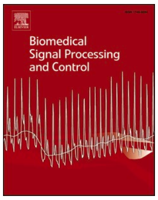
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Heart rate prediction from facial video with masks using eye location and corrected by convolutional neural networks

Kun Zheng [a], Kangyi Ci [a], Hui Li [a], Lei Shao [b,*], Guangmin Sun [a], Junhua Liu [a], Jinling Cui [a,*]

[a] *Faculty of Information Technology, Beijing University of Technology, No.100, Pingleyuan, Chaoyang District, Beijing 100124, China*
[b] *Department of Investigation, Sichuan Police College, No.186, Longtouguan Road, Jiangyang District, Luzhou, Sichuan 646000, China*

ARTICLE INFO

Keywords:
Eye location
Heart rate
Measurement consistency
Residual network
rPPG

ABSTRACT

Remote photoplethysmography (rPPG), which aims at measuring heart activities without any contact, has great potential in many applications. The emergence of novel coronavirus pneumonia COVID-19 has attracted worldwide attentions. Contact photoplethysmography (cPPG) methods need to contact the detection equipment with the patient, which may accelerate the spread of the epidemic. In the future, the non-contact heart rate detection will be an urgent need. However, existing heart rate measuring methods from facial videos are vulnerable to the less-constrained scenarios (e.g., with head movement and wearing a mask). In this paper, we proposed a method of heart rate detection based on eye location of region of interest (ROI) to solve the problem of missing information when wearing masks. Besides, a model to filter outliers based on residual network was conceived first by us and the better heart rate measurement accuracy was generated. To validate our method, we also created a mask dataset. The results demonstrated that after using our method for correcting the heart rate (HR) value measured with the traditional method, the accuracy reaches 4.65 bpm, which is 0.42 bpm higher than that without correction.

1. Introduction

Heart rate (HR) is an important physiological parameter reflecting a person's health condition. Traditional heart rate detection mainly includes two ways: electrocardiograph (ECG) and contact photoplethysmography (cPPG) based on sensors. PPG, an optical method for detecting the blood volume pulse (BVP) from the skin, rely on the principle that blood absorbs more light than surrounding tissues, thus the change of blood volume can affect the transmittance or reflectivity of light accordingly.

The cPPG sensors, such as fingertip pulse oximeter, require the equipment to make physical contact with the subject. However, this measuring method needs the skin to closely fit with the equipment without relative movement. Due to the limitations of cPPG methods, it is particularly important to study a non-contact HR detection method. Verkruysse et al. [1] proposed that rPPG (remote photoplethysmography) signal can be obtained from the face video collected by the camera under ambient light for the first time. The rPPG has been proven to be superior because it is non-intrusive. It may be suitable for continuous measurement of heart rate (HR) in many cases, such as neonatal ICU monitoring [2], driver status assessment [3] and online learning [4]. Since then, many scholars conducted research on how to measure vital sign remotely from facial videos.

The emergence of covid-19 helps to increase the possibility of using non-contact techniques in detecting patients' vital signs [5]. A patient's heart rate, respiratory rate, blood oxygen and other physiological parameters can be detected through rPPG techniques. Besides, the risk of infection among healthy workers can also be reduced by using non-contact methods.

At present, most algorithms detect vital signs of the human body through computer recorded videos. However, the widespread use of smart phones with video recording function provides an opportunity for the integration of rPPG methods to create non-invasive vital signs assessment (e.g., via an app download) [6]. Estimating the HR through rPPG methods by the front camera of smartphone device exist now [7,8]. This greatly facilitates the needs of people to detect physiological information anytime and anywhere.

So far, the traditional methods and deep learning methods have made outstanding contributions in solving illumination changes and motion artifacts. Their common goal is to obtain better rPPG signals for

* Corresponding authors.
E-mail addresses: zhengkun@bjut.edu.cn (K. Zheng), cikangyi@emails.bjut.edu.cn (K. Ci), lihui@bjut.edu.cn (H. Li), shaolei@scpolicec.edu.cn (L. Shao), gmsun@bjut.edu.cn (G. Sun), liujunhua@bjut.edu.cn (J. Liu), cuijnling@bjut.edu.cn (J. Cui).

https://doi.org/10.1016/j.bspc.2022.103609
Received 19 September 2021; Received in revised form 16 February 2022; Accepted 27 February 2022
Available online 9 March 2022

HR detection. But the video for testing is required to include the whole face. Under this framework, it has strong physical constraints on subjects, making it not suitable for HR detection under the condition of lack of face information. However, in the light of recent research, combining face mask detection method [9] and efficient skin segmentation models [10] can also be used to detect HR when wearing a mask. Studies have shown that region of interest (ROI) from forehead and both cheeks contain good rPPG SNR signal quality [11]. Thus, we believe that the rPPG signal quality obtained by using the skin pixels of the forehead is higher than that obtained by using all the skin pixels of the upper part of the face. At present, there is no algorithm to directly detect the forehead region, but there are algorithms that can be used to detect eyes, such as Viola-Jones (VJ) [12]. When we detect the HR of a person wearing a mask, we can use the eyes to locate the forehead region and then carry out subsequent detection steps.

The existing public datasets (PFF [13], VIPL [14], UBFC-RPPG [15], COFACE [16], PURE [17], etc) for physiological signal detection are diverse in terms of recording scene, equipment, environment and video format. However, the complete information of the face is all included. That means there is no lack of information such as wearing a mask in these datasets. In order to study how to detect the HR at the scene of face information missing, we add masks to the published datasets.

A major concern in heart rate estimation today is to continue to improve the accuarcy. There are several reasons for causing detection error. The first reason is that the HR obtained by the contact device itself has slight errors. If the HR is compared with the measured value as an accurate value, it will cause the error. Up to now, the methods based on rPPG generally collect signals as input for a period of time, and the output result has only one value as HR. However, the real HR will fluctuate for a period of time. Only one value represents all possible heart rates in this time period, which is the second factor leading to error. In addition, none of the rPPG algorithm developed is perfect and sometimes the HR values measured in two adjacent time windows are quite different. The detection inconsistency is the third factor leading to the error. Existing methods prefer to use power spectrum to filter outliers, such as set threshold [18] and consistency check [19]. However, they cannot eliminate the noise in the heart rate frequency band. Therefore, we believe that it is not the best way to choose frequency for HR prediction, but it can be realized by means of self-learning through the network. A combination of deep learning and traditional method is realized by using the non end-to-end pattern developed recently. It has been shown that the results in non-contact HR measurement are promising. As far as we know, the current non end-to-end deep learning methods aim to create a HR estimator through the network. But these methods are established the mapping from the rPPG signal (image) to HR directly, without judging the quality of input signal. In the case of poor signal quality, the heart rate measured by traditional methods is likely to be an abnormal one. Therefore, we designed a model to judge whether the input signal is correct, which is used to filter the measured outliers. In summary, our contributions are:

1) We propose a HR detection method using the human eyes to locate ROI with skin segmentation method to solve the problem of lacking face information.

2) We propose a model to filter outliers based on residual network to improve the consistency of HR estimation.

3) A method for creating mask datasets is designed.

