Remote Heart Rate Estimation Based on Variational Mode Decomposition and Lip-Motion-Guided Artifact Removal

Boxiang Liu[1,2], Xiujuan Zheng[1,2,*] and Yue Ivan Wu[3]

[1]College of Electrical Engineering, Sichuan University, Chengdu 610065, China.
[2]Key Laboratory of Information and Automation Technology of Sichuan Province, Chengdu 610065, China.
[3]College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China

Abstract

Remote photoplethysmography (rPPG) enables non-contact heart rate (HR) monitoring and holds great promise for applications in health monitoring, emotion recognition, and beyond. However, motion introduces substantial noise within the HR frequency band, which is a primary cause of rPPG performance degradation. To enhance the robustness of rPPG in complex scenarios, we propose a signal-processing-based framework for remote HR estimation. First, variational mode decomposition (VMD) is introduced to achieve effective separation of noise and pulse components in the rPPG signal. Then, the main sources of motion artifacts are analyzed, and motion information is derived based on the positional variations of the lip landmark in the video frames. By leveraging time-delay analysis, motion-induced noise components in the rPPG signals are accurately removed. Finally, principal component analysis (PCA) is applied to reconstruct the heartbeat component from the remaining signal set, and HR estimation is further refined by exploiting the temporal continuity of physiological parameters. Using the proposed method, we achieved third place in the 4th Remote Physiological Signal Sensing (RePSS) Challenge.

Keywords

Remote photoplethysmography, HR estimation, signal separation, noise removal, pulse signal reconstruction, HR refinement

1. Introduction

HR is a fundamental indicator of human physiological status. Traditional HR monitoring methods mainly include electrocardiography (ECG) and photoplethysmography (PPG), both of which require direct contact with the skin. However, long-term use of such contact-based methods may cause discomfort or skin irritation, making them unsuitable for sensitive populations such as infants or patients with skin damage. As a result, rPPG technology, which enables non-contact HR measurement, has attracted increasing attention in recent years.

Human heartbeat causes fluctuations in blood volume within blood vessels, and these fluctuations affect the light absorption characteristics of the vessels, which manifest as rhythmic changes in skin color. The rPPG technology extracts the pulse signal based on changes in skin color. This technique enables unobtrusive and long-term monitoring of varied physiological parameters[1, 2], making it suitable for daily personal health management. However, a major obstacle to the practical deployment of rPPG technology is its high sensitivity to illumination changes and human motion, which significantly degrades its measurement accuracy.

In the early stages of rPPG technology development, commonly used methods included blind source separation (BSS), signal decomposition, and color space transformation. Poh et al. [3] employed independent component analysis (ICA) to separate the pulse component from RGB channel signals. Song et al. [4] applied ensemble empirical mode decomposition (EEMD) to decompose rPPG signals from several facial sub-regions. Wang et al. [5] introduced the Plane-Orthogonal-to-Skin (POS) method to suppress illumination variations and specular reflection interference. Nevertheless, these traditional methods generally overlooked the importance of accurate signal decomposition and explicit extraction

*Corresponding author.
liubx@stu.scu.edu.cn (B. Liu); xiujuanzheng@scu.edu.cn (X. Zheng); y.i.wu@ieee.org (Y. I. Wu)
0009-0007-8620-5769 (B. Liu); 0000-0002-4703-9530 (X. Zheng); 0000-0001-5480-1741 (Y. I. Wu)

