Multi-hierarchical Convolutional Network for Efficient Remote Photoplethysmograph Signal and Heart Rate Estimation from Face Video Clips

Panpan ZHANG, Bin LI, Jinye Peng, Wei JIANG

1 School of Information Science and Technology, Northwest University, Xi’an, People’s Republic of China.

zhangpanpan1k@163.com(P.Z.); lib@nwu.edu.cn(B.L.); pjy@nwu.edu.cn(J.P.); 2402582732@qq.com(W.J.)

* Corresponding Author, E-mail: lib@nwu.edu.cn

Abstract: Heart beat rhythm and heart rate (HR) are important physiological parameters of the human body. This study presents an efficient multi-hierarchical spatio-temporal convolutional network that can quickly estimate remote physiological (rPPG) signal and HR from face video clips. First, the facial color distribution characteristics are extracted using a low-level face feature Generation (LFFG) module. Then, the three-dimensional (3D) spatio-temporal stack convolution module (STSC) and multi-hierarchical feature fusion module (MHFF) are used to strengthen the spatio-temporal correlation of multi-channel features. In the MHFF, sparse optical flow is used to capture the tiny motion information of faces between frames and generate a self-adaptive region of interest (ROI) skin mask. Finally, the signal prediction module (SP) is used to extract the estimated rPPG signal. The experimental results on the three datasets show that the proposed network outperforms the state-of-the-art methods.

Keywords: Remote heart rate estimation; rPPG; multi-hierarchical feature fusion; spatio-temporal convolution.

1. Introduction
Physiological signals caused by the periodic activity of the heart contain a variety of important vital signs, such as heart rate (HR), blood oxygen, heart rate variability (HRV), which are considered to be important indicators of human health and considerably significant in various fields such as medical treatment, sports, and psychological testing. Two common methods of measuring these indicators are electrocardiogram (ECG) and photoplethysmography (PPG). The ECG measures the electrical activations that lead to the contraction of the heart muscle, using electrodes attached to the body, usually at the chest, it is widely used in the medical field. PPG uses a small optical sensor combined with a light source to measure changes in light absorption by blood vessels caused by the cardiac cycle. This method is usually used for wearable monitoring devices such as sports bracelets. However, both methods are based on contact measurement, which is inconvenient to users and unfriendly to people with extremely sensitive skin such as newborns or burn patients. Thus, many researchers have focused on developing non-contact measurement methods.

Non-contact physiological signal estimation approaches are mainly divided into two categories: (1) cardiography [1] based on periodic head movement, which aims to extract heartbeat information from the subtle head movement caused by the heart periodically injecting blood into the large blood vessels, and (2) remote photoplethysmography (rPPG), which aims to obtain heartbeat information by analyzing periodic changes in the skin’s absorption of light caused by changes in blood volume and oxygen saturation in the blood vessels. The latter is more robust to changes in head movement and illumination. With the development of imaging technology, commercial cameras have met the video capture requirements for remote physiological signal measurement. Early research on HR measurement based on periodic skin changes [2-6] has achieved good results under normal ambient light conditions, which can be summarized as follows: (1) detection of regions of interest (ROIs) by
face detection with a feature tracker, and calculation of raw signals by averaging the respective color channel of all pixels contained in the ROI of the frame; (2) acquisition of the plethysmographic signal using signal analysis methods such as filtering and dimensionality reduction; (3) estimation the average HR using frequency analysis and peak location. However, this change in light absorption by skin tissue due to blood volume is too weak to be observed by the eyes; it is also easily affected by head movement and light changes. Traditional feature extraction and filtering processes may lose important information and are unfriendly to data. In the era of big data, some data-driven methods such as deep learning have shown powerful modeling capabilities and have attracted increasing attention. Deep learning-based remote HR estimation has also achieved good results [7-12, 20-23, 27, 31, 32, 34]. However, all the aforementioned methods have one or the other of the following shortcomings: (1) design based on a two-dimensional (2D) spatial neural network, neglecting time information in video-based remote HR estimation; (2) not an end-to-end method, as the input is not the original facial video; (3) ROI detection is unstable when there is a large range of head movements in the video sequence; (4) longer measurement time, generally requiring 30 s of facial video to calculate the average HR.[9, 20, 22, 47]

In this study, we propose an end-to-end three-dimension (3D) spatio-temporal convolutional network to efficient recover the rPPG signal and average HR value from the original video clips. We add a multi-hierarchical feature fusion (MHFF) based attention module to make the network to focus on rPPG-related features and reduce the impact of head movement and background noise on signal recovery.

The main contributions of our work include the followings:

1. An efficient end-to-end 3D spatio-temporal convolutional network with an MHFF based
attention model is proposed. The skin map label generated based on sparse optical flow effectively solves the influence of background noise and head movement.

2. Only 15 s face video clips are needed for efficient reconstruction of rPPG signal and accurate estimation of HR.

3. The experiments are conducted on three datasets to verify the effectiveness of the proposed network.

4. A new face video physiological parameters dataset with annotated PPG and HR signal is presented, which contains 300 videos from 300 subjects.

The remainder of this paper is organized as follows. Section 2 reviews video-based remote physiological measurement algorithms in recent years. In Section 3, the proposed method is described in detail. The experimental results and discussion are presented in Section 4. Finally, Section 5 summarizes the study.

2. Related work

Previous methods for rPPG signal measurement and spatio-temporal network frameworks are briefly reviewed in the following two subsections.

2.1 Remote Photoplethysmograph Signal Measurement

The mainstream rPPG-based remote HR measurement methods can be divided into two types: the traditional method based on signal analysis and the data-driven method based on deep learning.