2. Related work

2.1. Existing rPPG methods

To suppress the influence of illumination variations, one possible way is to separate illumination variation signals from the pulse signals. For instance, Chen et al. [20] applied ensemble empirical mode decomposition (EEMD) algorithm to the green channel for separating environmental noise freed heart rate variation. Another solution to this problem is complementary adopting the RGB and the NIR domains from the cameras. Kado et al. [21] takes an RGB-NIR face video as an input and used spatial-spectral-temporal fusion method to improve the robustness of HR estimation against illumination fluctuations.

To solve the influence of motion artifact on measuring heart rate, traditional methods can be divided into two categories, called BSS-based (blind source separation) and model-based methods [22].

Independent component analysis (ICA) is a classical algorithm for BSS methods. The premise of its application is that the pulse wave signal and motion noise signal are independent and uncorrelated. Poh et al. [18] used ICA to separate the pulsation information from RGB signals. Then the joint blind source separation (JBSS) method was applied to adapt to the situation that multiple facial sub regions are used as ROI [3]. Later, Qi et al. [23] proved that incorporate the data from different facial sub-regions can improve the remote measurement performance. Besides, principal component analysis (PCA) was proposed by Lewandowska et al. [24] to extract the periodic pulse signal and they demonstrated that PCA has similar accuracy with ICA. Project_ICA [25] was also used to robustly extract the pulse signal in real scene such as poor illumination and face movements.

The common point of the model-based method is to eliminate the specular reflection component that independent of blood volume changes [26]. De Haan and Jeanne developed a chrominance-based approach, which is an optical skin model to reduce the influence of head movement [27]. Subsequently, De Haan and Leest derived the relative pulsatilities from the RGB channel of the camera and proposed PBV method for improving the motion robustness [28]. Another mathematical model combines the optical and physiological characteristics of skin reflections was present in [26], in which the POS tone in the temporally normalized RGB space was proposed.

The development of rPPG technology emphasizes the importance of fair comparison of different algorithms, and promotes the research of repeatability. Thus, Boccignone et al. [29] provided a systemic open framework, allowing to assess eight traditional rPPG algorithms (ICA [18], PCA [24], GREEN [1], CHROM [27], POS [26], SSR [30], LGI [31], PBV [28]), by setting parameters and metaparameters.

After denoising the raw signals by the traditional methods, Fast Fourier Transform (FFT) is usually applied to the purified signal for finding frequency that corresponding to the HR. Most traditional methods usually have certain assumptions and work based on a specific environment, and their performance may be invalid in the real scene. Thus, only use FFT may get a noise contaminated spectral distribution that leads to inaccurate measurement results. However, these traditional methods have potential to remove part of the noise from the raw signal. Using these methods as preprocessing tools can simplify the complexity in deep learning methods [32].

Besides, using deep learning (DL) approaches in the rPPG field have emerged and yielded excellent results in recent years. Ni et al. [33] and Cheng et al. [34] provided a comprehensive review of DL-based methods. Since training network needs a large amount of data collected in various real scenes, ensures the robustness and flexibility of the DL method for practical application. Up to now, DL methods can be divided into two categories, end-to-end and non-end-to-end.

The end-to-end methods learn features by themselves and output the HR or rPPG signal. Chen and McDuff [35] presented the first end-to-end convolutional attention network DeepPhys, which used normalized frame difference as input and rPPG signals as output. HR-CNN [36], proposed by Špetlík, used sequence of images of a subject's face as input and heart rates as output. Huang et al. [37] proposed PRnet, a one-stage spatio-temporal framework, to estimate the pulse rate from a stationary facial video. Another method from Yu et al. fed face images with RGB channels to spatio-temporal networks (STVEN + rPPGNet) [38] and PhysNet [39] to recover rPPG signals for HR and HRV features measurement.

In contrast, non-end-to-end methods utilize DL techniques at various stages. After extracting the rPPG signal, DL methods can be used for HR estimation. For such methods, features need to be generated firstly. The rPPG signals can be converted into images through a specific way [13,32]. As a preprocessing step, creating feature images for rPPG signals plays a key role, with its quality directly affect the following training model and even the prediction results. Hsu et al. [13,40] generate 2D TFR maps through different preprocessing approach for training the Convolutional Neural Network (CNN). Qiu et al. [41] extracted feature images by using spatial decomposition and temporal filtering, which was then cascaded with CNN for HR prediction. Articles [14,32,42,43] all utilized ResNet-18 for mapping the spatial–temporal maps to HR value. In Reference [44], POS-STMap followed by NAS network, was used for HR estimation.

Besides, DL methods can also be used at other steps in HR estimation. For example, PulseGAN [45] was designed for denoising the rough rPPG signals, and rPPG signals can be converted to contact BVP signals using dedicated deep learning solution [46].

However, if the signal includes a lot of noise, it will increase the difficulty of network learning and reduce the accuracy of HR prediction.

2.2. *Face detection*

Face detection is a crucial preprocessing step for the traditional rPPG methods and non end-to-end deep learning methods to measure HR. As a previous step in determining ROI, its accuracy has a direct impact on the accuracy of HR detection. At present, three mainstream methods VJ [12], Histogram of Oriented Gradients (HOG) [47] and MTCNN [48] are often used for face detection. The purpose of VJ algorithm is to detect face, but it can also be used to detect eyes. The VJ algorithm utilizes Harr-like features and AdaBoost algorithm to construct a cascade classifier. Dlib face detection method is based on HOG feature descriptor and linear SVM classification. HOG constructs the feature by calculating the gradient direction histogram of the local area of the image. MTCNN is a multi-task cascaded CNN based framework, which consists of three stages for joint face detection and alignment. Besides, face detection is also the basis of face segmentation and face swapping technologies [49].

However, few articles considered detecting heart rate in the absence of face information. When people turn their heads at a large angle or wear masks, it is impossible to detect faces. Thus, the HR cannot be detected. The face-eye location method was proposed for the first time by Zheng et al. [4] to detect HR when wearing a mask. Koen and Marnix [50] tested rPPG's accuracy on most likely visible body parts like wrist, hand palm and calf. This shows that people aim to explore some new robust rPPG methods to be used in more realistic scenarios in future. And this article is also based on human eyes for ROI locating to estimate HR when face information is missing.

2.3. *Consistency check for HR estimate*

There are techniques that have been developed to solve inconsistent of heart rate measurement. In other words, outliers can occur at certain times. Poh et al. [18] proposed a threshold correction method after using FFT. If the pulse rate measured at time T differs by more than ± 12 bpm from the pulse rate measured at time T-1, the pulse rate measured at time T is rejected and the frequency corresponding to the second highest power that met the constraint is found. However, if the pulse rate measured at time T differs significantly from the real value, it is likely that the subsequent correction results will be invalid. Demirezen et al. [19] proposed History-based consistency check (HBCC) algorithm to improve the consistency of heart rate estimation. Although the highest power magnitude and the SNR value and consistency were taken into consideration, spectrum analysis was still a problem that could not distinguish the noise in the heart rate band. When the noise signal overwhelms the pulse signal, the measurement is still inaccurate.