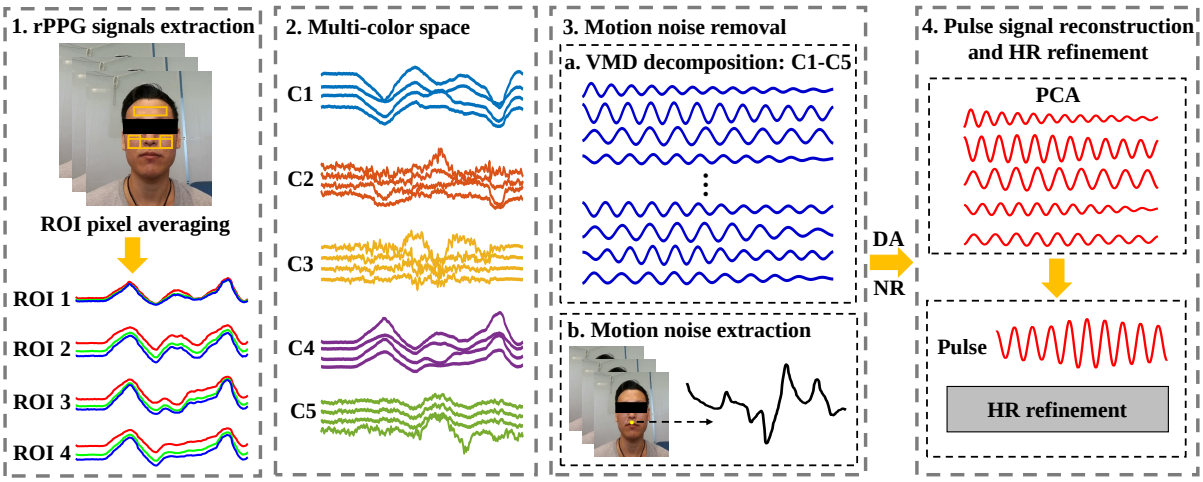

Figure 1: Overall framework of the proposed algorithm. C1–C5 represent the NIR-RGB, CHROM, POS, Green, and NIR color channels, respectively. DA denotes delay analysis, and NR indicates noise removal.

of noise source information, making them susceptible to noise interference and performance degradation in practical applications.

In recent years, deep learning methods have rapidly advanced and have been increasingly studied in the field of rPPG. Bousefsaf et al. [6] employed a 3D convolutional neural network (3D CNN) to extract features directly from raw video and used a multilayer perceptron to regress HR. Lokendra et al. [7] incorporated facial action units (AUs) along with temporal signals from various triangular ROIs into a multi-channel temporal convolutional network (TCN) to denoise and enhance estimation accuracy. However, these deep learning–based approaches rely on large-scale annotated datasets and suffer from limited interpretability.

We propose an algorithmic framework that integrates multi-color space signal decomposition, noise removal, pulse reconstruction, and HR refinement. The main contributions of this work are as follows:

1) We introduced VMD into the rPPG field, which significantly improved the accuracy of rPPG signal decomposition and laid the foundation for subsequent accurate noise removal and pulse reconstruction.

2) For the first time, motion reference noise is extracted based on the positional variations of lip landmarks, and motion-induced noise is accurately identified in complex scenarios by exploiting the common source characteristics shared between the motion components in rPPG signals and the reference noise.

3) A temporal continuity constraint of physiological parameter variation is incorporated into the HR estimation process, further enhancing the stability of the estimated HR.

2. Method

The proposed algorithm consists of four main steps, as illustrated in Fig. 1. First, the regions of interest (ROIs) are segmented, and the raw rPPG signals are extracted. Second, the ROI signals are transformed into multiple color spaces. Third, each color-space signal is decomposed using VMD. Time-delay analysis is then performed between the decomposed components and the reference noise to eliminate motion artifacts. Finally, PCA is applied to reconstruct the pulse waveform, and a temporal continuity constraint is used to refine the HR estimation.

2.1. Regions of Interest Segmentation and Raw rPPG Signal Extraction

We employed the open source MediaPipe Face Mesh [8] to obtain facial landmarks. Four ROIs were then defined on areas of the face rich in capillaries, as shown in Fig. 1. We perform spatial pixel averaging within each ROI of every frame to obtain the corresponding raw rPPG signals.

2.2. Color Space Conversion

In this study, we employed five color spaces that are favorable for pulse extraction, including CHROM [9], POS [5], Green, and NIR. In addition, we proposed a fusion scheme that integrates NIR and RGB.

The variations in light intensity at the facial ROI locations are simultaneously reflected in both the visible and NIR channels. Based on the derivation in Ref. [5], we propose projecting the normalized RGB and NIR data onto a direction orthogonal to the illumination variation, as shown in Eq. (1).

$$S_{\mathrm{f}} = 3 \times NIR_{\mathrm{n}} - R_{\mathrm{n}} - G_{\mathrm{n}} - B_{\mathrm{n}} \tag{1}$$

where NIR_{n}, R_{n}, G_{n}, and B_{n} represent the normalized NIR and RGB channel signals, respectively, and S_{f} denotes the fused signal. Based on Eq. (1), the RGB and NIR data can be effectively fused while suppressing illumination variations.

2.3. Motion Noise Removal

2.3.1. Variational Mode Decomposition

VMD is a fully non-recursive signal decomposition method that decomposes a signal by formulating and solving a constrained variational problem. It decomposes the input signal into k intrinsic mode functions (IMFs), and it effectively avoids mode mixing. The details of the VMD algorithm can be found in Ref. [10]. VMD decomposes the signals from multiple color spaces into a set containing both pulse and motion noise components. Therefore, it is necessary to remove the motion components.

