A Transformer-based method for non-contact heart rate estimation from facial videos recorded in realistic environment[⋆]

Yi Hu[1], Xuenan Liu[1,*], Xinlong Rao[1] and Bojing Li[2]

[1] *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, Anhui, China*
[2] *School of Software, Hefei University of Technology, Hefei, 230601, Anhui, China*

Abstract

The Video Photoplethysmography (VPPG) technique, while increasingly popular due to its convenience and cost-effectiveness, faces challenges in handling continuous head movements and vigorous motion interferences encountered in real-life scenarios. In this paper, we present a Transformer-based approach aimed at enhancing the robustness of heart rate estimation from facial videos. Leveraging the self-attention mechanism inherent in Transformers, our method adeptly captures both temporal dependencies and spatial information, thereby elevating the accuracy and resilience of heart rate estimation, even in challenging conditions. Through extensive experiments conducted on real-world face video datasets, we illustrate the effectiveness of our approach. Our results demonstrate significant improvements over existing methods in mitigating motion artifacts and enhancing the reliability of non-contact heart rate estimation in practical environments.

Keywords

Transformer, Video Photoplethysmography, heart rate detection, non-contact type

1. Introduction

The pulse is one of the physiological rhythms in our body, and much information about the state of the body can be obtained by observing its frequency, regularity and intensity. An abnormal pulse may indicate underlying health problems such as arrhythmia, tachycardia or bradycardia. While VPPG technology (video Photoplethysmography) monitors pulse and heart rate through optical sensors [1, 2] and has a wide range of applications including health monitoring devices, clinical diagnostics, exercise physiology and biometrics. Using changes in optical signals, cardiovascular health information is obtained in real time, providing an effective tool for medical treatment, health tracking, and identity verification.Although VPPG technology is widely used, it suffers from shortcomings such as accuracy being affected by the environment, being less precise than ECG, and complex data processing.

The 3rd Vision-based Remote Physiological Signal Sensing(RePSS)Workshop

[⋆] This work is supported by Young Teachers' Research and Innovation Programme (JZ2024HGQA0113).
[*] Corresponding author.
[†] These authors contributed equally.

✉ 2020213430@mail.hfut.edu.cn (Y. Hu); 2024800001@hfut.edu.cn (X. Liu); 2021214654@mail.hfut.edu.cn (X. Rao); 2021215259@mail.hfut.edu.cn (B. Li)
0000-0002-5357-3000 (Y. Hu); 0000-0002-6804-964X (X. Liu); 0009-0004-7870-0186 (X. Rao); 0009-0001-2449-1778 (B. Li)

Existing research work mainly focuses on the motion interference problem, which can be summarised into the following categories. **1.Optimal facial region selection.** The face is divided into several sub-regions, and the optimal detection region is evaluated by analysing the intensity of skin colour changes in different regions [3] and the degree of influence by motion [4]. This type of method can remove local motion interference such as speech and expression changes, but it is difficult to effectively deal with global motion interference such as head bobbing. **2.Spatial decomposition of pulse representation.** Starting from the principle of skin reflection and transmitted light, the decomposition models of pulse signal and motion signal in orthogonal chromaticity space are investigated, including CHROM [5], 2SR [6] and POS [7]. These ideal models have limited ability to cope with the complex mixing of pulse and motion signals. **3.Pulse signal filtering.** According to the range and characteristics of human pulse rate variation, band-pass filtering [8], wavelet decomposition [9], minimum mean square error filtering [10], etc. are used to suppress noise signals other than heartbeat frequency, but it is difficult to separate the interference components with similar frequency characteristics. **4.Blind source separation of pulse signals.** According to the time-domain statistical properties of pulse signals, methods such as independent component analysis [11] and sparse representation [12] are used to construct pulse substrates and fit them to reconstruct distorted pulse signals. Due to the limited descriptive ability of such substrates, the separation of motion interference signals is not obvious.