Besides, the reliability of using contact PPG sensors to measure heart rate has been proved. Article [51,52] assessed the accuracy of the wearable sensor in measuring HR in different conditions. Their experimental results show that the wearable sensor provides accurate HR measures compared to gold-standard equipment.

3. Methods

The goal of our method is to enable a non end-to-end measurement of HR when missing information. Fig. 1 is an outline of our method. We first locate the frontal ROI through human eyes' location and generate spatio-temporal feature images. Then, the handmade feature images are fed into CNN to train a model to judge the authenticity of them. We choose ResNet-18, initialized with ImageNet pretraining, for feature learning. After that, we use the model to predict the authenticity of the input image. Our method is described in detail below.

3.1. *Generating mask dataset*

Due to the lack of mask dataset, we first generated one. We added masks to the three data sets of PFF, PURE, and UBFC. Face detection and 68-point facial landmarks are applied to find the cheek area. Then the area is turned into black by image processing technology to simulate the situation of wearing a mask.

Pulse From Face (PFF) database [13] contains 13 subjects, each subject has five scenarios. The five scenarios are Bright-Static (BS), Bright-Moving (BM), Dark-Static (DS), Dark-Moving (DM), Bright-Riding (BR). Each video clip was recorded for 3 min with frame rate 50 Hz and resolution 1280 × 720 pixels. The subjects were sitting in front of the camera with an average distance of 0.5 m. A total of 24 videos were selected from Normal Light Static (NS) and Dark Static (DS) static scenes without obvious bangs.

The PURE database [17] consists 10 subjects, and each subject recorded head image sequences in six different setups. The six setups are Steady (01), Talking (02), Slow Translation (03), Fast Translation (04), Small Rotation (05), Medium Rotation (06). A total number of 60 sequences of 1 min each were recorded. We synthesized the image sequences to videos which have a frame rate of 30 Hz and resolution of 640 × 480 pixels. The test subjects were placed in front of the camera with an average distance of 1.1 m. We chose 17 videos without obvious bangs in static and conversation scenes.

The UBFC-Phys dataset [53] is a public multimodal dataset dedicated to psychophysiological studies. 56 participants followed a three-step experience that involved a rest task T1, a speech task T2 and an arithmetic task T3. Each video is 3 min long and the frame rate is 35 fps. This dataset also gives the BVP signals measured by contact equipment during the three tasks. Through Fourier transform and spectrum analysis of BVP signal, we selected 33 videos with less noise interference and no obvious bangs under T1 task.

Three datasets specifications are list in Table 1.

We obtained a total of 74 videos from these three datasets. Dlib face detector is used to detect the subject's face firstly. Then, the prediction model providing 68-point facial landmarks (https://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2)is applied to the detected front face to obtain the (x, y) coordinates of the facial in different postures. We selected the rectangle surrounded by points 1, 8 and 15 to add masks, the pixel value of the region is changed to 0, and the result is shown in Fig. 2.

Some videos in the mask dataset are shown in Fig. 3.

3.2. *Frontal ROI localization and skin segmentation*

The ROI in the forehead was determined through the location of the human eyes, as shown in the yellow area in Fig. 4. The human eyes were first detected with VJ eye detector and then the forehead is selected as ROI through a fixed proportion.

In order to exclude irrelevant areas such as eyebrows and hair, we obtained the following optimal proportion through experiments. Assuming that the point in the upper left corner of the human eyes re-

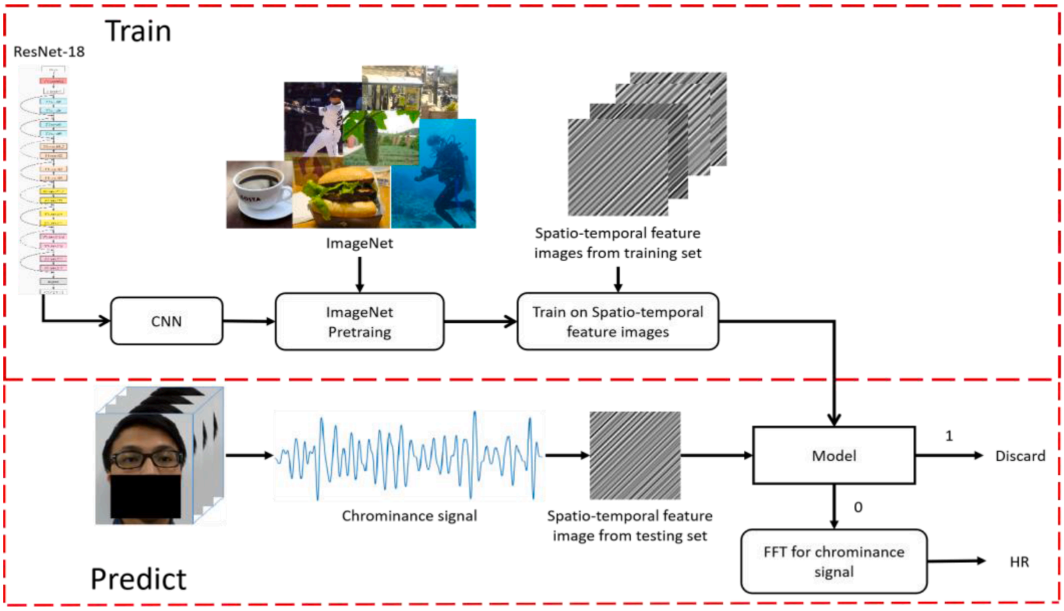

Train
ResNet-18
ImageNet
Spatio-temporal feature images from training set
CNN
ImageNet Pretraing
Train on Spatio-temporal feature images
Chrominance signal
Spatio-temporal feature image from testing set
Model
1
Discard
0
FFT for chrominance signal
HR
Predict

Fig. 1. Algorithm flowchart. Firstly, we use ResNet-18 network to train the spatio-temporal feature images generated by the video in the training set to obtain a model that can judge the signal quality (shown in the upper part of the figure), and then we use the model to judge the quality of the spatio-temporal feature image generated in the testing set (shown in the lower part of the figure).

Table 1
Datasets Parameters.

	Camera	Fps (Hz)	Resolution	Time (min)	Scenarios
PFF	Nikon D5300	50	1280 × 720	3	BS, BM, DS, DM, BR
PURE	Eco 274CVGE	30	640 × 480	1	01,02,03, 04,05,06
UBFC-Phys	EO-23121C	35	1024 × 1024	3	T1,T2,T3

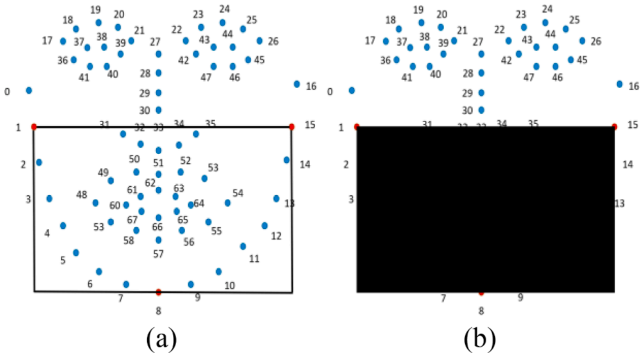


Fig. 2. (a) Points 1, 8 and 15 are marked in red. (b) Schematic diagram of adding mask.

gion is $A(x_1, y_1)$, the length and width of the human eyes region are w_1 and h_1, the point in the upper left corner of the forehead ROI is $B(x_2, y_2)$, the length and width of the ROI are w_2 and h_2. The coordinates of the upper left corner of the ROI and the length and width are deduced according to (1).

$$\begin{cases} x_2 = x_1 + 0.125*w_1 \\ y_2 = y_1 - 1.5*h_1 \\ w_2 = 0.75*w_1 \\ h_2 = h_1 \end{cases} \tag{1}$$

Then we used a skin segmentation algorithm from the iPhys Matlab toolbox [54] to further extract skin pixels in ROI.

