Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.8 No.3

Remote Heart Rate Estimation using Swin Transformer V2 and Wrapping Temporal Shift Modules

Hyunduk Kim[1,*], Sang-Heon Lee[1], Myoung-Kyu Sohn[1], Jungkwang Kim[1]
[1] Division of Automotive Technology, DGIST, Daegu, Republic of Korea.

Abstract. Remote heart rate (rHR) estimation, which aims to measure heart activities without any physical contact with the subject, is performed using remote photoplethysmography (rPPG) and has great potential in many applications. In this paper, we introduce a remote heart rate (rHR) estimation algorithm using Swin Transformer V2 and Wrapping Temporal Shift Modules (WTSM). Moreover, we apply difference layer to reduce the lighting and motion noise and shallow stem to extract coarse local spatio-temporal features. Finally, we apply linear layer to project features to 1D rPPG signal and estimate heart rate using FFT. To evaluate the performance of the proposed algorithm, we train and test on the public UBFC-rPPG and PURE dataset. The experimental results show that the proposed algorithm achieve better accuracy than CNN based methods.

Keywords; Remote heart rate estimation; Vision Transformer; Temporal Shift Module

1. Introduction

Heart rate is an essential physiological signal that reflects the physical state of a person and is widely applied to many application areas, such as sport, fitness, and healthcare. Heart rate is usually obtained by electrocardiogram (ECG) or photoplethysmogram (PPG) signals. However, these signals are measured from skin-contact sensors, which may be inconvenient and discomfort. To solve this problem, many researchers have introduced remote photoplethysmograph (rPPG) estimation algorithms from face video, which estimate PPG signal remotely and without any contact. In detail, the rPPG methods analyze the change of skin surface color recorded from a

* Corresponding author: hyunduk00@dgist.ac.kr
Received: Aug 3, 2024; Accepted: Sep 11, 2024; Published: Sep 30, 2024

remote video camera. In a former study, most rPPG estimation algorithms used handcrafted features. These approaches mainly consist of multiple stages, such as the region of interest (ROI) detection, skin segmentation, color space transformation, signal decomposition, and filtering step. Finally, heart rate is estimated using frequency analysis, such as FFT, or peak detection. However, these approaches achieved low accuracy for lighting changes, head motion, and changes in facial appearance [1].

Figure 1. Overall architecture for remote heart rate estimation

Recently, many researchers have proposed deep learning based-rPPG estimation algorithms to robust these variations. Most of these methods include 3D spatio-temporal convolution to treat time series analysis and skin-based attention mechanism to focus on areas of a face most suitable for estimating PPG signal. However, these methods still require additional skin segmentation step for training and inference. Hence, vision transformer is suitable for remote PPG estimation because vision transformer includes self-attention mechanism.

2. Methodology

In this paper, we propose remote heart rate estimation using Swin Transformer v2 and Wrapping Temporal Shift Module (WTSM). The original videos are firstly preprocessed to crop the face to get the face video. Then we fed 180 face sequences to Preprocessing Module, which consists of a different layer and a shallow stem. The difference layer computes the first forward difference along the temporal axis of the raw video frames, by subtracting every two adjacent frames and the stem extract coarse local spatio-temporal features. And then we extract spatiotemporal representation using WTSM and Swin Transformer V2 and estimate the PPG signal using Linear layer. Finally, we calculate heart rate using FFT. Figure 1 shows the structure of the proposed remote heart rate estimation algorithm.

A. Wrapping Temporal Shift Module

The temporal Shift Module [2] First splits the input tensor into three chunks, shifts the first chunk to the left by one place (advancing time by one frame) and shifts the second chunk to the right by one place (delaying time by one frame). However, TSM

fail to build robust temporal representations because of the proportion of zeroed-features increases. While traditional TSM shifts out and zero-pad channel-folds from the first and last time-step, WTSM [3] resolves the problem of zeroed features by wrapping the shifted-out folds to fill the previously zero-padded folds.

B. Swin Transformer V2

Recently, visual transformers for image and video understanding have achieved outperform than convolutional neural network. Moreover, attention mechanism is important to capture suitable feature for estimating PPG signal. Hence, we apply visual transformer to capture spatial-temporal information However, video-based visual transformers require high computational complexity and 2D visual transformer is only able to learn spatial feature. So, we add Wrapping Temporal Shift Module (WTSM) before Swin Transformer v2 [4] block to facilitate information exchange across the temporal axis.

