TURNIP: TIME-SERIES U-NET WITH RECURRENCE FOR NIR IMAGING PPG

Comas, Armand; Marks, Tim; Mansour, Hassan; Lohit, Suhas; Ma, Yechi; Liu, Xiaoming

TR2021-099 September 10, 2021

Abstract

Near-Infrared (NIR) videos of faces acquired with active illumination for the problem of estimating the photoplethysmogram (PPG) signal from a distance have demonstrated improved robustness to ambient illumination. Contrary to the multichannel RGB-based solutions, prior work in the NIR regime has been purely model-based and has exploited sparsity of the PPG signal in the frequency domain. In contrast, we propose in this paper a modular neural network-based framework for estimating the remote PPG (rPPG) signal. We test our approach on two challenging datasets where the subjects are inside a car and can have a lot of head motion. We show that our method outperforms existing model-based methods as well as end-to-end deep learning methods for rPPG estimation from NIR videos.

IEEE International Conference on Image Processing (ICIP) 2021

TURNIP: TIME-SERIES U-NET WITH RECURRENCE FOR NIR IMAGING PPG

Armand Comas[*2] *Tim K. Marks*[1] *Hassan Mansour*[1] *Suhas Lohit*[1] *Yechi Ma*[*3] *Xiaoming Liu*[4]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Northeastern University, Boston, MA, USA
[3]Princeton University, Princeton, NJ, USA
[4]Michigan State University, East Lansing, MI, USA

ABSTRACT

Imaging photoplethysmography (iPPG) is the process of estimating the waveform of a person's pulse by processing a video of their face to detect minute color or intensity changes in the skin. Typically, iPPG methods use three-channel RGB video to address challenges due to motion. In situations such as driving, however, illumination in the visible spectrum is often quickly varying (e.g., daytime driving through shadows of trees and buildings) or insufficient (e.g., night driving). In such cases, a practical alternative is to use active illumination and bandpass-filtering from a monochromatic near-infrared (NIR) light source and camera. Contrary to learning-based iPPG solutions designed for multi-channel RGB, previous work in single-channel NIR iPPG has been based on hand-crafted models (with only a few manually tuned parameters), exploiting the sparsity of the PPG signal in the frequency domain. In contrast, we propose a modular framework for iPPG estimation of the heartbeat signal, in which the first module extracts a time-series signal from monochromatic NIR face video. The second module consists of a novel time-series U-net architecture in which a GRU (gated recurrent unit) network has been added to the passthrough layers. We test our approach on the challenging MR-NIRP Car Dataset, which consists of monochromatic NIR videos taken in both stationary and driving conditions. Our model's iPPG estimation performance on NIR video outperforms both the state-of-the-art model-based method and a recent end-to-end deep learning method that we adapted to monochromatic video.

***Index Terms*—** Human monitoring, vital signs, remote PPG, imaging PPG, deep learning.

1. INTRODUCTION

Driver monitoring technology, including methods for monitoring driver attention via gaze and pose estimation as well as monitoring vital signs such as heart rate and breathing, has the potential to save the lives of drivers and of others on the road.

In this paper, we focus on the problem of estimating the heart rate (HR), and more generally the hearbeat waveform, of the driver from a distance with the help of a camera installed inside the car. Measuring the pulse signal remotely via a camera, known as imaging PPG (iPPG), is more convenient and less intrusive than contact-based methods. However, iPPG while driving presents a host of challenges such as head pose variations, occlusions, and large variations in both illumination and motion. In recent work [1], we have demonstrated that narrow-band active near-infrared (NIR) illumination can greatly reduce the adverse effects of lighting variation during driving, such as sudden variation between sunlight and shadow or passing through streetlights and headlights, without impacting the driver's ability to see at night. However, NIR frequencies introduce new challenges for iPPG, including low signal-to-noise ratio (SNR) due to reduced sensitivity of camera sensors and weaker blood-flow-related intensity changes in the NIR portion of the spectrum.