3.3. Generating Spatio-temporal feature images and divide into dataset

Raw RGB signal generated from skin with the chrominance-based signal processing [3] can obtain pure pulse signal. After that, we used

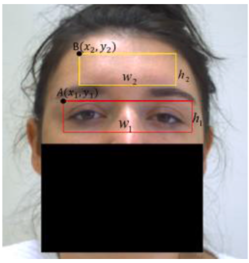


Fig. 4. ROI location.

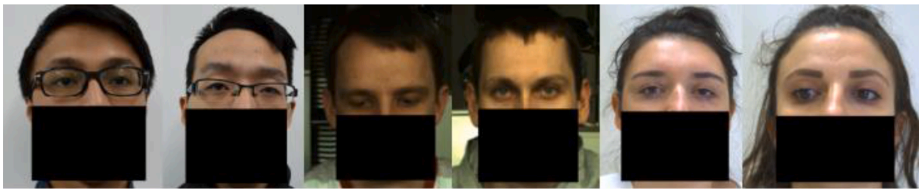


Fig. 3. Video display of mask dataset.

the method of Song et al. [32] to construct a spatio-temporal feature images for each pulse wave signal in a time-delay manner. Fig. 5(a)–(e) displays a pipeline to generate spatio-temporal feature images from the input video. The window length is set to 20 s, increasing every second, as shown in Fig. 6. Then the test dataset was divided by spatio-temporal feature images. We referred to the ± 12 bpm threshold [18]. If the difference between the calculated heart rate and the real value after Fourier transform of the chrome signal at time T is$<$12 bpm, the spatio-temporal feature images at time T is classified as correct, otherwise, it is wrong.

3.4. Training a model for judging signal quality

After obtaining the spatio-temporal feature image, we use the deep learning method to judge its quality. Here, we selected ResNet-18 as the CNN model. Resnet-18 consists of a convolution layer, four residual structures, an average pooling layer and a full connection layer. We used transfer learning. After loading the weight, we replaced the fully connected layer of the original network to predict the quality of the images. Cross entropy was defined as loss function to measure the difference between feature images and labels. The cross-entropy loss function is in the following equation:.

$$Loss = -(p_{True}(x_i)log(q_{True}(x_i)) + p_{False}(x_i)log(q_{False}(x_i))) \tag{2}$$

where $p_{True}(x_i)$ represents the probability that the i^{th} sample is divided into true class, $q_{True}(x_i)$ represents the probability that the i^{th} sample is predicted to be true class, $p_{False}(x_i)$ represents the probability that the i^{th} sample is divided into false class, and $q_{False}(x_i)$ represents the probability that the i^{th} sample is predicted to be false class.

As shown in the prediction part of Fig. 1, the model is used to judge the quality of the input picture in the test. When the signal quality is good, the model outputs 0, continue to perform FFT processing on the chrominance signal to predict the HR. When the signal quality is bad, the model discards it and outputs 1.

4. Experiments AND results

4.1. Metrics

We selected root mean square error (RMSE), Pearson correlation coefficient (R) and 5% and 10% of the effective measurements of each video to measure the proposed method.

RMSE is defined as the mean square root of the square difference between the measured value and the real value. Its formula is:.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(HR_{predict^i} - HR_{true^i})^2} \tag{3}$$

where $predict^i$ represents the average of all measured values of the i^{th} video, $true^i$ represents the average of all given real values of the i^{th} video, and n represents the total number of videos. Pearson correlation coefficient formula is shown in (4).

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{4}$$

where X_i represents the measured average value of the i^{th} video, Y_i represents the real average value of the video, and —X represents the average value of the vector X.

4.2. Face detection result

We carried out several sets of experiments to test the effectiveness of our mask adding method. VJ, Dlib and MTCNN face detection methods were used to detect the faces of 74 videos respectively. The proportion of face detection in each scene are shown in Table 2. We also computed the scene average detection rate (SADR) of three face detection algorithms and the average detection rate (ADR) of each algorithm. PURE 01 refer to the scenario in the PURE dataset. The rest are the same.

4.3. Within-database testing

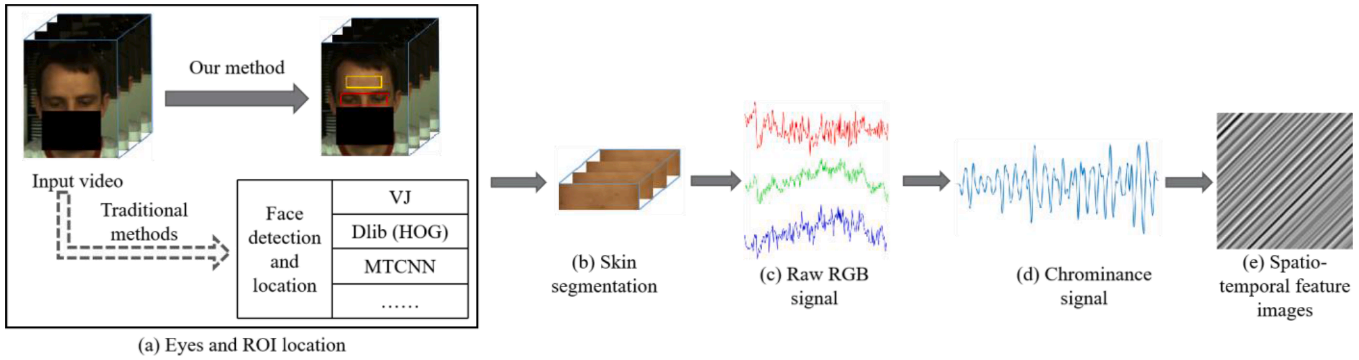
We perform the within-database testing with our self-made mask dataset.

From 74 videos processed with masks, we selected 53 videos for training and 21 for testing. According to the set signal processing method, video and BVP data are processed using a sliding window with a length of 20 s and an increment of 1 s each time. A total of 6766 spatio-temporal feature images were generated from our self-made mask dataset, including 6091 for training and 675 for validation. All the feature images were sampled to 300 × 300 before being input to the network. The learning rate was set as 0.0001 for the first 10 epochs and then 0.00001 for the next 3 epochs. The final accuracy on the validation set is 92%. The within-database estimation results are summarized in Table 3. It lists the measured average value of HR before correction (MABC), which refer to use forehead region only, and after correction (MAAC), which refer to use forehead region and our correction method on 21 test videos. The ground-truth average value (GT), and the proportion of the measured value in the range of 5% and 10% above and below the GT are also list in Table 3.

