WPPG Net: A Non-contact Video Based Heart Rate Extraction Network Framework with Compatible Training Capability

Weiyu Sun, Xinyu Zhang, Ying Chen, Yun Ge, Chunyu Ji, Xiaolin Huang

Nanjing University

Abstract. Our facial skin presents subtle color change known as remote Photoplethysmography (rPPG) signal, from which we could extract the heart rate of the subject. Recently many deep learning methods and related datasets on rPPG signal extraction are proposed. However, because of the time consumption blood flowing through our body and other factors, label waves such as BVP signals have uncertain delays with real rPPG signals in some datasets, which results in the difficulty on training of networks which output predicted rPPG waves directly. In this paper, by analyzing the common characteristics on rhythm and periodicity of rPPG signals and label waves, we propose a whole set of training methodology which wraps these networks so that they could remain efficient when be trained at the presence of frequent uncertain delay in datasets and gain more precise and robust heart rate prediction results than other delay-free rPPG extraction methods.

Keywords: Remote heart rate estimation, rPPG, signal processing, end-to-end learning

This is the little-adjusted submitted version, we will update the improved one soon.

1 Introduction

There are many “secrets” hidden in our body’s images, containing precious physiological information that could be utilized for medical treatment. Video-based physiological measurement has therefore gained increasing popularity in recent years. It generally relies on Photoplethysmography (PPG) technology. When pulsatile blood flows beneath our skin, it causes the subtle color change containing cardiovascular characteristics, known as PPG signals. Primitive physiological measurement involves wearable devices, which is accurate and reliable with comprehensive details. It is still often used in routine physical examinations and clinical diagnosis. However, there are many limitations such as discomfort and inconvenience, hindering its application in many scenarios.

The initial non-contact physiological measurement mainly uses professional devices such as Doppler radar to collect signals. In 2000, T.Blazek et al. presented
the first video-based evaluation system simply consisting of computer, camera and light source [27], which thereby spawned the development of video-based remote photoplethysmography (rPPG) signal measurement. As a non-invasive optical technology, rPPG process is simple and convenient, which is conducive to widespread application.

In the inception, solutions on rPPG signal extraction mainly focused on the selection of skin areas and the demodulation on waves of pixel values on different color channels [4, 6, 18, 22]. These solutions could be easily affected by motion and ambient optical field. With the revival of deep learning, many works have been done [3, 11–13, 30, 31]. The related datasets usually consist of videos on human faces as inputs and Blood Volume Pulse (BVP) or electrocardiography (ECG) signals of human in videos as labels [11]. BVP is the variation in the volume of blood vessels indicating vital signs, while ECG is the change of the action potential of cardiomyocytes during the heartbeat cycle whose periodicity is consistent with the heartbeat.

![Fig. 1. The specific process of different heart rate extraction method. The ECG signals are collected through electrodes stick on body, they reflect the heart activity directly. The BVP signals are generally collected through finger clips which sensors, they reflect the contraction of finger blood vessels. Similarly, the color change curves on face contain information on the blood activity of face area. Because of the time cost of blood flowing and the unknown time delay through the signal processing units of BVP/ECG collecting devices, finally BVP/ECG signals and real facial color change from video slices could hence have volatile temporal deviation with each other.](image-url)
from different part of our bodies, there exists the pulse transit time (PTT)\textsuperscript{20}——the unknown temporal delay and phase deviation. The delay not merely varies from person to person, but also can be changeable on certain individual under different circumstances.

For another thing, both BVP and ECG signals have gone through a set of processing and transformations before monitored by datasets-maker, and the time consumption during this period is uncertain as well. Although in theory the frequency of both ought to be the same, there exists volatile phase delay between ground truth waves and real rPPG signals in time dimension. The uncertain delay more or less adds difficulties to temporal alignment between them. If signals with phase mismatch take up too much of the datasets, the training of networks will no doubt be influenced, which will further impact the accuracy of the result prediction.

Generally, two methods\textsuperscript{3, 11, 13, 30, 31} are adopted to blunt the impact mentioned above. One of the methods is that we can avoid using ground truth waves directly when training networks. Models such as Rhythmnet\textsuperscript{13} use heart rate values calculated from label waves as ground truth and output predicted heart rate values. By doing this, the temporal delay could be ignored because the frequency of label wave is sufficient enough for heart rate estimation. However, this may miss too many temporal context characteristics related to heart beat which hence increase the difficulty of network training.

Another strategy is that manual temporal alignment between ground truth waves and real rPPG signals is performed during the data pre-processing stage and train ”Frames-to-waves”\textsuperscript{3, 30} network on them. In this way ground truth waves are generated and the rPPG waves of input videos could be predicted. But apparently the alignment task is laborious and time-consuming. Moreover, it generally could not deal with all videos of datasets because they are not robust enough under the circumstance of participants’ fierce head motions and the effect of ambient light. Videos that fail to generate standard waves have to be discarded.

