MANet: a Motion-Driven Attention Network for Detecting the Pulse from a Facial Video with Drastic Motions

Xuenan Liu
Hefei University of Technology
Anhui Province Key Laboratory of
Industry Safety and Emergency Technology
Hefei, 230601, Anhui, China
xuenanliu@mail.hfut.edu.cn

Xuezhi Yang
Hefei University of Technology
Anhui Province Key Laboratory of
Industry Safety and Emergency Technology
Hefei, 230601, Anhui, China
xzyang@hfut.edu.cn

Ziyan Meng
Hefei University of Technology
Anhui Province Key Laboratory of
Industry Safety and Emergency Technology
Hefei, 230601, Anhui, China
2020111045@mail.hfut.edu.cn

Ye Wang
Hefei University of Technology
Anhui Province Key Laboratory of
Industry Safety and Emergency Technology
Hefei, 230601, Anhui, China
1403532447@qq.com

Jie Zhang
The First Affiliated Hospital of USTC
Hefei, 230601, Anhui, China
951860401@qq.com

Alexander Wong
University of Waterloo
Waterloo, N2L3G1, Ontario, Canada
a28wong@uwaterloo.ca

Abstract

Video Photoplethysmography (VPPG) technique can detect pulse signals from facial videos, becoming increasingly popular due to its convenience and low cost. However, it fails to be sufficiently robust to drastic motion disturbances such as continuous head movements in our real life. A motion-driven attention network (MANet) is proposed in this paper to improve its motion robustness. MANet takes the frequency spectrum of a skin color signal and of a synchronous nose motion signal as the inputs, following by removing the motion features out of the skin color signal using an attention mechanism driven by the nose motion signal. Thus, it predicts frequency spectrum without components resulting from motion disturbances, which is finally transformed back to a pulse signal. MANet is tested on 1000 samples of 200 subjects provided by the 2nd Remote Physiological Signal Sensing (RePSS) Challenge. It achieves a mean inter-beat-interval (IBI) error of 122.80 milliseconds and a mean heart rate error of 7.29 beats per minute.

1. Introduction

The pulse is a rhythmical throbbing of arteries resulting from heartbeat. It provides a wealth of information about the cardiovascular system, and plays an important role in disease prophylaxis and cardiac rehabilitation. Previous researches [1,2] have shown that the pulse signal can be remotely detected from a facial video using an optical biomonitor technique named video Photoplethysmography (VPPG). The color in facial skin changes subtly in response to the pulses due to the varying light absorption of the blood flowing beneath the skin. The VPPG technique captures the skin color changes in a video using a consumer-level camera, followed by detecting the pulse from the changes. This technique can be implemented on mobile devices such as cell phones and laptops, and frees users from the firm contact between the skin and sensors. With the rapid development of the computer vision and artificial intelligence, it is becoming increasingly practical and popular. Up to now, it has been used for detecting heart rate [3], oxyhemoglobin saturation [4], atrial fibrillation [5,6], etc.

However, the VPPG technique is not sufficiently robust to unsteady factors in the realistic environments such as face motions, weak illumination, and diverse skin tones. A va-

riety of methods have been proposed to improve its robustness, and they can be summarized as follows. (1) *Blind source separation* (including principal component analysis (PCA) [7], independent component analysis (ICA) [8] and singular spectrum analysis [9]) linearly decomposes the observed color signal into several basis vectors, from which the vectors with pulsating features are selected to constitute the pulse signal. (2) *Signal filtering* (including bandpass filtering [10], least mean square adaptive filtering [11] and homomorphic filtering [12]) separates the pulse signal out of the observed color signal based on the pulse prior information such as frequency range and temporal periodicity. (3) *Multi-region analysis* [13] extracts color signals from different facial areas, followed by combining the color signals with high signal-noise ratios into the pulse signal. (4) *Chromatic model* transforms the face video from RGB space to another chromatic space (including CHROM [14], SR [15], and POS [16]) where the pulse signal is furthest (orthogonal) to noise. (5) *Deep learning models* (including DeepPhys [17], STVEN-rPPGNet [18], RhythmNet [19]) learn the relationship between the skin color changes and the underlying pulse signal from a large number of samples, which have achieved noticeable progress with today's data explosion.