Verkruysse et al. [6] proposed an early rPPG-based remote physiological signal measurement method that requires the objects to remain stationary under ambient light. In these recordings, the ROIs were manually selected through the bounding box. From the pixels contained in the ROIs, the average value of the red-green-blue (RGB) color channels was calculated per frame as the original
signal, and a band-pass filter was used to remove the noise. It was found that the green channel contains the strongest PPG signal; this was used in the initial work [2]. To improve the manual bounding box, Poh et al. [14] used a face detector to track the object's face frame by frame, and used a 30s moving window and a bounding box containing the face as the ROI to achieve continuous measurement. Weighting and merging the color space produced better results than the green single-channel signal. CHROM [4] uses a standardized skin color configuration to perform white balancing operations on video frames and obtain a linear combination of chromaticity signals. Poh, McDuff, and Picard [14, 15] introduced the blind source separation (BSS) method to estimate the rPPG signal as a linear combination of three channels, this improves the signal-to-noise ratio (SNR) and achieves the purpose of dimensionality reduction. They used independent component analysis (ICA) to identify the major components related to HR. In contrast to the ICA method applied to the entire ROI, Lam and Kuno [13] applied ICA at the patch level for HR estimation. In addition to ICA, principal component analysis (PCA) [16] and constrain ICA (CICA) [29] are also used to estimate the HR signal. Lan et al. [19] compared various linear and nonlinear techniques of BSS and found that Laplacian feature mapping could produce the best results. After the main components of the HR were obtained, fast Fourier transformation (FFT) was needed to obtain the spectrum diagram; the frequency with the highest response is estimated to be the corresponding HR frequency. To reduce the noise spectrum contained in the FFT, Kumar et al. [30] combined HR signals with different ROIs using frequency features as weights. All these methods require the assumption of a feasible frequency band for human HR. The common choice of frequency band is [0.7 Hz, 4 Hz], which corresponds to an HR of 42 to 240 beats per minute (bpm) [33]. In addition, there are also some methods based on reflection models, such as converting RBG channels into other color spaces
to obtain more robust rPPG signals. However, these methods rely on prior knowledge of researchers, and their performance largely depends on the effect of manual data preprocessing and whether the environment meets the hypothesis conditions. Complex constraint conditions and weak generalization ability hinder the achievement of good results in a non-constrained environment with large noise such as ambient light change and head movement, by traditional methods; therefore, these methods have certain limitations in practical applications.

In the era of big data, data-driven learning-based methods have rapidly developed and provided better results in complex environments.

Tulyakov et al. [23] divided the face into multiple ROI regions to obtain a time representation matrix, and used matrix completion to purify HR-related signals. Wang et al. [31] proposed an object-related method that uses spatial subspace rotation to compute HR signals by providing a complete continuous sequence of the object's face. Hsu et al. [32] combined the frequency domain features of RGB channels and ICA components to obtain the representation of HR signals, and used support vector regression to estimate HR. Base on this work, they proposed an improved method with better effect [10]. After extracting rPPG signals using the traditional CHROM method [4], they used short-time Fourier transform to construct 2D time-frequency representation (TFR) sequences, and 2D TFR was used as the input of the VGG-16 model and the majority voting based network to achieve HR estimation through. To solve the problem of insufficient training data, Niu et al. [7] proposed a transfer learning method called Synrhythm, which first uses the synthesized HR signal to pre-train the model, and then transfers the pre-trained model to the real HR estimation task. As head movement had a higher effect on brightness than on chrominance, in a later study, Rhythmnet [8] converted the video sequence from an RGB channel to a YUV channel and computed a
spatiotemporal map from ROIs as a representation of the HR signal. It then used a convolutional neural network (CNN)–recurrent neural networks (RNN) model incorporating a gating unit (GRU) to learn the mapping from the spatiotemporal map to HR. However, these methods still require a strict preprocessing stage, and are based on preset ROIs for signal extraction while neglecting other facial areas. End-to-end deep learning methods have also been proposed. Spetlik, Cech, and Vojtěch [12] proposed an HR-CNN network that includes the extractor and HR estimator. The extractor is a 2D CNN with frequency constraints that extracts non-contact reflection PPG (NrPPG) signals from facial images. The HR estimator estimates the HR by calculating different loss functions of the NrPPG signals. Considering the influence of head movement on rPPG signal extraction, DeepPhys [9] extracted the relevant physiological signals from a skin reflex-based motion representation model and a CAN-based appearance model, according to the different frequency characteristics of different signals to extract the corresponding physiological signals from a 30s facial video. Tang et al. [11] performed skin pixel segmentation through CNN and extracted the HR signals of the detected skin areas through PCA and ICA signal analysis. Such skin-based signal extraction is conducive to continuous HR measurement in challenging environments, such as neonatal intensive care units (NICUs). Tsou et al. [21] proposed the Siamese-rPPG Network, which is a weight-sharing mechanism to extract rPPG signals from the cheeks and forehead respectively, it added the two branch signals to obtain face rPPG signals to avoid signal loss when one ROI is occluded. To effectively restrain noise features while maintaining the integrity of periodic physiological features, Niu et al [22] processed the video sequence into MSTmap as the network input, and proposed a strategy of cross-verified feature disentangling (CVD) to extract physiological signals from interference signals through a 30 s facial video. In addition to data noise, such as environmental
factors, the network structure is also one of the main factors affecting the prediction results. AutoHR [27] composed of a searched backbone and time difference convolution, achieved remote HR measurement by finding an appropriate network backbone through neural architecture search. Although 2D CNN achieved a good effect in spatial feature extraction, it neglects the interframe motion information of the time dimension. Some works [4,25] first performed spatial pooling of RGB sequences and then projected them into a more representative color space. Thus, temporal context-based normalization was needed to avoid the effects of noise such as movement and environmental changes. The use of 3D CNN-based network can integrate these steps and regard the time dimension as the third dimension. Yu et al. [34] proposed a 3D network for regressing rPPG signals from facial videos, however this method did not consider the influence of compressed video on signal recovery. Yu et al. [20] proposed a two-stage network containing STVEN for video enhancement and rPPGNet for rPPG signal recovery, and accurately recovered rPPG signals from a 30 s highly compressed facial video.