3. Experiments

We evaluate the accuracy of the proposed algorithm by the PURE dataset and UBFC-RPPG dataset. PURE dataset contains 60 videos from 10 subjects with a diverse set of motion tasks such as steady, talking, head movements and head rotation. We used 80% videos for training and the others for validation. The UBFC-RPPG dataset Contains 42 videos from 42 subjects.

For preprocessing, we synchronize facial video and corresponding ppg signals using rppg-toolbox toolkit [5]. We also used RetinaFace [7] python package to detect face bounding box. All experiments are done on the NVIDIA RTX A5000 GPU using PyTorch. We use the AdamW as an optimizer with the learning rate of 0.0001 and the MSE loss. We set epoch to 30, image size to 128x128, and batch size to 4. We used continuous 190 frames of face images as input. To evaluate the performance of the proposed network on three public datasets we use three statistical metrics, such as the mean absolute error (MAE), the root mean square error (RMSE), the mean average percent error (MAPE), and the Pearson correlation coefficient (R). Table I and Fig.2 show the results of the proposed network and current state-of -the-art models. As shown in Table I, we can observe that the proposed algorithm achieves better than physformer [8] and efficientphys [9]. Moreover, as shown in Fig 2, we can also observer that the proposed algorithm estimates heart rate more accurately than other methods.

Table I. CROSS-DATASET EVALUATION WITH MODELS TRAINED ON PURE ONLY AND TESTED ON UBFC-RPPG

Model	MAE	RMSE	MAPE	R
PhysFormer [8]	1.946	4.160	2.052	0.973
EfficientPhys [9]	1.758	3.959	1.845	0.976
DGISTPhys	1.139	2.591	1.323	0.99

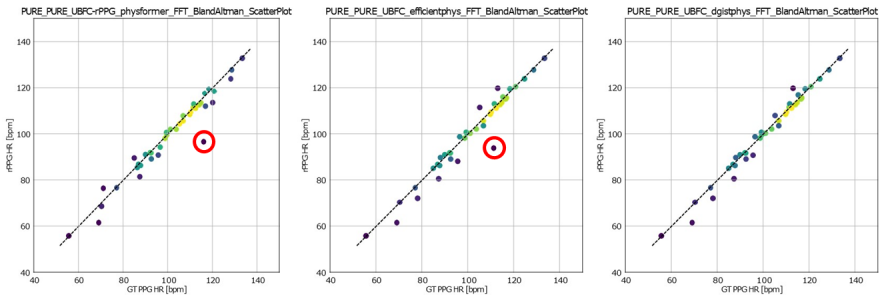

Figure 2. The Bland-Altman scatter plots

4. Conclusion

In this paper, we introduced remote heart rate estimation based on swin transformer v2 and wrapping temporal shift modules. We designed the proposed network by preprocessing module, which consists of different layer and stem, wrapping temporal shift modules, and swin transformer v2 blocks. The experimental results showed that the proposed network can estimates heart rate more accurately than previous methods. In fact, we used only two stage-swin transformer architecture because the rPPG dataset is a relatively small dataset compared to ImageNet dataset. And it is well-known that transformer need to large-scale dataset for stable training in general. Hence, in the future, we will apply light transformer architecture for stable training.

Acknowledgment

This work was supported by the DGIST R&D Program of the Ministry of Science and ICT (24-IT-01).

References

[1] A. Ni, A. Azarang, and N. Kehtarnavaz, “A review of deep learning-based contactless heart rate measurement methods,” Sensors, vol. 21, no. 11, 3719, 2021.

[2] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," In Proceedings of the IEEE/CVF international conference on computer vision, pp. 7083-7093, 2019.

[3] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, and S. Patel, "Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements," In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7914-7924, 2024.

[4] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,pp. 12009-12019, 2022.

[5] X. Liu, X. Zhang, G. Narayanswamy, Y. Zhang, Y. Wang, S. Patel, and D. McDuff, D, "Deep physiological sensing toolbox," arXiv preprint arXiv:2210.00716, 2022.

[6] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," arXiv preprint arXiv:1905.00641, 2019.

[7] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4186-4196, 2022.

[8] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 5008-5017, 2023.