Remote Estimation of Heart Rate Based on Multi-Scale Facial ROIs

Changchen Zhao, Weiran Han, Zan Chen, Yongqiang Li, Yuanjing Feng*
College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China
fyjing@zjut.edu.cn

Abstract

While most rPPG approaches extract the pulse signals based on single facial region of interest (ROI), this research proposes a new method to extract pulse signals from ROIs with multiple scales. The idea is that rich pulse features can be extracted by varying ROI scales and combining these features would contribute to the accuracy improvement. The proposed framework consists of three main steps: 1) constructing facial ROI pyramid with multiple scale levels, 2) blood volume pulse (BVP) signals extraction, and 3) signal fusion using convex combination with Gaussian and uniform priors, respectively. This paper also investigates how the commonly used algorithms perform under multi-scale ROIs. Experiments were conducted using one publicly available dataset and one self-collected dataset. The results show that the ROI with a size slightly smaller than the face boundary achieves on average higher measurement accuracy. The high-quality pulse signal appears not consistently in one scale level but rather in multiple levels according to measurement environments and motion statuses. Therefore, the fusion of multiple pulse signals is beneficial to the measurement accuracy improvement.

1. Introduction

Recent years have witnessed a rapid growth of remote photoplethysmography (rPPG), a technology that measures blood volume pulse (BVP) and heart rate in a non-contact way based on optical/physiological principles [33]. Compared to the conventional photoplethysmography (PPG) that utilizes a contact pulse oximeter under green and infrared light [4], rPPG utilizes a consumer-level digital camera under visible light, which broadens the pulse measurement application including incubator monitoring [5], fitness exercise [35], face anti-spoofing [10], sleep monitoring [25], etc. Despite its wide application, a large amount of scientific problems makes it an active research topic.

The choice of a facial ROI acts as the first key step of the system. First, the pulsatile signal strength varies at different locations on the face due to the distribution of capillaries beneath the skin surface. The location of an ROI has a direct impact on the quality of the raw rPPG measurement. Second, the shape of an ROI always leads to unnecessary inclusion of undesired pixels like eyes, mouth, hair, or background pixels, thus, introducing rigid/non-rigid motion artifacts. Third, the scale of an ROI determines the proportion of different face location pixels in the ROI, which also affects the shape of the raw rPPG measurement. The location, shape, and scale of an ROI are factors that directly determine the signal separation model used for pulse extraction. It is crucial to choose a good ROI to guarantee a higher measurement accuracy.

Prior researches have paid attention to the relationship between facial subregion and rPPG signal quality. Kwon *et al*. [6] divided the face into seven regions and evaluated the quality of the signal of each region. They found that a forehead and both cheeks have a potential to be good candidates for pulse extraction, while the signal quality from a mouth and a chin is relatively low. Poh *et al*. [17] proposed to select an ROI of 60% of the width of the full face and full height. Zhao *et al*. [37] used an ROI below the eye line that covers the skin region within the nose, mouse, and cheeks. Tulyakov *et al*. [24] divided the warped facial region into subregions and proposed self-adaptive matrix completion (SAMC) to dynamically select regions useful for robust heart rate estimation. Wang *et al*. [31] treated every facial subregion as an independent sensor for pulse measurement and proposed an algorithm to exploit the redundancy of an image sensor to distinguish the pulse signal from motion-induced noise. The signal strength variation is partially due to the physiological facts of the human anatomy, *e.g*., the cheek, lips, and chin region comprise a higher proportion of capillaries that results in higher absorption of light compared to other regions of the face. However, the region with higher pulsatile strength is not necessarily suitable for rPPG extraction because these regions may intervened by non-rigid motions like eye-blinking, talking, and smile, etc.

Image resolution and video compression also have an impact on rPPG measurement accuracy. Due to some recording device and transmission limitations, the resulting image is sometimes compressed and the ROI has low resolution, resulting in a decrease in signal-to-noise-ratio of the extracted pulse signal [14, 34]. To mitigate this limitation, McDuff [13] proposed to use a deep image super resolution networks prior to rPPG pulse extraction pipeline. Zhao *et al.* [36] discarded the blue and red channel signals that are polluted heavily by video compression and proposed singular spectrum analysis (SSA) decompression and spectral masking algorithm to refine the extracted pulse signal.

Eulerian video magnification (EVM), firstly proposed by Wu *et al.* [32] to reveal subtle color variations in the image, has shown to be effective for rPPG pulse extraction [3, 19]. Instead of using the ROI directly, EVM-based rPPG approaches first construct a Gaussian pyramid and then the highest level images are used for rPPG extraction. Essentially, EVM acts as a feature extraction model that converts to image representations to other feature space for further learning of rPPG related features.

This paper introduces multi-scale image processing techniques into rPPG signal extraction, which has not been fully investigated in the published literature. The basic idea of this paper is to increase the diversity of the BVP signal by building multi-scale facial ROIs, from which complementary pulse features can be extracted. Combining the candidate features would contribute to the final signal quality improvement. To this end, we propose multi-scale facial ROIs by scaling up and down the original ROI, resulting in multiple averaged rPPG measurements (traces) that vary in waveform. rPPG features are extracted from these traces in each level, respectively, and fused to obtain the final pulse signal. The subregion selection and partial ROI based approaches, together with the EVM-based approaches in essence deal with single-scale ROI. On the contrary, by varying the scale of an ROI, the number of candidate pulse signals is increased, which mitigates the limitation of some core rPPG algorithms that at most $n-1$ independent distortions can be suppressed by linearly combining n source signals.

The contributions of this paper are summarized as follows:

1) We propose a novel rPPG pulse extraction framework based on multi-scale feature extraction and fusion. To the best of our knowledge, this is the first attempt to investigate simultaneous pulse signal extraction on multi-scale facial ROIs.

2) We analyze some (linear combination) rPPG approaches' response to multi-scale facial ROIs, revealing the distribution of signal quality between multiple scale levels under various motion statuses and recording setups.