Table 4 lists the statistical measurement results of 21 video RMSE, $<$5% and$<$10%, as well as the Pearson correlation coefficient.

To further analyze the estimation consistencies by individual approach, we also draw the Bland-Altman plots and regression plots of the within-database testing shown in Fig. 7, Fig. 8 respectively. It can be seen that filtering out outliers by residual network can narrow the gap between estimated HR and ground-truth.

Moreover, 10-fold cross-validation was used to evaluate the



(a) Eyes and ROI location
(b) Skin segmentation
(c) Raw RGB signal
(d) Chrominance signal
(e) Spatio-temporal feature images

Fig. 5. A pipeline for showing how to generate spatio-temporal feature images from the input video. When the face cannot be detected, traditional methods cannot continue to detect HR, but our method uses the eyes and ROI location to continue to detect HR.

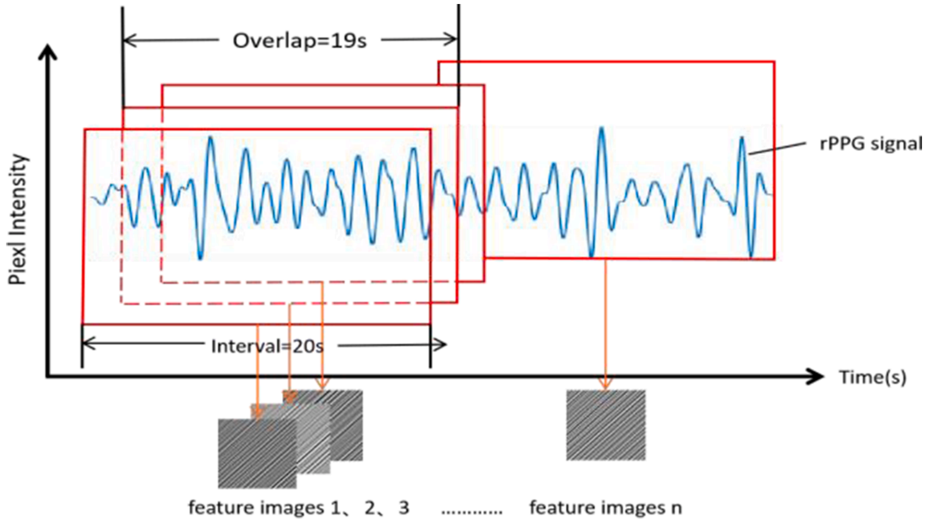


Fig. 6. Signal processing process.

Table 2
Mask video face detection result using different face detection alforithms.

	VJ	Dlib	MTCNN	SADR
PURE 01	0.90%	0.98%	9.85%	3.91%
PURE 02	4.46%	11.79%	0.51%	5.59%
PFF NS	0.38%	10.75%	21.94%	11.02%
PFF DS	4.50%	27.46%	31.93%	21.30%
UBFC-Phys T1	2.57%	10.12%	15.66%	9.45%
ADR	2.56%	12.22%	15.98%	

Table 3
Test Video MEASUREMENT RESULTS USING our mask dataset: A WITHIN-DATABASE CASE (MABC: measured average value of HR before correction, MAAC: measured average value of HR AFTER correction, GT: ground-truth average value).

Video Number	MABC	MAAC	GT	Before correction (CHROM)		After correction (CHROM + DL)	
				<5%	<10%	<5%	<10%
1	96.13	95.36	100.66	0.44	0.68	0.44	0.71
2	68.23	68.18	68.51	0.99	1.00	0.99	1.00
3	90.16	89.94	87.97	0.41	0.58	0.41	0.59
4	79.49	79.47	80.19	0.81	0.94	0.80	0.94
5	105.14	105.04	108.35	0.85	0.93	0.84	0.93
6	100.09	101.03	108.89	0.57	0.70	0.59	0.72
7	89.45	89.38	89.76	0.75	0.97	0.73	0.96
8	87.39	87.33	89.68	0.91	0.99	0.91	0.99
9	74.87	74.87	71.00	0.54	0.75	0.54	0.75
10	87.39	87.16	85.00	0.75	0.90	0.77	0.92
11	94.51	94.58	94.00	1.00	1.00	1.00	1.00
12	60.73	60.73	53.00	0.27	0.39	0.27	0.39
13	67.75	67.75	67.00	0.73	0.98	0.73	0.98
14	84.45	84.45	84.00	0.82	0.94	0.82	0.94
15	103.12	102.19	87.00	0.00	0.02	0.00	0.04
16	94.81	94.71	94.00	0.96	1.00	0.95	1.00
17	76.63	76.71	78.00	0.80	0.98	0.80	0.98
18	74.25	74.25	73.00	0.83	1.00	0.83	1.00
19	101.73	101.47	101.00	0.88	0.97	0.86	0.96
20	103.29	103.25	104.00	0.88	1.00	0.89	1.00
21	104.43	81.25	76.00	0.02	0.04	0.50	0.75

generalization ability of the model. For the convenience of calculation, 70 videos were randomly selected from the self-made mask dataset and randomly divided them into 10 groups. All the results are provided in Table 5.

Table 4
Statistical measurement results: a within-database case.

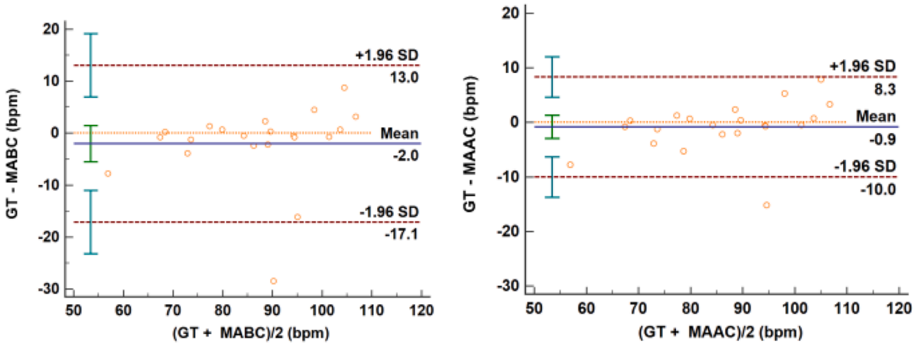
	Before correction	After correction
RMSE	5.07	4.65
<5%	0.68	0.70
<10%	0.80	0.84
R	0.85	0.95

From the results, we can see that the highest accuracy is 96.7%, the lowest accuracy is 73.7%, and the average accuracy is 84.42%. This indicates that the selection of different training and testing data has a great impact on the accuracy of the training model. The difference between the highest accuracy and the lowest accuracy is 23%, which also shows the importance of randomly selected data for deep learning. Theoretically, the model accuracy obtained from other random choice method should be between 73.7% and 96.7%.