Previous work on iPPG estimation from monochromatic NIR video uses straightforward linear signal processing or simple hand-designed models with very few learned parameters, such as optimizing sparsity of the PPG signal in the frequency domain [1]. In contrast, we propose a deep-learning-based approach in a modular framework. Our first module uses automatic facial landmark detection and signal processing to extract a multi-dimensional time-series of average pixel-intensity variations from salient face regions. Our second module is a deep neural network (DNN) with a U-net architecture that inputs the region-wise time series that were extracted by the first module and outputs an estimate of the one-dimensional pulsatile iPPG signal. Our U-net is unusual in that it processes a multidimensional time series rather than an image or image sequence, and unique in that it introduces a network of gated recurrent units (GRUs) to the passthrough (copy) connections of the U-net. The system is trained on narrow-band single-channel NIR videos of multiple subjects using our recently released MERL-Rice Near-Infrared Pulse (MR-NIRP) Car Dataset [1], in which the ground-truth pulse signals were provided by a contact PPG sensor (fingertip pulse oximeter).

Recent work on iPPG estimation from RGB videos has used end-to-end DNNs [2, 3, 4, 5, 6]. In fewer cases, a modular approach is used [7, 8]. Specifically, [7] uses an approach similar in spirit to ours. However, previous DNN methods can't handle NIR monochromatic videos and are not tested on real-world in-car data, which is the focus of our work. The main contributions of this work include:

1. A modular approach for iPPG. Our first module uses face processing and signal processing to extract region-wise time series from NIR facial video. For the second module, we designed a time series U-net to estimate the heartbeat signal, by recovering the PPG signal.
2. A novel time-series U-net architecture, with a GRU subnetwork included in the pass-through connections.
3. The proposed framework outperforms prior work on the challenging MR-NIRP Car Dataset, a public monochromatic NIR dataset where lighting and motion are large nuisance factors.

2. RELATED WORK

Using one vs. multiple color channels for iPPG: Most recent iPPG algorithms are designed for RGB cameras, and rely on combinations of the three color channels to isolate the PPG signal from

[*] AC and YM worked on this project when they were interns at MERL.

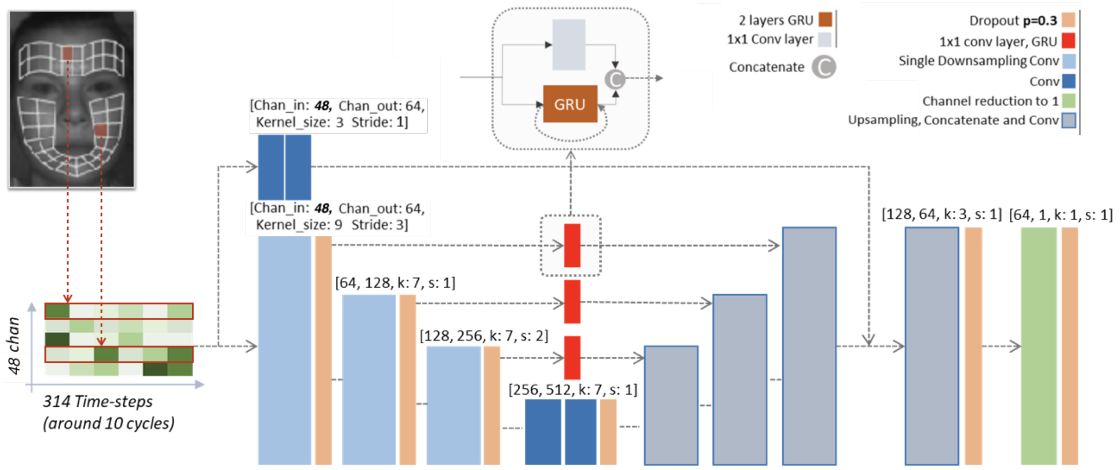

Fig. 1. Overview of the proposed modular approach for rPPG estimation from NIR videos. The first module extracts the 48 relevant facial regions from each frame and computes the average of the pixel intensities for each region, thus generating a 48-dimensional time series. The time series is fed into a time-series U-Net with GRUs which maps it to the desired PPG signal.

noise. Linear combinations of color channels can be used to separate the heart rate signal from noise [9, 10, 11], and using multiple color channels improves the robustness to lighting and motion [12, 13]. However, RGB cameras do not work well in the dark (*e.g.*, night driving) or when ambient light varies drastically. Van Gastel *et al.* [14] used three NIR cameras, each fitted with a different narrow-band filter, to achieve robustness to both motion and light, but such a system could be prohibitively expensive or unwieldy for many applications.