3) We demonstrate the effectiveness of the proposed algorithm on two benchmarking datasets in comparison with state-of-the-art methods. Our approach achieves higher accuracy than state-of-the-art rPPG methods.

2. Related work

2.1. Remote photoplethysmography

The models used for rPPG pulse extraction include: 1) blind source separation (BSS) based model, *e.g.*, principle component analysis (PCA) [15, 8] and independent component analysis (ICA) [18]; 2) skin reflection model, which is based on the optical/physiological principles [2, 26, 29, 28]; and 3) deep learning based models [20, 22, 1, 16]. This paper is mostly related to the skin reflection model, more specifically, CHROM [2] and POS [26]. These algorithms extract pulse signal by projecting the traces to axes related to specular and diffusion reflectance, followed by a fine-tuning step. These algorithms extract the pulse signal by linearly combining the trace. The trace averaged over a single-scale ROI always has three color channels. At most two independent sources signal can be eliminated. However, in the real world situation, the trace contains many more independent sources [30]. Therefore, diversity of the trace is desired for better signal separation performance.

2.2. Multi-scale image processing

Multi-scale image processing plays an important role in many computer vision tasks such as image compression, image denoising, image restoration, image enhancement, and super-resolution. Learning a discriminative image representation is one of the main objectives of a visual processing system. Usually, multi-scale image transforms may contribute to a good representation that captures the scale of an object in the real world. Image pyramid is one of multi-scale image representations that transforms image with repeated smoothing and subsampling. Image pyramid is widely used in keypoint detection [11], image classification [7], image segmentation [9], etc.

The advantage of multi-scale image processing lies in the fact that some features that are not significant at certain scale may become significant in other scales. Multi-scale image representation generates rich feature diversity in scale space, which is beneficial to the related vision task. Based on this idea, we introduce multi-scale image processing technique to the field of rPPG pulse extraction, in an attempt to exploit the effectiveness of the combination for accuracy improvement.

3. The proposed method

3.1. Overview

Fig. 1 depicts the overview of the proposed rPPG signal extraction method, which consists of three main steps: 1) video pyramid establishment, 2) BVP signal extraction, and 3) multi-scale signal fusion. Each frame is cropped by a set of multi-scale facial ROIs, resulting in a video pyramid with multiple scale levels. The images are spatially averaged to

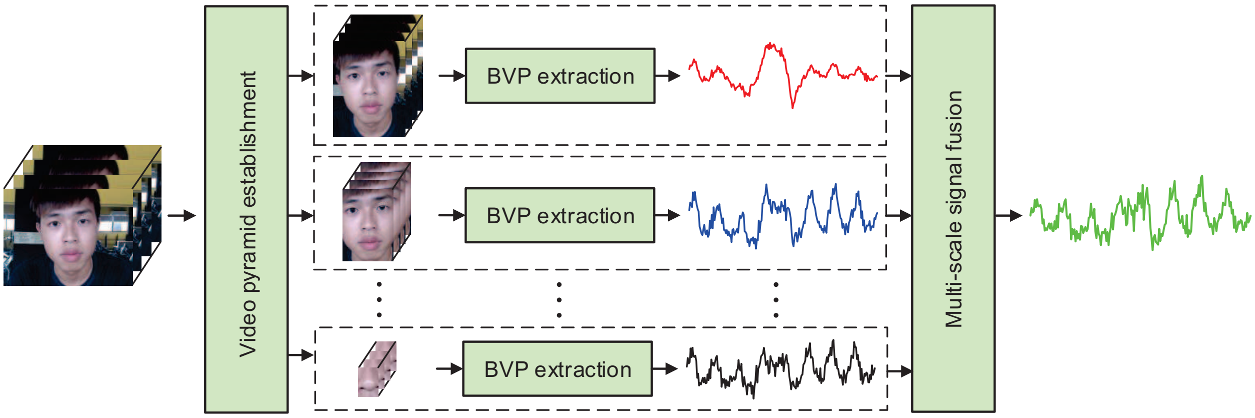

Figure 1. Overview of the proposed rPPG signal extraction framework.

obtain the rPPG traces. Each trace is then fed into a BVP signal extraction algorithm to obtain the initial pulse signal. They are then fused together to obtain the final pulse signal.

3.2. Video pyramid establishment

The key to establishing a video pyramid is the multi-scale facial ROIs. Let w_l, h_l be the width and height of the facial ROI of level l, the multi-scale facial ROI are defined as follows:

$$\begin{aligned} w_l &= w_0 \cdot \left(\frac{1}{2}\right)^l \\ h_l &= h_0 \cdot \left(\frac{1}{2}\right)^l \end{aligned} \tag{1}$$

for $l = -1, 0, 1, ...$, where w_0, h_0 are the width and height of the ROI in level 0. The multi-scale facial ROIs share the same center (c_x, c_y). We designate the ROI in level 0 as the bounding box that tightly encompasses the boundary of the whole face, which can be obtained by a commonly used face detector and tracker. To construct the multi-scale facial ROIs, on one hand, we half the initial ROI size, *i.e.*, $l = 1, 2, ...$. These ROIs mainly cover the skin region, *i.e.*, no background pixels are involved. As these ROIs cover different facial regions, the color variations are different accordingly. For example, some ROIs involve eye-blinking or talking artifact while others do not. By this means, some artifacts are initially separated to some extent. On the other hand, we double the initial ROI size, *i.e.*, $l = -1$. The purpose is to involve some background pixels because it is beneficial to consider the behavior of background when applying the signal extraction algorithm, *i.e.*, the specular reflection and the background color variation are closely related to non-rigid motion artifacts. Fig. 2 illustrates the multi-scale facial ROIs of a given frame.

