effect of variations in resolutions and frame rates (fps) on HR estimation through facial videos. The results indicated that a lower resolution led to a higher signal-to-noise ratio. This research introduced the GREEN method, which is commonly employed and utilizes a region of interest that is manually selected by the user. The GREEN method calculates a raw signal from the selected ROI pixels by calculating the mean value for each RGB channel. Then, a 4-th order band-pass filter was used to exclude certain cut-off frequencies. This paper facilitated further exploration as the general feasibility of rPPG was established. It's called GREEN because fast Fourier transformation (FTT) determined the power-spectral density which showed that the green channel contains the strongest PPG signal. This also corresponds with the fact that haemoglobin exhibits the highest absorption rate, therefore it's also the preferred method for BVP extraction due to the fact that the green channel contains more reflective information.

Further strides were made by Poh et al. which introduced BSS and Bland-Altman correction for extracting rPPG signals. They developed a novel automatic face tracking which extracted the ROI frame by frame, with a moving window of 30 seconds to achieve continuous measurement (Poh, McDuff, & Picard, 2010b). The BSS (ICA) algorithm composes the three colour channels of the RGB video through a linear mixture of the source channels. Subsequently, Poh et al. improved their algorithm by addressing the selection of the component with the highest power spectrum instead of always choosing the second component (Poh et al., 2010b). However, the

methods mentioned above still suffer from issues with specular and motion artefacts. Therefore, a more robust approach was proposed by researchers which introduced a chrominance-based signal-extracting processing method, in short CHROM. The CHROM method was developed by a research group at Phillips where they introduce temporal normalization of colour differences, thereby improving robustness to non-white illumination and reducing noise from motion (De Haan & Jeanne, 2013). Although this method improves upon BSS-based methods it still suffers from issues with reflectance from specular noise.

So other researchers suggested a new approach named plane orthogonal to skin-tone (POS) which incorporates a main features from spatial subspace rotation (2SR) which is another existing algorithm (Wang, Den Brinker, Stuijk, & De Haan, 2016). 2SR offers an advantage as the core idea of the algorithm leverages temporal rotations of skin pixels by integrating a subspace of the facial pixels and subsequently determining the rotation angles (Wang, Stuijk, & De Haan, 2015). The POS approach uses physiological-based reasoning projection of axes, thus resulting in a more robust approach as it's less influenced by noisy face masks. Besides these conventional techniques, rPPG pulse extraction has also seen promise with machine and deep learning methods. A deep learning approach was suggested by Špetlík et al., where the HR was predicted as a single scalar value by maximizing the signal-to-noise ratio (SNR) in a 2D convolutional neural network. Their algorithm was validated on three public datasets which contained different motion and lighting conditions (Špetlík, Franc, & Matas, 2018). However, these end-to-end deep-learning approaches require an enormous amount of data for training and optimization. So, such models need further development for clinical applications to be viable.

IV. rPPG Process

The following section provides a comprehensive overview of the most important steps of a general rPPG approach for HR measurement. Each step of the rPPG process for HR measurement will be discussed in detail, including the methods, processing and techniques used, the strengths and limitations of each step.

1) Face Extraction: After the video input $v(t)$ is obtained, the face region is detected and extracted from each video frame, $t = 1, 2..., T$ represents each frame This is an important step as it helps in accurately locating the face in the video.

2) ROI Selection: The ROI selection step involves selecting a region of the face where the blood vessels are prominent and the pulsation of blood can be easily observed. This is usually the region around the cheeks, forehead or temple as these pixels contain PPG related information.

3) RGB computation: In the RGB computation step, the raw signal is extracted from the RGB channels of the video frames. This step involves converting the colour information of the pixels in the ROI into intensity values for each of the RGB channels.

4) Signal Processing: The signal processing step involves the extraction of the rPPG signal from the raw signal. This involves several processing techniques such as filtering, de-noising and baseline correction to improve the SNR, thus increasing the quality of the extracted rPPG signal.

5) rPPG Method: The rPPG method step involves the estimation of HR from the processed rPPG signal. This is done by computing the frequency components of the signal and selecting the peak corresponding to the HR.

6) HR Estimation: The HR estimation step involves the calculation of the HR from the estimated frequency components.

Figure 10 shows a visual representation of the most important steps in a general rPPG process for HR measurement. It showcases the crucial components involved in the process, starting from face extraction, followed by ROI selection, signal processing, rPPG method, and finally HR estimation. The diagram highlights the importance of each step and the relationship between them.

A. Face Extraction

The first step in the process is the acquisition of facial video from the subject. This is usually done by a webcam, smartphone or other video-capturing devices. Given a facial video, the process starts by extracting a portion of the face from each frame. This ROI selection aims to collect the signals with the most information for estimating HR. The ROI is the area within the video frame that contains the raw signal for the algorithm. Research shows that ROI selection is of major influence on the quality of the extracted PPG signal (Kim et al., 2021). It's imperative for the successful implementation of rPPG that the facial region is precisely localized within each video frame. Even slight misalignments of the facial region across frames can result in substantial variations in colour due to the presence of background pixels, leading to substantial noise in the pulse signal.(Kim et al., 2021). This is usually an intermediate step so that later a more precise measurement can be taken. In early research, the bounding box was manually selected by the researchers (Verkruysse et al., 2008). Currently, the most frequent approach that is implemented for selecting the bounding box is the Viola-Jones method (Viola & Jones, 2001). This has mainly to do with the popularity of the OpenCV library in Python, as it's open source and updated frequently. The selection of the bounding box was based on a Haar cascade-face detector. Another popular face extraction approach is the Histograms of Oriented Gradients (HOG) which is also available in the OpenCV library. These approaches also provide facial landmark points e.g. lips, eyebrows, eyes etc. However, it is important to note that these approaches are insufficient when dealing with spatial or appearance-distorted faces, such as those resulting from movement under realistic conditions. With the emergence of deep learning, many algorithms have been able to tackle the face extraction problem so that partial face can be extracted (Qian, Sezan, & Matthews, 1998; Taigman, Yang, Ranzato, & Wolf, 2014).

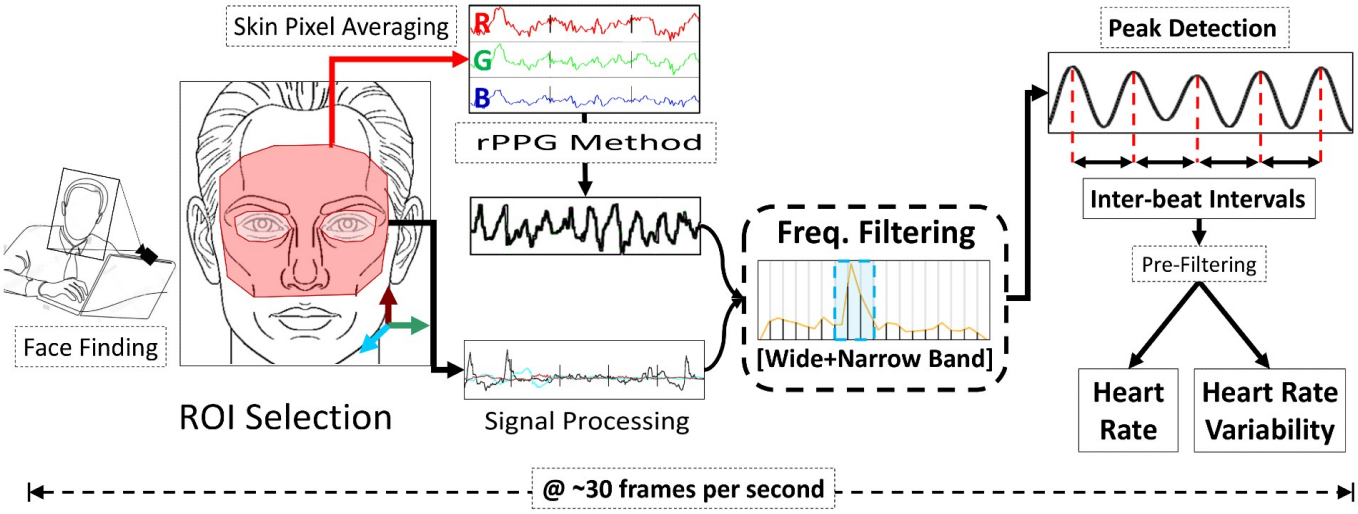

Fig. 2. Schematic overview of rPPG pipeline. (Gudi, Bittner, & van Gemert, 2020)