2.2 Spatio-temporal Network Frameworks

With the continuous development of machine vision, traditional 2D networks can no longer satisfy many video-based learning tasks (such as action recognition and video subtitles) because 2D CNN cannot capture the time context information of video media. Researchers have focused on the spatiotemporal network structure which can capture both temporal and spatial information and extract the local spatial features of the image while maintaining the time continuity of the video. At present, mainstream spatiotemporal network frameworks are divided into two categories. The first is a spatio-temporal network framework that combines a CNN and an RNN, for example, long-term short-term memory (LSTM) [45] and convolutional LSTM [46], CNN feed-forward network is used
for spatial modeling, and LSTM is used for temporal modeling. However, this structure causes difficulties in end-to-end optimization. The second category is the 3D CNN-based spatiotemporal network framework. Du et al. [41] first proposed a 3D convolutional network (C3D) that could simultaneously perform temporal and spatial modeling. In the subsequent development, various 3D CNN-based spatiotemporal network frameworks such as pseudo 3D [42], inflated 3D [43] and separable 3D [44] have been derived, which are widely used in video understanding and are more suitable for video-based remote physiological signal estimation depending on small facial skin tone changes.

**Fig. 1** The architecture of proposed method. The top is the overall structure of the proposed network. The bottom is the structure of MHFF that include a C_feature extractor and a skin generator.
In contrast to previous works, considering the comfort of the test subjects in practical applications, we further accelerated the signal reconstruction speed on the basis of ensuring the prediction accuracy. We rapidly and accurately reconstructed the physiological signals through the short-time facial video, and calculated the 15 s average HR. Feature maps at different levels focus on different information: low-level feature maps have rich spatial information, whereas high-level feature maps focus on semantic information. To consider both spatial and semantic information, we propose a multi-hierarchical convolutional network that integrates multi-level feature maps.

3. Methodology

The proposed method aims to reconstruct rPPG from RGB facial videos by designing a 3D spatio-temporal convolutional network with multi-hierarchical fusion. As shown in Fig. (1), the proposed network includes four modules: low-level face feature generation (LFFG), 3D spatio-temporal stack convolution (STSC), MHFF, and signal predictor (SP).

3.1. Data Preparation

As the variation range of facial skin’s light absorption is very weak, some highly reflective backgrounds may affect the predicted results, so we used [39] cascading the ensemble of regression trees (ERT) algorithm [37] to approximately crop the redundant background of the video.

\[ CV_i = \text{crop}(V_i) \quad (1) \]

where \( V_i \) is the original video frames, \( CV_i (t=1,2,...,T) \) is the coarse-cropped video frames. Then normalized video frames to \( W \times H \) and \( T \) frames are aggregated into a video stream as the network input \( CV_{t,t} \in \mathbb{R}^{C \times T \times W \times H} \), \( C \) is the number of channels (\( C=3 \) means three R,G,B channels),
T is the number of input frames, and W and H are the spatial sizes of input frames.

3.2. Low-level Face Feature Generation

Low-level feature maps have rich spatial information that is beneficial for obtaining the facial color distribution in the image. Therefore, we propose a three-layer convolutional network model, LFFG, for superficial facial feature extraction. The input of LFFG is an RGB face video clip $C_{V_{t:t+T}} \in \mathbb{R}^{C \times T \times W \times H}$ with T frames. The output of the LFFG is as follows:

$$f_i = \text{conv}_{2:3} \left( \text{Max} \left( \text{conv}_1 \left( C_{V_{t:t+T}} \right) \right) \right)$$

(2)

where $\text{conv}_1(\cdot)$ is the spatial convolution with kernel size of $1 \times 5 \times 5$, $\text{Max}(\cdot)$ is the spatial max pool with $1 \times 2 \times 2$ kernel, $\text{conv}_{2:3}(\cdot)$ is the 3D spatio-temporal convolution with kernel size of $3 \times 3 \times 3$. After each convolution layer, there is a batch normalization and rectified linear unit (ReLU) activation layer. $f_i \in \mathbb{R}^{3T \times W \times H}$ are the low-level feature maps of facial video clips.

3.3. Spatio-temporal Stack Convolution)

Temporal feature acquisition and multi- hierarchical spatiotemporal feature fusion are very important for rPPG signal recovery from facial video. Rich shallow spatial features obtained by LFFG are input into STSC module for deeper feature extraction. STSC is composed of two serial convolutional blocks with similar structures, as shown in Fig. (2).

$$f_h = STC_i(f_i)$$

(3)

where $STC_i(\cdot) (i = 1, 2)$ is a convolutional block. Each convolution block consists of two 3D spatiotemporal convolution layers and a down sample layer. Kernel size of 3D spatio-temporal convolution layers are $3 \times 3 \times 3$. After each convolution layer, there is batch normalization and ReLU activation layer. To extract the temporal context information more effectively and reduce the redundancy and noise of the time dimension, the down-sample layer subsamples the time
dimension and the space dimension through the kernel size of $2 \times 2 \times 2$ and 2 steps. The output of the STSC is a high-level feature map $f_n \in \mathbb{R}^{64 \times T \times W \times H}$.

Fig. 2 The architecture of STSC. Includes two convolution blocks of similar structure.