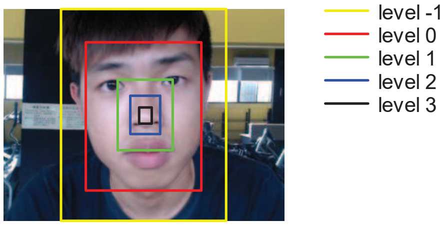

Figure 2. Multi-scale facial ROIs.

The pixels within the multi-scale facial ROIs are converted to the raw rPPG traces by spatial averaging, which is calculated by the following equation:

$$I_{l,c}(t) = \frac{1}{\text{Area}(R_l(t))} \sum_{x,y \in R_l(t)} I_c(x, y, t) \tag{2}$$

for $l = -1, 0, 1, ...$, where $I_c(x, y, t)$ denotes the pixel intensity at coordinate (x, y) of the t-th frame, $c \in \{R, G, B\}$ denotes the color channel, $R_l(t)$ denotes the ROI of level l and the t-th frame, Area$(R_l(t))$ is calculated as the total number of pixels within $R_l(t)$. The ROIs in every frame are determined by a face tracker. The raw rPPG trace is calculated by concatenating the averaged pixel intensity $I_{l,c}(t)$ over all t, which is denoted as $\boldsymbol{I}_l(t)$.

The multi-scale facial ROIs increase the diversity of the raw traces. An illustration can be seen in Fig. 3. In the first video (Rows 1 and 2), the eye-blinking artifacts are more significant in trace level 0 (red) than in trace level 2 (green). This is because ROI level 0 covers the entire eyes region while the percentage of eye-pixels is relatively low in ROI level 2. For the second video (Rows 3 and 4), two traces are anti-phase due to the fact that ROI level 0 covers background pixels while ROI level 3 does not. The increased diversity of rPPG traces leads to rich pulse features extracted by the BVP signal extraction algorithm. Some of the pulse features are complementary and thus contribute to the improvement of the final pulse signal quality.

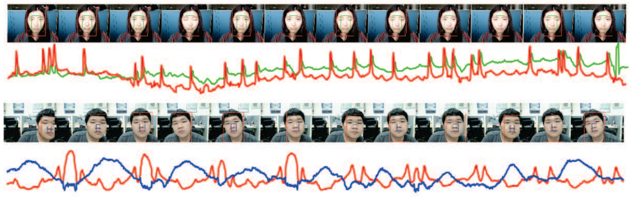

Figure 3. Illustration of the raw traces extracted from multi-scale facial ROIs. First row: example video 1, in which the subject is blinking eyes; second row: rPPG traces extracted from video 1, ROI level 0 (red) and level 2 (green); third row: example video 2, in which the subject moves his body horizontally; last row: rPPG traces extracted from video 2, ROI level 0 (red) and level 3 (blue).

3.3. BVP signal extraction

The BVP signal needs to be extracted from the multi-scale rPPG traces. Numerous BVP extraction approaches have been proposed. In this paper, we employ Plane-Orthogonal-to-Skin (POS), which is first proposed by Wang *et al*. [26] and widely adopted by various rPPG researches. POS is based on the skin reflection model. POS defines a projection plane orthogonal to the vector $[1, 1, 1]^T$ in order to eliminate the dependency of skin tone. The raw traces are projected on two vectors on this plane in order to separate the BVP signal and the motion artifacts. The projected signals are further fused by α-tuning to obtain the final BVP signal. We apply POS to the raw trace $\boldsymbol{I}_l(t)$ for all scale levels. Therefore, in this section, we temporally omit the level index l for the raw traces $\boldsymbol{I}_l(t)$ and denote it simply as $\boldsymbol{I}(t)$. Specifically, the raw trace is first processed by the temporal normalization,

$$\boldsymbol{I}_n(t) = \boldsymbol{N} \cdot \boldsymbol{I}(t) \tag{3}$$

where $\boldsymbol{N} \in \mathbb{R}^{3\times3}$ is a diagonal matrix whose i-th diagonal gives the reciprocal of mean value of the i-th row of $\boldsymbol{I}$, *i.e.*,

$$\boldsymbol{N}_{ii} = 1/\mu(\boldsymbol{I}_i) \tag{4}$$

The temporally normalized trace is then projected on two vectors defined by a projection matrix, $\boldsymbol{P}_p = [0\ 1\ -1; -2\ 1\ 1]$ where each row denotes a projection axis orthogonal to each other. The projected signals can be written as follows:

$$S_1(t) = I_{nG}(t) - I_{nB}(t) \tag{5}$$

$$S_2(t) = I_{nG}(t) + I_{nB}(t) - 2I_{nR}(t) \tag{6}$$

In order to separate the specular and pulsatile components, the projected S_1 and S_2 need to be processed by α-tuning,

$$x(t) = S_1(t) + \alpha S_2(t) \tag{7}$$

where $\alpha = \sigma(S_1(t))/\sigma(S_2(t))$, and $\sigma(\cdot)$ denotes the standard deviation. $x(t)$ is the extracted BVP signal, in this paper, we also call it the POS feature. We apply POS algorithm to the trace of each level, resulting in a total number of L BVP signals, where L denotes the number of levels in the video pyramid. Hereafter, we denote $x_l(t)$ as the POS feature $x(t)$ extracted in level l.

Applying POS to multi-scale traces is a generalization of the original POS algorithm. When we set $l = 0$, the proposed algorithm becomes POS. The purpose of applying multi-scale POS feature extraction is to facilitate easy pulse extraction. In the conventional (single-scale) POS extraction, all the motion artifacts are combined with the BVP signal. A linear POS operation has limited strength to extract clear BVP signal at all motion circumstances and recording environments. Instead, with the help of multisscale facial ROIs, the motion artifacts are partially separated. The resulting traces in each level contains fewer signal sources, which eases the pulse extraction.

The skin reflection model does not consider the background changes. One issue may arise when applying POS to level $l = -1$, where background pixels are involved. We argue that the POS algorithm is applicable to level $l = -1$ as long as the skin pixels dominate the ROI. This condition is mild and the following cases can satisfy: the subject keeps stationary or the motion is not vigorous.