In this article, we demonstrate a novel network framework and relative training process that allows us to save the efforts of manual wave alignment. Our method mainly focuses on the intrinsic self similarity of ground truth waves and real rPPG signals instead of their specific state over time, which contains sufficient rhythm and periodicity information for network to learning efficiently. By wrapping traditional ”Frames-to-waves” rPPG network\textsuperscript{3} using our methodology, we create a convenient training environment which is free from uncertain delays in rPPG datasets for wrapped network and thus maximize its potentials without too much temporal information loss. We compare our result on datasets with uncertain delays with other previous rPPG signal estimation solutions to prove the effectiveness of our method.
2 Relative works

The typical rPPG approach obtains cardiovascular characteristics by detecting variations in volume of blood vessels. The process could be concluded into several steps. First, face detection is performed and facial pixels are selected. Second, the pixel values in different color channels undergo a series of transformations to form a preliminary wave. Finally, some filters are applied and the final curve is predicted with one single frequency. HR-related signal extraction is often followed by a fast Fourier transform (FFT) to obtain a spectral distribution, the peak value of which is considered to be the HR frequency. Common practices include chrominance-based method [6], independent component analysis (ICA) [18], pulse blood vector [4], adaptive matrix completion [22], etc. The commonly used camera sensors are RGB sensors, and Gastel et al. [5] also explored the feasibility of near-infrared images. Experiments show that the RGB camera outperforms the infrared camera, especially in the case of the head movement considered, which can be attributed to the existence of parallel parallax.

It is worth noting that most of the above models demand manual selection of the target region of interest (ROI) [16]. Facial regions are determined based on individual experience and knowledge, which may not correspond with the fact. Handcrafted feature extraction is also empirical and labor-consuming. Additionally, they are put forward on the assumption that the contribution of each skin pixel is equal, ignoring the weight of elements in different facial areas.

The upsurge of machine learning and deep learning has provided more possibilities for solving image and video processing problems. The face detection algorithm proposed by Viola and Jones [24] is widely used in face ROI selection. On this basis, face tracking and skin selection are completed, and finally the heart rate value is predicted. Qiu, Y. et al. [19] combines convolutional neural network (CNN) with Eulerian Video Magnification (EVM) [28] to magnify facial color contrast, and then measures HR from spatiotemporal feature maps, which greatly improves model performance. Radim Špetlík et al. [21] proposed an end-to-end CNN model featured with a frequency-limited 2D convolutional network to output the PPG signal and a 1D convolutional layer for result evaluation.

W. Chen et al. [3] proposed an end-to-end “DeepPhys” system consisting of a motion model and a appearance model. The former features a focus mechanism inputting normalized frame differences and outputting PPG signals; the latter derives an attention mask by generating a probability map. Zhan et al. [31] explored the essence and limitations of CNN in this area based on the “DeepPhys” model, and probed whether the appearance of the tested skin affects the performance of CNN, and whether the evaluation of moving objects is robust.

In 2018, Niu et al. proposed the ”Synrhythm” model [12], which pre-trains a HR regression model on ImageNet and synthetic feature maps then ameliorates it with real-life spatiotemporal maps, the trained network directly converts the spatiotemporal features into heart rate values. In 2020 [13], the advanced ”RhythmNet” network was given. Face alignment is performed with face tracking and landmark detection, spatiotemporal maps are derived as feature representation subsequently. The final outcome is the mean value of all predicted values
from each divided video slices. Considering the fact that there are few public datasets currently available, which cannot satisfy the actual training needs, the team also introduced a more comprehensive and accurate dataset [11].

Yu et al. [31] designed the first end-to-end HR estimation solution to highly compressed videos, composed of video enhancement block “STVEN” and PPG recovery module “rPPGNet”. A skin-based module and partition constraints are applied to generate physiological information on both HR and HRV levels. They further proposed an end-to-end “PhysNet” network [30], providing two options of framework, 3D convolution and RNN-based LSTM relatively, to promote the robustness of the network. The two are recommended rather than traditional 2D convolutional network structure for richer details in time dimension.

3 Approach

Facial color changes (rPPG signals), BVP signals and ECG signals have a common intrinsic connection, they all result from cardiac activities. Although they generally present different waveform patterns (such as QRS complex of ECG signals and quadratic harmonic peak of BVP signals), they share the same rhythm and frequency which could reflect the information of heart rate, according to which we could cast attention on the inherent law of their rhythms and then design methods to extract these rhythm features. In this chapter, we introduce our proposed preprocessing method on ground truth waves (BVP signals and ECG signals), network framework, designs of corresponding loss functions and the calculation of heart rate. These four sections consist our methodology, and we name it as WPPG (Wrapper remote Photoplethysmography) net.