In 3D spatio-temporal convolution, the value of position $p$ $y^p$ in each feature map is obtained by convolving a stack of local receptive fields $x^p$ of $n$ successive frames to capture the interframe variation information of the video. As shown in Fig. (3), we set $n = 3$.

$$y^p = \sum_{t=1}^{n} conv(x^p_t)$$

Fig. 3 3D Spatio-temporal Convolution

### 3.4. Multi-hierarchical Feature Fusion

In order to retain more effective information and make the network focus on feature extraction related to rPPG signal, we proposed the MHFF structure to generate and fuse To retain more
effective information and make the network focus on feature extraction related to the rPPG signal, we propose the MHFF structure to generate and fuse multi-hierarchical features, as shown in Fig. (1(b)). MHFF includes a parallel C-feature extractor and a skin map generator based on the residual structure.

**C_feature extractor:** Because different R, G, and B channels carry different physiological signals, to find more stable channels for rPPG signal reconstruction and HR estimation, we send a high-level feature map $f_h$ generated by STSC to the C_feature extractor for channel-wise feature extraction.

$$f_c = \partial \left( \xi \left( \text{Max} \left( f_h \right) \right) + \xi \left( \text{Avg} \left( f_h \right) \right) \right)$$

(5)

where $\xi(\cdot)$ includes two 2D convolution layers with $1 \times 1$ kernel and a ReLU activation layer, $\text{Max}(\cdot)$ is the max pool with an output size of 1, $\text{Avg}(\cdot)$ is the average pool with an output size of 1, $\partial(\cdot)$ includes a spatial average pool and a sigmoid classifier. The output of the C_feature extractor is a channel-wise feature map $f_c \in \mathbb{R}^{T \times W \times H}$.

**Skin Map Generator:** To further emphasize the spatial ROI of the feature map, a skin map generator is proposed to perform spatial-wise feature extraction on the low-level feature map $f_i$ that includes rich spatial information to obtain the skin map $f_s$:

$$f_s = \sigma \left( \text{conv} \left( \phi \left( f_i \right) \right) \right)$$

(6)

where $\phi(\cdot)$ is a convolutional block with a residual structure. This includes two backbone convolution layers with $1 \times 3 \times 3$ kernels and a residual convolution with a $1 \times 1 \times 1$ kernel. $\text{conv}(\cdot)$ is a spatial convolution with kernel of $1 \times 3 \times 3$, $\sigma(\cdot)$ includes a sigmoid classifier and two down-sample layers of $2 \times 2 \times 2$ kernels with two steps. The output of the skin map generator is a skin map $f_s \in \mathbb{R}^{T \times W \times H}$.
Then, the C-feature extractor generates a weight mask by feature fusion of the low-level skin map $f_S$ and high-level channel-wise feature map $f_C$:

$$M_w = \psi(f_C + f_S)$$  \hspace{1cm} (7)

where $\psi(\bullet)$ is a softmax classifier and the weight mask is $M_w \in \mathbb{R}^{T \times W \times H}$. Finally, the high-level feature map $f_h$ is multiplied by the weight mask $M_w$ by channels.

$$f_{roi}^{i} = f_{h}^{i} \ast M_w$$ \hspace{1cm} (8)

where $f_{h}^{i} \in \mathbb{R}^{T \times W \times H}$ ($i = 1, 2, ..., 64$) is the high-level feature map of the $i$-th channel, $f_{roi}^{i}$ is the hybrid feature map of $i$-th channel, the final output hybrid feature map of the MHFF is $f_{roi} \in \mathbb{R}^{64 \times T \times W \times H}$.

3.5. Signal Predictor

To reduce the redundancy of the time dimension, we inserted two subsamples in the feature extraction stage. Before feature aggregation, two consecutive up-samplings of the time dimension on the feature map are performed to restore the original time length.

$$f_{roi}^{up} = \rho(f_{roi})$$ \hspace{1cm} (9)

where $\rho(\bullet)$ is a transposed convolution with a kernel size of $4 \times 1 \times 1$. The stride is set as $2 \times 1 \times 1$ and the result of up-sampling is $f_{roi}^{up} \in \mathbb{R}^{64k \times T \times W \times H}$.

After multiple convolution operations, a multichannel spatio-temporal feature representation stream is formed. Spatial global average pooling is performed on $f_{roi}^{up}$ to aggregate the features of the spatial dimension and introduce an independent channel-wise convolution filter to project the multichannel spatio-temporal representation stream into the rPPG signal space.

$$rPPG = conv_c\left(\text{Avg}_G\left(f_{roi}^{up}\right)\right)$$ \hspace{1cm} (10)

where $conv_c(\bullet)$ is the channel-wise convolution with a $1 \times 1 \times 1$ kernel, $\text{Avg}_G(\bullet)$ is a spatial
global average pooling with an output size of $T \times 1 \times 1$. The output of the SP is an rPPG signal.

The rPPG signal contains a variety of physiological information. HR is a powerful predictor of the most common major medical events and is widely used in clinical medicine and daily health monitoring. The rPPG-based HR estimation is mainly divided into time-domain and frequency-domain methods. In the case of a short video sequence, we choose the frequency-domain method with higher accuracy. The frequency spectrum corresponding to the HR can be obtained through a frequency analysis of the rPPG signal. The HR can be obtained using the following formula:

$$HR = 60 \omega_{hr}$$

(11)

where $\omega_{hr}$ is the frequency corresponding to the index with the highest spectral power.