3.4. Multi-scale signal fusion

The final pulse signal is computed by fusing the candidate POS features extracted from multi-scale traces. Due to the fact that the candidate POS features are assumed to be complementary, we cast the signal fusion problem as feature combination rather than feature selection. To this end, the convex combination is employed:

$$p(t) = \sum_{l=-1}^{L} \lambda_l \cdot x_l(t) \tag{8}$$

where λ_l denotes the weight of level l and is subject to the constraint $\sum_l \lambda_l = 1$ to prevent intensity augmentation. The key step is to determine the weight for each level. A natural thought is to compute a pulsatile measure and assign a larger weight to the level with a higher score. However, we tested two pulsatile measures, *i.e.*, autocorrelation that measures the periodicity of the candidate pulse [12] and signal-to-noise-ratio (SNR) that measures the ratio between the spectral power in and out of the common heartbeat frequency range [27]. Unfortunately, the results are discouraging.

As an alternative, we notice that the POS features have different pulsatile energy in different levels. Larger weights should be assigned to those levels with stronger pulsatile energy. We exploit two priors to determine the weights. The first is Guassian prior,

$$\lambda_l(\mu_0, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{(l-\mu_0)^2}{2\sigma_0^2}\right) \tag{9}$$

where μ_0 and σ_0 denote the center and standard deviation of the level index, respectively. The Gaussian prior is based on the observation that the pulsatile strength in middle-levels is stronger than that of the lower and higher levels.

The second is uniform prior,

$$\lambda_l = \frac{1}{L+2} \tag{10}$$

for $l = -1, 0, 1, ..., L$. The uniform prior assigns equal weights to all the POS candidates.

The operations discussed above are within the range of a time window. For long time monitoring, we concatenate the windowed output to obtain the long-time pulse signal. Specifically, given a video of length N, we first divide the sequence into segments of length T, apply the proposed algorithm to obtain the windowed output, and apply overlap adding [2] to concatenate them to obtain the final output.

4. Experimental setup

4.1. Datasets

We employ two datasets for the evaluation of the performance of the proposed algorithm, one publicly available dataset PURE [23] and one self-collected dataset Self-RPPG.

PURE [23]: A benchmark video dataset involving 10 healthy subjects (8 male, 2 female). The video sequences are of size 640 × 480 pixels, 30 Hz, and stored in PNG format. 6 head motions are performed: 1) **steady** (sitting still, no movement), 2) **talking**, 3) **slow translation** (head movements parallel to the camera plane with slow speed), 4) **fast translation** (head movements parallel to the camera plane with fast speed), 5) **small rotation** (subjects rotate their heads with angles of 20°), and 6) **medium rotation** (subjects rotate their heads with angles of 35°). The rich head motion types make this dataset suitable for this research to investigate their impact to the multi-scale pulse signals.

Self-RPPG: The PURE dataset is restricted to head motions. In order to simulate real world situations, we collected a more challenging dataset. Fore categories are considered: 1) **stationary** (sitting still, no movement), 2) **head translation** (head movement horizontally with medium and fast speed), 3) **recovery** (heart rate recovery after a 3-minute running on a treadmill), and 4) **biking** (subjects perform exercises on a biking machine). The dataset contains 13 healthy subjects (10 male, 3 female). 78 video sequences were recorded of 30 fps, 640 × 480 pixels, 1-minute duration, and stored in uncompressed AVI format. Ground-truth pulse waveforms are recorded simultaneously using a pulse oximeter (Model CMS50E, Contec Medical).

Fig. 4 shows some snapshots of the datasets.

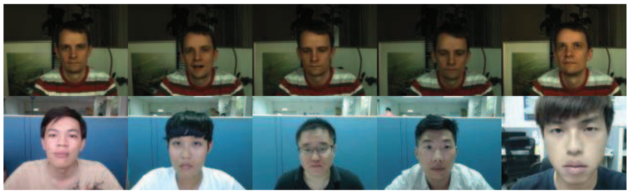

Figure 4. Sample images of the datasets. First row: PURE, second row: Self-RPPG.

4.2. Evaluation metrics

Three commonly used metrics are employed to evaluate the performance of the algorithms.

Signal-to-noise-ratio (SNR) was first defined by De Haan *et al.* [2] and is widely adopted in rPPG field to measure the quality of the estimated pulse signal compared with the ground truth, which is defined as follows:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{f=0.8}^{5} U(f) P^2(f)}{\sum_{f=0.8}^{5} (1 - U(f)) P^2(f)} \right) \tag{11}$$

where $P(f)$ denotes the power spectrum of the extracted pulse waveform, f denotes the frequency in Hz, and $U(f)$ denotes a template separating signal and noise, which is defined as:

$$\hat{U}(f) = \begin{cases} 1, & f_r - \frac{\gamma}{2} \leq f \leq f_r + \frac{\gamma}{2} \\ 1, & 2f_r - \frac{\gamma}{2} \leq f \leq 2f_r + \frac{\gamma}{2} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where f_r denotes the ground-truth heart rate calculated by transforming the pulse signal to the Fourier domain by discrete Fourier transform (DFT), and γ denotes the spectral window length.

We calculate one heart rate value HR for each windowed output and compare it with the ground truth HR$_r$. The mean absolute error (**MAE**) and root mean squared error (**RMSE**) are used to evaluate the accuracy,

$$\text{MAE} = \frac{1}{M} \sum_{t=1}^{M} |\text{HR}(t) - \text{HR}_r(t)| \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{t=1}^{M} (\text{HR}(t) - \text{HR}_r(t))^2} \tag{14}$$

where M denotes the total number of heart rate estimations.