Above-mentioned methods can deal with certain kinds of face motions such as local expression changes and brief head motions, but fail to work effectively for drastic disturbances such as continuous head twisting. Furthermore, some disturbances with similar features to the pulse complicate the problem. For example, the motion signal resulting from the head twisting overlap with the underlying pulse signal in both the time and space domains, which is hard to separate by deep learning models or filters with an empirical setting. Thanks to the opportunity to participate in the 2nd Remote Physiological Signal Sensing (RePSS) Challenge, we propose a motion-driven attention network (MANet) for detecting the pulse signal from a facial video, with a focus on addressing the drastic motion disturbances in the testing videos provided by the organizers. In addition to skin color signal, the proposed model also takes the nose moving trail into account, and uses the nose motion signals to counteract the motion artifact in the skin color signal. The rest of the paper is structured as follows. The proposed model is elaborated in Section II. The experiments for evaluating the model are introduced in Section III, followed by a conclusion about this paper in Section IV.

2. Method

2.1. Pre-processing

In a recorded video, the face area is tracked in order to eliminate its rigid movements. The face tracking method introduced in [11] is used in this work. At first, the face rectangle on the first frame of the video is detected using the Viola-Jones face detector. Then, 66 facial landmarks inside the face rectangle is found through Discriminative Response Map Fitting (DRMF) method. Following the movement trajectory of the landmarks, the face rectangle is tracked through the rest of the video.

The chromatic space of the video is transformed from the RGB space to the CHROM space [14] to highlight the color changes due to the pulse. For each pixel, two color signal are computed as $\boldsymbol{X} = 3\boldsymbol{R} - 2\boldsymbol{G}$, and $\boldsymbol{Y} = 1.5\boldsymbol{R} + \boldsymbol{G} - 1.5\boldsymbol{B}$. Two signals are filtered in a bandpass (0.7-4.0 Hz) manner, and then combined into a signal as $\boldsymbol{Z} = \boldsymbol{X} - \alpha\boldsymbol{Y}$, where $\alpha = \sigma(\boldsymbol{X})/\sigma(\boldsymbol{Y})$ and σ refers to the standard deviation.

The cheek is chosen as the region of interest (ROI) as it is less affected by hair and talking. By connecting four facial landmarks around the cheek with straight lines, the ROI in each frame is marked off, where all the pixels are globally averaged. Thus, a temporal sequence which reveals the skin color changes can be constructed.

The temporal sequence is interpolated linearly into a color signal with 300 elements, for the purpose of signal length consistency. The color signal is then processed by wavelet decomposition method to remove noise outside the heart rate band. In this work, the raw color signal is decomposed into an approximate component a_5 and 5 detail components $d_1 \sim d_5$ using the Meyer wavelet, from which the fourth detail component d_4 (whose frequency band is around 0.7-2.5 Hz) is selected as the color signal that contains pulse information.

2.2. Pulse signal detection with MANet

In a video with facial movements, the color signal $\boldsymbol{c}(t)$ is a combination of the pulse signal $\boldsymbol{p}(t)$ and the motion signal $\boldsymbol{m}(t)$. The pulse signal can be obtained by removing the motion signal out of the color signal. Although the motion signal is unknown, it can be estimated from the moving trajectory of the nose. Specifically, after detecting the nose rectangle in the first frame of a tracked face video using the Viola-Jones nose detector, we calculate the location of the central point of the nose rectangle $\widetilde{m}_1$. Subsequently, the motion signal is estimated as the Euclidean distance between the central point of the first frame and those of the following frames, i.e., $\widetilde{\boldsymbol{m}}(t) = [\Delta\widetilde{m}_1, \Delta\widetilde{m}_2 ... \Delta\widetilde{m}_T]$, where $\Delta\widetilde{m}_t = |\widetilde{m}_t - \widetilde{m}_1|$. The estimated motion signal $\widetilde{\boldsymbol{m}}$ differs from the actual motion signal $\boldsymbol{m}$ in waveform, but they are similar in frequency (i.e., the frequency of head movement). Therefore, after computing the frequency spectrum of the estimated motion signal $\widetilde{\boldsymbol{M}}(f)$ and the color signal $\boldsymbol{C}(f)$ by Fourier transform, we plan to remove the motion signal under the guidance of the estimated motion signals from the frequency domain perspective.