3.6. Loss Function

To accurately recover the rPPG signal curve, we calculated both the time domain and frequency domain losses. Because both the ground truth PPG signal (from devices on fingers) and the predicted rPPG signal (from facial video) measure the blood volume changes using the optical method, the PPG and rPPG signals are similar. Because the ranges of PPG and rPPG signal values are not the same, compared to commonly used point-wise intensity errors such as mean square error, the Pearson product-moment correlation coefficient (Pearson correlation) can better guide the network to fit the curve tendency to accurately locate the peak position. The Pearson correlation ranges between $-1$ and 1. When the linear relationship between two variables is positively correlated, the correlation coefficient approaches 1, whereas it tends to $-1$ when the linear relationship between two variables is negatively correlated. If the correlation coefficient is equal to 0, there is no linear correlation between the variables. To make the predicted rPPG signal curve more similar to the
ground truth, we used a negative Pearson correlation to calculate the loss in the time domain:

\[
L_i = 1 - \frac{T \sum_{i=1}^{T} x_i y_i - \sum_{i=1}^{T} x_i \sum_{i=1}^{T} y_i}{\sqrt{\sum_{i=1}^{T} x_i^2} \sqrt{\sum_{i=1}^{T} y_i^2} - \left( \sum_{i=1}^{T} x_i \right)^2}
\]  

(12)

where T is the length of video clips, \( x_i \) is the \( i \)-th predicted rPPG signal, \( y_i \) is the \( i \)-th ground truth PPG signal.

Inspired by SNR loss, we treated HR estimation as a classification task through frequency transformation and introduced cross entropy loss to calculate the loss in the frequency domain:

\[
L_f = CE(PSD(X), HR_{gt})
\]

(13)

where \( PSD(X) \) represents the power spectral density of the predicted rPPG signal, \( HR_{gt} \) is the average HR value of the ground truth, and \( L_f = CE(X, Y) \) is used to calculate the cross-entropy loss between the predicted value and the ground truth.

To avoid the influence of non-face areas, we also introduced the binary cross-entropy loss of the skin map \( f_s \) and binary skin label \( S_l \):

\[
L_s = BCE(f_s, S_l)
\]

(14)

where \( BCE(X, Y) \) calculates the binary cross entropy loss of the predicted value X and the ground truth Y.

When obtaining binary skin labels, we introduced an optical flow field to obtain a stable ROI. First, we obtained the landmarks and ROI in \( Cv_1 \) (the first frame) and used sparse optical flow [28] to track the trajectories of 29 peripheral facial landmarks \( l = \{l_0, \ldots, l_{28}\} \). We constructed the L-layer face image pyramid as shown in Fig. (4).
In Fig. (4), $Cv_{t-1}$ and $Cv_t'$ represent $l$-layers of the frames $t-1$ and $t$, respectively; $l = (0, 1, ..., L)$ is the number of pyramid layers; when $l = 0$, it is the original image; $u' = [u'_x, u'_y](i = 0, ..., 28)$ is the $i$-th face coordinate that needs to be tracked in $Cv_{t-1}$; and $u'' = [u''_x, u''_y] = \frac{u'}{2^L}$ is the coordinate of the corresponding point on the $L$-layer $g^0$ in $Cv_{t-1}$. The optical flow of the highest layer $L$ to $g^L = [0, 0]^T$ is initialized. Then the optimal value of the $L$ layer is iteratively calculated, providing the initial value of the optical flow for the $L - 1$ layer $g^{L-1} = 2 \times [g^L + d^L]$ according to [28]. In sequence, successive iterations are performed from $L$ to 0 layers to obtain the optical flow $g^0$ and the optimized value $d^0$ of the lowest layer $Cv_0$. Using the optical flow vector $d = g^0 + d^0$, the coordinate point $u'$ of the previous frame $Cv_{t-1}$ in the current frame $Cv_t$ is obtained as $v' = u' + d$.

There is a relative movement between the target and the background when the face is offset. The effective face ROI can be updated through sparse optical flow and can also reduce the background noise and capture the tiny inter-frame motion information. The skin labels were obtained to assist the skin map generator to generate a high-quality skin map, as shown in Fig. (5). The algorithm flow is listed in Table 1.
Fig. 5 (a) 81 landmarks of RGB facial frames. (b) Binary skin label

Table 1 Framework of Optical Flow based Skin Label

Algorithm 1 Framework of Optical Flow based Skin Label

Input: RGB video frames $C_v[C_{v_0}, C_{v_1}, ..., C_{v_T}]$

Output: skin label $S_l$

1: 81 landmarks detection in the first frame and select 29 landmarks $l = \{l_0, ..., l_i, ..., l_{28}\}$ for tracking

2: for $t = 1$ to $T$ do

3: for $i = 0$ to $28$ do

4: \[ v^i = u^i + d \]

5: \[ l[i] = v^i \] // tracking landmarks with sparse optical flow

6: end for

7: for $k = 1$ to $W \times H$ do:

8: if $k$ in $l$ :

9: \[ p[k] = 255 \]

10: else $p[k] = 0$ // binaryzation

11: end if

12: end for

13: return $S_l$
The final overall loss is calculated as:

$$\text{Loss} = \alpha L_r + \beta L_f$$

(15)

where $\alpha$ and $\beta$ are the weight coefficients to balance the loss.

4. Experiment results

We trained and tested proposed method on our dataset, and the ratio of training set and testing set is 4:1. And cross-tested our method on the UBFC-RPPG [23] dataset and the COHFACE [26] dataset.

4.1 Datasets

Our dataset contained a total of 300 VIS videos with a frame rate of 30 fps from 300 objects at the age of 18-26 years. The length of each video was 1 min with pixel resolution of 1920×1080. These videos were collected by an Honor v30 mobile phone in a well-lit environment. Physiological signals were collected by BIOPAC MP160, including the average HR, respiratory rate, SpO\textsuperscript{2}, ECG signal, and blood volume pulse (BVP) wave of each subject. The physiological signal sampling rate was 1000 Hz.

The UBFC-RPPG database [23] contained 42 videos from 42 subjects. The videos were recorded using a simple low-cost webcam (Logitech C920 HD Pro) at 30 fps with a resolution of 640×480 pixels in an uncompressed 8-bit RGB format. A CMS5OE transmissive pulse oximeter was used to obtain the ground truth PPG waveform and PPG HRs. During the recording, the subject sat in front of the camera (approximately 1 m away from the camera) with his/her face visible. All experiments were conducted indoors with varying amounts of sunlight and indoor illumination.