4.3. Compared methods

Two state-of-the-art rPPG core algorithms, CHROM [2] and POS [26], are chosen for comparison because they achieved very good performance in most of the rPPG applications. Due to the fact that they are designed for single scale ROI, we apply them to the traces in each level individually. By doing this we can investigate how they perform with varying ROI scales, which has not be studied before in the published literature. In addition, the proposed algorithm with Gaussian and uniform priors are also compared.

4.4. Implementation details

The face bounding box of the initial scale ($l = 0$) is manually selected for the first frame and tracked by Kanade-Lucas-Tomasi (KLT) tracker [21] for the following frames. We have a total number of 5 scale levels ($L = 3$), window length $T = 1.6s$. We use $\boldsymbol{\lambda} = [0.05, 0.2, 0.5, 0.2, 0.05]^T$ for Gaussian prior and $\boldsymbol{\lambda} = [0.2, 0.2, 0.2, 0.2, 0.2]^T$ for uniform prior. We use a $10s$ time window and $1s$ step size when calculating the heart rate.

5. Results and discussion

The performance of the proposed algorithm in comparison with the compared methods on two datasets are reported in Tables 1 and 2.

5.1. Response to ROI scale variation

The results in Tables 1 and 2 show that POS and CHROM in levels $l = 0, 1$ have better accuracy (*e.g.*, SNR, MAE, and RMSE). The performance gets worse when l increases or decreases. The trend can be seen in Fig. 5, where the trace and pulse of a video in Self-RPPG dataset are plotted. Compared with conventional single-scale POS or CHROM, the changes brought about by the proposed multi-scale facial ROIs can be summarized in the following three aspects. First, the traces in higher levels, *e.g.*, $l = 3, 2$, exhibit a large amount of camera quantization error (Fig. 5, last row, column 4-5). This is because the spatial size of the ROI is small and the camera quantization error cannot be fully eliminated by spatial averaging. Second, the BVP signals are the most prominent in middle levels, *e.g.*, $l = 1, 0$, which is the most effective range for the skin reflection model (Fig. 5, last row, column 3). This is in agreement with previous research where ROIs that are slightly smaller than or tightly bound the face boundary were widely used. Third, the lowest level contains significant motion artifacts (Fig. 5, last row, column 1-2), which is caused by non-skin pixels that are not considered in the skin reflection model.

The results imply that CHROM and POS have their suitable ROI scales, *i.e.*, ROIs that are too large or too small will make the algorithms ineffective. The results also imply the effectiveness of the multi-scale facial ROIs in generating diverse pulsatile features that cannot be processed by state-of-the-art core rPPG algorithms.

Category	01	02	03	04	05	06
SNR (dB)						
CHROM$_{l_{-1}}$	9.96	3.42	7.04	5.58	7.43	6.65
POS$_{l_{-1}}$	10.90	4.38	7.36	6.03	8.30	6.62
CHROM$_{l_0}$	12.13	5.45	9.12	6.37	8.66	7.66
POS$_{l_0}$	13.21	6.22	9.56	7.90	9.93	8.69
CHROM$_{l_1}$	13.44	5.13	**12.82**	**12.19**	11.18	10.30
POS$_{l_1}$	13.73	5.53	12.11	11.21	**11.79**	10.74
CHROM$_{l_2}$	10.08	4.84	9.02	8.58	7.94	7.80
POS$_{l_2}$	9.96	4.97	9.21	8.48	8.20	7.69
CHROM$_{l_3}$	5.18	2.68	5.08	4.34	3.85	4.04
POS$_{l_3}$	5.65	2.82	5.22	4.53	4.11	4.49
proposed_g	**13.79**	**6.58**	12.37	11.42	11.77	**11.09**
proposed_u	11.90	6.47	10.64	9.60	9.57	9.16
MAE (bpm)						
CHROM$_{l_{-1}}$	1.93	3.35	2.72	7.35	2.05	3.40
POS$_{l_{-1}}$	1.98	9.31	2.20	4.16	2.74	2.57
CHROM$_{l_0}$	1.87	4.92	1.66	4.28	2.53	2.92
POS$_{l_0}$	**1.74**	3.68	1.61	2.06	1.51	1.68
CHROM$_{l_1}$	1.89	4.18	**1.60**	1.81	**1.45**	1.65
POS$_{l_1}$	1.65	3.58	1.61	1.78	1.48	**1.58**
CHROM$_{l_2}$	1.81	3.54	1.67	1.91	1.51	1.62
POS$_{l_2}$	1.85	4.35	1.70	2.02	1.47	1.62
CHROM$_{l_3}$	2.82	8.95	2.80	6.03	5.04	6.89
POS$_{l_3}$	2.78	8.71	1.94	6.46	5.12	4.85
proposed_g	**1.74**	**2.99**	1.63	**1.77**	1.48	**1.58**
proposed_u	1.76	3.22	1.65	1.81	1.48	1.61
RMSE (bpm)						
CHROM$_{l_{-1}}$	2.95	15.65	5.13	10.22	3.82	5.65
POS$_{l_{-1}}$	3.15	13.02	3.35	6.78	5.00	4.60
CHROM$_{l_0}$	2.86	8.50	2.14	6.62	4.92	5.28
POS$_{l_0}$	2.49	6.25	**2.11**	3.21	2.02	2.21
CHROM$_{l_1}$	2.90	7.56	**2.11**	2.66	**1.95**	2.11
POS$_{l_1}$	**2.14**	5.72	**2.11**	2.35	1.99	**2.03**
CHROM$_{l_2}$	2.74	5.93	2.25	3.06	2.03	2.08
POS$_{l_2}$	2.76	7.40	2.31	3.41	**1.95**	2.09
CHROM$_{l_3}$	5.67	15.29	5.15	8.42	7.68	9.35
POS$_{l_3}$	6.13	16.55	3.17	8.42	8.21	7.84
proposed_g	2.51	**4.22**	2.15	**2.33**	1.97	2.04
proposed_u	2.53	5.41	2.15	2.38	1.98	2.08

Table 1. Results on PURE dataset. 01-Steady, 02-Talking, 03-Slow translation, 04-Fast translation, 05-Small rotation, 06-Medium rotation. The best scores of each column are highlighted in boldface.