While the application of deep learning methods has become the main direction of current pulse signal extraction research, Contrast-Phys [13] used an unsupervised learning method to generate multiple rPPG signals from different spatio-temporal locations in each video using a 3D convolutional neural network (3DCNN) model trained with contrast loss, Contrast-Phys+ [14] used a 3DCNN model to generate multiple spatio-temporal rPPG signals and incorporates a priori knowledge of rPPG into the contrast loss function, Privacy-phys [15] A new approach based on pre-trained 3D convolutional neural networks for modifying rPPG in facial videos for privacy preservation, MTTS-CAN [16] combines a temporal displacement module, an attentional module, and a multitasking mechanism to improve accuracy and efficiency, PhysNet [17] uses a deep spatio-temporal convolutional network to recover remote photovoltaic volumetric pulsogram (rPPG) signals from face videos, which can reveal the potential separability of pulse signals from motion signals driven by training data.

The attention mechanism of Transformers excels in handling noisy signals. It enables the network to establish better connections between different parts of the signal, effectively distinguishing noise and preserving essential features of the pulse signal.

2. Method

2.1. Pre-processing

Face detection and tracking is crucial since slight head movements of the subject are inevitable in practical applications. In the recorded video, the face region is tracked in order to eliminate the rigid motion of the face region. In this paper, we use the facial tracking method introduced in [18]. We used the SDK provided by MediaPipe to implement this facial tracking functionality.

The chromaticity space of the video is converted from RGB space to CHROM space [5] to

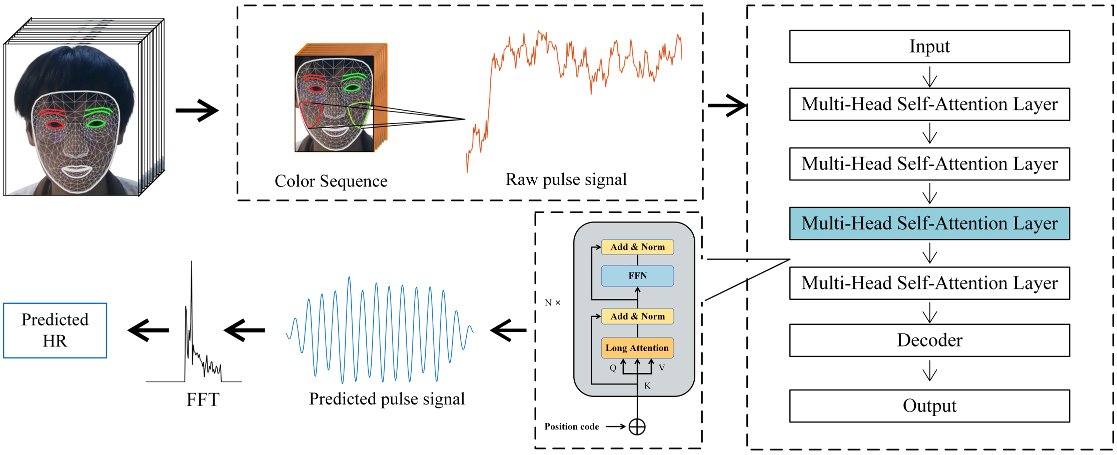


Figure 1: Our method aims to accurately estimate heart rate from facial videos in real environments. We begin by recording the video and conducting preprocessing, which involves noise and blur removal, as well as facial detection and tracking. Subsequently, we extract facial features and construct a Transformer model, incorporating a self-attention mechanism and a feed-forward neural network. Following this, we derive an output estimate by training the model to minimize the heart rate estimation error. Finally, we perform additional processing to obtain the final heart rate values.

highlight the colour changes due to impulses. For each pixel, two colour signals were computed X = 3R - 2G and Y = 1.5R + G - 1.5B. The two signals were filtered in a band-pass (0.7-4.0Hz) manne and then combined to form the $Z = X - \alpha Y$ signal, where $\alpha = \sigma\left(X\right)/\sigma\left(Y\right)$ and σ is the standard deviation.

Defining the ROI follows two rules: the first rule is to exclude the eye region because blinking may interfere with the estimated HR frequency; the second is to indent the ROI boundary with the face boundary. Therefore, the cheeks were chosen as the region of interest (ROI), which is less affected by hair and speech. The ROI is labelled in each frame by connecting the four facial marker points around the cheeks with straight lines, where all pixels are globally averaged. Thus, a time series showing changes in skin colour can be constructed.

