Computers in Biology and Medicine 151 (2022) 106307

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Heart rate estimation network from facial videos using spatiotemporal feature image

Kokila Bharti Jaiswal [*,1], T. Meenpal [2]

Department of ECE, National Institute of Technology, Raipur 492010, India

ARTICLE INFO

Keywords:
Remote health monitoring
Heart rate
Remote photoplethysmography (rPPG)
Spatiotemporal image
Convolutional neural network

ABSTRACT

Remote health monitoring has become quite inevitable after SARS-CoV-2 pandemic and continues to be accepted as a measure of healthcare in future too. However, contact-less measurement of vital sign, like Heart Rate(HR) is quite difficult to measure because, the amplitude of physiological signal is very weak and can be easily degraded due to noise. The various sources of noise are head movements, variation in illumination or acquisition devices. In this paper, a video-based noise-less cardiopulmonary measurement is proposed. 3D videos are converted to 2D Spatio-Temporal Images (STI), which suppresses noise while preserving temporal information of Remote Photoplethysmography(rPPG) signal. The proposed model projects a new motion representation to CNN derived using wavelets, which enables estimation of HR under heterogeneous lighting condition and continuous motion.

STI is formed by the concatenation of feature vectors obtained after wavelet decomposition of subsequent frames. STI is provided as input to CNN for mapping the corresponding HR values. The proposed approach utilizes the ability of CNN to visualize patterns.

Proposed approach yields better results in terms of estimation of HR on four benchmark dataset such as MAHNOB-HCI, MMSE-HR, UBFC-rPPG and VIPL-HR.

1. Introduction

The SARS-CoV-2 (COVID-19) epidemic is changing the landscape of global healthcare [1,2]. This transformation may be witnessed by the abrupt change in the medical appointments via telehealth. To deliver high quality patient care and lower the risk of COVID-19, telehealth emerges as a solution. Performing primary care visits to the patient's home lowers the risk of infection, improves visit efficiency, and makes treatment more accessible to persons who live in distant areas or are unable to travel. Tele-health provides diagnosis based on self reported symptoms and visual observation. But they are unable to objectively analyze the patient's physiological state in the majority of cases. Special attention must be given to cardiovascular protection while treatment, as suggested by experts [3]. By the development of more precise and efficient non-contact cardiopulmonary measurement technologies, remote physicians would have access to data allowing them to make more informed decisions. It is significant that the telehealth will be imbibed in the health care system of future generation.

Non contact cardiopulmonary measurement technology such as Photoplethysmography(PPG) traps subtle changes in the reflected light due to physiological activity. HR estimation is possible by color variations that are synchronized with blood volume. In recent years there has been an increased interest in exploring methods for rPPG measurement [4]–[5], which allows HR estimation from the skin, i.e., the face area, without contact with a person.

Verkruysse et al. [4] proposed an early approach for the detection of rPPG signal depending upon the blood volume, using a low cost camera. Since its inception, researchers have made tremendous progress in camera-based HR estimation. Since the pulse signal is very weak it can be easily affected by noise due to presence of body movement or variation in illumination. Many researchers simplified this noise by making certain assumptions and successfully derived HR from rPPG signal. For example, sources are considered to be statistically independent in Blind Source Separation(BSS) methods such as Principle Component Analysis (PCA) [6] and Independent Component Analysis (ICA) [7]. Similarly, model based methods such as Chrominance (CHROM) [8] and Plane-Orthogonal-to-Skin (POS) [5] have made certain presumptions and are based on simple skin reflection model. However, the noise present in realistic

[*] Corresponding author.
E-mail address: kbjaiswal.phd2018.etc@nitrr.ac.in (K.B. Jaiswal).

[1] Research Scholar

[2] Assistant Professor

https://doi.org/10.1016/j.compbiomed.2022.106307
Received 2 June 2022; Received in revised form 5 November 2022; Accepted 6 November 2022
Available online 9 November 2022

situation is far more complex and varied. Under such situation it is difficult to get the reliable measurement using conventional methods.

Recently, from last few decades, the invasion of deep learning in prominent research fields such as computer vision and natural language processing, results in a substantial improvement. Motivated by these, researchers intrigued to test the applicability of deep learning techniques in the field of rPPG. Several researches [9–12] have used CNN architectures and able to provide much better rPPG estimation than conventional methods. As suggested in [13], representation of HR signals plays a significant role to train a deep HR estimator. Inspired by these we have proposed a new deep learning based rPPG estimation method using spatiotemporal representation of videos. In particular, a unique feature extraction strategy for efficiently presenting a sequence of frames to CNN is devised. We create a video abstraction method for transforming important spatial and temporal information of a video in a 2D representation called STI. The STI will be submitted to CNN for estimation of HR.

1.1. Article contribution

- A new paradigm is proposed for the estimation of HR under realistic noise conditions.
- We proposed a spatio-temporal map to highlight the signal related to heart rate while suppressing other uncorrelated signals. Such spatiotemporal image forces the CNN to specifically learn informative signal related to heart rhythm for the estimation of final HR.
- A new method is proposed to form a 2D STI from 3D videos, for better representational learning by CNN. STI utilizes the property of wavelets to work at multiple scales and resolution.
- The construction of STI are based on wavelets, which is robust to uniform motions. Under extreme noise circumstances and even if there are some occlusions, tracking using wavelet transform is still feasible.
- Cropping and rescaling STI are used to solve the problem of variation in the size of input video frames.
- We run thorough intra- and cross-dataset tests to demonstrate that the proposed method outperforms state-of-the-art methods.

2. Related works

In this section, we discuss previously proposed techniques for both conventional and deep learning-based rPPG measurement.

2.1. Conventional methods

Verkruysse et al. [13] initially assessed the possibilities to measure heart rate remotely using face recordings. Since then, numerous researchers have committed themselves to rPPG research. The conventional methods of rPPG based HR estimation are majorly divided into two categories namely 1. Blind Source Suppression(BSS) based algorithms 2. Model based algorithms. BSS based methods assumes sources to be statistically independent and non Gaussian. It extracts pulse signal from mixed signals of pulse and noises without having the prior knowledge of mixing. Most commonly used BSS methods are ICA [7] and PCA [6].

Empirical mode decomposition method proposed by [14] separates signal into number of intrinsic modes, and one clean signal is selected which resembles pulse signal. These above mentioned methods becomes inefficient in the presence of head movement, facial deformations, and illumination variations.