5.2. Performance of the proposed algorithm

The results in Table 1 show that the proposed algorithm achieves comparable performance to the compared methods

Category	Stationary	Translation	Recovery	Biking
SNR (dB)				
CHROM$_{l_{-1}}$	7.21	2.23	4.49	-3.16
POS$_{l_{-1}}$	8.40	3.30	4.78	-3.56
CHROM$_{l_0}$	9.58	4.11	6.44	2.33
POS$_{l_0}$	10.87	5.49	7.47	2.34
CHROM$_{l_1}$	9.98	7.58	7.51	2.45
POS$_{l_1}$	11.27	8.38	7.95	2.20
CHROM$_{l_2}$	8.45	6.13	5.89	2.26
POS$_{l_2}$	9.12	6.13	5.99	2.14
CHROM$_{l_3}$	6.34	4.64	4.20	1.77
POS$_{l_3}$	6.64	4.73	4.20	1.43
proposed_g	**11.94**	**9.08**	**8.63**	**3.98**
proposed_u	11.47	8.38	8.31	3.57
MAE (bpm)				
CHROM$_{l_{-1}}$	1.15	11.90	3.05	33.19
POS$_{l_{-1}}$	1.46	9.73	1.83	37.49
CHROM$_{l_0}$	0.70	6.32	1.25	10.69
POS$_{l_0}$	0.66	4.19	1.32	11.87
CHROM$_{l_1}$	**0.66**	2.67	0.92	5.10
POS$_{l_1}$	0.77	1.19	0.91	8.91
CHROM$_{l_2}$	0.72	2.10	1.00	6.36
POS$_{l_2}$	0.86	1.76	1.01	6.79
CHROM$_{l_3}$	0.76	2.06	1.38	8.79
POS$_{l_3}$	1.00	1.70	1.47	8.52
proposed_g	0.74	**0.70**	0.82	4.82
proposed_u	0.79	0.88	**0.79**	**3.16**
RMSE (bpm)				
CHROM$_{l_{-1}}$	2.82	14.57	7.91	39.52
POS$_{l_{-1}}$	3.03	12.51	4.56	42.81
CHROM$_{l_0}$	1.87	8.91	3.61	15.02
POS$_{l_0}$	1.80	6.06	3.37	16.51
CHROM$_{l_1}$	**1.77**	4.33	2.63	8.58
POS$_{l_1}$	2.07	2.40	2.68	12.68
CHROM$_{l_2}$	1.83	3.47	3.22	9.01
POS$_{l_2}$	2.24	2.86	2.76	9.61
CHROM$_{l_3}$	1.88	3.27	3.38	12.98
POS$_{l_3}$	2.59	3.67	4.01	11.65
proposed_g	2.03	**1.36**	2.47	7.32
proposed_u	2.11	1.54	**2.30**	**5.03**

Table 2. Results on Self-RPPG dataset. The best scores of each column are highlighted in boldface.

on PURE dataset, *e.g.*, the proposed algorithm has better SNR score in 'Steady', 'Talking', and 'Medium rotation' categories, while the POS and CHROM have better SNR score in 'Slow translation', 'Fast translation', and 'Small rotation' categories. The MAE and RMSE results exhibit similar tendency to SNR. The results in Table 2 show that the proposed algorithm outperforms the compared methods in SNR by a large margin, *e.g.*, increased by 0.67 dB in 'Stationary', 0.7 dB in 'Translation', 0.58 dB in 'Recovery', and 1.53 dB in 'Biking', respectively. The results imply that by combining POS features extracted in multi-scale ROIs are beneficial to signal quality improvement, which means that the multi-scale POS features are complementary to each other.

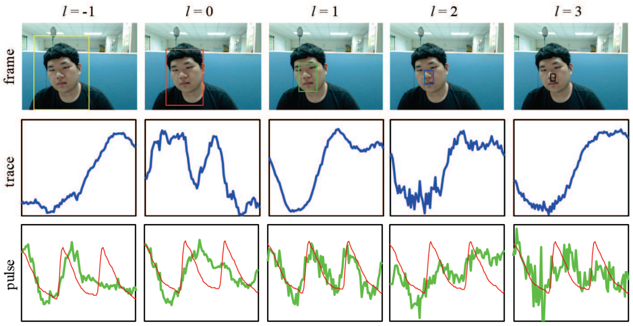

Figure 5. Trace and extracted pulse plot on different scale levels. First row: snapshot of a video, in which the subject moves his head horizontally. Second row: traces averaged over corresponding ROIs. Third row: pulse signals (green line) extracted by POS in comparison with the ground truth PPG signal (red line).

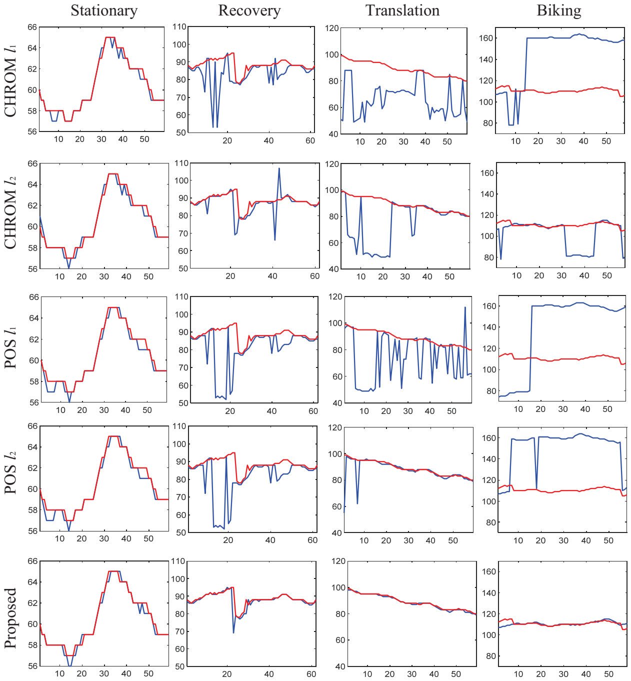

Figure 6. Estimated heart rate (blue lines) in comparison with the ground truth heart rate (red line) on Self-RPPG dataset. Vertical axis denotes the heart rate in beats per minute (bpm) and horizontal axis denotes time in second of a video.