There are a few algorithms for iPPG that use only one color channel. Most of them use just the green channel of an RGB video, since variations in skin appearance due to blood volume variation are most pronounced in this channel [15, 16, 17]. Two recent algorithms for monochromatic PPG, SparsePPG [18] and AutoSparsePPG [1], perform well on single-channel NIR videos by leveraging the fact that iPPG signals are sparse in the frequency domain and low-rank across facial regions. These two methods show impressive results when combined with face alignment and region-wise signal extraction, and we compare our method with them in Sec. 4.

Deep-learning based methods: Although iPPG estimation for monochrome videos has not received much attention from deep learning methods, several deep learning methods were recently developed for RGB videos. Chen and McDuff [4] propose a two-stream approach. Their motion model stream inputs a motion representation obtained by normalized frame differencing. Their appearance model stream processes video frames to obtain soft attention masks per frame, which are then applied to feature maps in the other stream. McDuff [8] presents a deep-learning based super-resolution preprocessing step for iPPG estimation from low-resolution videos. Yu *et al.* [3] design a two-stage, end-to-end method to counter video compression loss in highly compressed videos.

PhysNet [5, 6] is an end-to-end learning framework using a spatio-temporal DNN with three different versions: PhysNet-3DCNN and PhysNetLSTM were introduced in [6]; the original version of PhysNet [5], which we call PhysNet-ST, has spatio-temporal (ST) blocks that contain alternating spatial and temporal convolution layers. Recently, Niu *et al.* [7] propose an architecture to disentangle non-physiological noise from the intensity variations caused by blood flow. They first identify a small number of face regions with facial landmarks, then extract a multi-dimensional time series using multiple color channels from each region. Pairs of such time-series are fed into the disentangling module, whose output is fed to another DNN to determine the iPPG signal and the heart rate.

3. IPPG ESTIMATION FROM NIR VIDEOS

The main noise factors contaminating the signal are motion and illumination changes. With this knowledge, we propose to divide the method into two main components: a hand-crafted feature extraction module that accounts for motion, and a deep network PPG estimator. Figure 1 provides a graphical representation of the architecture.

3.1. Time series extraction module

From each monochromatic NIR video, we extract a 48-dimensional time series corresponding to the pixel intensities over time of 48 facial regions, similar to [1]. We localize 68 facial landmarks in each frame using OpenFace [19], then smooth these locations using a 10-frame moving average. The smoothed locations are used to extract 48 regions of interest (ROIs) located around the forehead, cheeks and chin (the face areas containing the strongest PPG signals [17]). In each video frame, we compute the average intensity of the pixels in each region. This averaging reduces the impact of quantization noise of the camera, motion jitter due to imperfect landmark localization, and minor deformations due to head and face motion.

The time series are temporally windowed and normalized before being fed to the PPG estimator described in Sec. 3.2. The windowed sequences are of 10 seconds duration (300 frames at 30 fps), with a 10-frame stride at inference (we experimented with strides from 10–60 frames in training, since longer strides are more efficient for the larger driving dataset). To each window, we add a preamble of 0.5 seconds by adding the 14 additional frames immediately preceding the start of the sequence. While the heartbeat signal is locally periodic, its period (heart rate) changes over time—the 10 s window is a good compromise duration for extracting a current heart rate. For each 10-second window of the ground-truth PPG signal, we filter its spectrum according to the natural cardiac frequency range of $[42, 240]$ beats per minute (bpm) [11]. Finally, the 10-second sequences are normalized as: $\hat{y}_i = (y_i - \mu_i)/\sigma_i$, where μ_i and σ_i are, respectively, the sample mean and standard deviation of the temporal sequence y_i corresponding to the ith face region.