We utilize the self-attention mechanism which focuses on the internal relative periodicity on rhythms of facial color changes and ground truth waves instead of the situation of them in specific time stamps. Therefore, our method could exclude the effect of uncertain delay between ground truth waves and real facial color changes and still ensure the consistency of label and network output.

3.1 Preporcessing of Datasets

The preprocessing on datasets is shown in Fig.2. First, we apply Butterworth filter within [0.65, 4] Hz and cwt wavelet filter [1] on ground truth wave \( r(t) \) in datasets to extract the core frequency on heart rate. The frequency of filtered wave \( f(t) \) has become single, thus \( f(t) \) could be described as (1):

\[
f(t) = B(t) \cos (2\pi f_s t + \varphi)
\]

where \( \varphi \) is uncertain delay between real facial color change curve \( r(t) \), \( f_s \) is the frequency of heart rate in \( r(t) \), and \( B(t) \) is the envelope of the signal depends on the amplitude of \( r(t) \), whose frequency is much lower than \( f_s \). The form of \( f(t) \) in (1) to some extent resembles the modulated signals in signal transmission, we only cast attention on \( 2\pi f_s t + \varphi \) in \( f(t) \) because the envelope \( B(t) \) is independent
of facial condition. Therefore, we can apply Hilbert transform to demodulate \( f(t) \) and extract \( 2\pi f_s t + \varphi \) and \( B(t) \) separately. The demodulation with Hilbert transform could be demonstrated as follows:

\[
F(t) = f(t) + i\hat{f}(t) = \Re e^{i\varphi(t)}
\]

where \( \Re = \sqrt{f(t)^2 + \hat{f}(t)^2} \), \( \varphi(t) = \arctan(\hat{f}(t)/f(t)) \)

Among equation (2), \( \hat{f}(t) \) is the result Hilbert transformation of \( f(t) \). Considering the condition that the frequency of envelope \( B(t) \) is much lower than \( f_s \), the \( \varphi(t) \) of the analytical signal \( F(t) \) could be approximated to:

\[
\varphi(t) \approx \arctan\left(\frac{B(t)\sin(2\pi f_s t + \varphi)}{B(t)\cos(2\pi f_s t + \varphi)}\right) = 2\pi f_s t + \varphi
\]

Thereby we gain the instantaneous phase \( \varphi(t) \) of \( f(t) \). After integrating \( \varphi(t) \) along time and wrapping it between \( 0 \sim 2\pi \), we thereby get the temporal phase change \( A(t) \). To this step, the uncertain delay \( \varphi \) still exists. To uproot the influence of \( \varphi \), we cast attention on the difference \( \Delta t \) between frames by calculating the cosine value of \( \Psi_{\Delta t} \), which could be described as:

\[
\Psi_{\Delta t} = A(t + \Delta t) - A(t) = ((2\pi f_s (t + \Delta t) - 2k\pi + \varphi) - (2\pi f_s t + \varphi))
\]

where \( 2k\pi \) is the wrapped period in \( A(t) \). After the subtraction, the public component \( \varphi \) in \( A(t + \Delta t) \) and \( A(t) \) is neutralized, hence we exclude the uncertain delay \( \varphi \). We calculate \( \cos(A(t_1) - A(t_2)) \) on every combinations of time stamps \( t_1, t_2 \in \{0, 1, 2, \ldots, n-1\} \), where \( n \) is the length of \( r(t) \). Thereby we gain the target label matrix \( R_{n \times n} \), in which every element \( r_{ij} = \cos(\Psi_{i-j}) \). The visualization of \( R_{n \times n} \) is is shown in Fig.2, from which we could observe the periodical rhythm information.