It is implemented by MANet in this work, whose archi-

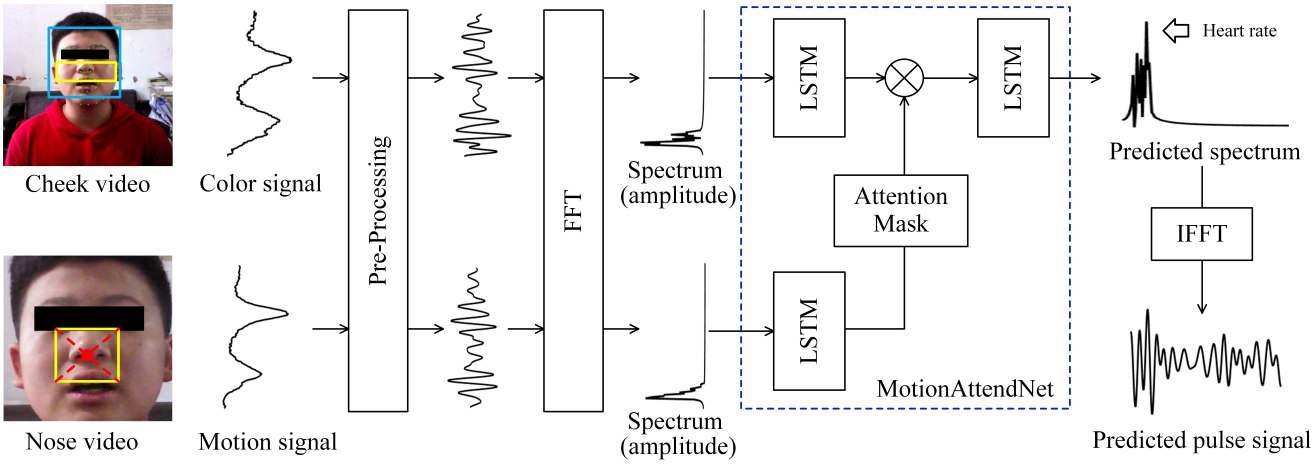


Figure 1: Framework of the proposed method. A color signal and a motion signal are first extracted from the cheek and noise regions in a face video respectively. Two signals are then pre-preprocessed and transformed into the frequency domain to acquire their spectrum. The spectrum of two signals is map into a latent feature space by MANet, where the motion features of two signals cancel each other out by an attention mechanism. The spectrum predicted by MANet is finally transformed back to the time domain to construct the pulse signal.

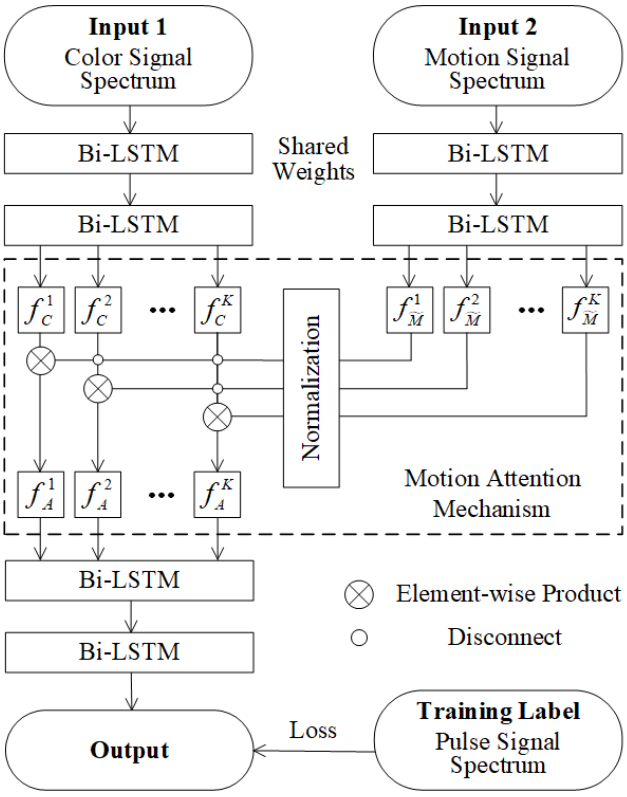


Figure 2: Architecture of MANet.