2.3.2. Motion Noise Extraction

Human motion can be categorized into rigid and non-rigid movements: rigid motion is associated with large-scale head movements, while non-rigid motion relates to deformations of local facial regions. From a signal processing perspective, if reference noise correlated with these movements can be identified, it becomes possible to accurately remove motion components from the rPPG signal.

Non-rigid movements are typically associated with mouth motion, while rigid movements can also be reflected in the displacement of the mouth. Therefore, we extract facial motion information based on the relative positional changes of lip landmarks in the image. In this study, we select the landmark located at the center of the upper lip. For a given video sequence, a time series related to the distance between the landmark and the origin can be obtained, which we refer to as reference noise.

For motion noise in the rPPG signal, it shares a common source with the reference noise. Therefore, their waveform variations exhibit a high degree of similarity, i.e., they have a small time delay. Based on this, we analyze the time delay τ between each VMD-decomposed signal and the reference noise to identify motion-related components. Due to noise interference or system errors, the time delay τ between the motion component in the rPPG signal and the reference noise may exhibit slight deviations around zero. Based on experiments, this paper considers decomposed components with cross-correlation peaks satisfying $|\tau| \leq 0.2$ s as motion noise and removes them accordingly.

2.4. Pulse Signal Reconstruction and HR Refinement

After motion noise removal, the signal set $\boldsymbol{S}_{\mathrm{r}}$ is considered to primarily contain pulse components along with some random noise. PCA is capable of extracting the common signal components shared across multiple channels into the first principal component, while suppressing weakly correlated random noise. Based on this, we apply PCA to $\boldsymbol{S}_{\mathrm{r}}$ to reconstruct the dominant pulse component. For the reconstructed pulse signal, the HR frequency f_{hr} can be calculated using the FFT, and then multiplying f_{hr} by 60 yields the HR value.

To further improve the accuracy of HR estimation, we developed a refinement scheme based on the continuity of HR variation. Since physiological changes in the human body occur gradually, the HRs estimated from three consecutive temporal segments are expected to exhibit relatively small differences.

Assuming the estimated HRs for these three segments are hr_1, hr_2, and hr_3, the condition specified in Eq. (2) should be satisfied.

$$\forall i \in \{1,\ 2\}, \quad |hr_{i+1} - hr_i| < th_1 \tag{2}$$

In the proposed algorithm, the interval between two signal segments is 0.2 seconds, and the threshold th_1 is set to 10 beats per minute (bpm). In addition, if the HR exceeds 122.6 bpm or falls below 35.6 bpm, we introduce an additional constraint for HR refinement. Specifically, the three signal segments are slid with a step size of one frame. If the HR difference between each segment and its corresponding initial segment is less than 5 bpm, the obtained three signal segments will be used for the final HR estimation.

3. Experiments

3.1. Datasets

The training set for this challenge is a subset of the VIPL-HR dataset [11], containing data from 42 subjects. To match the test set, we selected modality (b) and modality (c) for our experiments, which consist of paired RGB and NIR videos recorded using a RealSense F200 camera. The RGB images have a resolution of 960×540, while the NIR images are 640×480; both are recorded at 25 fps. We used 10 seconds as the signal length for HR estimation in the training set.

The test set includes portions of the VIPL-HR and OBF datasets [12]. The VIPL-HR test set consists of RGB and NIR videos from 100 subjects, with an average video length of 9.74 seconds. The OBF dataset also includes RGB and NIR videos from 100 subjects, with each video segment lasting 10 seconds.

3.2. Performance Metrics

We use mean absolute error (MAE), root mean square error (RMSE), mean error (ME), and Pearson correlation coefficient (PCC) to evaluate the performance of the algorithm.

4. Results

We conducted experiments on the challenge training set. Taking the motion scenario as an example, the raw rPPG signal extracted from the face is presented in the upper subfigure of Fig. 2(a). The original rPPG signal is severely interfered by motion noise. Then, the NIR and RGB channels are fused using Eq. (1), and the filtered signal is shown in the lower subfigure of Fig. 2 (a). This signal is decomposed using VMD and compared with the results of EEMD and wavelet decomposition, as shown in Fig. 2(b), (c), and (d). In Fig. 2(d), wavelet decomposition is disturbed by high-frequency noise (IMF1 at 11.3 Hz). Obvious mode mixing occurs in IMF2, and IMF3 contains only 1.2 Hz noise. This method fails to separate the 1.4 Hz pulse component. In Fig. 2(c), EEMD shows severe mode mixing in IMF2, with multiple signal components blended together, while IMF1 and IMF3 correspond to 1.2 Hz noise. In contrast, VMD avoids mode mixing, with each mode displaying a distinct, sharp frequency peak that enables accurate separation of noise and pulse signals.

