Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.9 No. 1

Remote Heart Rate Estimation using RGB-NIR Fusion

Hyunduk Kim), Sangheon Lee, Myoungkyu Sohn, and Jungkwang Kim*

Division of Automotive Technology, DGIST, Daegu, Republic of Korea

Abstract. Remote photoplethysmography (rPPG) has emerged as a promising method for contactless heart rate estimation using video sequences. In this study, we propose CrossSTSPhys, which incorporates a cross-attention mechanism between dual streams of video inputs: the original RGB stream and the NIR stream. This dual-path structure enhances the network's ability to exploit complementary features from the two input modalities. The CrossSTSPhys architecture adopts Spatial-Temporal SwiftFormer blocks and integrates cross-attention layers at multiple hierarchical levels to exchange and refine information across the two streams. Experimental results show that CrossSTSPhys achieves superior heart rate estimation accuracy on benchmark datasets, outperforming the baseline STSPhys model and existing state-of-the-art methods.

Keywords; remote heart rate estimation; RGB-NIR fusion, visual transformer

Cite this paper as : Hyunduk Kim, Sangheon Lee, Myoungkyu Sohn, and Jungkwang Kim (2025) "Remote Heart Rate Estimation using RGB-NIR Fusion", Journal of Industrial Information Technology and Application, Vol. 9. No. 1, pp. 1070 - 1074

*Corresponding author: hyunduk00@dgist.ac.kr
Received: Nov. 24. 2024 Accepted: Feb. 17. 2025 Published: May. 31. 2025

1. Introduction

Heart rate estimation through remote photoplethysmography (rPPG) has gained significant attention in healthcare and wellness applications. Unlike traditional contact-based devices, rPPG utilizes subtle changes in pixel intensities caused by blood flow to estimate heart rate, offering a non-invasive and unobtrusive solution. Recent advancements in deep learning have accelerated progress in this field, with models like Spatial-Temporal SwiftFormer (STSPhys) achieving substantial improvements by effectively capturing spatial and temporal dependencies in video sequences [1-3]. Despite its success, STSPhys relies solely on RGB video inputs, which may limit its ability to generalize under varying lighting conditions, skin tones, and motion artifacts. In this paper, we introduce CrossSTSPhys, a novel dual-stream framework designed to overcome these limitations by leveraging complementary information from an additional grayscale-enhanced stream. The cross-attention mechanism embedded within the Spatial-Temporal SwiftFormer blocks enables the model to dynamically exchange information between the RGB and grayscale streams, thus improving its capacity to capture subtle physiological signals under diverse conditions.

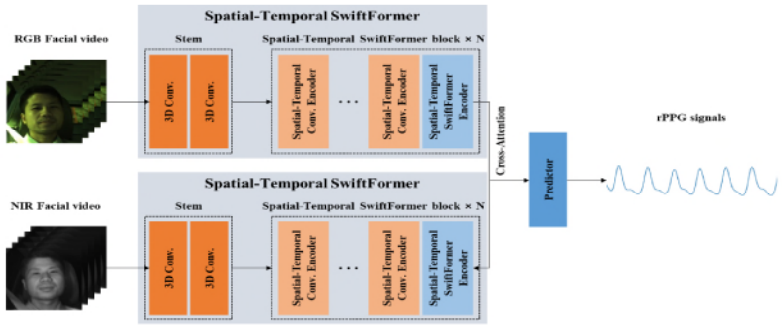

Figure 1. Overview of the proposed CrossSTSPhys architecture.

2. Methodology

A. Overview of STSPhys

STSPhys [3] is a state-of-the-art model designed for remote photoplethysmography (rPPG), leveraging the Spatial-Temporal SwiftFormer (STS) architecture to extract and model spatial and temporal features from video sequences efficiently. Its backbone includes multiple Spatial-Temporal SwiftFormer blocks, which are designed to process both spatial and temporal patterns using lightweight convolutional layers and self-attention mechanisms. The pipeline begins with a 3D

convolutional stem that captures short-term temporal features from video frames. These features are subsequently passed through a series of Spatial-Temporal Conv-Encoder and SwiftFormer blocks to refine spatial and temporal dependencies over a longer duration. The final feature maps are processed by a predictor module to estimate the physiological signal, such as heart rate, from subtle variations in pixel intensities. Despite its effectiveness, STSPhys relies on single-stream RGB video input, which can limit its generalization under challenging conditions such as motion artifacts, varying lighting, or diverse skin tones.

B. Proposed CrossSTSPhys Framework

CrossSTSPhys extends the STSPhys architecture by introducing a dual-stream design to overcome the limitations of single RGB input. This framework incorporates an additional grayscale-enhanced stream alongside the RGB input, allowing the model to exploit complementary information from both modalities. Each stream processes the input independently through Spatial-Temporal SwiftFormer backbones, which consist of multiple hierarchical blocks for spatial and temporal feature extraction.