### 3.2 Proposed network framework

Our network framework mainly consists of two parts, as shown in Fig.3. The first part is traditional ”Frames-to-waves” network wrapped by our methodology. The network wrapped in experiments is the adjusted version from Deepphys [3]. The inputs of adjusted Deepphys are facial video slices \( \{\text{frame}_0, \text{frame}_1, \text{frame}_2, \ldots, \text{frame}_N\} \) and the difference of adjacent frames \( \{\text{frame}_1 - \text{frame}_0, \text{frame}_2 - \text{frame}_1, \ldots, \text{frame}_N - \text{frame}_{N-1}\} \), where \( N + 1 \) is the length of input videos. The difference of adjacent frames provides information on the color changes of facial areas and facial video slices help judge the information weight on skin areas on faces. This combination of inputs has been proved efficient in Deepphys in original 2-D CNN layers version, therefore we make adjustments on it and extend an extra temporal dimension on original convolution kernels. The output of the
Fig. 2. Preprocessing on datasets (BVP signal in this figure). When demodulating $f(t)$, we utilize Hilbert transform to generate $\hat{f}(t)$ (yellow curve in complex space), and we focus on the helical line (green circles) projected by $F(t)$ on the real−imag surface. The helical line in figure is transformed in order to explicitly show the instantaneous phase $\varphi(t)$. Purple line are $A(t)$ intergrated by $\varphi(t)$ and it is wrapped between 0 ∼ 2π, every index in whose x-axis reflects a special time stamp recorded in ticks (let all waves in figure have totally n ticks). Then all pairs of time stamps will be calculated $\cos(\Psi(\Delta t))$ and inserted into the phase map $R$ according to their indexes. Because of the even symmetry property of cos function, the phase map $R$ is symmetrical according to diagonal. And the final visualization of phase map shows the periodicity and rhythm properties of ground truth wave.

First part network is a serial of feature maps $\{f_{m_0}, f_{m_1}, f_{m_2}, \cdots, f_{m_N-C-1}\}$ (considering the temporal convolution consumption is $C$, for we do not set temporal paddings), each feature map conveys information on the context of facial color change trend on a certain time stamp. The second part of network divides these features into different slices $\{s_0, s_1, \cdots, s_{N-C-L}\}$ using sliding window with a fixed length $L$ (according to the frame rate, we adopt 11 when fps is 30), every slice records a section of rhythm pattern of facial color change curve. Then we apply FCN layer on these slices and turn them into vector sets $\{v_0, v_1, \cdots, v_{N-C-L}\}$. In the end, we calculate cosine similarity on all combinations of $(v_i, v_j)$, where $i, j \in [0, N - C - L]$, and thus we get the output matrix $\hat{R}$, in which element $\hat{r}_{i,j}$ represents the cosine similarity value between $v_i$ and $v_j$. The periodicity of facial color changes causes the repeatability of local rhythm patterns, which is recorded in generated vector sets of network. Therefore, by calculating cosine similarity among generated vectors, we could observe this periodicity through each row and column in $\hat{R}$ which swings with the increasing of distance between two vectors in time scale, as shown in Fig.3.

### 3.3 Loss function and regularization

As designed, the result of preprocessed datasets $R$ and the output of our network $\hat{R}$ reflect the common internal relative periodicity on rhythms caused by cardiac activities, we can utilize $R$ serve as the label of our network, thus with the pro-
The first part (Network wrapped is altered Depphys)

Residual Module

Skin-based Attention Mask

Feature Mask Generation

Appearance Module

Skin Mask Generation

Global Average Pool

Mutual cosine Similarity Calculation

The second part

Feature map slices

Phase-based Attention Map Generation

INPUT

Raw Frame r(t)

Conv 1×8×8, 3
AvgPool 1×2×2
Conv 1×5×5, 32
Conv 3×3×3, 16
AvgPool 1×2×2
Conv 3×3×3, 32
Conv 2×8×8, 3
AvgPool 1×2×2
Conv 1×5×5, 64
Conv 3×3×3, 32
AvgPool 1×2×2
Conv 3×3×3, 64
Conv 2×8×8, 3
AvgPool 1×2×2
Conv 1×5×5, 64
Conv 3×3×3, 32
AvgPool 1×2×2
Conv 3×3×3, 64

Residual Frame

r(t+1)-r(t)

Skin Mask

Feature Mask

FCN

(Projection)

FCN

Input

Conv 5×3×3, 64
AvgPool 1×2×2
Conv 7×3×3, 32
Conv 5×3×3, 64
AvgPool 1×2×2
Conv 7×3×3, 32

Fig. 3. The whole framework of WPPG net, which includes two parts of the network and the postprocessing (calculate loss function and predicted heart rate) on output of network. To facilitate expression, the length of input videos \((N + 1)\) is 71 frames (the length we used when training network), \(C\) which is 16 represents the length consumption because of the absence of temporal paddings in our wrapped network, \(L\) which is 11 is the length of sliding window, the final shape of Phase Map and Phase map is \((n + 1, n + 1)\), where \(n = N - L - C\) is 43. The length of \(Seq_R\) is either \(n + 1\). The first part of network refers to adjusted Depphys \([3]\) as mentioned above. The output matrix \(\hat{R}\) and \(Seq_R\) derived from it are calculated loss function with \(R\) and \(Seq_R\) generated using methods shown in Fig.1 from ground truth waves. And we calculate predicted heart rate from \(Seq_R\) in equation (10). There is an additional loss function on facial area selection involved in the first part of network, which depends on the design of wrapped network.
In progress of training, the network could learn about the capability on extracting rhythm features from input facial videos. And our loss functions between labels and network outputs are as follows.