In order to overcome noise due to motion, model based methods is proposed. Chrominance method (CHROM) [8] uses a method that utilizes a skin reflection model and combines RGB channels in a linear fashion. Dependence on skin reflection model, easily separates the pulse, and noise occurs due to motion. POS [5] utilizes the same skin reflection model but uses different projection for the separation of pulse and motion induced noise. With the prior knowledge of signature blood volume pulse vector, robustness of PBV [15] for noise is provide better than CHROM and POS. By using clearly defined skin mask another method called spatial subspace rotation (2SR) [16] is proposed to measure a spatial subspace of skin-pixels and measure its temporal rotation for pulse extraction. The advantage of this method is that, it does not require preliminary information about skin tone or pulse. All of the methods presented above are based on calculating the spatial mean of the entire face, assuming that each pixel's contribution to rPPG estimate is equal. Such an assumption is noise-sensitive, making it impossible to use in realistic circumstances. For HR estimation in naturalistic condition, [17] proposed a method which automatically discards noisy features and selects only features corresponding to HR value. This method used a matrix completion theory involving self adaption strategy. In this method the author is trying to localize perturbations in chrominance features due to face movements. However the methods stated above are not sufficiently robust for unsteady environmental conditions such as low illumination, abrupt motion etc. This constraint motivates the adoption of deep learning techniques to achieve robustness.

2.2. Deep learning-based remote HR estimation

Technologically enhanced results are driving force behind using deep learning techniques for real world problems. HR measurement from facial videos can no longer be untouched. There are two possible ways of applying deep learning technique, end-to-end or as a feature decoder. The first end-to-end method [18] provides spatio-temporal visualization of physiological signal via attention mechanism. The main aim of the proposed network is to discriminate motion from different sources and output target motion signal. Motion analysis is a very important step in deep learning based video processing. Motion representation such as frame difference or optical flow are calculated manually to fed into the convolutional neural networks Deepphys [18] exploits motion difference information followed by attention networks for better estimation of rPPG signals. AutoHr [19] proposed an end-to-end network with strong Neural Architecture Search(NAS) method. End-to-end method proposed in [11] learns features through a feature extractor CNN network, which is further presented to HR estimator. The end-to-end methods results in a mysterious black box model, which is often difficult to understand.

In contrast non end-to-end methods first extracts handcrafted features followed by deep learning network for HR estimation [9,10,12, 20–22]. Spatiotemporal features of video are extracted, and then these uniquely arranged spatiotemporal maps is applied to CNN network for HR estimation in [9,12,20]. After extracting the rPPG signals via traditional CHROM [8], Hsu et al. [23] formed the Time Frequency representation (TFR) of an image by directly accumulating the frequency component of processed time domain signal. VGG15 is used as a backbone network for HR regression. Qiu et al. [9] extracted features via Eulerian Video Magnification(EVM) and then CNN is employed to regress the HR value. Niu et al. [12,13,20] formed a spatiotemporal map, which is a representation of aggregated information from multiple ROIs. Next, ResNet18 is cascaded for HR prediction. However, these methods need strict preprocessing procedures and neglect the global clues outside the pre-defined ROIs. Performances of these methods is limited to availability of accurate ROIs. Although they offers substantial improvement in RMSE but at the cost of preprocessing overhead.

Methods stated above performs effectively for slight head movements and motion due to natural facial expression, but is less reliable for drastic head movements or continuous motion. Some motion signals occurs due to periodic head movements, closely aligned with the pulse signal, further complicate the problem. Such signals cannot be easily distinguished by deep learning methods or filters with empirical settings.

Table 1

A brief summary of the existing HR estimation networks based on spatiotemporal maps.

Method	Input signal	Feature extraction	Backbone network	Outcomes
Hsu et. al [23]	Green Channel Signal	Time frequency representation of an image	VGG15	Pioneering work for real-time rPPG measurement using deep learning framework
Qiu et. al [9]	RGB	EVM	regression CNN	Realtime HR estimation with very less processing time
Niu et. al [12]	CHROM	n temporal signals are concatenated row wise to form spatiotemporal map	Deep regression model with Gated Recurrent Unit(GRU)	HR estimation in general situations like head movements and bad illumination
Pulse GAN [22]	CHROM	Noisy rPPG signal	Conditional GAN	Noise-less realistic rPPG signal is generated
Song et. al [24]	CHROM	Feature map is constructed by arranging the peaks of the signal in a time delayed manner	ResNet18	Noise-less feature images are produced which improves the prediction accuracy of HR
Wu et. al [25]	RGB	Spatiotemporal feature map generation similar to Song et. al with equivalent padding	ResNet18	Able to generate spatiotemporal maps which compensates missing frames in unstable situations.
Proposed	RGB	Spatiotemporal feature map generation using wavelets, for better motion estimation	ResNet18	Motion robust HR estimation is possible under realistic situations

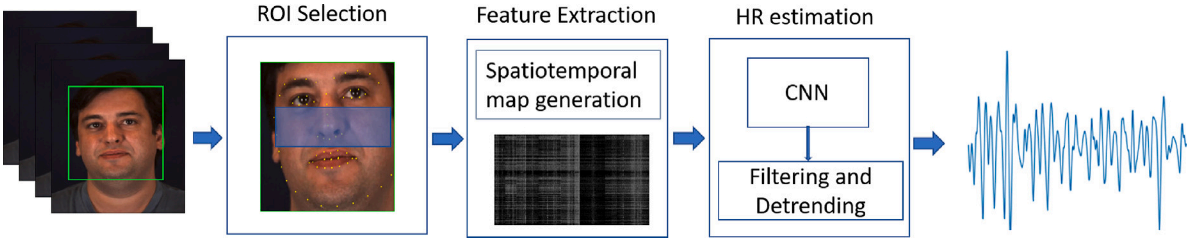

Fig. 1. Overview of our system for HR estimation. Three modules: ROI selection, Feature extraction, HR estimation are present in our system.

The theme of deep learning methods is the quest of strong representations to capture salient features for a given task, which enables improved performance. Spatial correlations between the CNN representations can be captured with the help of learning mechanisms, which increases the effectiveness of the network (see Table 1).