A cross-attention mechanism is employed between the corresponding blocks of the RGB and grayscale backbones, enabling bidirectional feature refinement. Specifically, the RGB features serve as the query while the grayscale features provide the key and value, and vice versa. This mechanism ensures dynamic exchange of information, allowing the model to capture subtle physiological signals that may not be evident in a single modality. After cross-attention, the refined features from both streams are concatenated and passed through a shared predictor module, which generates the final heart rate signal. By incorporating the dual-stream design and cross-attention mechanism, CrossSTSPhys achieves significant performance improvements over the original STSPhys model, demonstrating enhanced robustness against noise, artifacts, and environmental variations.

3. Experiments

We conducted our experiments on the MR-NIRP Car dataset [4], which is the first publicly available dataset capturing pulse signals during actual driving. It comprises 190 videos from 19 subjects recorded under various weather and time conditions, using both broadband RGB and narrow-band NIR. Pulse oximeter readings serve as ground truth, and subjects perform a variety of motions (e.g., looking around, checking mirrors, talking, and remaining still), making this dataset a highly realistic benchmark for rPPG-based heart rate estimation. We compared our proposed CrossSTSPhys framework against three existing rPPG methods, PhysFormer [1],

EfficientPhys [2], and STSPhys [3], under three different input configurations: (1) RGB-only, (2) NIR-only, and (3) RGB+NIR fusion. For preprocessing, we synchronize facial videos and the corresponding PPG signals using the rppg-toolbox toolkit [5]. We detect face bounding boxes using RetinaFace [6] and crop each frame to 1.5 times the detected bounding box size. We employ Adam as the optimizer with a learning rate of 0.0001 and the MSE loss function. The number of epochs is set to 10, the input image size is set to 128×128, and the batch size is set to 4. We use continuous segments of 160 face frames as input. All experiments were performed using PyTorch on an NVIDIA RTX A5000 GPU.

Table 1 presents the performance of each method in terms of mean absolute error (MAE), root mean square error (RMSE), mean average percent error (MAPE), and Pearson correlation coefficient (R). Across RGB-only inputs, STSPhys [3] achieves the lowest error (MAE 8.05) and MAPE (10.60), with PhysFormer [1] displaying similar errors but a slightly more negative correlation (R = –0.14). EfficientPhys [2] exhibits higher errors (MAE 19.67) but a mildly positive correlation (R = 0.08).

Table 1. Performance comparision on the MR-NIRP car dataset.

	Model	MAE	RMSE	MAPE	R
RGB	PhysFormer [1]	8.39	10.30	11.33	-0.14
	EfficientPhys [2]	19.67	20.66	26.65	0.08
	STSPhys [3]	8.05	10.72	10.60	-0.01
NIR	PhysFormer [1]	10.69	12.31	14.44	-0.10
	EfficientPhys [2]	20.84	22.07	27.60	0.21
	STSPhys [3]	11.67	14.05	15.60	-0.24
RGB+NIR	CrossSTSPhys	9.16	11.66	12.06	0.02

Under NIR-only, PhysFormer yields the best MAE (10.69), while EfficientPhys shows larger errors (MAE 20.84) yet obtains the highest correlation (R = 0.21). By fusing RGB and NIR, the proposed CrossSTSPhys achieves an MAE of 9.16 and a modest correlation (R = 0.02), illustrating that leveraging both modalities improves overall robustness compared to single-input methods.

4. Conclusion

In this paper, we introduced CrossSTSPhys, a dual-stream framework that fuses RGB and near-infrared (NIR) inputs for remote photoplethysmography (rPPG) heart

rate estimation. By integrating the Spatial-Temporal SwiftFormer architecture with a cross-attention mechanism, CrossSTSPhys effectively captures complementary features from both modalities, leading to improved robustness against varying lighting conditions and subject motion. Experiments on the MR-NIRP Car dataset, which reflects realistic driving scenarios, show that CrossSTSPhys achieves balanced performance with reduced estimation errors and stronger correlation with ground truth signals compared to single-stream methods. These results highlight the potential of multimodal fusion and cross-attentive architectures in developing more accurate and reliable rPPG solutions for practical, real-world applications.

Acknowledgment

This work was supported by the DGIST R&D Program of the Ministry of Science and ICT (25-IT-01).

References

[1] 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.

[2] 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.

[3] H. Kim, S. H. Lee, M. K. Sohn, J. Kim, and H. Park, “STSPhys: Enhanced Remote Heart Rate Measurement with Spatial-Temporal SwiftFormer,” IEEE Signal Processing Letters, vol 32, pp. 521-525, 2024.

[4] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, “Near-infrared imaging photoplethysmography during driving,” IEEE transactions on intelligent transportation systems, vol. 23, no.4, pp.3589-3600, 2020.

[5] X. Liu et al., "rppg-toolbox: Deep remote ppg toolbox," Advances in Neural Information Processing Systems, vol. 36, pp. 68485-68510, 2024.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5203–5212, 2020.