B. ROI Selection

Obviously, the Viola-Jones method also comes with some drawbacks as it doesn't always select the correct ROI for each frame. It still suffers from issues with a background border around the face area. Furthermore, it's also computationally intensive to run this for each frame as the method has high computational complexity. This is especially troublesome when real-time measurements need to be performed. In addition to this, the facial region contains eyebrows, eyes, glasses etc. which can confound the extracted rPPG signal. Lee et al. proposed an algorithm where skin-like pixels were selected based on a neural network classifier (K.-Z. Lee, Hung, & Tsai, 2012). The neural network takes colour RGB values as input and subsequently passes them through the network which results in a regression-based output. Then by setting a threshold value for $T_{skin} \in [0, 255]$, each pixel can be classified as skin or not. Skin detection can be done by machine learning but also alternative algorithms exist. Kolkur et al. proposed a novel algorithm that threshold skin pixels based on RGB, Hue saturation values (HSV), and YCbCr colour values (Kolkur, Kalbande, Shimpi, Bapat, & Jatakia, 2017). The different colour spaces are necessary as RGB values combine colour (chrominance) and intensity (luminance) information with non-uniform characteristics. While HSV and YCbCr approaches discriminate under uneven lighting conditions. By applying a threshold to the Hue. Cb, and Cr values skin pixels are selected within the bounding box. A threshold can be manually chosen by the researchers or an adaptive threshold can be applied. Adaptive thresholds are selected per frame based on the highest density of pixel distribution. The key idea behind this is that the majority of pixel values will belong to the skin. Therefore, thresholds should exclude less common pixels describing non-skin areas (eg. beards, hairs, eyes). Nonetheless, each face has its own features in terms of skin pigmentation, thus thresholds should exhibit adaptive behaviour. Despite the accurate performance, studies have found that skin detection for darker skin tones remains a challenging task (Wang et al., 2016; Aarts et al., 2013). The difficulty arises due to the decreased SNR as pigmentation levels increase. This can be attributed to the absorption of light by the epidermal layer containing melanin, leading to a reduction in the amount of light reflected by the underlying blood vessels.

Recent research has shown promise in normalizing the RGB based magnitude, which has demonstrated reliable performance across a broad range of skin types (De Haan & Jeanne, 2013). The benefit of skin detection over other approaches is that informative pixels from the shoulder and neck can be included in the evaluation. However, as discussed in the paper a drawback of this approach is that objects with similar colour as the skin will add additional noise (K.-Z. Lee et al., 2012). ROI selection is crucial for extracting accurate rPPG signals. Although some researchers simply use the provided bounding boxes from the Viola-Jones or manual face detection. Other researchers address these issues by combing face extraction and additional skin coordinates selection in the face region to exclude pixels which are non-informative. But these still face challenges with darker skin tones and interference from objects with similar colours as skin. Temporal noise could render the extracted rPPG signal obsolete for use of HR estimation. So, the implementation of ROI tracking aims for the rPPG signal to be invariant from subject motions (Qian et al., 1998). A simple approach is already discussed with the Viola-Jones method as it selects the ROI per frame. But, this is still challenging in real-time situations due to high computational complexity. As a result, other approaches have been proposed such as landmark tracing. An example of a tracking algorithm is the speeded-up robust features (SURF), which traces identified facial points (Bay, Tuytelaars, & Van Gool, 2006). Other researchers utilized kernels to

compensate for the subjects motion and update ROI selection at high FPS (Henriques, Caseiro, Martins, & Batista, 2014). These methods are based on a tracking algorithm that extends the rPPG pipeline which reduces computational complexity and noise attributed to motion artifacts. The landmark tracing helps in accurately defining a ROI to detect the pixel that belongs to facial skin. It can also aid in eliminating areas that aren't of interest such as the eyes, and mouth. The ROI most frequently selected by researchers are the forehead and cheeks as these areas are less susceptible to movement from facial expression (Tasli, Gudi, & Den Uyl, 2014).

C. Signal Processing

After ROI selection from the frames, the raw signal needs to e extracted. The raw signal is a time series signal of the colour channels, $s(t) \in [R, G, B]$. The raw pulse signal usually mixed contains large noises from a variety of sources, including misalignment in face tracking, noises in camera sensors, and illumination changes arising from camera's automatic adjustment (e.g. auto white balance, auto focus, etc.). Therefore the raw signal is pre-processed. The signal values are calculated by spatial averaging of all skin pixels in the ROI video frame. This is a simple way to cope with spatial noise which shows an improved SNR (Verkruysse et al., 2008). Besides the spatial noise, the raw signal contains other unwanted noises depending on illumination and other factors. By applying filters with knowledge about frequencies from the noise sources and other dependent factors the extracted rPPG signal can be improved. As a result, the SNR is increased which in turn improves the quality of the extracted rPPG signal. Various studies differ between when the filtering of the signal is applied either before or after dimensions reduction. All the signal processing techniques discussed are applied at different stages in the pipeline, and various studies use them at different stages. A common filtering method is centralizing the raw signal by subtracting the mean μ_s from the raw signal $s(t)$. An additional step can be done for normalization by dividing it by the σ_s standard deviation of the raw signal. Another filtering process that is frequently implemented is a band-pass filter which suppresses the frequency bandwidth (0.7 Hz to 5 Hz) components outside of the heart rate (40 to 220 bpm) (Boccignone et al., 2020). Thereby decreasing the noise in the rPPG signal. The use of moving average filtering is also frequently utilized in order to reduce the high-frequency components of a signal. It's an effective method for minimizing motion artifacts and noise. The filter is applied by calculating the average value of the input signal over a specified temporal window. Then, the raw signal is substituted for the average of the samples in the window which smooths the raw signal and reduces the high-frequency noise. Another filtering technique that can be applied to raw signals is the hamming windows. The hamming window widely used signal-processing technique. It's applied to the raw signals to reduce spectral leakage and improve the accuracy of the rPPG. The hamming window is used in conjunction with the Fourier Transform to obtain a more accurate representation of the frequency content of the rPPG signal.

The finite impulse response (FIR) and infinite impulse response (IIR) filters are also used in rPPG process. FIR filters have a fixed, finite impulse response and are characterized by a stable and predictable response, making them suitable for applications that require a consistent output. On the other hand, IIR filters have an infinite impulse response and are more efficient, but also more complex. The Butterworth IIR filter is often used in rPPG processing due to its flat passband and a sharp transition from the passband to stopband. Researchers have been exploring new techniques for reducing noise in rPPG signals (Boccignone et al., 2020). For example, they have been eliminating outliers in the signals and using the background illumination as a reference. A simple approach to deal with noise from light is utilizing the background pixels as a reference to estimate the (Deng & Kumar, 2020). In recent studies, researchers have been applying an adaptive filter to remove illumination noise. An adaptive bandpass filter is a novel component that dynamically changes the cutoff frequencies based on previously estimated HR to produce consistent HR estimates. It helps with isolating a narrow frequency band of interest within the larger bandwidth signal generated by rPPG. This technique uses an adaptive algorithm to dynamically modify the filter coefficients and track changes in the frequency content of the input signal.

V. Overview of rPPG methods

In this section, various algorithms are discussed that enhance the robustness and applicability of the rPPG technology to less constrained conditions. Most rPPG processing methods use a raw signal that consists of a multidimensional temporal signal (e.g. RGB). It is assumed that the raw signals contain a dimensional plethysmographic signal $p(t)$, which can be represented as a linear combination of these raw signals using a weighted sum. Estimating the weights for this combination has proven difficult and is one of the most debated issues in the literature on rPPG (Rouast, Adam, Chiong, Cornforth, & Lux, 2018). A full overview of these methods can be seen in the table below, which briefly describes their features. The methods are divided into conventional image-processing approaches and deep-learning approaches, depending on the type of algorithm used. Most conventional methods for remote HR measurement follow a similar framework as shown in Figure 10. While deep learning differs from this as there are various forms from end-to-end frameworks to hybrid models. The following sections discuss studies on rPPG signal processing, and algorithm, highlighting the contributions of the current literature. Table I shows the various methods from the literature and highlight the unique features and characteristics of each method. The last step of HR is an estimation by further post-processing will also be addressed, which typically involves frequency analysis and peak detection.