The time series was linearly interpolated into a 300-element colour signal to achieve signal length consistency. The colour signal is then processed using wavelet decomposition methods to remove noise outside the heart rate band. In this paper, the Meyer wavelet is used to decompose the original colour signal into an approximate component a_5 and five detail components $d_1 \sim d_5$, from which the fourth detail component d_4 as the colour signal containing the pulse information.

2.2. Pulse signal detection with Transformer

The pre-processed pulse signal contains noise, in order to extract the pulse wave signal accurately we use Transformer network. The function of this network is to receive the preprocessed pulse signal as input and output the pulse wave signal after removing the noise.The Transformer network is able to efficiently capture the long range dependencies in the signal and improve the accuracy and generalisation of the signal extraction, which enables us to analyse the pulse signal more reliably.

Since pulses are periodic and consistent, whereas noisy signals lack such consistent characteristics, Transformer's attention mechanism can be advantageous when dealing with noise-containing signals. This mechanism allows the network to better establish connections between different parts of the signal, thus effectively distinguishing noise and preserving important features of the pulse signal.

Transformer has achieved remarkable success through its unique self-attention mechanism and positional coding. This model mainly consists of encoder and decoder. In Transformer encoder, multi-head attention and fully connected feed forward network layer are the main components.

The attention mechanism first maps the feature vectors to different linear spaces to obtain three different vectors: the Queries vector (Q), the Keys vector (K), and the Values vector (V), and then obtains the attention vector according to Equation (1):

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V \tag{1}$$

where d_k denotes the dimension of the vector (K).

The multi-head attention mechanism feeds the input vectors to multiple parallel attention mechanisms for computation, splices the output vectors, and then maps them back to the space of the original input vectors to obtain the final attention vectors. The specific calculation is shown in the following equations (2) and (3):

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_h)W^0 \tag{2}$$

$$\mathrm{head}_i(Q,K,V) = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where Concat denotes the splicing of multiple matrices in a certain dimension, h denotes the number of parallel Attention operations, head_i denotes the computation of the i Attention, and W_i^Q , W_i^K and W_i^V are the mapping matrices in the i head.

Without introducing an attention mechanism, we first encode the pulse signal to obtain a feature vector V and then decode this feature vector V to generate the optimised pulse signal.

After introducing the attention mechanism, we first encode the pulse signal to obtain the feature vector V. Next, we multiply V with the attention score A computed from the product of Q and K to obtain the weighted feature vector AV. Finally, we decode AV to generate the optimised pulse signal. Since the attention mechanism assigns a higher weight Ai to the segment Vi with periodicity, while the distorted segment Vj is assigned a lower weight Aj, the pulse signal obtained by decoding AV is usually better than the one obtained by decoding V directly.

2.3. Heart Rate Measurement

An interpolation Fourier transform (IFT) [19] is implemented on the reconstructed iPPG signals to obtain its high-resolution frequency spectrum, from which the average heart rate can be detected using a peak detecting procedure. This process can be formulated as follows,

$$HR_{\mathrm{Hz}} = \arg\max_{f} \theta(f) \tag{4}$$

where $\theta(f)$ stands for IFT of the reconstructed iPPG signal. Finally, HR_{HZ} is multiplied by 60 to obtain HR_{bpm}, a heart rate measurement in the standard unit.

3. Experiments

3.1. Dataset

The model was trained on the UBFC-rPPG dataset, which records video at 30 frames per second, 640x480 resolution, in uncompressed 8-bit RGB format, while reference data, such as PPG waveforms and heart rate, were recorded using a CMS50E Transmissive Pulse Oximeter.

For testing, the model was evaluated on two datasets provided by the challenge (OBF and VIPL-HR-V2).The OBF dataset contains 500 videos of 100 subjects, 10 seconds each, with a resolution of 1080p and a frame rate of 30 frames per second, all recorded in static scenes, with the main challenge being the difference in subjects' skin colour. The VIPL- HR-V2 dataset also contains 500 videos of 100 subjects, each video is 10 seconds long, with a resolution of 720p and a frame rate of 13-25 fps, the videos are recorded in dynamic scenes where subjects perform actions such as talking and shaking their heads.