Using the method described in Section 2.3.2, reference noise is extracted from the lip landmark, and the filtered result is shown in Fig. 3(a). This reference noise has the same frequency as IMF3 in Fig. 2(b). Then, the cross-correlation function between the reference noise and IMF3 is calculated, as shown in Fig. 3(b). The highest correlation occurs at a time delay close to zero, approximately -0.2 s. Therefore, IMF3 is identified as motion noise and removed. PCA is then used to reconstruct the pulse signal, and the result is shown in Fig. 3(c). The reconstructed pulse signal exhibits a rhythm consistent with the ground truth, and the estimated HR is 84 bpm, matching the ground truth.

HR estimation was performed on the entire training set and compared with conventional methods (EEMD and wavelet) and the end-to-end method from the relevant challenge, as shown in Table 1. Conventional methods decompose the green channel signal and estimate HR using the component with

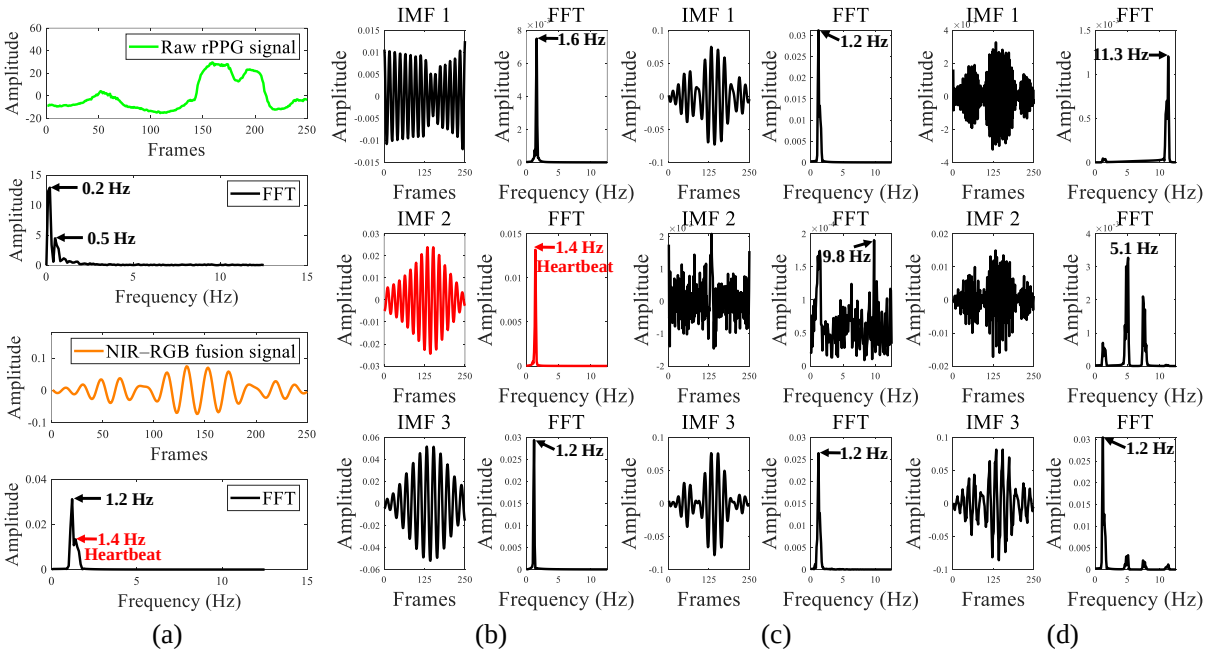

Figure 2: (a) The raw rPPG signal and the NIR-RGB fusion signal. (b), (c), and (d) show the decomposition results of VMD, EEMD, and wavelet, respectively. FFT represents the result of the Fast Fourier Transform applied to the corresponding time-domain signal.