First, as $R$ and $\hat{R}$ share the same range, mean square error $L_{MSE, map}$ is a direct and efficient metric which could measure the distance between them. Second, as mentioned above, the value of cosine similarity among vectors depends mainly on the temporal distance between vectors. Every row and column in $\hat{R}$ describes the swing of cosine similarity between a certain vector and all other vectors in time scale, which share the same periodicity feature with corresponding row and column in $R$. Thus, we could calculate negative Pearson correlation between each corresponding row and column on $R$ and $\hat{R}$ and calculate the average value on them, and thus we get $L_{Pear, map}$ as

$$L_{Pear, map} = \frac{\sum_{n=0}^{N-C-L} Neg_{pearson}(\hat{R}_n, R_n)}{N-C-L+1}$$

(5)

where $\hat{R}_n$ and $R_n$ represent the $n_{th}$ row in $\hat{R}$ and $R_n$ respectively. Third, we collect all elements $\hat{r}_{ij}$ which satisfy condition:

$$i - j = c, c \in [0, N - C - L]$$

(6)

into groups $\{G_1, G_2, ..., G_{N-C-L}\}$. The temporal distances of vectors when calculating cosine similarities in the same group are consistent, hence values of elements in the same group approach each other, according to which we add a regularization item which could be described as:

$$Reg = \frac{\sum_{n=0}^{N-C-L} std(G_n)}{N-C-L+1}$$

(7)

which could help constrict the range of values in the same group during the training of network. Meanwhile, we calculate the average value of elements per group and arrange them into $Seq_{\hat{R}}$, each element $m_i$ in which could be described as:

$$m_i = mean(G_i)$$

(8)

which contains all cosine similarity values made from vectors in a certain temporal distance $i$. The periodicity of $Seq_{\hat{R}}$ along temporal distance reflects the frequency of heart rate and $Seq_{\hat{R}}$ summarizes all information on $\hat{R}$. Similarly, we could generate $Seq_R$ from $R$ in the same method, and we calculate MSE $L_{MSE, seq}$ and negative Pearson correlation $L_{Pear, seq}$ between $Seq_{\hat{R}}$ and $Seq_R$. In addition, we calculate MSE between the face map generated in the skin mask generation step of the first part (shown in Fig.3) and ground truth facial skin area segmented from input video [16] as loss function $L_{MSE, mask}$, which is designed according to the property of the wrapped network. Thus, our final loss function could be described as:

$$Loss = \alpha L_{MSE, map} + \beta L_{Pear, map} + \gamma L_{MSE, seq} + \lambda L_{Pear, seq} + \xi L_{MSE, mask} + \mu Reg$$

(9)

where $\alpha, \beta, \gamma, \lambda, \xi, \mu$ are the weights of the corresponding loss function items.
3.4 Calculation on heart rate

Finally, we calculate heart rate depending on $\text{Seq} \hat{R}$, for it covers all information on $\hat{R}$ and its value becomes more reliable after averaging along groups. After filtering $\text{Seq} \hat{R}$ through a Butterworth bandpass filter within $[0.65, 4]$ Hz and a cwt filter, we collect peak points of filtered $\text{Seq} \hat{R}$ as $\{p_0, p_1, \cdots, p_{M-1}\}$, where $M$ is the number of peaks in filtered $\text{Seq} \hat{R}$. Then we calculate the temporal distances between two adjacent peaks, and hence we gain $\{\Delta p_0, \Delta p_1, \cdots, \Delta p_{M-2}\}$. Thereby the heart rate could be described as:

$$hr = \frac{fps \cdot 60}{\sum_{n=0}^{M-2} \frac{\Delta p_n}{M-1}}$$

4 Experiments and results

4.1 Dataset involved

The datasets used in this paper is VIPL-HR [11]. The ground truth waves of VIPL-HR are BVP signals, the datasets contain 107 participants and each participant is involved in nine tasks. More detailed information on different tasks could be found in the attached document of VIPL-HR. Each task on each participant is recorded through 4 different devices, they are marked from source1 to source4. Among them, the frame rate of videos in source2 is stable, which is 30 fps. Videos in Source1 and Source3 have volatile frame rates because of the compression of video, which could get checked from the attached documents of VIPL-HR. The videos in Source4 are recorded by infrared camera and they are not involved in our experiments. We select task from v1 to v7 on all 107 participants in source2 and select feasible video slices according to confidence interval of BVP signals (totally 457 videos, some videos are discarded as their length or ground truth waves are not qualified).