4.4. Cross-database testing

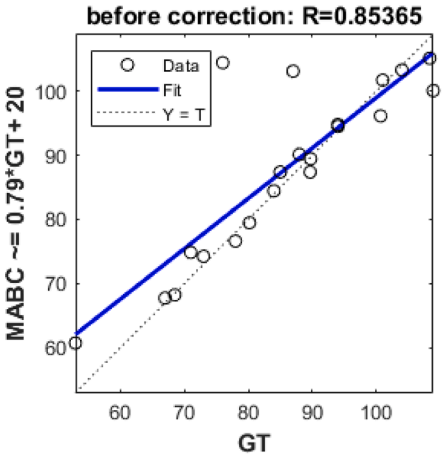
In order to validate the generalization ability of our model, we also conduct cross-database experiments on the COHFACE [16] database. This dataset is composed of 160 videos in two different illumination settings (studio lighting and natural lighting) collected from 40 healthy individuals. Their physiological signals are taken by a BVP sensor and a respiration belt, distributed using standard HDF5 containers. Studio lighting results in a uniform distribution of light on the face, while natural lighting results in a half bright and half dark on the face. Each client has a total of 4 videos, 2 of which are in studio lighting condition and 2 of which are in natural lighting condition. The heartpy, a python Heart Rate analysis toolkit, was applied to estimate HR using the signal processing method mentioned in Section III.C. We only selected 60 videos with computable HR and uniform facial illumination to do experiments.

The statistical measurement results are summarized in Table 6. The Bland-Altman plots and regression plots of the cross-database testing are also given in Fig. 9, Fig. 10 respectively. We can clearly observe that the results of all evaluation metrics decreased significantly. A mean reason for this phenomenon is the videos are heavily compressed, so noise artifact was unavoidably added [34,55]. McDuff et al. [56] also find that a considerable drop in SNR between raw and compressed videos through experiments. Thus, it is not surprising that RMSE is very large. The comparison results demonstrate that after correction, the RMSE decreases slightly from 13.16 bpm to 13.06 bpm. At the same time, the



(a) (b)

Fig. 7. Bland-Altman plots between GT and estimated HR values for a within-database case: (a) for MABC (b) for MAAC.



(a) (b)

Fig. 8. The regression plots for (a) MABC, (b) MAAC for a within-database case.

proportion<5% and 10% increases slightly, from 26.63% to 27.04%, from 37.94% to 38.64% respectively. But Pearson correlation coefficient almost unchanged.

Table 5
10-fold cross-validation RESULTS on self-made mask datasets.

Number	Accuracy (%)
1	87.0
2	88.0
3	83.1
4	84.6
5	73.7
6	78.8
7	82.9
8	96.7
9	91.2
10	78.2
Average	**84.42**

Table 6
Statistical measurement results: a cross-database case.

	Before correction	After correction
RMSE	13.16	13.06
<5%	26.63	27.04
<10%	37.94	38.64
R	0.31	0.31

5. Discussion

In the previous section, we evaluated the effectiveness of the mask algorithm and effectiveness of using deep learning method to remove outliers.

The results in Table 2 show that our mask algorithm has a certain inhibitory effect on face detection. However, the suppression ability is significantly different in different scenes for different algorithms. For the PURE dataset with dark light, the detection rate is small, up to 5.59%, while the detection rate of PFF and UBFC with strong light is up to 21.30%. This shows that light is very important for face detection. In addition, the detection capabilities of VJ, Dlib and MTCNN are also different. VJ has the weakest detection ability, only 2.56% of faces can be detected, and MTCNN has the strongest detection ability, 15.98% of faces can be detected. Due to MTCNN is time consuming, most rPPG based algorithms use Dlib and VJ for face detection. Our mask algorithm has a good suppression effect on these two face detection methods. In the process of face detection on the video with masks, we found that the position of masks will be different due to the involuntary shaking of the face up and down. When the subject looks up, the nose area is not covered by the mask, which will cause the face to be detected. When the subject lowers his head, the mask will cover half of the eyes, so that the eye cannot be detected in the following step. Therefore, whether the

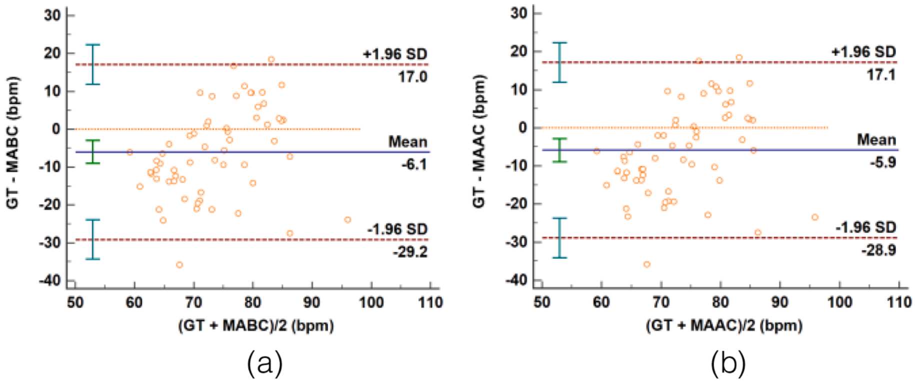


Fig. 9. Bland-Altman plots between GT and estimated HR values for a cross-database case: (a) for MABC (b) for MAAC.

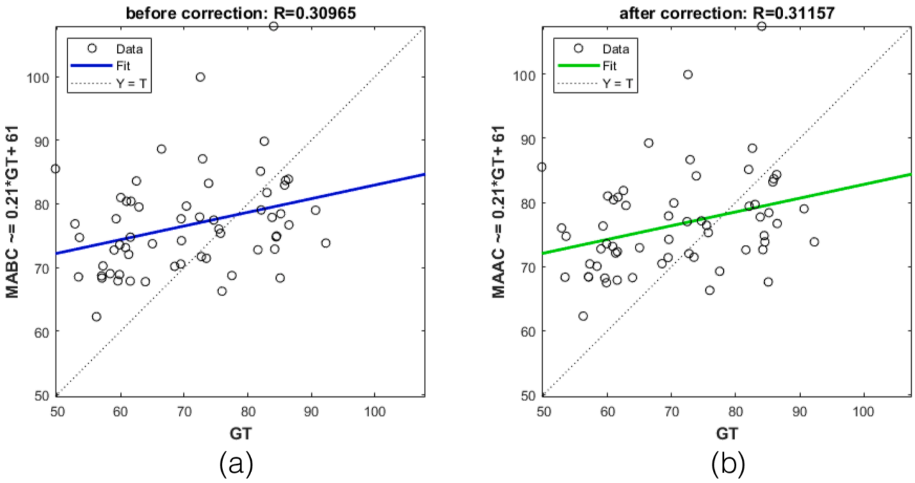


Fig. 10. The regression plots for (a) MABC, (b) MAAC for a cross-database case.

mask can effectively shield the face detection algorithm largely depends on the pose of the original video subject.