3. Proposed method

The proposed method is divided into three steps, ROI detection and tracking, STI generation and HR estimation using CNN. All the blocks of proposed scheme as shown in Fig. 1 is described briefly in following sections:

3.1. ROI detection and tracking

The subtle changes in the reflected light from the skin region is an important clue for extracting pulse rate. In facial image, there is no contribution of non-skin regions(hair, eyes, eyebrows) to pulse rate. But its presence definitely effects SNR. Therefore ROI selection is done to filter out the unwanted facial region. It is performed as a preprocessing step majorly in all state-of-the-art methods. For detecting faces in video frames we have used 68 landmark detector [26]. The detected ROI is tracked using Kanade–Lucas–Tomasi (KLT) algorithm [27]. In Eq. (1), eight points of the 68 facial landmarks are used to accurately define the ROI.

$$\begin{aligned} X_c &= X_{13} \\ Y_c &= max(Y_{40}, Y_{41}, Y_{46}, Y_{47}) \\ W &= X_{16} - X_{13} \\ H &= min(Y_{50}, Y_{52}) - Y_c \end{aligned} \tag{1}$$

Where X_c and Y_c are the coordinates of top left corner. X_n and Y_n represents coordinates of point n ($n = 1, 2, 3, 4, 5......68$), W stands for width of ROI block and H stands for height of ROI block. ROI defined in this way always exclude eyes and mouth region (see Fig. 2). Hence it reduces the impact of natural facial movements such as blinking of eyes and lip movements while talking.

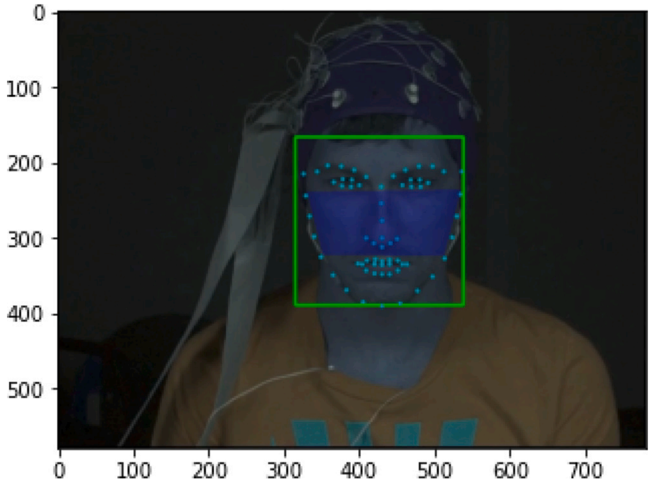

Fig. 2. The green dots represents 68 landmark points. The region highlighted in blue is the ROI subjected to further processing for HR detection.

3.2. Spatio-temporal map generation

Spatiotemporal networks possess a very important position in many video based detection systems such as action detection, action recognition etc. Inspired by this, in this paper, we have developed a novel feature extraction method to form STI. This feature extraction facilitates

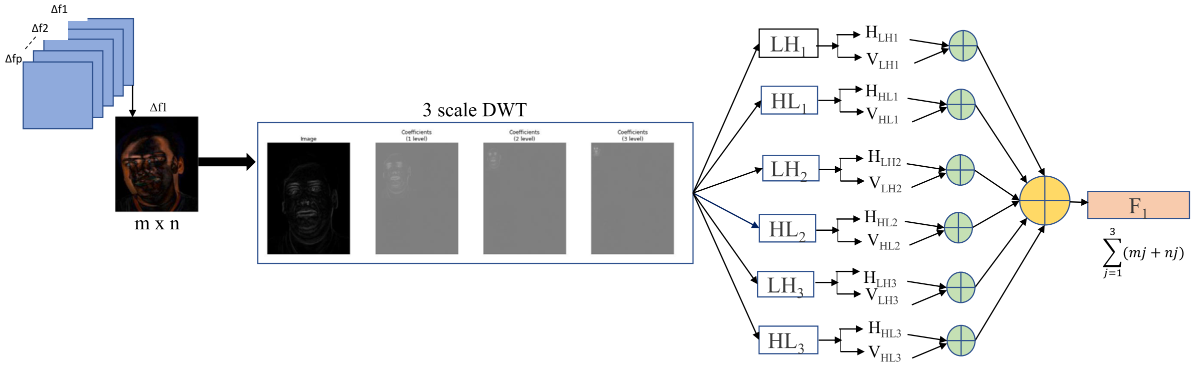

Fig. 3. Block diagram STI generation (HL and LH shows subbands of wavelet and ⊕ sign shows the vector concatenation.).

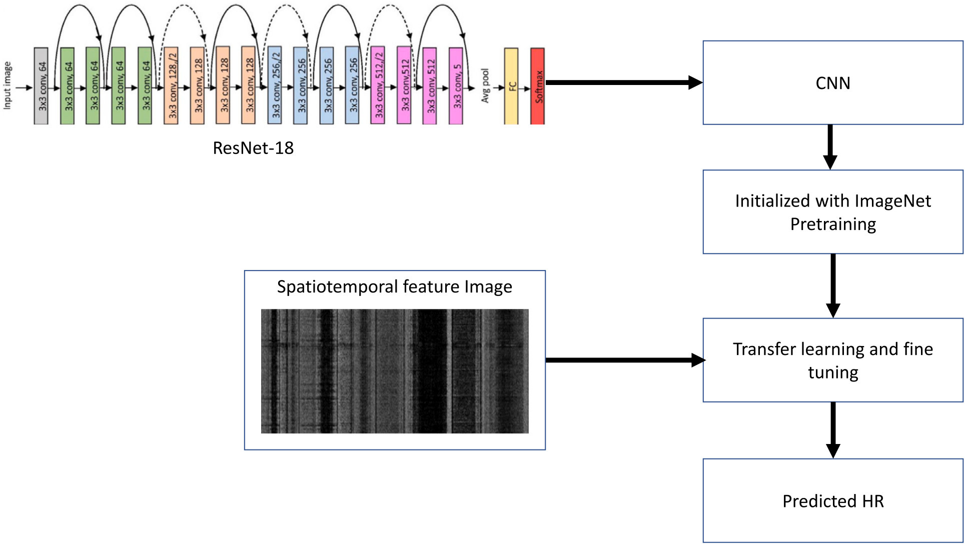

Fig. 4. A diagram of heart rate estimator using a spatiotemporal Image representation and transfer learning.

an efficient representation of input frames to be provided as input to CNN. Here a compressed 2D representation of a video is generated by decreasing spatial data redundancy while maintaining temporal dynamics of video. ROIs of input frames are subtracted consecutively to remove any static properties. Then three-scale discrete Daubechies' wavelet transform is applied to the difference ROI in order to obtain multiple spatial frequency bands. Gradient, texture and edge information can be extracted from coefficients of wavelet transform. The horizontal and vertical projections of wavelet coefficients results in a feature vector representing spatial information. HH subband is left, as there is no spatial correlation. Only HL and LH subbands are used for computing horizontal and vertical projection. Horizontal and vertical projection is computed as:

$$V_{rj}(y) = \sum_{x=0}^{q_j} (r^j(x,y)),\ \ y \in 1\ldots\ldots q_j \tag{2}$$

$$H_{rj}(x) = \sum_{y=0}^{p_j} (r^j(x,y)),\ \ x \in 1\ldots\ldots p_j \tag{3}$$

where, $r \in HL, LH$ and $j \in 1,2,3$ where subbands are represented by r^j in jth scale, $H_{r_j}(x)$ and $V_{r_j}(x)$ shows horizontal and vertical projection of subbands, respectively. p_j represents count of rows and q_j represents column count in subbands in jth scale.