Fig. 6 depicts the heart rate comparison on Self-RPPG dataset. We only depict $l = 1, 2$ results because their performances are better than other scale levels. One can see the accuracy by comparing the overlap between the blue line (estimated heart rate) and the red line (ground truth heart rate). The results show that all the methods achieve accurate results for the stationary case, while the performance degrades for other cases that involve motion. The accuracy in ‘Biking’ is on average lower than that in ‘Translation’, which means that the accuracy drop becomes larger when the motion gets vigorous. Nevertheless, the proposed algorithm achieves the highest accuracy.

To visualize the response of the compared methods to the scale levels, we draw the SNR map as is shown in Fig. 7. Let $p_e^t \in \mathbb{R}^T$ be the estimated pulse signal in the time window of length T, $p_g^t \in \mathbb{R}^T$ be the corresponding ground truth PPG signal. We compute the SNR of $p_e^t \in \mathbb{R}^T$ according to Equ. (11), where the ground truth heart rate is estimated based on $p_g^t \in \mathbb{R}^T$ using DFT. We set window length $T = 5s$ (150 frames) and step size to be 1 frame. The region with brighter color represents the place where the extracted BVP signal has higher similarity to the ground truth PPG. From Fig. 7 one can see that the presence of the clean pulse is not consistent at a certain level. It depends on the motion types and is time-varying. Therefore, the feature combination strategy is beneficial to the accuracy improvement.

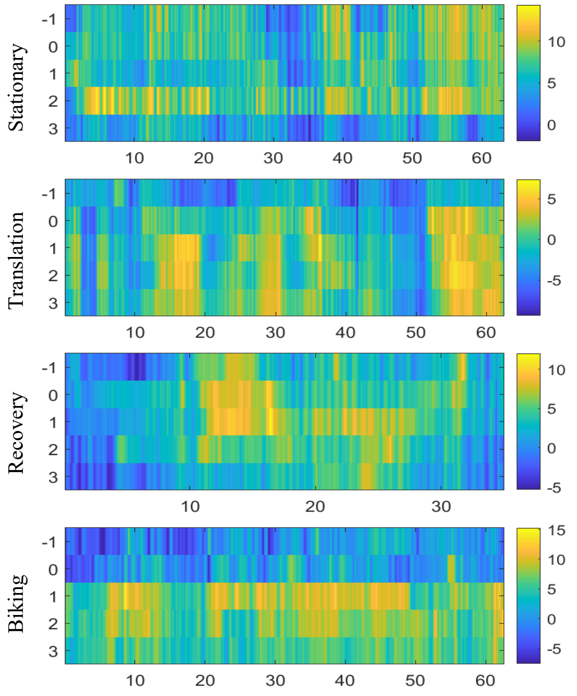

Figure 7. SNR map of a sample video. Vertical axis denotes scale levels and horizontal axis denotes time of a video.

5.3. Comparison between priors

In order to investigate the fusing effectiveness of the two priors, we give the box plot of SNR results on two datasets, as is shown in Fig. 8. The results show that the SNR accuracy of two fusion priors has the same tendency between motion types, *i.e.*, they have higher accuracies on stationary cases and lower accuracies on tougher cases like biking or talking. T-test results show that the SNR results of two priors have no significant difference except for Biking in Self-RPPG ($p = 0.2287 > 0.05$) and Talking in PURE dataset ($p = 0.7397 > 0.05$).

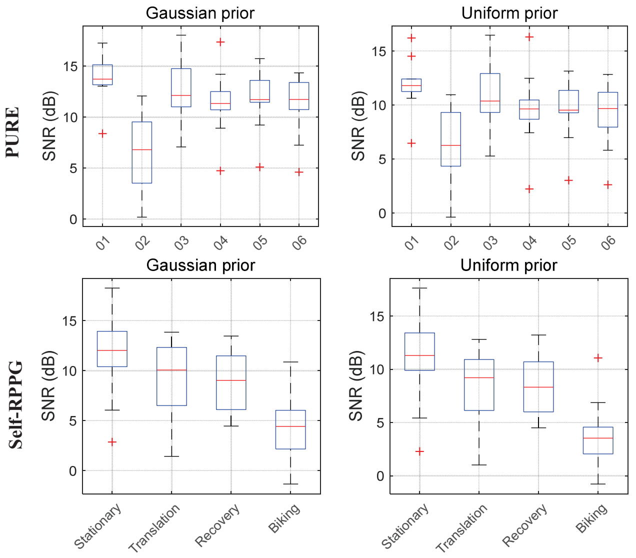

Figure 8. Box plot of SNR results of two fusion priors on two datasets.

6. Conclusions

The basic idea of this paper is to increase the diversity of the BVP signal by building multi-scale facial ROIs, from which complementary pulse features can be extracted. Combining the candidate features contributes to the improvement of the final signal quality. Experimental results demonstrate that varying the scale of facial ROI has a detrimental effect on state-of-the-art rPPG approaches, but the final pulse signal quality can be improved by combing multi-level pulse candidates. The reason can be explained that clean pulse signal appears not consistently in one level but rather depends on the subject status, recording setups, etc. This research demonstrates the effectiveness of varying ROI scale for rPPG pulse extraction. Moreover, it is known that the forehead and cheeks have the strongest rPPG signal. Therefore, to vary both the position and scale of the ROIs would further improve the measurement accuracy.