tecture is shown in Figure. 2. This model takes the frequency spectrum of the color signal $\boldsymbol{C}(f)$ and the estimated motion signal $\widetilde{\boldsymbol{M}}(f)$ as the inputs, and outputs the frequency spectrum of the pulse signal $\widehat{\boldsymbol{P}}(f)$. It is worth noting that the frequency spectrum of a vector is a complex vector which cannot be processed by common neural networks. Considering that the signal frequency information is mainly contained in the amplitude of the frequency spectrum, we divide the frequency spectrum into amplitude spectrum (the modulus of the frequency spectrum) and phase spectrum (the angles or orientations of the frequency spectrum), and use the amplitude spectrum only in the input and output of the model. For each input, two layers of Bidirectional Long Short Term Memory (Bi-LSTM) networks are used to extract its features at different frequency bands. The Bi-LSTM networks for two inputs have shared weights for the purpose of feature similarity. Each Bi-LSTM layer outputs a total of K hidden states. The features of the estimated motion signal spectrum $\boldsymbol{f}_{\widetilde{\boldsymbol{M}}} = \left[f^1_{\widetilde{M}}, f^2_{\widetilde{M}} \ldots f^K_{\widetilde{M}}\right]$ are then used as an attention mask overlaying on the features of the color signal spectrum $\boldsymbol{f}_{\boldsymbol{C}} = \left[f^1_C, f^1_C \ldots f^K_C\right]$, aiming to let the network assign higher weights for the part of the color signal spectrum that has less motion signal components. It can be accomplished by the following steps. (1) The features of the estimated motion signal $\boldsymbol{f}_{\widetilde{\boldsymbol{M}}}$ are reverse-normalized into a soft mask using a negative Sigmoid function,

$$\boldsymbol{f}_{\text{mask}} = \frac{-1}{1 + exp^{\boldsymbol{f}_{\widetilde{M}} - Ave\left(\boldsymbol{f}_{\widetilde{M}}\right)}} \tag{1}$$

where Ave refers to the mean, and $\boldsymbol{f}_{\text{mask}}$ is the soft mask. The weights in the mask range from 0 to 1, with higher weights indicating weaker motion features. (2) The features of the color signal spectrum are multiplied by the attention

mask in a element-wise manner,

$$\boldsymbol{F_A} = \boldsymbol{f_C} \otimes \boldsymbol{f}_{\text{mask}} \tag{2}$$

where $\otimes$ refers to the Hadamard product, and $\boldsymbol{F_A}$ is the masked features of the color signal spectrum. Through the attention mechanism, the masked color features have less motion components compared to the previous version.

The masked color features are fed into two layers of Bi-LSTM networks to predict the amplitude spectrum under the supervision of a label amplitude spectrum. The label is transformed from a clean PPG signal recorded in sync with the face video. The loss function for training the model is designed as the minimum mean square error (MSE) between the outputs and the labels,

$$Loss = \frac{1}{N}\sum\nolimits_{n=1}^{N}\left(\boldsymbol{P}_{\text{PPG}}\left(f\right) - \widehat{\boldsymbol{P}}\left(f\right)\right)^2 \tag{3}$$

where $\boldsymbol{P}_{\text{PPG}}\left(f\right)$ stands for the label amplitude spectrum of the PPG signal. One remaining problem is that the pulse signal cannot be reconstructed with the predicted amplitude spectrum only. In view of this, we combine the predicted amplitude spectrum with the separated phase spectrum of the color signal into the frequency spectrum, and then transform it back to the pulse signal $\hat{\boldsymbol{p}}\left(t\right)$ using inverse Fourier transform.

3. Experiments

3.1. Dataset

The model was trained on a self-collected dataset, and was tested on two datasets (OBF and VIPL-HR-V2) offered by the challenge organizers. The self-collected dataset was established by our team in collaboration with the first affiliated hospital of USTC (Anhui Provincial Hospital). It contained 470 subjects (including healthy individuals and patients with coronary heart disease, hypertension, and atrial fibrillation). For each subject, there were a 10-second static video where the subjects kept still and a 10-second dynamic video where the subjects performed typical face motions such as head twisting, speaking and nodding. All the videos were recorded at 30 frames per second (fps) with a resolution of 1080p using a Logitech C930 camera. Meanwhile, synchronous PPG signals viewed as the ground truth were detected by a pulse oximeter (ZJE PWS-20D) worn on the fingertips.

The OBF testing dataset included 100 subjects. For each subject, there were five 10-second videos with a resolution of 1080p at 30 fps. All the videos were recorded in static scenarios where the subjects had no facial motions. The challenge from the this dataset was the diverse skin tones. The VIPL-HR-V2 testing dataset included 100 subjects. Five 10-second videos with a resolution of 720p were recoded for each subject. All the videos were recorded in dynamic scenarios where the subjects talked and moved heads continuously. Furthermore, nearly half the videos were recorded in dim or uneven lighting, which increased the difficulty of the pulse signal detection.