The resultant feature vector is constructed by joining horizontal and vertical projection of subbands in different scales.

$$FV_i = (H + V)_{rj} \tag{4}$$

where i represents frame count, $r \in HL, LH$ and $j \in 1,2,3$. The computed FV_i is of the size $\sum_{j=1}^{s} 2(m_j + n_j)$, where s represents number of scales of Wavelet transform. Each pixel of the FV_i is globally averaged. As a result, a temporal sequence is created that illustrates how skin color changes (see Fig. 3). This spatiotemporal image contains the signal corresponding to the flow of blood, which ultimately exhibits HR value. Since our technique is built on a fully convolutional architecture, potentially face sequences of any size, both spatially and temporally, is viable input. Moreover, the formation of STI using wavelets, provides robustness towards uniform motions. Under extreme noise circumstances, even if there are some occlusions due to low light, tracking using wavelet transform is still feasible. To demonstrate the spatiotemporal image generation more precisely a spatiotemporal images corresponding to six frames of size 780 × 580 is shown in Fig. 6. The feature vector is calculated using formula given in the Eq. (4) and found to be of size 2382 × 19800. The dimensions of the spatiotemporal image depends on the number of frames and the size of feature vector. The size of spatiotemporal image plays a vital role in determining the number of layers of convolutional network, size of kernels and stride.

From spatiotemporal map shown in Fig. 6 it can be clearly seen that the videos with less motion gives same column values and the brightest column value corresponds to the peak value from which HR can be calculated. Such representation makes learning of videos very convenient to CNN. Moreover converting 3D videos to 2D representation also reduces computational complexity to much extent. The brightest column in the STI corresponds to the highest value of HR which can be passed to CNN for learning. The temporal sequence obtained from

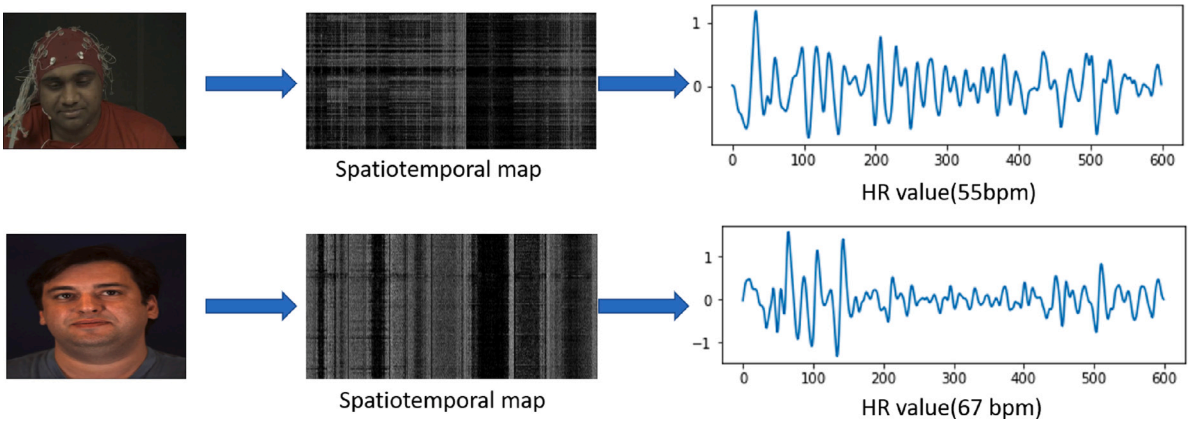

Fig. 5. Example frames from different dataset, corresponding spatiotemporal image and predicted HR value using CNN.

each frame of a video are arranged in form of rows for spatiotemporal representation (see Fig. 5). Conversion of videos into STI diminishes the effect of information loss due to motion such as change in scale, rotation or appearance which eventually reduces noise.

```
Algorithm 1: Feature Extraction
Input  : A raw RGB video with frames [$I_1, I_2, I_3 ...... I_n$], wavelet
         decomposition levels l
Output: Spatiotemporal image $S$
$S = [\ ]$
for each frame $I$ do
    Wavelet decomposition with level $l$
    Consider HL and LH subbands
    Compute vertical projection $H_{r_j}(x)$ using Eq. (2)
    Compute horizontal projection $V_{r_j}(y)$ using Eq. (3)
    Concatenate to form feature vector $FV_i$
end
$S.append(FV_i)$
```

3.3. Convolutional neural network

Feature extraction is performed to extract only the desired information as appropriate features from any given input. To utilize the potential of pre-trained deep neural networks for efficient representation of input, feature extraction is the quickest way. Using a pre-trained model for learning small database is a highly successful strategy. Pre-trained model, trained on original large database provides features, as effective as generic model when trained on small database. (see Fig. 4) It is worth noticing that Resnet18 model has fewer filters and lower complexity than VGG nets. Also the convergence of Resnet18 is faster and occurs at early stage then any other plain net. This motivates us to select Resnet18 [28] for estimation of label (i.e, HR value). ResNet-18 accepts input of size 224, therefore preprocessing of raw image is done to get the desired shape, before applying to the network. The ResNet-18 model then portrays an input hierarchically, with deeper layers contains high-level features. Since we are using ResNet for predicting a single value, therefore the output layer is modified as a fully connected dense layer having only one node with linear activation function. We have used linear activation function here. The model is trained with a mean square error loss function. Adam optimizer [29] is used for optimization. Learning rate of 0.0005 is used for training. Maximum training epoch is 70 for all the database. Same as other CNN regression tasks [30], L_1 loss function is used to minimize the difference between regression output label and ground truth. To avoid overfitting we use dropout layer with dropout ratio of 0.6. Dropout enhances the network generalization capability. The loss function is described in Eq. (5)

$$Loss = \sum_{i=1}^{N} |HR_i - \widetilde{HR_i}| \tag{5}$$

where, HR_i is actual HR value and $\widetilde{HR_i}$ is predicted HR value.

4. Result analysis

4.1. Experimental settings

We have conducted all the experiments on publicly available databases. All the databases MAHNOB-HCI [31], MMSE-HR [32], UBFC-rPPG [33] and VIPL-HR [34] are available online and contains facial videos of a subject with respective physiological signals.