TABLE I
OVERVIEW OF RPPG METHODS AND THEIR DESCRIPTIVE CHARACTERISTICS

Method	Description
Conventional techniques	
GREEN	Green channel extraction as it contains more reflective information from haemoglobin compared to the blue and red channels.
PCA	A blind source separation technique which extracts uncorrelated components
ICA	Other blind source separation technique to obtain independent components form temporal RGB signal
CHROM	Chrominance-based method implements normalization of colour differences to reduce non-white illumination and motion artifacts
POS	Plane orthogonal to skin leverages temporal normalization of the RGB space.
SSR	The SSR or 2SR method uses subspace rotation and temporal rotation for rPPG pulse extraction.
PBV	PBV utilize knowledge from blood volume changes in different wavelengths to distinguish between pulse signal changes and movement noise.
Deep Learning	
2D CNN	2D convolutional neural networks use end to end frameworks to estimate HR
3D CNN	3D convolutional neural networks use spatiotemporal networks to analyze the temporal information in the video frames to estimate the HR.
RNN	Recurrent neural networks used temporal networks to propagate spatial features from 2D CNN for rPPG signal extraction with LSTM and attention mechanism.
Hybrid Models	Deep learning techniques are applied in some parts of the rPPG pipeline for optimization,extraction or HR estimation.

A. GREEN Method

The GREEN method was first reported by Verkuysse et al. in 2008 (Verkruysse et al., 2008). The Green method is based on the observation that the green channel provides the strongest photoplethysmography signal, corresponding to an absorption peak by oxyhaemoglobin (Verkruysse et al., 2008). The absorption of green light by oxyhaemoglobin is the primary factor that causes the green channel to provide the strongest rPPG signal. This makes the green channel method a reliable and effective way to measure the changes in blood volume and oxygenation in the skin. The blue and red colour channels also contain photoplethysmographic information, but they diffuse less reflective information and thus provide a weaker signal. Wu et al. further illustrated this in their study where they attempted to display the pulse changes over time by maximizing the green channel (Wu et al., 2012)

B. PCA Method

A popular algorithm for BSS is PCA which is a commonly used technique in signal processing and machine learning fields. Leandowska et al. first introduces PCA for rPPG extraction and which was later used by others (Lewandowska et al., 2011; Balakrishnan et al., 2013). PCA is a statistical technique that reduces the dimensionality of the multi-dimensional signals while preserving the maximizing of variance and minimizing the covariance. PCA separates the raw multi-dimensional signals into linearly uncorrelated components and orders them based on variance. PCA calculates the covariance matrix of the multi-channel temporal signals. The covariance matrix is calculated as $C = \frac{1}{m-1}X^T X$, where $X \in \mathbb{R}^{n \times m}$ is the data matrix, with n being the number of channels and m being the number of time samples. Subsequently, the eigenvectors of the covariance matrix are calculated. The eigenvectors corresponding to the principal components of the raw multi-dimensional signal. It"s calculated by $Cv_i = \lambda_i v_i$, where λ_i is the eigenvalue corresponding to the eigenvector v_i. The multi-channel temporal signals are projected onto the principal components. The rPPG signal can be extracted by projecting the data onto the eigenvectors corresponding to the first few PCs, represented as $Y = Xv_1, Xv_2, \ldots, Xv_k$, where k is the number of PCs used for the projection. PCA helps identify the underlying structure in the multi-channel temporal signals of rPPG and extract the most important information.

C. ICA Method

Another popular BSS is ICA. ICA is also a statistical technique, but it aims at decomposing a linear mixture of sources under the assumption of non-Gaussianity and independence. Poh et al. introduced ICA for deriving rPPG signals from three RGB colour channels. The calculation process of ICA for a raw multi-dimensional temporal signal $x(t) \in \mathbb{R}^{m \times 1}$ can be done by transferring the signal to a new coordinate system, where m is the number of channels (Poh et al., 2010b). ICA splits the multidimensional signal into multiple components $x(t) \in \mathbb{R}^{m \times 1}$, where m is the number of channels. The first step is t get the whitening matrix $W \in \mathbb{R}^{m \times m}$ by using a Jacobian rotation. The raw multi-dimensional temporal signal $x(t)$ is transformed into a white signal $y(t)$ by calculating $y(t) = Wx(t)$. The independent components are obtained by calculating $s(t) = Wy(t) = W^2 x(t)$. Poh et al. used the joint approximate diagonalization of eigenmatrices (JADE) method to separate the mixed signal $x(t)$ into four independent components $s(t)$. The rPPG signal was empirically determined by selecting the second signal. This is explained by the authors as the second component usually is the most periodic signal. However, in later studies, Poh et al. invalidated this assumption, as theoretically the order of the ICA is random. An improved algorithmic version employs selection criteria to determine the rPPG. An example of selection criteria is to choose the independent component with the highest peak in the frequency spectrum (Poh et al., 2010a). Another simple criterion that can be done is choosing the highest periodicity according to the percentage of spectral power. The criteria for

component selection are equally applicable to PCA. In Figure 3 an example of ICA is illustrated. The figure displays the decomposition of a multidimensional signal into its independent components. The selection of the independent component was also addressed by machine learning techniques. They proposed a k-nearest neighbour algorithm which outperformed the other manual selection approach and the criterion mentioned before (Monkaresi, Calvo, & Yan, 2013).

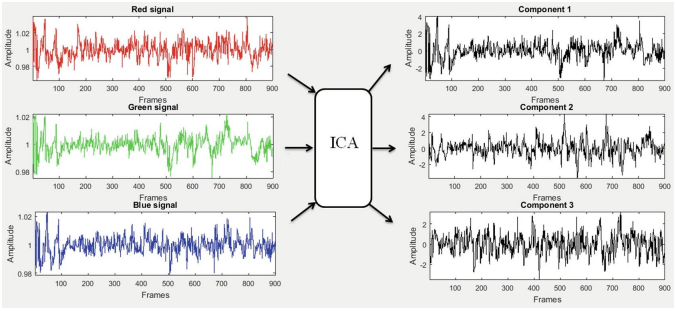

Fig. 3. Schematic overview of ICA (Waqar et al., 2021)

D. CHROM Algorithm

The CHROM method has been proposed to address the weakness of other methods, as it enhances against motion robustness. The CHROM method leverages the unique characteristic of facial skin colour changes, which are caused by alterations in the cardiac cycle, and can be more pronounced than variations in intensity. The variation in colour exists out of two components. First, the diffuse reflection is related to blood flow, and second, the specular reflections, constitute the colour of the light source and don't contain any pulse signal. Under realistic conditions, the contribution of these two reflections is dependent on the angle between the camera, skin, and light source (De Haan & Jeanne, 2013). Thus, the previously mentioned methods are less robust against motion artifacts as they don't eliminate the additive specular component. So, CHROM tackles this by removing the specular component resulting in enhanced motion robustness. After pre-processing with centralization of the raw signal $x(t)$, the values are projected into two orthogonal chrominance vectors X_{CHROM} and Y_{CHROM} (Moço, Stuijk, & de Haan, 2016). The two vectors are calculated as follows, with r,g, and b representing the respective channels:

$$X_{CHROM}(t) = 3x_r(t) - 2x_g(t) \tag{1}$$

$$Y_{CHROM}(t) = 1.5x_r(t) + x_g(t) - 1.5x_b(t) \tag{2}$$

The final rPPG signal is then calculated as:

$$s(t) = X_{CHROM}(t) - \alpha Y_{CHROM}(t) \tag{3}$$

where $\alpha = \frac{\sigma(X_{CHROM}(t))}{\sigma(Y_{CHROM}(t))}$, and $\sigma(\cdot)$ is the standard deviation. The parameter α accounts for imprecision in skin-tone standardization.