The COHFACE dataset [26] contained 160 videos with high compression rates from 40 subjects (12 women and 28 men); each of the subjects contributed four one-minute videos: two videos in
well-lit conditions, and the other two captured under natural light. The videos were recorded using a Logitech HD C525 with a resolution of 640×480 pixels and a frame rate of 20 fps. Each subject wore a contact PPG sensor to obtain the BVP data.

4.2 Training settings

The facial videos and corresponding physiological signals were synchronized before the training. We set a variety of video sequence lengths \( T = \{150,300,450,600\} \) to obtain the best results, and the corresponding video durations were 5 s, 10 s, 15 s, and 20 s. The spatial dimensions of the coarse-cropped video frames were all standardized as 112×112. Our proposed method was implemented using the Pytorch framework and performed on an Intel(R) Core(TM) i7-9700 computer with 32 GB RAM and an NVIDIA GeForce RTX 2080Ti GPU. Adam optimizer [38] with an initial learning rate of 0.0001 was used for training. The maximum epoch number for training was set to 80 for the experiments on our dataset.

4.3 Evaluation Metrics

We use the following four evaluation metrics to measure the performance of the proposed method:

(1) **Pearson correlation coefficient** \( (r) \): The Pearson correlation coefficient varies between \(-1\) and 1. Changes in the position and scale of the two variables do not cause changes in the coefficient. When the correlation coefficient approaches 1, it indicates that the two variables have a strong positive correlation, which can be used to measure the correlation between the predicted HR and ground truth.

\[
    r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (16)
\]

(2) **Mean absolute error** (MAE): MAE is a linear score in which all individual differences have
the same weight on the average value. It is used to measure the average value of the absolute error between the predicted HR and HR ground truth.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|$$  \hspace{1cm} (17)

(3) **Root mean square error (RMSE):** The RMSE represents the sample standard deviation of the difference between the predicted value and the ground truth (called residual). The penalty for high errors is higher.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}$$  \hspace{1cm} (18)

(4) **Error standard deviation ($SD_e$):** $SD_e$ is used to measure the dispersion degree of the absolute error, which can more intuitively reflect the stability of the method.

$$SD_e = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (err_i - \bar{err})^2}$$  \hspace{1cm} (19)

Among the aforementioned evaluation metrics, $x_i$ represents the predicted HR, $\bar{x}$ represents the average predicted HR of all predicted HRs, $y_i$ represents the HR ground truth, $\bar{y}$ represents the average HR ground truth of all ground truths, $err$ represents the absolute error between the predicted HR and ground truth, and $\bar{err}$ represents the average absolute error.

**4.4 Intra-dataset Testing**

We first performed an intra-dataset testing on our dataset and compared the prediction accuracy with the state-of-the-art methods, as shown in Table 2.

Table2. The HR estimation results by the proposed method and several state-of-the-art methods on our dataset.
| Method          | MAE↓   | RMSE↓   | SD$_e$↓ | r↑   |
|-----------------|--------|---------|---------|------|
| Green [6]       | 6.59   | 11.91   | 9.92    | 0.53 |
| ICA [14]        | 6.19   | 9.86    | 7.67    | 0.45 |
| POS [25]        | 5.95   | 10.56   | 8.72    | 0.54 |
| CHROM [4]       | 4.15   | 8.03    | 6.88    | 0.65 |
| PhysNet [34]    | 2.41   | 4.39    | 3.91    | 0.88 |
| Proposed Method | **1.75** | **3.43** | **2.95** | **0.93** |

It can be seen from the results that our proposed method achieved the expected results with an MAE of 1.75 bpm, RMSE of 3.43 bpm, SD of 2.95 bpm, and r of 0.93. The results of all four indicators were better than the results of traditional and deep-learning-based state-of-the-art methods. These results also indicate that on larger datasets, deep learning-based methods (PhysNet, proposed method) can achieve better performance. By reducing the MAE, the error distribution was also more concentrated, which is reflected in the smaller SD$_e$. As shown in Fig. (6), the peaks and troughs of the rPPG signal we predicted are well aligned with the ground truth PPG signal, which meets the prediction requirements of multiple physiological signals such as respiratory rate and HRV with considerable application potential. To intuitively reflect the impact of facial video length on HR prediction, Fig. (7(a)) plots the RMSE for different video lengths. It can be seen from the figure that the 15 s video achieved the best results.

From the scatter plot of the predicted HR and the ground truth, shown in Fig. (7(b)), it can be seen that these parameters are highly positively correlated.

We further checked the error distribution of the proposed method. As shown in Fig. (8), most
samples (88.5%) had an absolute error of less than 3 bpm; 92.8% of samples had an absolute error of less than 5 bpm, indicating that our proposed method has strong stability.

Fig. 6 An example from our dataset. (a) The original video frame in our dataset. (b) Example waves of the estimated rPPG and the ground truth BVP signal. The blue wave is ground truth PPG signal, the red wave is predicted rPPG signal.

Fig. 7 (a) Evaluation of the HR prediction with different video lengths. (b) The scatter plot of the ground truth and the predicted HR.
Fig.8 The estimation error distribution of our proposed method.

4.5 Cross-dataset Testing

To evaluate the generalization ability of the proposed method, we constructed a cross-dataset testing on the small-scale datasets UBFC-RPPG and COHFACE. We trained our method on the collected RGB color facial videos. Due to the large difference in data distribution, we fine-tuned and tested the pre-trained models of learning-based methods (HR-CNN [12], PhysNet [34], MSTmap+CVD [22], and proposed method) on UBFC-RPPG and COHFACE. The test results of the proposed method and state-of-the-art methods are presented in Tables 3 and 4.