4.2 Experiments and metrics

In our experiments, we use Adadelta [33] to optimize network on Tesla V100, the learning rate is set as 1e-3 and total training epochs are set as 50. Before inputing video slices into network, we perform face detection on the first frame using “mmod\_human\_face\_detector” [7] from python package “dlib”, and enlarge the detected rectangle containing human face until it covers the whole face area. Then, only areas in the rectangle are selected in video slices and we thereby gain a facial sequence which is the real input of network. When training networks, the length of input facial videos is set as 70 frames and the batch size of inputs is 6. Because of the length consumption which results from the absence of temporal paddings and the presence of sliding windows performed in our network, the length of output would reduce by 24 frames. To compensate, when generating phase maps from datasets, we cut 12 frames on both ends of the ground truth of waves. When testing performance on test set, we select input facial videos
with 300 frames on fixed positions of original videos, and we generate predicted results through five-fold subject-exclusive cross-validation on datasets. The ground truth heart rate for performance metrics is calculated in equation (10) on filtered ground truth waves \( f(t) \) in Fig.2.

Four accuracy metrics between predicted heart rate (bpm, beat per minute) and ground truth heart rate (bpm) are used to judge the performance of different methods: mean, mean absolute error (MAE), root mean square error (RMSE) and standard deviation (std), which are typical accuracy metrics applied in previous researches \[3,13\].

| Method Type      | Method       | Mean(bpm) | Std(bpm) | MAE(bpm) | RMSE(bpm) |
|------------------|--------------|-----------|----------|----------|-----------|
| Traditional      | Tulyakov2016 | 10.8      | 18.0     | 15.9     | 21.0      |
|                  | POS [25]     | 7.87      | 15.3     | 11.5     | 17.2      |
|                  | CHROM [6]    | 7.63      | 15.1     | 11.4     | 16.9      |
| Non Frame-to-wave| RhythmNet    | 0.73      | 8.11     | 5.30     | 8.14      |
|                  | ST-Attention | -3.10     | 7.99     | 5.40     | 7.99      |
|                  | CVD [14]     | -         | 7.92     | 5.02     | 7.97      |
|                  | Dual-GAN [10]| -         | 7.63     | 4.93     | 7.68      |
| Frame-to-wave    | I3D [2]      | 1.37      | 15.9     | 12.0     | 15.9      |
|                  | PhysNet [30] | 1.68      | 14.9     | 10.8     | 14.8      |
|                  | DeepPhys [3] | -2.60     | 13.6     | 11.0     | 13.8      |
|                  | AutoHR [29]  | -         | 8.48     | 5.68     | 8.68      |
|                  | PhysFormer [32]| -          | 7.74     | 4.97     | 7.79      |
|                  | WPPG (Ours)  | 0.88      | 7.69     | 4.96     | 7.73      |

Table 1. Performance on VIPL-HR dataset, the results of previous methods share the same protocol as claimed in \[?32\]. Our method applied the same protocol on VIPL-HR, so we can use these data directly for performance comparison! The data we referred comes from the latest paper \[32\].

**Fig. 4.** Scatter plots of prediction results (bpm) and target results (bpm) in different methods
| Method Type          | Method          | Mean(bpm) | Std(bpm) | MAE(bpm) | RMSE(bpm) |
|---------------------|-----------------|-----------|----------|----------|-----------|
| Traditional Method  | Poh2011 [17]    | 2.04      | 13.5     | -        | 13.6      |
|                     | CHROM [16]      | -         | -        | 13.49    | 22.36     |
|                     | Li2014 [9]      | -3.30     | 6.88     | -        | 7.62      |
|                     | Tulyakov2016 [23]| 3.19      | 5.81     | 4.96     | 6.23      |
| Non Frame-to-wave   | SynRhythm [12]  | 0.3       | 10.88    | -        | 11.08     |
|                     | RhythmNet [13]  | 0.41      | 3.99     | -        | 3.99      |
| Frame-to-wave       | HR-CNN [21]     | 1.37      | -        | 7.25     | 9.24      |
|                     | rPPGNet [31]    | -         | 7.82     | 5.51     | 7.82      |
|                     | DeepPhys [13]   | -         | 4.75     | 3.97     | 5.10      |
|                     | AutoHR [29]     | -         | 4.9      | 3.01     | 3.68      |
|                     | Meta-rPPG [8]   | -         | 3.87     | 3.25     | 3.97      |
|                     | PhysFormer [32] | -         | 5.15     | 3.85     | 5.25      |
|                     | WPPG (Ours)     | -1.06     | 5.15     | 3.85     | 5.25      |