E. POS Method

The POS algorithm was developed by Wang et al. to extract a pulse signal (Wang et al., 2016). The POS algorithm addressed the same problem that's also tackled by CHROM, which relates to the reduction of specular noise. The POS algorithm uses a plane orthogonal to the skin tone in the temporally normalized RGB space for rPPG extraction. It's similar to the CHROM method but changes the order of colour distortion reduction. The study tested the new approach on subjects with different skin tones and activity levels, in a laboratory setting. The published results showed that the POS algorithm outperforms CHROM and PCA/ICA algorithms (Wang et al., 2016). The POS algorithm performs the following steps from a raw extracted ROI signal $x(t)$. First, a temporal normalization step is performed to derive two components $X_{POS}(t)$ and $Y_{POS}(t)$:

$$X_{POS}(t) = x_g(t) - x_b(t) \tag{4}$$

$$Y_{POS}(t) = x_g(t) + x_b(t) - 2x_r(t) \tag{5}$$

Similar to CHROM, the last step is accomplished to tune the projection direction within the bounded region defined by the temporal normalization step:

$$s(t) = X_{POS}(t) + \alpha Y_{POS}(t) \tag{6}$$

Here, α represents the same as the CHROM method. The POS approach differs slightly from the CHROM method, because in the latter the two projected signals are anti-phase, while the POS algorithm directly finds two projection axes giving in-phase signals (Wang et al., 2016).

F. SSR Method

Wang et al. also developed the SSR or 2SR algorithm that observes a subspace of skin pixels over time and measures their "rotation" for rPPG extraction (Wang et al., 2015). The paper of Wang et al. proposed a data-driven algorithm, that has the advantage of extending the pulse amplitude and reducing the distortion by the light reflection. The SSR algorithm was introduced to overcome two problems related to skin-tone or pulse related priors of previously discussed rPPG methods. The SSR method consists of two steps. First, the subspace of skin pixels needs to be constructed. Then the rotation angle of the subspace can be computed for subsequent frames. The skin pixels in an RGB space can be characterized by the eigenvectors obtained through the eigenvalue decomposition of the RGB representation of the skin pixels (Wang et al., 2015). The SSR algorithm starts with creating a matrix X of skin-pixel vectors, where each row is a single pixel and the columns are the RGB channels, with dimensions N x 3 (N being the number of pixels). Then a correlation matrix C can be computed as the matrix product of X transpose and X divided by N:

$$C = \frac{X^T X}{N} \tag{7}$$

The rotation between the vector $\mathbf{u}t1$ and orthonormal plane $\mathbf{u}\tau2$, $\mathbf{u}\tau3$ is used to compute the subspace rotation, which is given by $\mathbf{V}0 = (\mathbf{u}t1)^T \cdot (\mathbf{u}\tau2, \mathbf{u}\tau3)$. The decomposition of $\mathbf{C}$ also gives a scale/energy change of the subspace, represented by $E = \sqrt{\frac{\lambda t1}{\lambda_{\tau 2}} \cdot \frac{\lambda_{t1}}{\lambda_{\tau 3}}}^T$ (Boccignone et al., 2020).

To obtain the time-consistent $\mathbf{EV}$ over multiple strides, the rotation and scaling need to be combined and back projected into the original RGB space, represented as $\mathbf{EV}0 = \sqrt{\frac{\lambda_{t1}}{\lambda_{\tau 2}}} \cdot \mathbf{u}t1^T \cdot \mathbf{u}\tau2 \cdot \mathbf{u}\tau2^T + \sqrt{\frac{\lambda_{t1}}{\lambda_{\tau 3}}} \cdot \mathbf{u}t1^T \cdot \mathbf{u}\tau3 \cdot \mathbf{u}_{\tau 3}^T$.

Finally, multiple $\mathbf{EV}_0$ between the reference frame and succeeding frames are estimated and concatenated into a 3-dimensional trace $\mathbf{EV}$ in a single stride. To derive a pulse signal, the anti-phase traces $\mathbf{EV}_1$ and $\mathbf{EV}_2$ are combined as $\bar{\mathbf{p}} = \mathbf{EV}_1 - \frac{\sigma(\mathbf{EV}_1)}{\sigma(\mathbf{EV}_2)}\mathbf{EV}_2$, and a long-term pulse-signal is estimated from subsequent strides by using overlap-adding as $\bar{\mathbf{P}}(t-l) = \bar{\mathbf{P}}(t-l) - (\bar{\mathbf{p}} - \mu(\bar{\mathbf{p}}))$, where μ denotes the averaging operator. The final output is represented as $s(t) = \bar{\mathbf{P}}(t)$ (Boccignone et al., 2020).

The study of the SSR algorithm had participants with varying skin tones and under various illumination and activity conditions. The SSR algorithm outperformed previous BSS methods (PCA, ICA) and CHROM. However, the subspace axes constructed by SSR are data-driven without physiological considerations. This results in a limitation of the performance when the spatial measurements are unreliable, an example of this is when the skin mask is noisy or poorly chosen.

G. PBV Method

The pulse blood volume (PV) method was suggested by researchers from Phillips and ASML to alleviate problems with motion noise. PBV uses the signature of blood volume changes in different wavelengths to explicitly distinguish the pulse-induced colour changes from motion noise in RGB signals (De Haan & Van Leest, 2014). PBV vector is calculated as follows:

$$P_{bv}(t) = \frac{\sigma(X_c)}{\sqrt{\sigma^2(X_r) + \sigma^2(X_g) + \sigma^2(X_b)}} \tag{8}$$

where $X = X_r$, X_g, X_b and $x(t)$ is the pre-processed signal obtained from ROI selection and filtering. $\sigma()$ is the standard deviation operator. To compute the final output of the rPPG signal a projection needs to be done with the orthogonal matrix M, whereby k represents a normalizing factor:

$$s(t) = Mx(t) \tag{9}$$

$$M = kP_{bv}(XX^T)^{-1} \tag{10}$$

H. BKF Method

The bounded Kalman filter (BKF) is a method that aims at minimizing motion artifacts such as blurring and noise caused by head movement and facial expressions. The BKF is a kinematic estimation that implements the tracking of regions of interest in a facial video. It is an extension of the existing Kalman filter. BKF models the predicted feature point locations of a frame as a function of the velocity of the feature points from previous frames, minimizing the errors caused by drift and instantaneous movements. The model uses a cubic spline interpolated function to extrapolate the next possible feature point locations and eliminates drift errors by incorporating a boundary kernel in the Kalman filter prediction. BKF consists of three primary phases: predict phase, the update phase, and the boundary comparison phase. It uses kinematic equations to calculate the predicted feature point locations, process noise, and acceleration of each feature point being tracked. The predicted feature point locations are stored in a matrix.

I. End-to-End Deep Learning Methods

The further section will go into detail about end-to-end deep learning approaches for rPPG. End-to-end methods are deep learning methods that take a video input and generate a physiological signal as the output. Deep learning methods are indisputably great tools due to their straightforward model optimization process. Deep learning methods typically work well with sufficient training data and when the validation, testing data and training data are of similar distribution. However, imbalances in the data set can lead to biases and overfitting. Further experimental research and data gathering need to be done to validate the translation for clinical applications.

J. 2D CNN

The algorithms discussed in previous sections are all based on conventional signal processing techniques with prior knowledge about the HR range. rPPG extraction can also be regarded as a regression problem and classification problem where the video data is the input and the ground truth is the output. Advanced deep learning techniques, such as 2D convolutional neural networks utilize automatic feature selection and reduce the steps of the conventional rPPG pipeline. Back in 2018 Chen and Mcduff developed an end-to-end framework for HR and breathing rate (W. Chen & McDuff, 2018). The network called DeepPhys is a deep learning-based 2D convolutional based on the VGG architecture. The VGG network is often used in the field of computer science as it can be utilized for transfer learning with pre-trained weights. VGG networks have an accurate performance in object detection and image recognition. The DeepPhys model takes normalized frames difference as input motion representation. It utilizes a two-stream method with motion representation and an attention mechanism using appearance information representation. The network learned soft-attention masks from the original video frames and allocated higher weights to skin areas with stronger signals. The implemented attention mechanism enables visualization of the spatiotemporal distribution of physiological signals, which can be seen in Figure 4. Figure 4 illustrates the DeepPhys architecture based on VGG, where the time–frequency spectra of PPG signals are used as the input for vital parameter estimation.