3.2. Set up

This work utilizes a Transformer model for training, with 20% of the training set reserved for validation. The model's input and output lengths are set to 150, capturing half of the signal spectrum. Each attention head of the Transformer has 15 hidden units, employing a multi-head self-attention mechanism with ReLU activation. Training parameters are optimized using Adam optimizer with an initial learning rate of 0.1 and updated via backpropagation over 100 epochs. Training iterations involve batches of 64 samples. The experimental setup utilizes Tensorflow 2.0 and Matlab 2022b for data processing.

3.3. Assessment of indicators

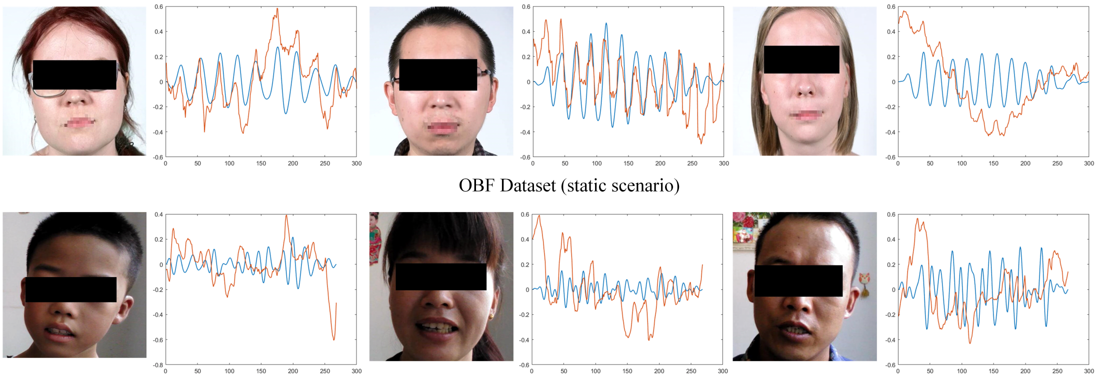
In the RePSS challenge, the performance of the proposed method is evaluated using the Root Mean Square Error (RMSE) as a metric to calculate the RMSE between the ground truth heart rate, y, and the measured value, y'.The RMSE reflects the extent to which the measured data is far away from the true value and measures the standard deviation of the residuals. The calculation is shown below:

$$rmse_{HR} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y - y')^2} \tag{5}$$

Where y is the true heart rate value and y' is the detected value of heart rate.

3.4. Comparative Test

Table 1 presents the experimental results on the test set provided by the organizer. Our method achieved a Root Mean Square Error (RMSE) of 11.77657 on the test set, which is 22.9% lower


OBF Dataset (static scenario)

VIPL-HR-V2 Dataset (dynamic scenario)

Figure 2: The predicted pulse signal from the method. The blue curve is the predicted pulse signal and the red curve is the original pulse signal.

Table 1
Comparison of the 13 teams in this challenge

Team	$rmse_{HR}$(bpm)	Team	$rmse_{HR}$(bpm)	Team	$rmse_{HR}$(bpm)
Face AI	8.50693	HFUT-VUT	8.85277	PCA_Vital	8.96941
Hash Brown	9.26198	AIIA	9.28902	SHDMIC	10.74201
HFUT-BCDH	11.77657	NeuroAI_KW	14.47930	NUIST	15.79680
SCUT_rPPG	15.88228	b7	19.06485	FulgenceWen	21.48006
Rhythm	24.02410				

than the 8th place. Fig. 2 illustrates the experimental outcomes with six examples. Clearly, the pulse signals, particularly in VIPL-HR-V2, exhibit enhanced regularity post-processing with our method. In the static dataset (OBF dataset), the subject remains stationary, and the raw pulse signals (red curve) show steady fluctuations. Our method accurately captures heart rate variations and extracts periodic heart rate signals from the stable signal. In the dynamic dataset (VIPL-HR-V2 dataset), movements such as head rotation and nodding induce disturbances in the predicted pulse signal, resulting in larger fluctuations in the observed raw pulse signal (red curve). However, our method effectively filters motion interference and accurately extracts the heart rate signal, as evident from the predicted pulse signal (blue curve) in the figure.