A window of 5 s is taken for processing of videos and their corresponding physiological data. In order to remove the variations in datasets, the ground truth is resampled at 30 Hz. The range of bandpass filter is taken as [0.7–3] Hz. Generated spatiotemporal images are resized to 224 × 224 before feeding to CNN.

4.1.1. Dataset

MAHNOB-HCI tagging database consists of video and corresponding physiological signals (EEG,ECG, respiration amplitude and skin temperature). Total of 527 videos from 15 female and 12 male participants is accumulated. This time duration plays a crucial role as very small duration results in loss of information and large duration videos results in smaller number of samples. Videos of 10 s duration is considered for the formation of spatiotemporal images, results in 527 × 10 = 5270 spatiotemporal images.

VIPL-HR is a large scale database consists of 2378 visible light videos and 752 near infrared videos(NIR) accumulated from 107 subjects. The congruency of VIPL-HR database environment to the real-world scenario benefits us in training CNN for challenging conditions. Taking 10 s video of each subject, total 2378 × 10 = 23780 spatiotemporal maps are generated from visible light videos only.

UBFC-rPPG dataset is specifically created for rPPG analysis. UBFC-rPPG is divided into two i.e, dataset-I(simple) and dataset-II(realistic). We have considered only 42 videos of dataset-II collected under realistic scenario. According to the same time setting of videos 42 × 10 =420 spatiotemporal images are generated.

4.2. Evaluation metrics

To evaluate the accuracy of the rPPG pulse extraction algorithms, we adopt metrices from recently published articles [35–38]. The evaluation metrics are standard deviation (HR_{sd}), the mean absolute HR error (HR_{mae}), the root mean squared HR error (HR_{rmse}), the mean of error rate percentage (HR_{mer}), and Pearson's correlation coefficients ρ.

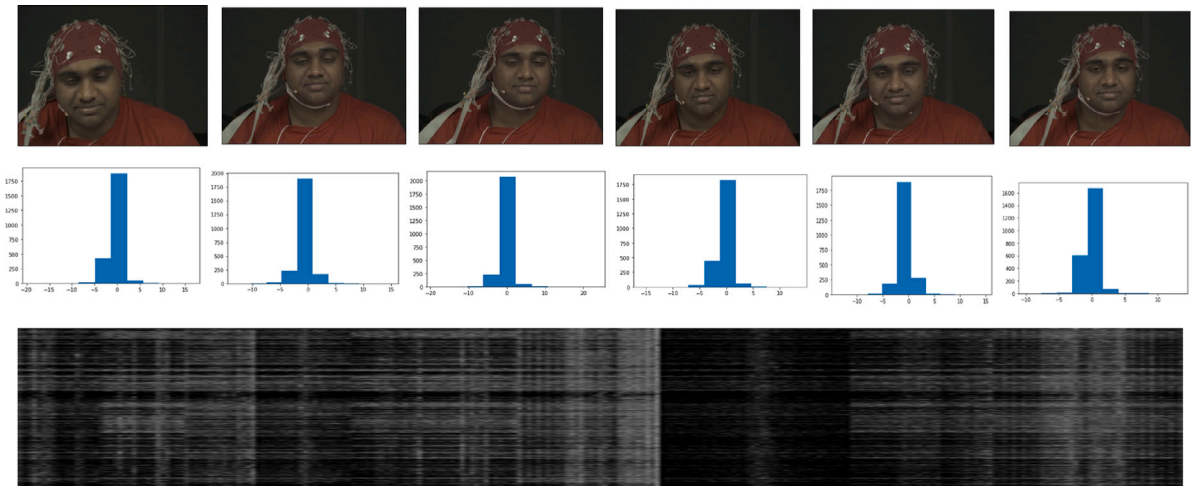

Fig. 6. Six consecutive frames of MAHNOB-HCI dataset with the corresponding wavelet feature vectors and spatiotemporal image.

1. Standard Deviation: Standard deviation quantifies the variation in the HR values.

$$HR_{sd} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(HR_e^i - \overline{HR_e}\right)^2} \tag{6}$$

where $HR_e = HR_{estimated} - HR_{gnd}$

2. Mean absolute error:

$$HR_{\text{mae}} = \frac{1}{n}\sum_{i=1}^{n} HR^i_{estimated} - HR^i_{gnd} \tag{7}$$

3. Root Mean Square Error:

$$HR_{\text{rmse}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(HR^i_{estimated} - HR^i_{gnd})^2} \tag{8}$$

RMSE is difference of squares of predicted and ground truth value. Lower values indicates strong correlation between data.

4. Mean Error Rate Percentage:

It represents accuracy as percentage.

$$HR_{\text{mer}} = \frac{1}{n}\sum_{i=1}^{n}\frac{HR^{(i)}_{estimated} - HR^{(i)}_{gnd}}{HR^{(i)}_{gnd}} \times 100 \tag{9}$$

difference of average of all HR estimated value and ground truth HR value provides mean absolute error.

5. Pearson's Correlation Coefficient:

To measure the correlation between estimated value and ground truth value Pearson correlation is used.

$$\rho = \frac{\sum_{i=1}^{n}\left(X^{(i)} - \overline{X}\right)\left(Y^{(i)} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X^{(i)} - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y^{(i)} - \overline{Y}\right)^2}} \tag{10}$$

where, X represents estimated HR and Y represents true HR.

4.3. Evaluation on MAHNOB-HCI:

Our proposed method is evaluated on widely used MAHNOB-HCI dataset for HR measurement. The video samples are of high compression rate and impulsive motions caused due to change of facial expression based on the movie scenes. MAHNOB-HCI is multimodal dataset, recorded basically for emotion recognition. It is one of the oldest dataset used for HR detection, hence can be profoundly found in existing literature. Table 2 shows the results on MAHNOB-HCI dataset. It shows the comparison of different state-of-the-art methods with or

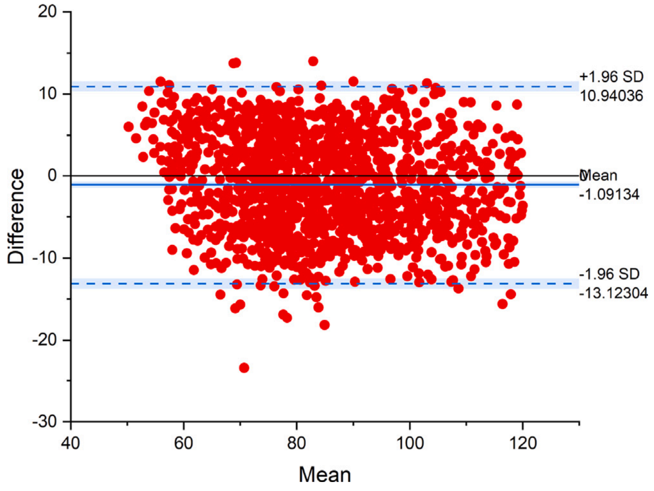

Fig. 7. Bland–Altman plot demonstrating the correlation between ground truth and predicted HR on the MAHNOB-HCI data set; the horizontal lines represent the mean and 95% limits of agreement.