Other authors made further improvements on top of DeepPhys and developed the MTTS-CAN model. The mode proposed by Liu et al. introduces a temporal shift module that uses a mechanism to replace 3D CNN without reducing accuracy while requiring less computational power (Liu, Fromm, Patel, & McDuff, 2020). This is achieved by shifting chunks of the tensor along the temporal axis. Also, the computational power was reduced by averaging multiple adjacent frames than using the original video frames.

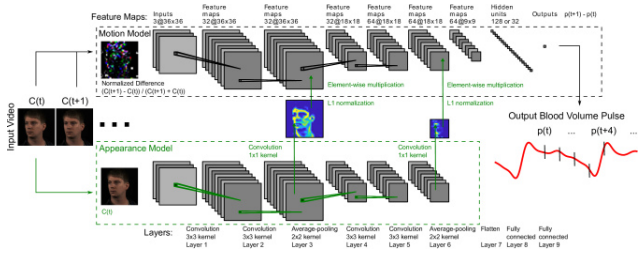

Fig. 4. Network architecture of DeepPhys (W. Chen & McDuff, 2018).

K. 3D CNN

Solely-based 2D CNN only takes partial dimensions into account, either spatial or temporal. Due to this limitation, researchers have proposed other frameworks that take spatial as well as temporal information into account. This developed network integrates more available information from the facial video to estimate the vital parameters of the subject by using 3D sliding kernels. Yu et al. developed a 3D CNN based called PhysNet, which uses spatiotemporal dimensions to estimate HR and HRV accurately (Yu, Li, & Zhao, 2019). PhysNet takes the RGB video frames as input and directly returns a rPPG signal as the final output. They implemented a loss function based on negative Pearson correlation in order to minimize problems with peak location errors (Yu, Li, & Zhao, 2019). This architecture with 3D CNN displayed better performance over another architecture the researchers proposed in their paper which incorporated recurrent neural networks (RNN).

Yu et al. also developed another two-stage end-to-end framework to overcome the problem of highly compressed facial videos. The aim of the spatiotemporal video enhancement network (STVEN) is to improve the quality of the video while retaining as much information (Yu, Peng, Li, Hong, & Zhao, 2019). The paper provided two independent methods, one for video enhancement and the other for rPPG signal recovery. STVEN is the first method proposed in the papers which is a video-to-video generator aimed at enhancing the quality of compressed videos. The STVEN model is optimized on two loss functions. First the mean squared error loss with L1 loss. In addition, a cycle loss is used to improve the generalisation of the proposed algorithm. The second proposed method rPPGNet contains three main components, spatiotemporal CNN, a skin-based attention module, and a partition constraint. This network also uses a minimization of negative Pearson correlation as a loss function. The skin-based module helps in refining the importance of features selection of the ROI, instead of skin pixel averaging like the conventional methods. The partition constraint helps with the features in the model by dividing the deep features into uniform parts and applying global average pooling. The paper of Yu et al. suggested that the rPPGNet is able to recover better rPPG signals with curves and peak locations for accurate HR and HRV estimation (Yu, Peng, et al., 2019). Figure 5 illustrates the two architectures of STVEN and rPPGNet modules in an end-to-end framework.

Further research developed was done by Yu et al. where they implemented evolutionary algorithms through neural architecture search (NAS) (Yu, Li, Niu, Shi, & Zhao, 2020). By using NAS the researchers suggested a new model with improved general performance. They developed a novel approach for temporal difference convolutions (TDC) that uses a 3x3x3 kernel with 1 channel to calculate the temporal differences. Furthermore, the researchers also provided a novel hybrid loss function that accounts for time and frequency constraints.

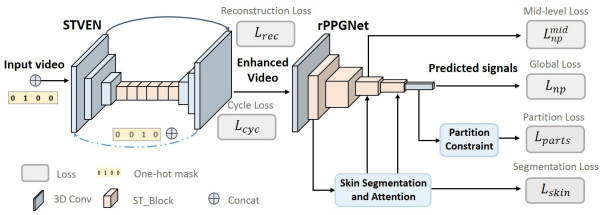

Fig. 5. The overall end-to-end framework of STVEN and rPPGNet (W. Chen & McDuff, 2018).

L. Recurrent Neural Networks

In the before mentioned study of Yu et al., a network architecture was proposed to extract rPPG with recurrence (Yu, Li, & Zhao, 2019). Recurrent neural networks (RNN) are model that uses recurrence in the network to use temporal information to make estimations, by processing the data in one step while retaining information from the previous step. The proposed model combined 2D CNN with RNN (LSTM, BiLSTM, ConvLSTM). The input and the output of the two PhysNet were similar. The RNN based on the input was processed into the 2D CNN to extract spatial features. Subsequently, the extracted features were further propagated in the temporal domain by the ability of recurrence in the model. However, after compares of the two models, the 3D CNN based PhysNet achieved an overall better performance that the recurrent based model. Other researchers also proposed an RNN based model that implemented the combination of 2D CNN and ConvLSTM networks and attention mechanism to extract an rPPG signal. This LSTM is an RNN architecture that offers an advantage as it can handle sequences of data points such as speech or video. It consists of two branches, the trunk branch and the mask branch. The trunk branch extracts features through a 2D CNN. The mask branch uses a max pooling layer and an attention mechanism to select and enhance critical parts of the feature maps and eliminate noise. The ConvLSTM is

present in the first and last two layers of the network to use the sequential information of feature maps at different scales and receptive fields. Figure 6 illustrates the architecture with the trunk branch and mask branch.

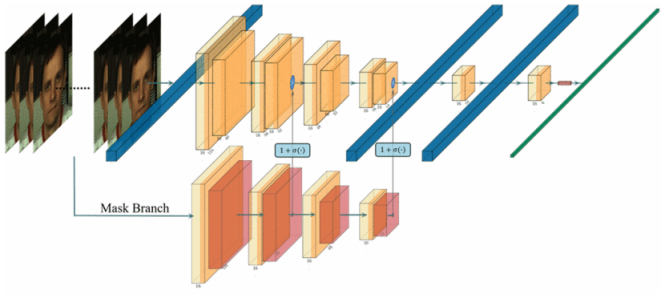

Fig. 6. The spatial temporal CNN (W. Chen & McDuff, 2018).

Additionally, other researchers published a paper that also used an end-to-end framework for rPPG estimation (E. Lee, Chen, & Lee, 2020). The model called meta-rPPG uses supervised learning with training data. To improve the robustness of the model the researchers employed an adaptive transductive meta-learner to cope with changes during testing through the use of weights adjustment based on unlabeled samples. The model consists of two major parts, the feature extractor and the meta-learner. The proposed end-to-end meta-learning framework showed substantial improvements on the MAHNOB-HCI and UBFC-rPPG datasets demonstrating state-of-the-art results. Overall, it displayed better accuracy than conventional methods like CHROM and ICA. It also improved the accuracy compared to other deep learning models DeePhys and PhysNet.

M. Hybrid Models

The following section discusses hybrid deep learning techniques that are only applied in some parts of the pipeline. These hybrid techniques aim to take advantage of both the traditional methods and deep learning methods to achieve a better performance compared to using one of them alone.