As shown in Tables 3 and 4, because video compression affects the video-based remote physiological signal measurement results, the performance of the methods on the highly compressed dataset COHFACE is significantly lower than that of the uncompressed dataset UBFC-RPPG. Even on unfamiliar datasets with different data distributions, after minimal data fine-tuning (the ratio of fine-tuned data to test data is 1:1), our model still shows stronger generalization ability than other comparison algorithms, and provides better results than state-of-the-art methods in terms of MAE and RMSE indicators. In particular, for the highly compressed dataset COHFACE, the overall
The performance of our method was significantly improved compared to the comparison algorithms. The MAE decreased from 7.8 bpm (CHROM) to 5.57 bpm, and the RMSE dropped below 10 bpm, indicating that our method is more robust in the case of strong equipment noise.

Table 3. The HR estimation results by the proposed method and several state-of-the-art methods on the UBFC-RPPG dataset

| Method           | MAE↓ | RMSE↓ | SD↓ | r↑  |
|------------------|------|-------|-----|-----|
| Green [6]        | 10.2 | 20.6  | 20.2| -   |
| ICA [14]         | 8.43 | 18.8  | 18.6| -   |
| POS [25]         | 4.12 | 10.5  | 10.4| -   |
| CHROM [4]        | 10.6 | 20.3  | 19.1| -   |
| PhysNet [34]     | 3.63 | 5.29  | 3.85| 0.94|
| MSTmap+CVD [22]  | 7.85 | 9.05  | 11.43| 0.79|
| Proposed Method  | **2.15** | **3.82** | **3.15** | **0.97** |

Table 4. The HR estimation results by the proposed method and several state-of-the-art methods on the COHFACE dataset

| Method          | MAE↓ | RMSE↓ | SD↓ | r↑  |
|-----------------|------|-------|-----|-----|
| CHROM [4]       | 7.80 | 12.45 | -   | 0.26|
| Li2014 [2]      | 19.98| 25.59 | -   | -0.44|
| 2SR [31]        | 20.98| 25.84 | -   | -0.32|
| HR-CNN [12]     | 8.10 | 10.78 | -   | 0.29|
| PhysNet [34]    | 8.82 | 11.85 | 7.9 | 0.47|
| Proposed Method | **5.57** | **9.50** | **7.69** | **0.75** |
4.6 Ablation Study

Table 5 shows our ablation studies on the function model and loss function in different settings on our dataset; the default input video duration was 15 s. Note that “No C_feature extractor” refers to skipped C_feature extractor model, “No Skin Map” refers to canceled skin map generator model, and “Proposed method” includes all the function models and loss functions.

Evaluation of function module: To verify the effectiveness of the multilevel feature fusion, we verified the influence of the C_feature extractor and skin map generator in MHFF on the prediction results. The different feature maps generated by the function models are shown in Fig. (9).

1. **No C_feature extractor**: To verify the effectiveness of the channel-wise high-level feature map, we skipped the C_feature extractor, which means that the high-level feature map generated by STSC was directly fused with the skin map, and then fed into the softmax classifier to obtain the weight mask $M_w$. As shown in Table 5, after the module was added, the MAE decreased from 3.78 bpm to 2.95 bpm with other settings being the same, and the performance of all other evaluation indicators was improved.

2. **No skin map**: To verify the effectiveness of the skin map, we canceled the skin map generator and directly sent the channel-wise feature map $f_c$ generated by the C_feature extractor into the softmax classifier to obtain the weight mask $M_w$. As shown in Table 5, the addition of skin map is important to accurately predict HR. After adding the skin map, the MAE decreased from 7.16 bpm to 2.95 bpm with other settings being the same, and the performance of all other evaluation indicators was considerably improved.

Evaluation of loss function: To evaluate the most effective combination of loss functions, we verified the influence of the three loss functions on the prediction results: negative Pearson
correlation loss ($L_r$), frequency domain cross entropy loss ($L_f$), and binary cross entropy loss ($L_s$). As shown in Table 5, it is sufficient to use only the time domain loss negative Pearson correlation to guide the network to accurately predict the average HR. The joint supervision of the time and frequency domains can enable better HR predictions. The addition of binary entropy loss can guide the skin map generator to predict facial ROI more efficiently and avoid ROI misdetection caused by background noise, which can considerably improve the efficiency of HR prediction.

| Method                        | MAE↓ | RMSE↓ | $SD_r$↓ | $r$↑  |
|-------------------------------|------|-------|---------|-------|
| No C_feature extractor        | 1.85 | 3.78  | 4.07    | 0.90  |
| No Skin Map                   | 5.50 | 7.16  | 9.03    | 0.51  |
| Loss ($L_r$)                  | 6.88 | 7.18  | 9.95    | 0.44  |
| Loss ($L_r + L_f$)            | 5.07 | 7.34  | 8.92    | 0.54  |
| Proposed Method ($L_r + L_f + L_s$) | **1.75** | **3.43** | **2.95** | **0.93** |

**Fig.9** Visualization of different feature maps on our dataset. (a) Coarse face area. (b) Low-level face features. (c) Multi-wise high level feature map. (d) Skin map.

5. Conclusion

In this study, we proposed an efficient end-to-end multi-hierarchical convolutional network for
rPPG signals and HR estimation, which only requires 15 s of face video. An MHFF model was built to avoid the interference of background noise and head movement. Moreover, a 3D spatio-temporal convolutional network baseline was applied to consider both the temporal and spatial characteristics of rPPG signals. We tested our dataset and common datasets (UBFC-RPPG and COHFACE). The experimental results show that the proposed method can accurately reconstruct the rPPG signal and average HR, and the position of each pulse peak is well aligned. The predicted rPPG signal can be used to estimate more physiological signals, which is one of our future research plans.