Table 2. Performance on MAHNOB-HCI dataset, the results of previous methods share the same protocol as claimed in [19, 32]. Our method applied the same protocol on MAHNOB-HCI, so we use these data directly for performance comparison.

| Case        | $L_{map}$ | $L_{seq}$ | $L_{MSE, mask}$ | $Reg$ | MAE(bpm) |
|-------------|-----------|-----------|-----------------|-------|----------|
| map-only    | ✓         | -         | -               | -     | 4.28     |
| seq-only    | -         | ✓         | -               | -     | 4.62     |
| map&seq     | ✓         | ✓         | -               | -     | 4.29     |
| w/o mask    | ✓         | ✓         | ✓               | -     | 4.52     |
| w/o reg     | ✓         | ✓         | -               | ✓     | 4.02     |
| default     | ✓         | ✓         | ✓               | ✓     | 4.23     |

Table 3. Mean Absolute Error on test set of VIPL-HR.

| SW length | Mean(bpm) | Std(bpm) | MAE(bpm) | RMSE(bpm) |
|-----------|-----------|----------|----------|-----------|
| 1         | 0.56      | 11.14    | 5.91     | 11.38     |
| 11        | 0.88      | 10.51    | 5.20     | 10.54     |
| 25        | -2.18     | 11.49    | 6.16     | 11.69     |

Table 4. Performance of different slide window’s lengths, fps30.

| Wrapped | Aligned | Mean(bpm) | Std(bpm) | MAE(bpm) | RMSE(bpm) |
|---------|---------|-----------|----------|----------|-----------|
| ×       | ×       | -         | -        | -        | -         |
| ×       | ✓       | 3.90      | 9.97     | 7.28     | 10.71     |
| ✓       | ×       | 1.68      | 5.85     | 4.78     | 7.56      |

Table 5. “Frame-to-wave” networks can’t converge directly on VIPL-HR: label waves must be aligned. Our method could train directly and gain better performance. “-” means failed.
Performances on our method and other previous solutions in 4 metrics are listed in Table 1 and Table 2. To further present the prediction consistency of our network, we plot the scatter plot (predicted heart rate and real heart rate) in Fig.4(a) according to the prediction results of our network, and we plot the scatter plots of prediction results from Zhang et al [7] and CHROM [6] in Fig.4(b) and Fig.4(c) respectively. Meanwhile, we make ablation experiments to verify the effect of our method, as shown in Table 3, Table 4 and Table 5. More details will be checked in supplement materials uploading soon.

5 Discussion

According to the result of experiments, our method gain good performance. According to Table 5 as to other ”Frames-to-waves” networks such as STVEN [30] and our adjusted Deepphys [3], they are not suitable for the datasets with uncertain delay we used if there is no extra preprocessing, they could not reach convergence when training. And other non ”Frame-to-wave” traditional solutions to varying degrees make non deep learning preprocessing on origin facial videos, and this preprocessing is not efficient. Much ambient noise would be entangled into preprocessed color channel signals which are the inputs of these solutions. Under this circumstance performances on these methods are limited. On the other hand, like RhythmNet, replacing network output with heart rate value could get rid of the effect of uncertain delay, but it increases the difficulty of network training in another degree. The compression from ground truth waves to heart rate values would discard some valuable information, which causes the decrease of efficiency of network training.

It is challenging to find out rPPG signals from facial videos since they are very weak, adding they generally are entangled with ambient noise. To better and robustly extract rPPG signals, we could utilize attention mechanisms of the wrapped network and focus on facial areas which contains more information on facial color changes caused by cardiac activities and thereby reduce the influence of noise, which could be feasible under the circumstance that labels of dataset are rational. ”Frames-to-waves” network structure is efficient when ground truth labels are reliable and temporal aligned, because they are more intuitive and they draw information directly from origin facial sequences, thereby ”Frames-to-waves” could get learned quickly and have stronger anti-noise ability. Our works in this paper maximize their potentials when uncertain delays are frequent in datasets. Any ”Frames-to-waves” network structures could be implemented in the first part of our network, like Deepphys [3] what we choose in our experiments, which could not reach convergence when training directly in our datasets.

Among our network framework shown in Fig.3, although our works are reflected mainly on the second part of network and additional pre and post processings, the first part of network (wrapped ”Frames-to-waves” network) is important as well because they do help provide real-time facial color change information, and the second part of network provide a buffer layer between the
first part of network and our preprocessed labels. By calculating cosine similarities among all slices generated from former network, the second part of network successfully turn the temporal information per time stamp into mutual relationships between temporal slices, and thereby get rid of the influence of uncertain delay. Therefore, when training networks, the former part of network get trained as like the uncertain delays do not exist. The output $\hat{R}$ of the second part and the generated phase map from ground truth signals $R$ both reflect the intrinsic rhythm and periodicity related to cardiac activities, therefore the correctness on the analytical results of networks could still be checked directly during the training of network. When the output of second part of network is proved reasonable through designed loss function, the first part of network would learn about that they submit the correct temporal information as well. Under this circumstance, the traditional ”Frames-to-waves” network (the first part) could still give full play to their full strength even though there exist uncertain delays.