N. Hybrid 2D CNN

Deep-HR is a 2D CNN that is used for HR estimation. The model consists of two components, the front end and the back end. The front end uses 2D CNN to learn the ROI of the face. The back end is a fully-connected NN trained on the back end output to predict HR (Sabokrou, Pourreza, Li, Fathy, & Zhao, 2021). The two components are trained independently which helps in better translation to other frameworks. Deep-HR also incorporates a GAN to generate rPPG and reduce noise. The ROI detector is optimised on an objective function. The signal extraction component distils the colour information of the RoI and provides the input to the BE component (Sabokrou et al., 2021). However, a disadvantage of the model is that it doesn't estimate rPPG signals but does direct HR estimation. Thus, the model doesn't translate well to other vital parameter estimations such as the respiratory rate or HRV.

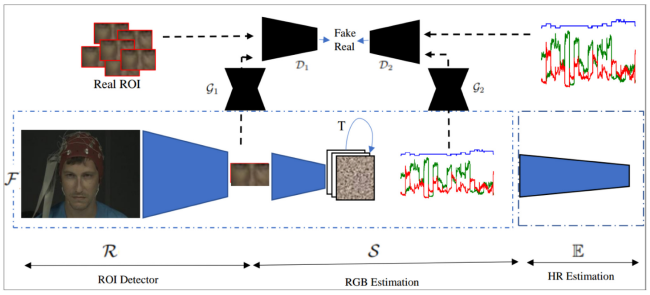

Fig. 7. The outline of the Deep HR method for HR estimation. On the left the front end and on the right the back end (W. Chen & McDuff, 2018).

O. Hybrid 3D CNN

The Siamese-rPPG is a hybrid deep learning network based on a Siamese 3D CNN architecture. The model is made of two branches with identical architecture. Those architectures contain multiple layers of 3D CNN and average pooling layers (Tsou, Lee, Hsu, & Chang, 2020). The model also has a weight-sharing mechanism which is implemented to improve the robustness in instances when noise is subjected to specific ROIs. The final layers use global average pooling and 1D CNN to collapse the feature maps into the rPPG output signal. Overall, the model displayed superior performance compared to conventional methods. However, it must be noted that the model was only evaluated on two data sets and probably only generalises well on the distributions of those datasets. The researchers also performed an ablation study to compare different regions of interest. They had three varying ROIs, the cheeks, forehead and the whole head (Tsou et al., 2020). Contrary to the conventional theory which implements the cheek and forehead region the overall performance was superior in the whole head condition (Poh et al., 2010a). Figure 8 illustrates the siamese-based network containing 6 layers and a final layer with global average pooling.

Another hybrid 3D CNN was presented by Bousefaf et al. (Bousefsaf, Pruski, & Maaoui, 2019). The CNN consists of a 3D convolutional layer with 32 filters of size 58 x 20 x 20, followed by a 3D max-pooling layer and a ReLU activation function. The final activations are then passed to fully connected layers. The model was trained using the backpropagation algorithm with the Adam optimizer and a categorical cross-entropy loss function. The research also incorporated synthetic data into the training of the model to improve generalisation. The training process also used an early-stopping criterion based on overfitting detection. The final model produced prediction maps for each group of pixels in the video stream, and the class with the highest score was saved and presented in the maps.

P. Hybrid RNN

The Long Short-Term Memory Deep-Filer (LSTM-DF) was introduced by Botina-Monsalve et al. to filter rPPG signals as an alternative to conventional signal processing techniques. The authors proposed a LSTM network to filter the rPPG

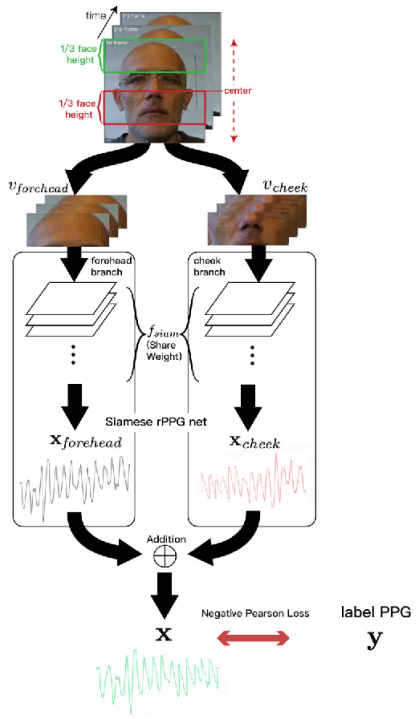

Fig. 8. The visual representation of the siamese-rPPG network (Tsou et al., 2020).

signal (Botina-Monsalve et al., 2020). The model was able to predict the rPPG signal and especially its temporal structure. This isn't possible with the usual signal processing based filtering methods. The results of this study show that the deep learning based filtering method outperforms the regular post-processing ones in terms of signal quality and accuracy of heart rate estimation. However, it must be mentioned that the authors only evaluate the model on one dataset. Therefore the model is limited in applicability to other data distributions.

Q. Hybrid General Adversarial Networks

The PulseGAN is a generative adversarial network designed to generate a realistic rPPG signal. The PulseGAN builds on top of the CHROM algorithm that aims for signal quality improvement similar to the reference PPG (Song et al., 2021). The basic structure of a GAN is made up of a generator and a discriminator. The generator network is designed as a denoising autoencoder with skip connections, while the discriminator network is composed of 1D convolutional layers and a fully connected layer. The generator is trained to minimize the error losses in both the time and frequency domains, which are defined by the loss functions of the generator and the discriminator. Adversarial learning is utilized to generate a target signal that is as close as possible to the reference signal (Song et al., 2021). The PulseGAN shows significant improvement in the accuracy of the CHROM method. This model also offers the advantage that it can be added on top of any existing conventional method. Figure 9 illustrates the discriminator and the generator components of the PulseGAN with skip connection layers.

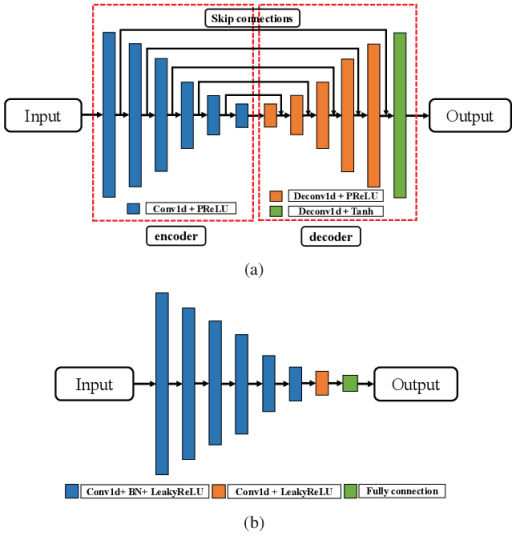

Fig. 9. Schematic overview of ICA (Song et al., 2021)

R. HR estimation using rPPG

In general two major methods are implemented for HR estimation after the rPPG is obtained. The first one that is often used is frequency analysis. FFT is the most commonly used for conversing to the HR frequency because the rPPG signal contains periodicity (Poh et al., 2010a). After it's translated to the frequency domain the HR with the highest spectral density is chosen as the HR. With peak detection, it's also possible to extract vital parameters such as the HRV. To improve peal detection cubic spline functions are used to interpolate the signal, by using a sliding window to identify the peaks (Rouast et al., 2018). The peaks are defined as the maximums within the signal, and their detection allows for a more accurate estimation of HR frequency.

VI. Discussion

In recent years, rPPG technology has gained momentum, and various methods have been suggested to address the challenges in remote HR measurement, such as illumination changes, motion artifacts, skin-tone variations, and video compression. This section will discuss a comparative analysis of the before mentioned studies. But it must be noted that due to the variations in data sets that contain various demographics and conditions which make a comparison between the studies challenging. The commonly used evaluation metrics in the research papers are mean average error (MAE), root mean squared error (RMSE) and Pearson correlation coefficient. The analysis and evaluation of conventional techniques on similar data sets were extensively explored in the paper of Boccignone et al. (Boccignone et al., 2020). The extensive comparison of the results shows that the earlier produced methods of GREEN, ICA, and PBV perform generally worse on the 15 datasets. The main conclusion of the authors shows that POS, CHROM,

PCA, and SSR are the superior conventional algorithms. Although CHROM and POS don't display significantly better performance on the evaluation metric than PCA and SSR. In general, the performance shows slightly better results for CHROM and POS. The extensive table of the MAE and PCC for the conventional algorithms and the respective datasets is in Appendix A.