It can be seen from Table 3 that the measurement accuracy is improved after correction. For the 8 videos 1, 2, 4, 5, 7, 8, 11 and 20, the measurement accuracy decreases slightly after correction, that is, the measurement average is farther from the real average. The measurement effects of the five videos 9, 12, 13, 14 and 18 remain unchanged, and the accuracy of the eight videos 3, 6, 10, 15, 16, 17, 19 and 21 is improved. A further novel finding is that video No. 21 has the largest increase of 23.18 bpm. This means that there is noise in the video that can not be removed by the chrome-based method alone, but the characteristics of this noise can be learned and eliminated through deep learning. In addition, only the proportion of No. 4, 5, 7, 16 and 19 video measurement errors within 5% decreased slightly, and the proportion of No. 7 and 19 video measurement errors within 10% decreased slightly. Together, the present findings confirm that although the measurement error of some videos decreases after correction, most outliers are filtered out, so that the number of values measured in the confidence interval increases.

Table 4 shows the RMSE, Pearson correlation coefficient and the proportion of measured values in the 5% and 10% confidence intervals of 21 videos. It can be found that the corrected root mean square error is increased by 0.42 bpm, the Pearson correlation coefficient is increased by 0.1, and the increase range in the 10% confidence interval is 0.04, which is 2 percentage points higher than that in the 5% confidence interval. This result is consistent with the conclusion in Table 3, that is, outliers can be effectively removed through the residual network.

Although we randomly selected 21 out of 74 videos for testing, different videos may have slightly different training results. We only selected the still video in the three datasets, because the video duration of other data sets is short or the video has been compressed. Moreover, the influence of the setting of window length on the training results is also worth studying. The most important thing is how to measure the heart rate if the subjects wear masks and their forehead is also covered by their hair. In the future, we will not be limited to the measurement of heart rate only under static conditions. In real scenes such as head rotation or horizontal rotation, subjects moving, and various special situations mentioned above, we will conduct further research.

6. Conclusion

In this paper, we proposed a HR detection method when missing information and a model to filter outliers based on residual network. We first used the human eyes to locate ROI, and then combined traditional methods with deep learning for heart rate detection to solve the problem of lack of facial information and unstable output. Besides, in order to test the effectiveness of our algorithm, we also designed a method to create mask dataset. Through experiments, we concluded that after adding masks to the video, the detection rate of different face detection algorithms is 15.98% at the highest and 2.56% at the lowest, which effectively suppresses face detection. In the test dataset, the RMSE of our proposed algorithm is reduced to 4.65 bpm, the proportion$<$5% and$<$10% are increased by 2% and 4% respectively, and the Pearson correlation coefficient is increased by 0.1. In conclusion, the results show the effectiveness of our method on test datasets. We hope this work can provide effective help in the biomedical field in the future.

Funding

This work was supported by Scientific Research Plan of Beijing Education Commission (SZ202110005002), and Beijing Natural Science Foundation (4192005), CN.

CRediT authorship contribution statement

Kun Zheng: Conceptualization, Methodology, Software, Writing – review & editing, Funding acquisition, Project administration. **Kangyi Ci:** Data curation, Software, Writing – original draft, Writing – review & editing. **Hui Li:** Writing – review & editing. **Lei Shao:** Conceptualization, Visualization, Investigation. **Guangmin Sun:** Methodology, Supervision. **Junhua Liu:** Visualization, Software. **Jinling Cui:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Verkruysse, L.O. Svaasand, J.S. Nelson, Remote plethysmographic imaging using ambient light, Opt. Express 16 (26) (2008) 21434–21445, https://doi.org/10.1364/OE.16.021434.

[2] L.A.M. Aarts, V. Jeanne, J.P. Cleary, C. Lieber, J.S. Nelson, S. Bambang Oetomo, W. Verkruysse, Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—A pilot study, Early Hum. Develop. 89 (12) (2013) 943–948.

[3] Z. Guo, Z. J. Wang, and Z. Shen, "Physiological parameter monitoring of drivers based on video data and independent vector analysis," in *Proc. ICASSP*, 2014, pp.4374-4378, DOI:10.1109/ICASSP.2014.6854428.

[4] K. Zheng, K. Ci, J. Cui, J. Kong, J. Zhou, Non-contact heart rate detection when face information is missing during online learning, Sensors 20 (24) (2020) 7021.

[5] W. Taylor, Q.H. Abbasi, K. Dashtipour, S. Ansari, S.A. Shah, A. Khalid, M.A. Imran, A review of the state of the art in non-contact sensing for covid-19, Sensors 20 (19) (2020) 5665.

[6] R. Sinhal, K. Singh and A. Shankar, "Estimating vital signs through non-contact video-based approaches: A survey," in *Proc. RISE*, 2017, pp. 139-141, DOI: 10.1109/RISE.2017.8378141.

[7] M. C. Li, Y. H. Lin, "A real-time non-contact pulse rate detector based on smartphone," in *Proc. IEEE Conf. ISNE*, 2015, pp. 1-3, DOI: 10.1109/ISNE.2015.7132025.

[8] A. Qayyum, A. S. Malik, A. N. Shuaibu and N. Nasir, "Estimation of non-contact smartphone video-based vital sign monitoring using filtering and standard color conversion techniques," in *Proc. IEEE Conf. LSC*, 2017, pp. 202-205, DOI: 10.1109/LSC.2017.8268178.

[9] S. Sethi, et al., Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread, J. Biomed. Inform. 120 (2021), https://doi.org/10.1016/j.jbi.2021.103848.

[10] S. Chaichulee, *et al.*, "Multi-task Convolutional Neural Network for Patient Detection and Skin Segmentation in Continuous Non-contact Vital Sign Monitoring," in *Proc. IEEE Conf. FG*, 2017, pp. 266-272. DOI: 10.1109/FG.2017.41.

[11] S. Kwon, J. Kim, D. Lee, *et al.* "ROI analysis for remote photoplethysmography on facial video," in *Proc. IEEE Conf. EMBC*, 2015, pp. 4938-4941, DOI: 10.1109/EMBC.2015.7319499.

[12] P. Viola, M.J. Jones, Robust real-time face detection, Proc. Conf. IJCV 57 (2004) 137–154, https://doi.org/10.1023/B:VISI.0000013087.49260.fb.

[13] G. S. Hsu, A. M. Ambikapathi, M. S. Chen, "Deep learning with time-frequency representation for pulse estimation from facial videos," in *Proc. IJCB*, 2017, pp. 383-389, DOI: 10.1109/BTAS.2017.8272721.

[14] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Trans. Image Process. 29 (2020) 2409–2423, https://doi.org/10.1109/TIP.2019.2947204.

[15] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognit. Lett. 124 (2019) 82–90, https://doi.org/10.1016/j.patrec.2017.10.017.

[16] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," *arXiv:1709.00962*, 2017.

[17] R. Stricker, S. Muller, H. M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. 23rd IEEE International Symposium on Robot & Human Interactive Communication,* 2014, pp. 1056-1062, DOI: 10.1109/ROMAN.2014.6926392.

[18] M.Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Opt. Exp. 18 (10) (2010) 10762–10774, https://doi.org/10.1364/OE.18.010762.

[19] H. Demirezen, C. Eroglu Erdem, Heart rate estimation from facial videos using nonlinear mode decomposition and improved consistency check, Signal Image Video Process. 15 (7) (2021) 1415–1423.

[20] D.-Y. Chen, J.-J. Wang, K.-Y. Lin, H.-H. Chang, H.-K. Wu, Y.-S. Chen, S.-Y. Lee, Image sensor-based heart rate evaluation from face reflectance using Hilbert-Huang transform, IEEE Sens. J. 15 (1) (2015) 618–627.