Acknowledgements

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

References

[1] G. Balakrishnan, F. Durand, and J. Guttag, Detecting pulse from head motions in video, in: Proc. IEEE CVPR. Jun. 2013, pp. 3430–3437.

[2] 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.

[3] 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.

[4] G. De Haan, V. Jeanne. Robust pulse rate from chrominance-based rPPG. IEEE Trans. Biomed.
[5] D. N. Tran, H. Lee, C. Kim. A robust real time system for remote heart rate measurement via camera, in: 2015 IEEE International Conference on Multimedia and Expo (ICME). 2015, pp.1–6.

[6] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. Opt. Express 16, 26(2008): 21 434–21 445.

[7] X. Niu, H. Han, S. Shan, Chen X. 2018. SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific. In 2018 24th International Conference on Pattern Recognition (ICPR). 3580–3585.

[8] X. Niu, S. Shan, H. Han, Chen X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Trans. Image Processing 29, 2409–2423 (2020)

[9] W. Chen, D McDuff. Deephys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 349–365, 2018.

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

[11] C. Tang, J. Lu, J. Li. Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera[C], in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE Computer Society, 2018.

[12] R. Spetlik, J. Cech, F. Vojtěch. Visual heart rate estimation with convolutional neural network[C], in: British Machine Vision Conference 2018, pp. 1309-1315
[13] A. Lam, Y. Kuno. Robust heart rate measurement from video using select random patches. In: IEEE International Conference on Computer Vision (ICCV). Washington, DC, USA, pp. 3640–3648 (2015). https://doi.org/10.1109/ICCV.2015.415

[14] 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

[15] M. Z. Poh, D. J. McDuff, R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. 58,1(2011): 7–11.

[16] M. Lewandowska, J. Ruminski, T. Kocejko, J. Nowak. Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity. In: 2011 federated conference on computer science and information systems (FedCSIS). IEEE, pages405–410.

[17] S. Kwon, H. Kim, K. S. Park. Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone, in: Proceedings of the 2012 IEEE Annual International Conference of the Engineering in Medicine and Biology Society. 2012, pp. 2174–2177

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

[19] L. Wei, Y. Tian, Y. Wang, T. Ebrahimi. Automatic webcam-based human heart rate measurements using Laplacian Eigenmap, in: Lecture Notes in Computer Science. 2013, pp. 281–292

[20] Z. Yu, W. Peng, X. Li, X Hong, and G. Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, in: Proceedings of the IEEE International Conference on Computer Vision. 2019, pp. 151–160

[21] Y. Tsou, Y. Lee, C. Hsu, S. Chang. Siamese-rPPG network: remote photoplethysmography
signal estimation from face videos[C], in: SAC ’20: The 35th ACM/SIGAPP Symposium on Applied Computing. ACM, 2020.

[22] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, G. Zhao. Video-based remote physiological measurement via cross-verified feature disentangling [M]. in: Proc. ECCV 2020.

[23] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognit Lett. 2017.

https://doi.org/10.1016/j.patrec.2017.10.017

[24] X. Niu, H. Han, S. Shan, X. Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video, in: Proc. ACCV, 2018, pp. 562-576

[25] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. IEEE Trans Biomed Eng 64, 7(2017): 1479–1491, 2017. 2, 4, 6, 7.

[26] G. Heusch, A. Anjos, S. Marcel. A reproducible study on remote heart rate measurement in: Proc. CVPR 2017

[27] 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[J]. IEEE Signal Process Lett 99 (2020): 1-1.

[28] J.-Y. Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm. Intel Corporation Microprocessor Research Labs.2000

[29] G. R. Tsouri, S. Kyal, S. A. Dianat, and L. K. Mestha. Constrained independent component analysis approach to nonobtrusive pulse rate measurements. J Biomed Opt 17(7, 2012): Art. no. 077011.

[30] M. Kumar, A. Veeraraghavan, and A. Sabharwal. DistancePPG: Robust non-contact vital signs monitoring using a camera. Biomed. Opt. Express 6, 5(2015): 1565–1588, 2015.
[31] W. Wang, S. Stuijk, and G. de Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. IEEE Trans Biomed Eng 63, 9(Sep. 2016): 1974–1984.

[32] Y. Hsu, Y.-L. Lin, and W. Hsu, Learning-based heart rate detection from remote photoplethysmography features, in: Proc. IEEE ICASSP, May 2014, pp. 4433–4437.

[33] P. V. Rouast, M. T. P. Adam. Remote heart rate measurement using low-cost RGB face video[J]. Frontiers of Computer Science, 2018, 12(5):858-872.

[34] Z. Yu X. Li and G. Zhao "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks" Proc. Brit. Mach. Vision Conf. pp. 1-12 2019.

[35] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan Amplitude-selective filtering for remote-ppg. Biomed. Opt. Express 8, 3, (2017): 1965–1980

[36] W. Wang, S. Stuijk, G. De Haan. Exploiting spatial redundancy of image sensor for motion robust rPPG. IEEE Trans Biomed Eng 62, 2(2015): 415–425

[37] V. Kazemi, J. Sullivan. c[C]. in: IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2014.

[38] D. P. Kingma, J. Ba. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[39] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05), (1), 2005: 886–893

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks, in: ICCV, 2015.

[41] Z. Qiu, T. Yao, T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks, in: Proceedings of the IEEE International Conference on Computer Vision. 2017, pp.
[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] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 305–321.

[44] S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural Comput, 9, 8(1997): 1735–1780.

[45] S.H.I. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems. 2015, pp. 802–810.

[46] E. Lee, E. Chen, C.Y. Lee. Meta-rPPG: Remote heart rate estimation using a transductive meta-learner, in: Proceedings of the European Conference on Computer Vision (ECCV). 2020.