A comprehensive overview of deep learning-based contactless heart rate measurement methods were also provided in the paper. It should be noted that the evaluation of the hybrid deep learning methods cannot be done by comparing individual components of the network, hence the discussion focuses on the overall performance of the end-to-end frameworks. One of the major advantages of using deep learning methods, such as convolutional neural networks (CNNs), in rPPG technology is that automatic feature selection can be done through backpropagation and fitting the training data. Ni et al. evaluated various deep learning networks such as STVEN-rPPGNet, MetarPPG, PhysNet, and iPPG. The results of the evaluation of rPPG extraction methods showed a minimal MAE and MSE for PhysNet. However, the black-box nature of deep learning is a barrier to applying such systems in healthcare. In addition, rPPG methods suffers from performance issues in low and high HR ranges.

Another notable weakness in the available datasets and algorithms is that they mainly focus on two major challenges, motion artifacts and illumination variations. Other challenges, such as skin-tone variations, multiple-person detection, and long-distance estimations, aren't addressed to meet the robust standards for real-world scenarios (Dasari, Prakash, Jeni, & Tucker, 2021). This is further confirmed in a recently published paper which investigated the biases of rPPG methods. The paper showed that the current state-of-the-art models for conventional as well as deep learning aren't robust enough. As the evaluation of a dataset with different demographic factors such as skin tone, age, gender and country of origin resulted in a substantial increase in error and standard deviation (Dasari et al., 2021). Thus, displaying less robustness in general. Therefore, future research needs to acquire high-diversity and high-quality datasets before it can evaluate the general robustness of new methods. and allow comprehensive training in supervised methods.

VII. Conclusion

This paper has provided a comprehensive overview of the relevant literature on contactless heart rate measurement methods. First, an overview of contact-based PPG and contactless PPG methods were covered. Then, the review focus on the rPPG pipeline, with the components of face extraction, ROI selection and signal processing. Next, conventional methods have been introduced in the literature for heart rate measurement using rPPG. Following this deep learning based methods for rPPG extraction have been explored. The rPPG methods were analyzed, and the capability of these methods to compare results on similar datasets was considered. methods were analyzed, and the capability of these methods to compare results on similar datasets was considered. The limitations of rPPG included factors such as motion artifact, skin pigmentation, and lighting conditions that can affect the accuracy and reliability of the signals obtained. While vital parameter estimation with rPPG technology has shown remarkable results, there are still many challenges to overcome, such as performance on different HR ranges, addressing biases, and the need for more comprehensive and high-quality datasets. Future research needs to address these challenges and further the development of novel methods and algorithms to improve the performance and accuracy of rPPG technology under real-life conditions. .

References

Aarts, L. A., Jeanne, V., Cleary, J. P., Lieber, C., Nelson, J. S., Oetomo, S. B., & Verkruysse, W. (2013). Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—a pilot study. *Early human development*, *89*(12), 943–948.

Ali, S. A., Arif, T. B., Maab, H., Baloch, M., Manazir, S., Jawed, F., & Ochani, R. K. (2020). Global interest in telehealth during covid-19 pandemic: an analysis of google trends™. *Cureus*, *12*(9).

Balakrishnan, G., Durand, F., & Guttag, J. (2013). Detecting pulse from head motions in video. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3430–3437).

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *Lecture notes in computer science*, *3951*, 404–417.

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, *3*(1), 1–9.

Boccignone, G., Conte, D., Cuculo, V., d'Amelio, A., Grossi, G., & Lanzarotti, R. (2020). An open framework for remote-ppg methods and their assessment. *IEEE Access*, *8*, 216083–216103.

Botina-Monsalve, D., Benezeth, Y., Macwan, R., Pierrart, P., Parra, F., Nakamura, K., ... Miteran, J. (2020, June). Long short-term memory deep-filter in remote photoplethysmography. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr) workshops.*

Bousefsaf, F., Pruski, A., & Maaoui, C. (2019). 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, *9*(20), 4364.

Burritt, M. F. (1998). Noninvasive and invasive sensors for patient monitoring. *Laboratory medicine*, *29*(11), 684–687.

Chen, J., Chang, Z., Qiu, Q., Li, X., Sapiro, G., Bronstein, A., & Pietikäinen, M. (2016). Realsense= real heart rate: Illumination invariant heart rate estimation from videos. In *2016 sixth international conference on image processing theory, tools and applications (ipta)* (pp. 1–6).

Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (eccv)* (pp. 349–365).

Cheng, C.-H., Wong, K.-L., Chin, J.-W., Chan, T.-T., & So, R. H. (2021). Deep learning methods for remote heart rate measurement: a review and future research agenda. *Sensors*, *21*(18), 6296.

Dasari, A., Prakash, S. K. A., Jeni, L. A., & Tucker, C. S. (2021). Evaluation of biases in remote photoplethysmography methods. *NPJ digital medicine*, *4*(1), 91.

De Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, *60*(10), 2878–2886.

De Haan, G., & Van Leest, A. (2014). Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, *35*(9), 1913.

Deng, Y., & Kumar, A. (2020). Standoff heart rate estimation from video: A review. *Mobile Multimedia/Image Processing, Security, and Applications 2020*, *11399*, 16–29.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and vision computing*, *28*(5), 807–813.

Gudi, A., Bittner, M., & van Gemert, J. (2020). Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation. *Applied Sciences*, *10*(23), 8630.

Gupta, P., Bhowmik, B., & Pal, A. (2018). Robust adaptive heart-rate monitoring using face videos. In *2018 ieee winter conference on applications of computer vision (wacv)* (pp. 530–538).

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, *37*(3), 583–596.

Hertzman, A. B. (1938). The blood supply of various skin areas as estimated by the photoelectric plethysmograph. *American Journal of Physiology-Legacy Content*, *124*(2), 328–340.

Kim, D.-Y., Lee, K., & Sohn, C.-B. (2021). Assessment of roi selection for facial video-based rppg. *Sensors*, *21*(23), 7923.

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

Kwon, S., Kim, H., & Park, K. S. (2012). Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone. In *2012 annual international conference of the ieee engineering in medicine and biology society* (pp. 2174–2177).

Lee, E., Chen, E., & Lee, C.-Y. (2020). Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxvii 16* (pp. 392–409).

Lee, K.-Z., Hung, P.-C., & Tsai, L.-W. (2012). Contact-free heart rate measurement using a camera. In *2012 ninth conference on computer and robot vision* (pp. 147–152).

Leibner, E. S., Stokes, S., Ahmad, D., & Legome, E. (2021). Practical protocols for managing patients with sars-cov-2 infection (covid-19) in the emergency department. *Emerg Med Pract*, *23*(Suppl 2), 1–38.

Lewandowska, M., Rumiński, J., Kocejko, T., & Nowak, J. (2011). Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 federated conference on computer science and information systems (fedcsis)* (pp. 405–410).

Li, X., Chen, J., Zhao, G., & Pietikainen, M. (2014). Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4264–4271).

Liu, X., Fromm, J., Patel, S., & McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, *33*, 19400–19411.

Liu, X., Wei, W., Kuang, H., & Ma, X. (2022). Heart rate measurement based on 3d central difference convolution with attention mechanism. *Sensors*, *22*(2), 688.

Lokendra, B., & Puneet, G. (2022). And-rppg: A novel denoising-rppg network for improving remote heart rate estimation. *Computers in biology and medicine*, *141*, 105146.

Moco, A. V., Stuijk, S., & De Haan, G. (2015). Ballistocardiographic artifacts in ppg imaging. *IEEE Transactions on Biomedical Engineering*, *63*(9), 1804–1811.

Moço, A. V., Stuijk, S., & de Haan, G. (2016). Motion robust ppg-imaging through color channel mapping. *Biomedical optics express*, *7*(5), 1737–1754.

Monkaresi, H., Calvo, R. A., & Yan, H. (2013). A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE journal of biomedical and health informatics*, *18*(4), 1153–1160.

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2010a). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, *58*(1), 7–11.