[21] S. Kado, Y. Monno, K. Yoshizaki, M. Tanaka, M. Okutomi, Spatial-spectral-temporal fusion for remote heart rate estimation, IEEE Sens. J. 20 (19) (2020) 11688–11697.

[22] X. Chen, J. Cheng, R. Song, Y.u. Liu, R. Ward, Z.J. Wang, Video-based heart rate measurement: recent advances and future prospects, IEEE Trans. Instrum. Measure. 68 (10) (2019) 3600–3615.

[23] H. Qi, Z. Guo, X. Chen, Z. Shen, Z. Jane Wang, Video-based human heart rate measurement using joint blind source separation, Biomed. Signal Process. Control 31 (2017) 309–320.

[24] M. Lewandowska, J. Ruminski, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam—a Non-contact method for evaluating cardiac activity," in *Proc. FedCSIS*, Sep. 2011, pp. 405-410.

[25] L. Qi H. Yu L. Xu R.S. Mpanda S.E. Greenwald Robust heart-rate estimation from facial videos using Project_ICA Physiological Measurement 40 8 2019 10.1088/1361-6579/ab2c9f 085007 085007.

[26] W. Wang, A.C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote PPG, IEEE Trans. Biomed. Eng. 64 (7) (2017) 1479–1491.

[27] G. de Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Trans. Biomed. Eng. 60 (10) (Oct. 2013) 2878–2886, https://doi.org/10.1109/TBME.2013.2266196.

[28] G. de Haan, A. van Leest, Improved motion robustness of remotePPG by using the blood volume pulse signature, Physiol. Meas. 35 (9) (2014) 1913–1926, https://doi.org/10.1088/0967-3334/35/9/1913.

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

[30] W. Wang, S. Stuijk, G. de Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE Trans. Biomed. Eng. 63 (9) (2016) 1974–1984, https://doi.org/10.1109/TBME.2015.2508602.

[31] C. S. Pilz, S. Zaunseder, J. Krajewski, *et al.*, "Local group invariance for heart rate estimation from face videos in the wild," in *Proc. IEEE/CVF Conf. CVPRW*, Jun. 2018, pp. 1335-1343, DOI: 10.1109/CVPRW.2018.00172.

[32] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, X. Chen, Heart Rate Estimation from Facial Videos Using a Spatiotemporal Representation with Convolutional Neural Networks, IEEE Trans. Instrum. Measure. 69 (10) (2020) 7411–7421.

[33] A. Ni, A. Azarang, N. Kehtarnavaz, A review of deep learning-based contactless heart rate measurement methods, Sensors 21 (11) (2021) 3719, https://doi.org/10.3390/s21113719.

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

[35] W. Chen, D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. ECCV*, 2018, pp. 356-373, DOI: 10.1007/978-3-030-01216-8_22.

[36] R. Špetlík V. Franc J. Cech Visual heart rate estimation with convolutional neural network Proc. BMVC 84 2018 1 12 Available https://bmvc2018.org/index.html.

[37] B. Huang, C.-L. Lin, W. Chen, C.-F. Juang, X. Wu, A novel one-stage framework for visual pulse rate estimation using deep neural networks, Biomed. Signal Process. Control 66 (2021) 102387.

[38] Z. Yu, W. Peng, X. Li, *et al.*, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. IEEE ICCV*, 2019, pp. 151-160, DOI: 10.1109/ICCV.2019.00024.

[39] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proc. Conf. BMVC*, 2019, pp. 1-12, DOI:10.5244/C.33.29.

[40] G.-S. Hsu, R.-C. Xie, ArulMurugan Ambikapathi, K.-J. Chou, A deep learning framework for heart rate estimation from facial videos, Neurocomputing 417 (2020) 155–166.

[41] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A.E. Saddik, EVM-CNN: Real-time contactless heart rate estimation from facial video, IEEE Trans. Multimedia 21 (7) (2019) 1778–1787, https://doi.org/10.1109/TMM.2018.2883866.

[42] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm: Learning a deep heart rate estimator from general to specific," in *Proc. IEEE 24th ICPR*, 2018, pp. 3580-3585, DOI: 10.1109/ICPR.2018.8546321.

[43] X. Niu *et al.*, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proc. IEEE FG*, 2019, pp. 1-8, DOI: 10.1109/FG.2019.8756554.

[44] H. Lu, H. Hu, "NAS-HR: Neural architecture search for heart rate estimation from face videos", *Virtual Real*, Intell. Hardw 3 (2021) 33–42, https://doi.org/10.1016/j.vrih.2020.10.002.

[45] R. Song, H. Chen, J. Cheng, C. Li, Y.u. Liu, X. Chen, PulseGAN: learning to generate realistic pulse waveforms in remote photoplethysmography, IEEE J. Biomed. Health Inform. 25 (5) (2021) 1373–1384.

[46] F. Bousefsaf, D. Djeldjli, Y. Ouzar, et al., iPPG 2 cPPG: reconstructing contact from imaging photoplethysmographic signals using U-Net architectures, Comput. Biol. Med. 138 (2021), 104860, https://doi.org/10.1016/j.compbiomed.2021.104860.

[47] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection," In *Proc. IEEE 2005 CVPR*, San Diego, CA, USA, 20-25 June 2005, pp. 886-893, DOI: 10.1109/CVPR.2005.177.

[48] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503, https://doi.org/10.1109/LSP.2016.2603342.

[49] Y. Nirkin, I. Masi, A. T. Tuan, *et al.*, “On face segmentation, face swapping, and face perception,” in *Proc. IEEE FG*, 2018, pp. 98-105. DOI: 10.1109/FG.2018.00024.

[50] K.M. van der Kooij, M. Naber, An open-source remote heart rate imaging method with practical apparatus and algorithms, Behav. Res. Methods 51 (5) (2019) 2106–2119, https://doi.org/10.3758/s13428-019-01256-8.

[51] A. Shcherbina, C. Mattsson, D. Waggott, H. Salisbury, J. Christle, T. Hastie, M. Wheeler, E. Ashley, Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort, J. Pers. Med. 7 (2) (2017) 3.

[52] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, M. Sarlo, Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions, Psychophysiology 56 (11) (2019), e13441, https://doi.org/10.1111/psyp.13441.

[53] R. Meziati Sabour Y. Benezeth P. De Oliveira J. Chappe F. Yang UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress 1 1.

[54] D. McDuff, E. Blackford, “iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox,” in *Proc. IEEE EMBC*, 2019, pp. 6521–6524. DOI: 10.1109/EMBC.2019.8857012.

[55] Y. Deng, A. Kumar, “Standoff Heart Rate Estimation from Video-A Review,” in *Proc. SPIE Defense + Commercial Sensing*, 2020, DOI: 10.1117/12.2560683.

[56] D. McDuff, E. Blackford, J. Estepp, “The Impact of Video Compression on Remote Cardiac Pulse Measurement Using Imaging Photoplethysmography,” in *Proc. IEEE FG,* 2017, pp. 63–70, DOI: 10.1109/FG.2017.17.