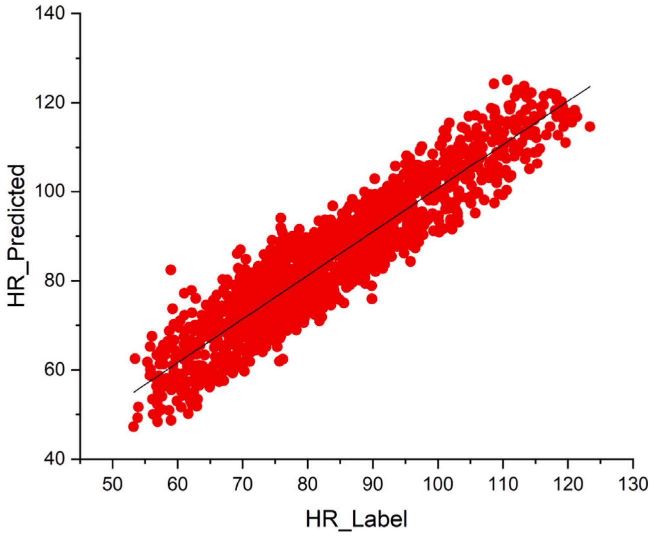

Fig. 8. Scatter plots between ground-truth HR and estimate HR on the MAHNOB-HCI data set.

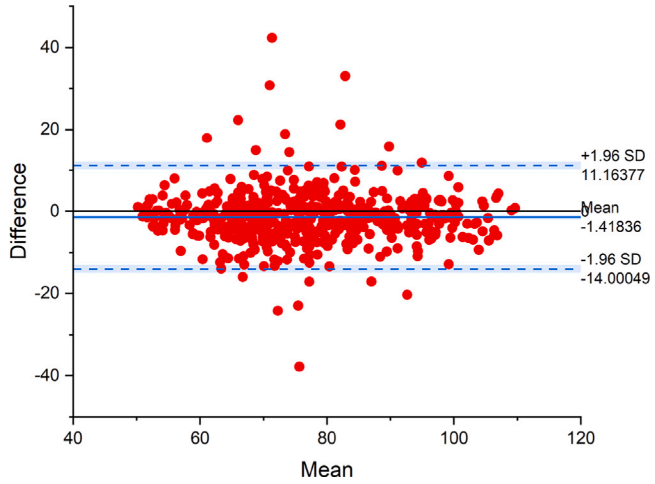

Fig. 9. Bland–Altman plots demonstrating the correlation between ground truth and predicted HR on the UBFC-rPPG dataset; the lines represent the mean and 95% limits of agreement.

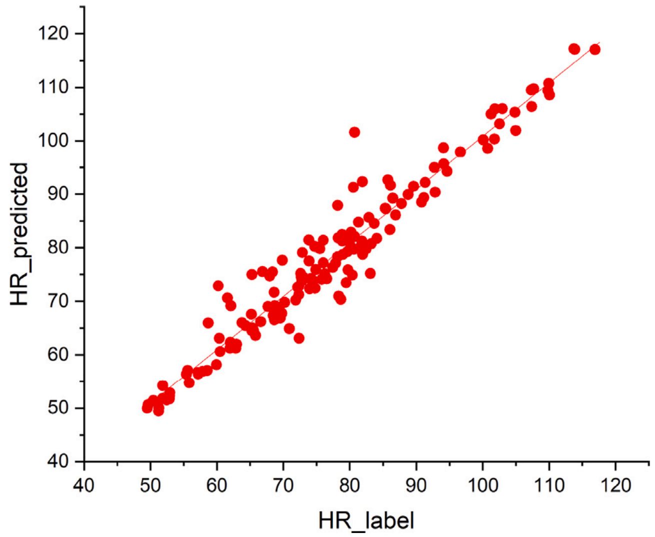

Fig. 10. Scatter plots between ground-truth HR and estimate HR on the UBFC-rPPG data set.

without deep learning. Firstly, the effectiveness of the preprocessing method can be seen from the projected results. In [17], the patch formation over facial region provides robust HR measurement for short duration signal, which results in significantly low errors (RMSE, MAE) compared to the methods utilizing full face region. Although the performance is good in [17] and [39], these methods involves intense preprocessing hence, not suitable for real time estimation of HR. On the other hand methods build over deep learning models provides better performance. The training protocols of all the deep learning based models is different, so the comparison is done at general level. It can be depicted from Table 2, the proposed model exhibit a decrease in MAE by a large margin when compared with the traditional and end-to end learning methods. The proposed method results is approximate to the results of existing end-to-end learning methods.

Table 2
Average HR estimation result on MAHNOB-HCI database using different methods.

Method	SD	MAE	RMSE	ρ
Poh [7]	24.3	25	25.9	0.08
Poh [6]	13.5	13.2	13.6	0.36
CHROM [8]	–	13.49	22.36	0.21
Li [39]	6.58	6.87	7.62	0.81
SAMC [17]	5.81	5.93	6.23	0.83
SynRhythm [40]	4.48	4.37	4.49	–
HR-CNN [11]	–	7.25	9.24	0.51
DeepPhys [18]	–	4.57	–	–
PhysNet [41]	7.8	5.96	7.88	0.76
Meta-rPPG [37]	4.9	3.01	3.68	0.85
EVM-CNN [9]	2.79	3.67	3.26	0.86
Auto-HR [19]	4.73	3.78	5.1	–
Rencheng [24]	5.57	4.61	5.70	0.86
Proposed	3.96	2.72	4.05	0.96

Table 3
Average HR estimation result on VIPL-HR database using different methods.