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2010b). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, *18*(10), 10762–10774.

Qian, R. J., Sezan, M. I., & Matthews, K. E. (1998). A robust real-time face tracking algorithm. In *Proceedings 1998 international conference on image processing. icip98 (cat. no. 98cb36269)* (Vol. 1, pp. 131–135).

Qiao, D., Ayesha, A. H., Zulkernine, F., Jaffar, N., & Masroor, R. (2022). Revise: Remote vital signs measurement using smartphone camera. *IEEE Access*.

Rouast, P. V., Adam, M. T., Chiong, R., Cornforth, D., & Lux, E. (2018). Remote heart rate measurement using

low-cost rgb face video: a technical literature review. *Frontiers of Computer Science*, *12*, 858–872.

Sabokrou, M., Pourreza, M., Li, X., Fathy, M., & Zhao, G. (2021). Deep-hr: Fast heart rate estimation from face video under realistic conditions. *Expert Systems with Applications*, *186*, 115596.

Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., & Chen, X. (2021). Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1373–1384.

Špetlík, R., Franc, V., & Matas, J. (2018). Visual heart rate estimation with convolutional neural network. In *Proceedings of the british machine vision conference, newcastle, uk* (pp. 3–6).

Sun, Y., & Thakor, N. (2015). Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE transactions on biomedical engineering*, *63*(3), 463–477.

Swinehart, D. F. (1962). The beer-lambert law. *Journal of chemical education*, *39*(7), 333.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1701–1708).

Taparia, N., Platten, K. C., Anderson, K. B., & Sniadecki, N. J. (2017). A microfluidic approach for hemoglobin detection in whole blood. *AIP Advances*, *7*(10), 105102.

Tasli, H. E., Gudi, A., & Den Uyl, M. (2014). Remote ppg based vital sign measurement using adaptive facial regions. In *2014 ieee international conference on image processing (icip)* (pp. 1410–1414).

Tohma, A., Nishikawa, M., Hashimoto, T., Yamazaki, Y., & Sun, G. (2021). Evaluation of remote photoplethysmography measurement conditions toward telemedicine applications. *Sensors*, *21*(24), 8357.

Tsou, Y.-Y., Lee, Y.-A., Hsu, C.-T., & Chang, S.-H. (2020). Siamese-rppg network: Remote photoplethysmography signal estimation from face videos. In *Proceedings of the 35th annual acm symposium on applied computing* (pp. 2066–2073).

Van Egmond, J., Hasenbos, M., & Crul, J. (1985). Invasive v. non-invasive measurement of arterial pressure: comparison of two automatic methods and simultaneously measured direct intra-arterial pressure. *BJA: British Journal of Anaesthesia*, *57*(4), 434–444.

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics express*, *16*(26), 21434–21445.

Villarroel, M., Guazzi, A., Jorge, J., Davis, S., Watkinson, P., Green, G., … Tarassenko, L. (2014). Continuous noncontact vital sign monitoring in neonatal intensive care unit. *Healthcare technology letters*, *1*(3), 87–91.

Viola, P., & Jones, M. (2001). 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* (Vol. 1, pp. I–I).

Wang, W., Den Brinker, A. C., Stuijk, S., & De Haan, G. (2016). Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1479–1491.

Wang, W., Stuijk, S., & De Haan, G. (2015). A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, *63*(9), 1974–1984.

Waqar, M., Zwiggelaar, R., & Tiddeman, B. (2021). Contact-free pulse signal extraction from human face videos: A review and new optimized filtering approach. *Biomedical Visualisation: Volume 9*, 181–202.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, *31*(4), 1–8.

Yu, Z., Li, X., Niu, X., Shi, J., & Zhao, G. (2020). Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, *27*, 1245-1249. doi: 10.1109/LSP.2020.3007086

Yu, Z., Li, X., & Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*.

Yu, Z., Peng, W., Li, X., Hong, X., & Zhao, G. (2019). Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In *2019 ieee/cvf international conference on computer vision (iccv)* (p. 151-160). doi: 10.1109/ICCV.2019.00024

APPENDIX

A. Evalution of conventional methods on 15 datasets

Dataset	MAE								CC							
	CHROM	GREEN	ICA	LGI	PBV	PCA	POS	SSR	CHROM	GREEN	ICA	LGI	PBV	PCA	POS	SSR
LGI-PPGI-talk	10.86	18.25	12.46	10.86	13.40	10.75	11.07	13.79	-0.03	-0.07	0.08	-0.04	-0.01	0.18	0.02	0.11
LGI-PPGI-gym	17.38	28.30	24.41	21.28	24.13	11.04	9.84	10.74	0.41	0.30	0.39	0.34	0.32	0.59	0.62	0.62
PURE-small-rot	1.05	4.25	3.03	1.36	4.18	1.27	1.01	3.28	0.78	0.45	0.46	0.67	0.39	0.75	0.79	0.59
LGI-PPGI-rotation	5.05	8.05	5.11	2.98	8.29	6.04	4.04	3.92	0.27	0.22	0.43	0.39	0.23	0.48	0.37	0.40
UBFC2	3.11	8.05	10.44	8.73	5.46	5.21	1.87	5.68	0.64	0.49	0.42	0.39	0.52	0.59	0.83	0.48
PURE-fast-trans	2.40	2.59	4.94	3.56	4.16	2.15	3.80	2.34	0.66	0.58	0.36	0.44	0.31	0.66	0.62	0.56
UBFC1	2.23	15.89	5.85	2.42	8.91	4.47	1.82	4.10	0.72	0.01	0.47	0.68	0.26	0.42	0.87	0.54
PURE-slow-trans	0.97	1.50	2.30	2.12	2.52	1.24	0.90	2.06	0.88	0.76	0.58	0.65	0.51	0.80	0.88	0.81
PURE-fast-rot	1.01	8.69	2.86	1.74	4.52	1.39	0.94	3.53	0.78	0.21	0.48	0.60	0.17	0.62	0.79	0.67
LGI-PPGI-resting	1.02	5.57	2.80	1.63	1.82	1.19	2.27	1.77	0.84	0.45	0.29	0.61	0.47	0.75	0.69	0.71
PURE-talking	3.27	10.19	9.38	7.29	4.14	3.87	3.50	3.43	0.56	0.33	0.20	0.24	0.53	0.58	0.62	0.52
PURE-steady	1.17	1.66	4.33	3.61	2.10	1.81	1.17	7.60	0.80	0.64	0.38	0.32	0.55	0.64	0.78	0.60
Median	**2.31**	**8.05**	**5.02**	**3.27**	**4.34**	**3.01**	**2.07**	**3.72**	**0.69**	**0.39**	**0.40**	**0.41**	**0.35**	**0.60**	**0.73**	**0.57**
IQR	**2.67**	**7.77**	**6.65**	**5.63**	**4.70**	**4.05**	**2.72**	**3.11**	**0.26**	**0.29**	**0.12**	**0.28**	**0.26**	**0.12**	**0.18**	**0.12**
COHFACE-naturalLight	13.89	14.69	11.99	13.70	14.05	11.29	14.02	13.89	0.03	0.04	0.18	0.05	0.04	0.22	0.04	0.00
COHFACE-studioLight	11.10	9.89	10.34	11.86	12.50	9.66	9.96	9.89	0.02	0.07	0.19	-0.02	0.05	0.31	0.03	0.07
MAHNOB	18.59	15.25	14.90	19.00	18.64	10.98	19.42	17.46	0.09	0.04	0.17	0.01	-0.01	0.21	0.02	0.06
Median (all)	**3.10**	**8.69**	**5.84**	**3.60**	**5.46**	**4.47**	**3.49**	**4.09**	**0.64**	**0.29**	**0.38**	**0.38**	**0.31**	**0.59**	**0.62**	**0.54**
IQR (all)	**9.86**	**10.06**	**7.53**	**9.08**	**8.79**	**8.60**	**8.40**	**6.96**	**0.60**	**0.41**	**0.24**	**0.45**	**0.37**	**0.28**	**0.58**	**0.35**

Fig. 10. MSE and PCC for conventional methods 15 datasets (Boccignone et al., 2020).