3.2. TURNIP: Deep neural architecture for PPG estimation

The input to our PPG estimator network is the multidimensional time series provided by the module of Sec. 3.1, in which each dimension is the signal from an explicitly tracked ROI. This tracking helps to reduce the amount of motion-related noise, but the time series still contains significant noise due to factors such as landmark localization errors, lighting variations, 3D head rotations, and deformations such as facial expressions. Our approach needs to recover the signal of interest from the noisy time series. Given the semi-periodic nature of the signal, we design our architecture to extract temporal features at different time resolutions. We present the Time-series U-net with Recurrence for NIR Imaging PPG (TURNIP), in which we apply a U-Net [20] architecture to time series data and modify the skip connections to incorporate temporal recurrence.

The 48 dimensions of the time sequence are fed to the network as channels, which are combined during the forward pass. For each 10-second window, our architecture extracts convolutional features at three temporal resolutions, downsampling the original time series by a factor of 3 and later an additional factor of 2. It then estimates the desired PPG signal in a deterministic way. At every resolution, we connect the encoding and decoding sub-networks by a skip connection. In parallel with the 1×1 convolutional skip connections, we introduce a novel recurrent skip connection. We utilize gated recurrent units (GRUs) to provide temporally recurrent features.

At each time scale, the convolutional layers of the U-net process all of the samples from the 10-second window in parallel. In contrast, the new recurrent GRU layers process the temporal samples sequentially. This temporal recurrence has the effect of extending the temporal receptive field at each layer of the U-net. After the GRU has run through all of the time steps in the 10-second window, the resulting sequence of hidden states is concatenated with the output of a standard pass-through layer (1×1 convolution). Note that the hidden state of the GRU is reinitialized for each 10-second window that is fed to the network. We show empirically that incorporating this GRU improves performance (see ablation study in Sec. 4).

Loss functions for training TURNIP: Denote by $\mathbf{y}$ the ground truth PPG signal and by $\overline{\mathbf{y}}(\theta)$ the estimated PPG signal in the time domain. Our objective is to find the optimal network weights θ^* that maximize the Pearson correlation coefficient between the ground truth and estimated PPG signals. Therefore, we define the training loss function $G(\mathbf{x}, \mathbf{z})$ for any two vectors $\mathbf{x}$ and $\mathbf{z}$ of length T as:

$$G(\mathbf{x}, \mathbf{z}) = 1 - \frac{T \cdot \mathbf{x}^\top \mathbf{z} - \mu_\mathbf{x} \mu_\mathbf{z}}{\sqrt{(T \cdot \mathbf{x}^\top \mathbf{x} - \mu_\mathbf{x}^2)(T \cdot \mathbf{z}^\top \mathbf{z} - \mu_\mathbf{z}^2)}}, \qquad (1)$$

where $\mu_\mathbf{x}$ and $\mu_\mathbf{z}$ are the sample means of $\mathbf{x}$ and $\mathbf{z}$, respectively. We experimented with two loss functions: temporal loss (TL) and spectral loss (SL). To minimize TL, find network parameters θ^* such that:

$$\theta^* = \arg\min_\theta G(\mathbf{y}, \overline{\mathbf{y}}(\theta)). \qquad (2)$$

For SL, the inputs to the loss function are first transformed to the frequency domain, and any frequency components lying outside of the $[0.6, 2.5]$ Hz band are suppressed because they are outside the range of heart rates in the dataset. In this case, the network parameters are computed to solve

$$\theta^* = \arg\min_\theta G(|\mathbf{Y}|^2, |\overline{\mathbf{Y}}(\theta)|^2), \qquad (3)$$

Table 1. HR estimation errors (mean ± std) in terms of PTE6 and RMSE on the MR-NIRP Car Dataset.