Method	SD	MAE	RMSE	ρ
SAMC [32]	18	15.9	21.0	0.11
POS [5]	15.3	11.5	17.2	0.30
CHROM [8]	15.1	11.4	16.9	0.28
13D [42]	15.9	12.0	15.9	0.07
DeepPhys [18]	13.6	11.0	13.8	0.11
RhythmNet [20]	8.11	5.30	8.14	0.76
Proposed	7.98	5.23	7.21	0.82

4.4. Evaluation on VIPL-HR:

To validate the effectiveness of the proposed method, we have tested it on VIPL-HR dataset. As shown in Table 3, the performance of all three traditional methods (Tulyakov2016 [32], POS [5] and CHROM [8]) is not on par for VIPL-HR dataset due to complex head motions and variable illumination. Moreover, low values of Pearson's correlation coefficient by end-to-end-learning based methods (e.g., 13D [42] and DeepPhys [18]) indicates non-reliability of predicted HR values. In contrast, our proposed model shows significant improvement and yielded promising results for the challenging dataset. The proposed model provides enhanced spatiotemporal features which improves the predictability of HR. As a result it can be seen in Table 3 that the proposed method achieves promising result with RMSE of 7.21 bpm. The proposed method also outperforms popular deep learning methods like [18,20] with MAE advantages of 5.62 and 0.13 respectively in HR estimation.

4.5. Evaluation on UBFC-rPPGNet:

We also evaluate our model in one of the widely used UBFC-rPPGNet data set. We have adopted the same comparison method as used in other datasets. The UBFC dataset is divided into training and validation sets using the ratio of 6:4 because it is a small dataset and the distribution is quite close, so that we can use all the information to its greatest potential. Till date many researchers have used this dataset for testing algorithms like [5,7,37,43]. Table 4 shows the comparison results using UBFC-rPPG dataset. It is observed from the table that the proposed method yielded quantitatively better results. The proposed method outperforms deep learning method like [43] and [37] with MAE advantage of 1.57 and 2.09 respectively in HR estimation. Results shown in Table 4 implies that proposed method performs best by providing MAE of 3.88%. The improvement in Pearson correlation coefficient indicates the robustness of model towards the small size of dataset with limited HR values range. It is observed that improving information learning through spatiotemporal images, enhances the performance of the model.

4.6. Evaluation on MMSE-HR:

MMSE-HR dataset is widely used, most popular dataset for HR estimation. From Table 5, it can be seen that the proposed methods outperforms most of the existing methods. However, EVM-CNN [9] obtains

Table 4
Average HR estimation result on UBFC-rPPG database using different methods.

Method	SD	MAE	RMSE
Verkruysse [4]	20.2	10.2	20.6
Poh [7]	18.6	8.43	18.8
CHROM [8]	19.1	10.6	20.3
POS [5]	10.4	4.12	10.5
3D-CNN [43]	8.55	5.45	8.64
Meta-rPPG [37]	–	5.97	7.42
Proposed	3.39	3.88	6.23

Table 5
Average HR estimation result on MMSE-HR database using different methods.

Method	SD	MAE	RMSE	ρ
Li [39]	20.2	14.6	19.95	0.38
CHROM [8]	14.08	12.2	13.97	0.55
SAMC [32]	12.24	10.8	11.37	0.71
POS [5]	–	5.77	–	0.82
DeepPhys [18]	–	4.72	8.68	0.82
EVM-CNN [9]	–	–	6.95	0.98
Proposed	6.63	6.4	6.82	0.95

best Pearson correlation coefficient. This result is due to over fitting of the proposed model as well as the fact that there are insufficient training data to cover all ranges of heart rate. Also we have found traditional methods such as POS outperformed some deep learning methods, which is a proof of excellent generalization ability of some traditional methods.

Results shown in Table 5 implies robustness of extracted features of proposed model by delivering lowest RMSE. The proposed model exhibits comparable results with existing deep learning methods, in terms of RMSE and Pearson correlation coefficient. The results imply that the proposed method is effective for not only one specific environmental condition but for most of the complex situations. Despite the varied facial expressions of the subject on the MMSE-HR dataset, our technique can deliver good results. This proves the robustness of the face modeling process used in our model.

4.6.1. Accuracy

The accuracy of the model can be improved only when the ROI selection method, the STI formulation and the backbone network used, works efficiently in collaboration. We have tested our model for within database case i.e, when testing and training are from the same dataset. For 10-fold cross validation, we have randomly selected 40 videos from each datasets and divided into 40 sets, 10 for each dataset. The average accuracy is found to be 87% for MAHNOB-HCI, 82% for MMSE-HR, 78% for UBFC-rPPG and 90% for VIPL-HR respectively.

4.7. Bland Altman plot

The Bland Altman plots comparing predictions on MAHNOB-HCI dataset and UBFC-rPPG dataset by our proposed method are shown in Figs. 7 and 9. This plots helps us to analyze the influence of the spatiotemporal Image formation on the prediction of HR through CNN. The range of estimated HR are within normal limits. Furthermore, from Figs. 7 and 9, it can be seen that the arrangement of HR is analogous and aligned. Analyzing scatter plots from Figs. 8 and 10 reveals that the model and the ground-truth HR holds linear relationship, which is gradually becomes stronger. In one statement, it can be said that the proposed model is more robust.

5. Discussion

The Covid-19 epidemic necessitates the adoption of digital healthcare. Non-contact techniques can lower the risk of infection and also enables patient to stay in a secure environment at home.

Spatiotemporal feature extraction has been predominantly used in [9,12,20] for estimation of HR using deep learning models. Based on the excellent performance gain by these methods, we have also endeavoured to build a model using STI. Improved performance measured in terms of evaluation metrics proves the reliability of the proposed method. In this study a spatiotemporal image is generated using wavelet transform. The motion estimation model using wavelet facilitates estimation of dense optical flow. Motion estimation using wavelet uses full resolution regions without blurring images. While concurrently optimizing the coarser and finer areas of optical flow, small facial movements and large motions can be accurately calculated by using large-to-small full-resolution zones without causing picture blur. However, accuracy of the model degrades when both artifacts i.e, motion and low illumination condition occurs together.

As one can see from the comparison tables that the lowest mean error rate percentage in obtained in MAHNOB-HCI database because our model is robust for low movements only. Whereas the Pearson correlation coefficient is also best for this database only. For rest of the databases like VIPL-HR the MAE is very close to the best. Looking at the advantages of formulating spatiotemporal images we have also formulated one spatiotemporal image which when applied to CNN performs best and shows significant improvement in HR measurement.

The cost effectiveness and simple architecture for STI generation proved advantageous over videos, to be directly presented to CNN. STI also possess resistance towards variation in scales. To experimentally verify this property the proposed method is tested on UBFC-rPPG database. Specifically, it can be said that STI map is the first level extracted features, further convolutional and pooling layers of CNN exhibits higher level features. Finally the features are fed to the fully connected layer for HR values.

Vertical and horizontal projection of subbands produces STI. We have also tried different combinations like diagonal directions, vertical and diagonal and horizontal and diagonal. But none of the combination gives improved results.