We focus mainly on the inherent law of ground truth waves and apply wave filters during the preprocessing of our methodology, so even signal sources with different patterns from different physical sources (such as BVP signals and ECG signals) could be generated into similar phase maps through the preprocessing. It means our method has excellent universality and it facilitates the training of network on across datasets with label waves of various types.

6 Conclusion

We analyze the uncertain delay phenomenon hidden in some rPPG signal estimation datasets and propose a whole set of methodology and corresponding network framework which could be trained efficiently when uncertain delays are frequent among datasets. Our proposed WPPG net outperforms other methods in datasets with uncertain delays, and we attribute this improvement to our self-similarity based information extraction method which could exploit sufficient periodical rhythm information caused by cardiac activities and the excellent capability of the ”Frames-to-waves” network under non-uncertain-delay situations wrapped in our framework. When training other ”Frames-to-waves” networks on datasets with uncertain delays, or training on mixed datasets whose ground truth signals are from different physical sources, our methodology could be applied to enhance the efficiency of network training and the performance on heart rate prediction.

References

1. Bousefsaf, F., Maaoui, C., Pruski, A.: Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. Biomedical Signal Processing and Control 8(6), 568–574 (2013)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: ECCV (2018)
4. De Haan, G., Van Leest, A.: Improved motion robustness of remote-ppg by using the blood volume pulse signature. Physiological Measurement 35(9), 1913–1926 (2014)
5. van Gastel, M., Stuijk, S., de Haan, G.: Motion robust remote-ppg in infrared. IEEE Transactions on Biomedical Engineering 62(5), 1425–1433 (2015). https://doi.org/10.1109/TBME.2015.2390261
6. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering 60(10), 2878–2886 (2013). https://doi.org/10.1109/TBME.2013.2266196
7. King, D.E.: Max-margin object detection. arXiv preprint arXiv:1502.00046 (2015)
8. Lee, E., Chen, E., Lee, C.Y.: Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In: European Conference on Computer Vision. pp. 392–409. Springer (2020)
9. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4264–4271 (2014)
10. Lu, H., Han, H., Zhou, S.K.: Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12404–12413 (2021)
11. Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video (2018)
12. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3580–3585 (2018). https://doi.org/10.1109/ICPR.2018.8546321
13. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing 29, 2409–2423 (2020). https://doi.org/10.1109/TIP.2019.2947204
14. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: European Conference on Computer Vision (ECCV) (2020)
15. Niu, X., Zhao, X., Han, H., Das, A., Dantcheva, A., Shan, S., Chen, X.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
16. Phung, S.L., Bouzerdoum, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. IEEE transactions on pattern analysis and machine intelligence 27(1), 148–154 (2005)
17. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering 58(1), 7–11 (2010)
18. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Transactions on Biomedical Engineering 58(1), 7–11 (2011). https://doi.org/10.1109/TBME.2010.2086456
19. Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H., Saddik, A.E.: Evm-cnn: Real-time contactless heart rate estimation from facial video. IEEE Transactions on Multimedia 21(7), 1778–1787 (2019). https://doi.org/10.1109/TMM.2018.2883866
20. Smith, R.P., Argod, J., Pépin, J.L., Lévy, P.A.: Pulse transit time: an appraisal of potential clinical applications. Thorax 54(5), 452–457 (1999)
21. Spetlik, R., Cech, J., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: BRITISH MACHINE VISION CONFERENCE (2018)

22. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2396–2404 (2016). https://doi.org/10.1109/CVPR.2016.263

23. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: 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. pp. 2396–2404 (2016)

24. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 1, pp. 511–518 (2001). https://doi.org/10.1109/CVPR.2001.990517

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

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

27. Wu, T., Blazek, V., Schmitt, H.J.: Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes. Proceedings of SPIE - The International Society for Optical Engineering pp. 62–70 (2000)

28. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A.: Video description generation incorporating spatio-temporal features and a soft-attention mechanism. Eprint Arxiv 53, 199–211 (2015)

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

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

31. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 151–160 (2019). https://doi.org/10.1109/ICCV.2019.00024

32. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. arXiv preprint arXiv:2111.12082 (2021)

33. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)

34. Zhan, Q., Wang, W., Haan, G.D.: Analysis of cnb-based remote-ppg to understand limitations and sensitivities pp. 1268–1283 (2019)