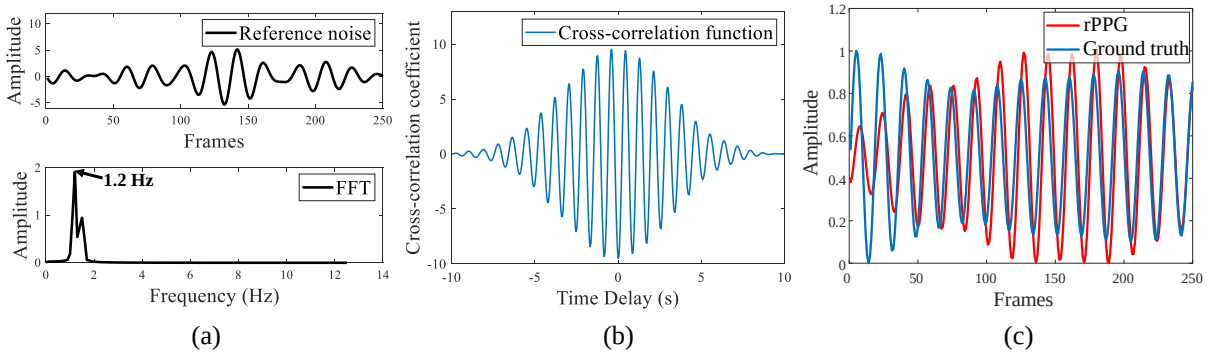

Figure 3: (a) Reference noise (b) Cross-correlation function between reference noise and IMF3 (c) Pulse signal reconstructed by the rPPG-based method (red) and the ground truth signal (dark blue).

the highest dominant frequency amplitude. The results verify that the proposed method effectively reduces HR estimation errors and improves PCC.

Table 1
HR estimation results of VIPL-HR training set.

Methods	MAE	ME	RMSE	PCC
End-to-end method [13]	10.97	–	13.24	0.06
EEMD	9.13	-5.32	11.64	0.07
Wavelet	9.24	-5.64	11.77	0.07
Proposed	7.50	-3.44	10.45	0.22

5. Conclusion

Methods based on EEMD and wavelet decomposition are prone to mode mixing, whereas VMD can more effectively separate noise and pulse components, making it more suitable for rPPG tasks. Conventional

methods often regard the lips as non-rigid motion regions to be excluded, overlooking the noise information related to the rPPG signal contained in this region. In this work, we analyze the sources of motion noise and establish a link between variations in lip landmark positions and motion noise in the rPPG signal. We propose a time-delay analysis-based method to identify motion noise, and successfully locate the motion component in the VMD-decomposed signals. This provides new insight for improving the signal-to-noise ratio (SNR) of the rPPG signal in related research. As shown in Table 1, the experimental results indicate that the proposed algorithm achieves better performance than both the end-to-end method and the conventional methods. Finally, our algorithm achieved third place in the 4th RePSS challenge with a score of 12.7079.

6. Declaration on Generative AI

During the preparation of this work, the authors used DeepSeek, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Liu, X. Zheng, Y. Ivan Wu, Remote heart rate estimation in intense interference scenarios: A white-box framework, IEEE Transactions on Instrumentation and Measurement 73 (2024) 1–14. doi:10.1109/TIM.2024.3419088.

[2] X. Zheng, B. Zou, C. Zhang, H. Tu, Remote blood pressure estimation using bvp signal features from facial videos, Pattern Recognition Letters 193 (2025) 122–127. URL: https://www.sciencedirect.com/science/article/pii/S0167865525001400. doi:https://doi.org/10.1016/j.patrec.2025.04.010.

[3] M.-Z. Poh, D. J. McDuff, R. W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation., Opt. Express 18 (2010) 10762–10774. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-18-10-10762. doi:10.1364/OE.18.010762.

[4] R. Song, J. Li, M. Wang, J. Cheng, C. Li, X. Chen, Remote photoplethysmography with an eemd-mcca method robust against spatially uneven illuminations, IEEE sensors journal 21 (2021) 13484–13494.

[5] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491. doi:10.1109/TBME.2016.2609282.

[6] F. Bousefsaf, A. Pruski, C. Maaoui, 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video, Applied Sciences 9 (2019) 4364. URL: https://www.mdpi.com/2076-3417/9/20/4364. doi:10.3390/app9204364.

[7] B. Lokendra, G. Puneet, And-rppg: A novel denoising-rppg network for improving remote heart rate estimation, Computers in Biology and Medicine 141 (2022) 105146. URL: https://www.sciencedirect.com/science/article/pii/S0010482521009409. doi:https://doi.org/10.1016/j.compbiomed.2021.105146.

[8] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, Real-time facial surface geometry from monocular video on mobile gpus, 2019. URL: https://arxiv.org/abs/1907.06724. arXiv:1907.06724.

[9] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg., IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886.

[10] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE Transactions on Signal Processing 62 (2014) 531–544. doi:10.1109/TSP.2013.2288675.

[11] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Transactions on Image Processing 29 (2020) 2409–2423. doi:10.1109/TIP.2019.2947204.

[12] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, M. Tulppo, G. Zhao, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 242–249. doi:10.1109/FG.2018.00043.

[13] C. Hu, K.-Y. Zhang, T. Yao, S. Ding, J. Li, F. Huang, L. Ma, An end-to-end efficient framework for remote physiological signal sensing, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 2378–2384. doi:10.1109/ICCVW54120.2021.00269.