We also provide insights for the robust heart rate measurements in realistic situations in future. The proposed method is able to measure HR in more real life situations exposed to uncontrolled environmental conditions. Overall, the proposed spatiotemporal pipeline can be considered as a generalized framework that can be used for other spatiotemporal tasks. For example, the proposed feature extractor model can be adopted for action recognition in videos. Another intriguing possibility for the future is to find ailments that are based on variations in heartbeats over time. Such a study would be very important to foretell cardiac conditions in people who would otherwise seems to be completely normal. Keeping in mind that most immediate and earlier diagnosis of the cardiac disorders is essential to the subject's survival. Such a study will be extremely important because of the advantages and ease of use of the measuring method. Unlike the traditional technologies like the electrocardiograph (ECG).

6. Conclusion

Telehealth and the SARS-CoV-2 pandemic have acutely highlighted the specific need for accurate and computationally efficient cardiovascular and pulmonary sensing. This article describes a neural network model for estimating HR. We constructed 2D STI from 3D videos using wavelet decomposition of each frames and projecting in different subspaces. These vertical and horizontal projections of subbands of wavelet are concatenated to form the feature vector. At the end, for the entire video, feature vectors of each frames are concatenated together to form STI. STI is subjected to CNN for better learning and extracting HR values. To prove the effectiveness, proposed approach is tested on publicly available datasets and comparable results are obtained.

Statements & declarations

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

CRediT authorship contribution statement

Kokila Bharti Jaiswal: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **T. Meenpal:** Writing – review & editing, Supervision.

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.

Data availability

The data that support the findings of this study are available from [31–34]. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors upon reasonable request and with the permission of [31–34].

References

[1] A.C. Smith, E. Thomas, C.L. Snoswell, H. Haydon, A. Mehrotra, J. Clemensen, L.J. Caffery, Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19), J. Telemedicine Telecare 26 (5) (2020) 309–313.

[2] X. Song, X. Liu, C. Wang, The role of telemedicine during the COVID-19 epidemic in China—experience from Shandong province, Crit. Care 24 (1) (2020) 1–4.

[3] Y. Zheng, Y. Ma, J. Zhang, X. Xie, COVID-19 and the cardiovascular system, Nat. Rev. Cardiol. (2020).

[4] W. Verkruysse, L.O. Svaasand, J.S. Nelson, Remote plethysmographic imaging using ambient light, Opt. Express 16 (26) (2008) 21434–21445.

[5] W. Wang, A.C. Den Brinker, S. Stuijk, G. De Haan, Algorithmic principles of remote PPG, IEEE Trans. Biomed. Eng. 64 (7) (2016) 1479–1491.

[6] M.-Z. Poh, D.J. McDuff, R.W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE Trans. Biomed. Eng. 58 (1) (2010) 7–11.

[7] M.-Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Opt. Express 18 (10) (2010) 10762–10774.

[8] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Trans. Biomed. Eng. 60 (10) (2013) 2878–2886.

[9] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A. El Saddik, EVM-CNN: Real-time contactless heart rate estimation from facial video, IEEE Trans. Multimed. 21 (7) (2018) 1778–1787.

[10] M. Hu, F. Qian, D. Guo, X. Wang, L. He, F. Ren, ETA-rPPGNet: Effective time-domain attention network for remote heart rate measurement, IEEE Trans. Instrum. Meas. 70 (2021) 1–12.

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

[12] X. Niu, H. Han, S. Shan, X. Chen, SynRhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th International Conference on Pattern Recognition, ICPR, 2018, pp. 3580–3585, http://dx.doi.org/10.1109/ICPR.2018.8546321.

[13] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, X. Chen, Robust remote heart rate estimation from face utilizing spatial-temporal attention, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, 2019, pp. 1–8, http://dx.doi.org/10.1109/FG.2019.8756554.

[14] J. Cheng, X. Chen, L. Xu, Z.J. Wang, Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition, IEEE J. Biomed. Health Inf. 21 (5) (2016) 1422–1433.

[15] G. De Haan, A. Van Leest, Improved motion robustness of remote-PPG by using the blood volume pulse signature, Physiol. Meas. 35 (9) (2014) 1913.

[16] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE Trans. Biomed. Eng. 63 (9) (2015) 1974–1984.

[17] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J.F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2396–2404.

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

[19] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Process. Lett. 27 (2020) 1245–1249.

[20] 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 (2019) 2409–2423.

[21] B. Lokendra, G. Puneet, AND-rPPG: A novel denoising-rPPG network for improving remote heart rate estimation, Comput. Biol. Med. 141 (2022) 105146.

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

[23] G.-S. Hsu, A. Ambikapathi, M.-S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in: 2017 IEEE International Joint Conference on Biometrics, IJCB, 2017, pp. 383–389, http://dx.doi.org/10.1109/BTAS.2017.8272721.

[24] 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. Meas. (2020).

[25] C. Wu, Z. Yuan, S. Wan, L. Wang, W. Zhang, Anti-jamming heart rate estimation using a spatial–temporal fusion network, Comput. Vis. Image Underst. 216 (2022) 103327.

[26] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3444–3451.

[27] J. Shi, et al., Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 1994, pp. 593–600.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.

[29] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, arXiv:1412.6980.

[30] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: A brief review, Comput. Intell. Neurosci. 2018 (2018).

[31] J. Lichtenauer, M. Soleymani, Mahnob-Hci-Tagging Database, London, 2011.

[32] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J.F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2396–2404.

[33] R. Macwan, Y. Benezeth, A. Mansouri, Heart rate estimation using remote photoplethysmography with multi-objective optimization, Biomed. Signal Process. Control 49 (2019) 24–33.

[34] X. Niu, H. Han, S. Shan, X. Chen, VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video, in: Asian Conference on Computer Vision, Springer, 2018, pp. 562–576.

[35] K. Zheng, K. Ci, H. Li, L. Shao, G. Sun, J. Liu, J. Cui, Heart rate prediction from facial video with masks using eye location and corrected by convolutional neural networks, Biomed. Signal Process. Control 75 (2022) 103609.

[36] Z. Qiu, J. Liu, H. Sun, L. Lin, Y.-W. Chen, CoSTHR: A heart rate estimating network with adaptive color space transformation, IEEE Trans. Instrum. Meas. 71 (2022) 1–10.

[37] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: European Conference on Computer Vision, Springer, 2020, pp. 392–409.

[38] G.-S. Hsu, A. Ambikapathi, M.-S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in: 2017 IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2017, pp. 383–389.

[39] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4264–4271.

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

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

[42] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[43] F. Bousefsaf, A. Pruski, C. Maaoui, 3D convolutional neural networks for remote pulse rate measurement and mapping from facial video, Appl. Sci. 9 (20) (2019) 4364.