3.5. Ablation Test

To verify the effectiveness of the model proposed in this article, we compared the test results obtained directly using Fourier transform after preprocessing with the results obtained by adding the model proposed in this article. As shown in Table 2, the proposed model method performs better in this experiment, confirming the effectiveness of the model.

Table 2
Ablation test of the proposed method in this paper

Preprocessing	$rmse_{HR}(bpm)$
Preprocessing-FFT	24.70581
Preprocessing-our model-FFT	11.77657

3.6. Limitation

(1) Nearly half of the videos in this test were recorded under dim or uneven lighting, which increases the difficulty of pulse signal detection and makes the detection accuracy of this paper's method somewhat compromised. The approaches described in [20] and [21] offer potential preprocessing steps to address light-related issues. In our future work, we will also integrate a module into the proposed model to mitigate light-induced interference.

(2) Although the model was trained using the UBFC-rPPG dataset, there may be a problem of insufficient dataset size. A smaller dataset size may cause the model to overfit and not generalise well to new, unseen data, making the root mean square error (RMSE) larger.

(3) Although the dataset includes healthy individuals and patients with different diseases, there may be insufficient data on some specific groups, such as people of different ages and ethnic backgrounds. This may limit the applicability of the model on these groups.

4. Conclusions

We propose a method to address the challenges of VPPG in the face of violent motion disturbances using Transformer technology. By feeding the features extracted by the VPPG into the Transformer model for sequence modelling, we are able to capture long distance dependencies between input sequences. This approach promises to suppress motion interference in real-time or offline scenarios, thereby improving the accuracy and stability of VPPG in detecting impulse signals in face videos.

References

[1] Y. Sun, N. Thakor, Photoplethysmography revisited: from contact to noncontact, from point to imaging, IEEE transactions on biomedical engineering 63 (2015) 463–477.

[2] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, Z. J. Wang, Video-based heart rate measurement: Recent advances and future prospects, IEEE Transactions on Instrumentation and Measurement 68 (2018) 3600–3615.

[3] M. Kumar, A. Veeraraghavan, A. Sabharwal, Distanceppg: Robust non-contact vital signs monitoring using a camera, Biomedical optics express 6 (2015) 1565–1588.

[4] R. Amelard, D. A. Clausi, A. Wong, Spectral-spatial fusion model for robust blood pulse waveform extraction in photoplethysmographic imaging, Biomedical optics express 7 (2016) 4874–4885.

[5] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rppg, IEEE transactions on biomedical engineering 60 (2013) 2878–2886.

[6] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE transactions on biomedical engineering 63 (2015) 1974–1984.

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

[8] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE transactions on biomedical engineering 58 (2010) 7–11.

[9] D. Wang, X. Yang, X. Liu, J. Jing, S. Fang, Detail-preserving pulse wave extraction from facial videos using consumer-level camera, Biomedical optics express 11 (2020) 1876–1891.

[10] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 4264–4271.

[11] W. Wang, S. Stuijk, G. De Haan, Exploiting spatial redundancy of image sensor for motion robust rppg, IEEE transactions on Biomedical Engineering 62 (2014) 415–425.

[12] X. Liu, X. Yang, J. Jin, A. 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 (2020) 3352–3362.

[13] Z. Sun, X. Li, Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast, in: European Conference on Computer Vision, Springer, 2022, pp. 492–510.

[14] Z. Sun, X. Li, Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[15] Z. Sun, X. Li, Privacy-phys: Facial video-based physiological modification for privacy protection, IEEE Signal Processing Letters 29 (2022) 1507–1511.

[16] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Advances in Neural Information Processing Systems 33 (2020) 19400–19411.

[17] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).

[18] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al., Mediapipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172 (2019).

[19] E. Aboutanios, B. Mulgrew, Iterative frequency estimation by interpolation on fourier coefficients, IEEE Transactions on signal processing 53 (2005) 1237–1242.

[20] X. Liu, X. Yang, D. Wang, A. 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 (2020) 1–15.

[21] 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. doi:10.1109/JSEN.2021.3067770.