3.2. Set up

In this work, 20% samples were separated from the training set, which was used as the validation to set the model hyperparameters. The lengths of the inputs and the output of MANet were 150 (half the length of the video) as only half of the signal spectrum was input. Each Bi-LSTM layer contained 15 hidden states in one direction, whose length was 10. The activation functions of all the Bi-LSTM layers were Relu. The parameters of MANet were trained by the Adam optimizer (with the initial learning rate at 0.1) based on the back propagation. The training epoch was 100, and a batch of samples (n=64) were used in each training iteration. The experimental data were processed by Tensorflow 2.0 and Matlab 2018A.

3.3. Evaluation Metrics

In the RePSS challenge, the performance of the proposed method was mainly evaluated by the inter-beat-interval (IBI) metrics [20] as follows,

1. *Mean of IBI error* (M_{IBI})

For the IBI curve of the estimated pulse signal $R\left(t\right)$ and the IBI curve of the ground truth $R_{\text{gt}}\left(t\right)$, the IBI error AE and its mean M_{IBI} can be computed as,

$$AE_{\text{IBI}} = \sum\nolimits_{t=0}^{T}\left|R\left(t\right) - R_{\text{gt}}\left(t\right)\right|, M_{\text{IBI}} = \frac{\sum_{k=0}^{K} AE_{\text{IBI}}\left(k\right)}{K} \tag{4}$$

where T is the time length of the IBI curves, and K is the number of videos.

2. *Standard deviation of IBI error* (SD_{IBI})

$$SD_{\text{IBI}}=\sqrt{\sum\nolimits_{k=0}^{K}\left[AE_{\text{IBI}}\left(k\right) - M_{\text{IBI}}\right]^2/K} \tag{5}$$

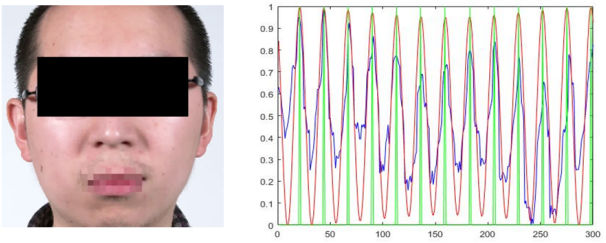
In addition, the heart rate metrics including *mean absolute error of heart rate* (MAE_{HR}), *root mean squared error of heart rate* ($RMSE_{\text{HR}}$), and *Pearson correlation coefficient of heart rate* (R_{HR}) were also used for reference.

3.4. Results

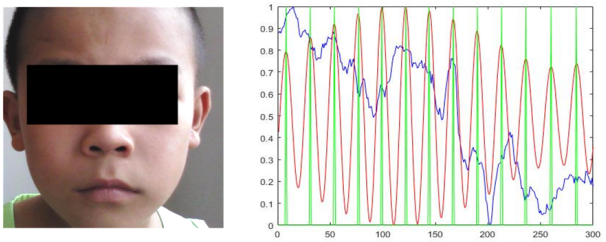
The experimental result on the testing set provided by the organizers was presented in Table I. Our team ranked 2nd out of 6 participating teams, with the mean IBI error M_{IBI} 4.7% larger than the best one, and 26.9% lower than the third one. Besides, our team achieved the second best SD_{IBI} and R_{HR}, and the best MAE_{HR} and $RMSE_{\text{HR}}$. Figure 3 shows the experimental result of MANet more visually with two examples. It can be clearly seen that two pulse signals (especially the one in VIPL-HR-V2) became more regular after being processed by MANet.

Team	M_{IBI} (ms)	SD_{IBI} (ms)	MAE_{HR} (bpm)	$RMSE_{\text{HR}}$ (bpm)	R_{HR}
TIME	**117.25**	**153.18**	7.31	11.44	**0.62**
Dr. L (Ours)	122.80	153.91	**7.29**	**11.05**	0.57
Anti-Spoofers	168.08	162.82	11.84	14.51	0.02
shankejinjiboy	224.41	163.98	15.44	18.75	-0.05
ZJUT-WTCrPPG	273.53	171.13	23.89	27.96	-0.03
ZJUT-ASTrPPG	295.70	175.24	29.24	33.69	-0.10

Table 1: Comparison among 6 teams in this challenge (best performance in bold).