Acknowledgement

This research was sponsored by National Natural Science Foundation of China under Grant No.s 61903336, 61976190, 61703369; Key Research and Development Program of Zhejiang Province under Grant No. 2020C03070.

References

[1] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision*, pages 349–365, 2018.

[2] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[3] Ennio Gambi, Angela Agostinelli, Alberto Belli, Laura Burattini, Enea Cippitelli, Sandro Fioretti, Paola Pierleoni, Manola Ricciuti, Agnese Sbrollini, and Susanna Spinsante. Heart rate detection using microsoft kinect: Validation and comparison to wearable devices. *Sensors*, 17(8):1776, 2017.

[4] Mohamed Abul Hassan, Aamir Saeed Malik, David Fofi, N M Saad, Babak Karasfi, Yasir Salih Ali, and Fabrice Meriaudeau. Heart rate estimation using facial video: A review. *Biomedical Signal Processing and Control*, 38:346–360, 2017.

[5] John H Klaessens, Marlies Van Den Born, Albert J Van Der Veen, Janine Sikkensvan De Kraats, Frank A Van Den Dungen, and Rudolf M Verdaasdonk. Development of a baby friendly non-contact method for measuring vital signs: first results of clinical measurements in an open incubator at a neonatal intensive care unit. *Proceedings of SPIE Advanced Biomedical and Clinical Diagnostic Systems XII*, 8935:57–62, 2014.

[6] Sungjun Kwon, Jeehoon Kim, Dongseok Lee, and K-wang Suk Park. Roi analysis for remote photoplethysmography on facial video. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2015, pages 4938–4941, 2015.

[7] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recoginition*, volume 2, pages 2169–2178, 2006.

[8] M Lewandowska, Jacek Ruminski, Tomasz Kocejko, and Jedrzej Nowak. Measuring pulse rate with a webcam ł a non-contact method for evaluating cardiac activity. pages 405–410, 2011.

[9] Yuxia Li, Bo Peng, Lei He, Kunlong Fan, and Ling Tong. Road segmentation of unmanned aerial vehicle remote sensing images using adversarial network with multiscale context aggregation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2279–2287, 2019.

[10] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] Richard Macwan, Yannick Benezeth, and Alamin Mansouri. Heart rate estimation using remote photoplethysmography with multi-objective optimization. *Biomedical Signal Processing and Control*, 49:24–33, 2019.

[13] Daniel Mcduff. Deep super resolution for recovering physiological information from videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recoginition Workshops*, pages 1367–1374, 2018.

[14] Daniel Mcduff, Ethan B Blackford, and Justin R Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 63–70, 2017.

[15] Daniel J. Mcduff, Ming Zher Poh, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–74, 2010.

[16] Xuesong Niu, Xingyuan Zhao, Hu Han, Abhijit Das, Antitza Dantcheva, Shiguang Shan, and Xilin Chen. Robust remote heart rate estimation from face utilizing spatial-temporal attention. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[17] Mingzher Poh, Daniel Mcduff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011.

[18] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, 2011.

[19] Ying Qiu, Yang Liu, Juan Arteagafalconi, Haiwei Dong, and Abdulmotaleb El Saddik. Evm-cnn: Real-time contactless heart rate estimation from facial video. *IEEE Transactions on Multimedia*, 21(7):1778–1787, 2019.

[20] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[21] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[22] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference*, pages 3–6, 2018.

[23] Ronny Stricker, Steffen Muller, and Horst Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *Proceedings of IEEE International Symposium on Robot and Human Interactive Communication*, 2014.

[24] Sergey Tulyakov, Xavier Alamedapineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recoginition*, pages 2396–2404, 2016.

[25] Tom Vogels, Mark Van Gastel, Wenjin Wang, and Gerard De Haan. Fully-automatic camera-based pulse-oximetry during sleep. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 1349–1357, 2018.

[26] Wenjin Wang, Brinker Bert Den, Sander Stuijk, and Haan Gerard De. Algorithmic principles of remote-ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

[27] Wenjin Wang, Albertus Cornelis Den Brinker, and Gerard De Haan. Full video pulse extraction. *Biomedical Optics Express*, 9(8):3898–3914, 2018.

[28] Wenjin Wang, Albertus Cornelis Den Brinker, and Gerard De Haan. Discriminative signatures for remote-ppg. *IEEE Transactions on Biomedical Engineering*, 2019.

[29] Wenjin Wang, Albertus Cornelis Den Brinker, and Gerard De Haan. Single-element remote-ppg. *IEEE Transactions on Biomedical Engineering*, 66(7):2032–2043, 2019.

[30] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Robust heart rate from fitness videos. *Physiological Measurement*, 38(6):1023C1044, 2017.

[31] Wenjin Wang, Sander Stuijk, and Gerard De Haan. Exploiting spatial redundancy of image sensor for motion robust rppg. *IEEE Transactions on Biomedical Engineering*, 62(2):415–425, 2015.

[32] Haoyu Wu, Michael Rubinstein, Eugene Shih, John V Guttag, Fredo Durand, and William T Freeman. Eulerian video magnification for revealing subtle changes in the world.

[33] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019.

[34] Changchen Zhao, Weihai Chen, Chunliang Lin, and Xingming Wu. Physiological signal preserving video compression for remote photoplethysmography. *IEEE Sensors Journal*, 19(12):4537–4548, 2019.

[35] Changchen Zhao, Chunliang Lin, Weihai Chen, Mingkun Chen, and Jianhua Wang. Visual heart rate estimation and negative feedback control for fitness exercise. *Biomedical Signal Processing and Control*, 56:101680, 2020.

[36] Changchen Zhao, Chunliang Lin, Weihai Chen, and Zhengguo Li. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1299–1308, 2018.

[37] Changchen Zhao, Peiyi Mei, Shoushuai Xu, Yongqiang Li, and Yuanjing Feng. Performance evaluation of visual object detection and tracking algorithms used in remote photoplethysmography. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2019.