OBF Dataset (static scenario)



VIPL-HR-V2 Dataset (dynamic scenario)

Figure 3: Pulse signals predicted by MANet. The blue curves were the observed color signal, the red curves were the predicted pulse signals, and the green binary signals indicate the locations of the signal peaks.

4. Conclusions

VPPG is an imaging technique for pulse signal detection from face videos, but its development is impeded by drastic motion disturbances. MANet is proposed in this paper to address this problem. It suppresses the motion components of the observed skin color signal in a spectral feature space using an attention mechanism driven by the nose moving signal, producing a pulse signal with less motion-induced distortions. The performance of MANet ranked 2nd out of 6 teams in the RePSS Challenge.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities of China (PA2021GDSK0071).

References

[1] Yu Sun and Nitish Thakor. Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE Transactions on Biomedical Engineering*, 63(3):463-477, 2016.

[2] Xun Chen, Juan Cheng, Rencheng Song, Yu Liu, Yu Liu and Z. Jane Wang. Video-Based Heart Rate Measurement: Recent Advances and Future Prospects. *IEEE Transactions on Instrumentation and Measurement*, 68(10):3600-3615, 2019.

[3] Xuenan Liu, Xuezhi Yang and Jin Jing. Self-adaptive signals separation for non-contact heart rate estimation from facial video in realistic environments, *Physiological Measurement*, 39(6):06NT01, 2018.

[4] Alessandro R. Guazzi, Mauricio Villarroel, Joo Jorge, Jonathan Daly, Matthew C. Frise, Peter A. Robbins and Lionel Tarassenko. Non-contact measurement of oxygen saturation with an RGB camera. *Biomedical Optics Express* 6(9)3320-3338, 2015.

[5] Bryan. P. Yan, William H. S. Lai, Christy K. Y. Chan, Stephen C. H. Chan, Lok-Hei Chan, *et al.* Contact-free screening of atrial fibrillation by a smartphone using facial pulsatile photoplethysmographic signals. *Journal of the American Heart Association*, 7:e008585, 2018.

[6] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio. Seppnen and Guoying. Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology* , 30(8):2781-2795, 2020.

[7] 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, February 2015.

[8] Ming-Zher Poh, Deniel J. Mcduff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using

video imaging and blind source separation. *Optics Express*, 18(10):10762-10774, 2010.

[9] Dingliang Wang, Xuezhi Yang, Xuenan Liu, Jing Jin, Shuai Fang. Detail-preserving pulse wave extraction from facial videos using consume-level camera. *Biomedical Optics Express*, 11(4), 2020.

[10] Ming-Zher 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.

[11] Xiaobai Li, Jie Chen, Guoying Zhao and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4264-4271, Columbus, USA, 2014.

[12] Xuenan Liu, Xuezhi Yang, Dingliang Wang and Alexander Wong. Detecting pulse rates from facial videos recorded in unstable lighting conditions: an adaptive spatiotemporal homomorphic filtering algorithm. *IEEE Transactions on Instrumentation and Measurement*, 70:1-15, 2021.

[13] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharwal. DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomedical Optics Express*, 6(5):1565-1588, 2015.

[14] Gerard D. Haan and Vincent Jeanne. Robust pulse rate from chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878-2886, 2013.

[15] Wenjin Wang, Sander Stuijk and Gerard D. Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974-1984, 2016.

[16] Wenjin Wang, Sander Stuijk, Gerard D. Haan, Bert D. Brinker. Algorithmic principles of remote PPG, *IEEE Transactions on Biomedical Engineering*, 64(7):1479-1491, 2017.

[17] Weixuan Chen and Daniel McDuff. DeepPhys: video-based physiological measurement using convolutional attention networks. *European Conference on Computer Vision (ECCV)*, pages 356-373, Munich, Germany, 2018.

[18] 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. *IEEE International Conference on Computer Vision (ICCV)*, pages 151-160, Seoul, Korea (South), 2019.

[19] Xuesong Niu, Shiguang Shan, Hu Han and Xilin Chen. RhythmNet: end-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29: 2409-2423, 2019.

[20] Xuenan Liu, Xuezhi Yang, Jing Jin and Alexander Wong. Detecting pulse wave from unstable facial videos recorded from consumer-level cameras: a disturbance-adaptive orthogonal matching pursuit. *IEEE Transactions on Biomedical Engineering*, 67(12):3352-3362, 2020.