	Driving		**Garage**	
	PTE6 (%) ↑	**RMSE (bpm)** ↓	**PTE6 (%)** ↑	**RMSE (bpm)** ↓
TURNIP (Ours)	$\mathbf{65.1 \pm 13.9}$	$\mathbf{11.4 \pm 4.1}$	$\mathbf{89.7 \pm 15.7}$	$\mathbf{4.6 \pm 4.8}$
PhysNet-ST-SL-NIR	53.2 ± 26.7	13.2 ± 7.0	88.8 ± 17.8	6.3 ± 6.7
AutoSparsePPG [1]	61.0 ± 5.2	11.6 ± 1.8	81.9 ± 5.9	5.1 ± 1.4
SparsePPG [18]	17.4 ± 3.4	> 15	35.6 ± 6.8	> 15
DistancePPG [17]	24.6 ± 2.3	> 15	37.4 ± 4.0	> 15

where $\mathbf{Y} := \text{FFT}(\mathbf{y})$ and $\overline{\mathbf{Y}} := \text{FFT}(\overline{\mathbf{y}})$, and $| \cdot |$ is the complex modulus operator. We have tested both loss functions, and we report our results using TL as it performs better with our method (see ablation study in Sec. 4).

4. EXPERIMENTAL RESULTS

Dataset: We use the MERL-Rice Near-Infrared Pulse (MR-NIRP) Car Dataset [1]. The face videos were recorded with an NIR camera, fitted with a 940 ± 5 nm bandpass filter. Frames were recorded at 30 fps, with 640 × 640 resolution and fixed exposure. The ground-truth PPG waveform is obtained using a CMS 50D+ finger pulse oximeter recording at 60 fps, which is then downsampled to 30 fps and synchronized with the video recording. The dataset features 18 subjects and is divided into two main scenarios, labeled *Driving* (city driving) and *Garage* (parked with engine running). Following [1], we evaluate only on the "minimal head motion" condition for each scenario. The dataset includes female and male subjects, with and without facial hair. Videos are recorded both at night and during the day in different weather conditions. All recordings for the garage setting are 2 minutes long (3,600 frames), and during driving range from 2 to 5 minutes (3,600–9,000 frames).

Data augmentation: The dataset consists of subjects with heart rates ranging from 40 to 110 bpm. However, the heart rates of test subjects are not uniformly distributed. For most subjects, the heart rate ranges roughly from 50 to 70 bpm. Examples in the extremes are infrequent. Therefore, we propose a data augmentation technique to address both (i) the relatively small number of subjects and (ii) gaps in the distribution of subject heart rates. At training time, for each 10-second window, in addition to using the 48-dimensional PPG signal that was output by the time series extraction module (see Sec. 3.1), we also resample that signal with linear resampling rates $1+r$ and $1-r$, where we randomly chose the value of $r \in [0.2, 0.6]$ for each 10-second window.

Training and test protocols: We trained TURNIP for 10 epochs, and selected the model after 5 training epochs to use for testing (results were similar across the range 3–8 epochs). We use the Adam optimizer [21], with a batch size of 96 and a learning rate of $1.5 \cdot 10^{-4}$ reduced at each epoch by a factor of 0.05. The train-test protocol is leave-one-subject-out cross-validation. At test time, we window the test subject's time-series as indicated in Sec. 3.1 and estimate the heart rate sequentially with a stride of 10 samples between the windows (we output one heart rate estimate every 10 frames).

Metrics: We evaluate the performance using two metrics. The first metric, PTE6 (percent of time the error is less than 6 bpm), indicates the percentage of HR estimations that deviate in absolute value by less than 6 bpm from the ground truth. The error threshold is set to 6 bpm as that is the expected frequency resolution of a 10-second window. The second metric, root-mean-squared error (RMSE) be-

Table 2. Ablation study. We compare TURNIP to its variants including: (i) removing data augmentation (DA), (ii) removing the GRU module, and (iii) changing the objective function from TL to SL.

	Driving		Garage	
	PTE6 (%) ↑	RMSE (bpm) ↓	PTE6 (%) ↑	RMSE (bpm) ↓
Ours	$\mathbf{65.1 \pm 13.9}$	11.4 ± 4.1	$\mathbf{89.7 \pm 15.7}$	$\mathbf{4.6 \pm 4.8}$
Ours, no DA	61.9 ± 22.3	$\mathbf{10.7 \pm 5.9}$	81.9 ± 31.0	5.9 ± 8.9
Ours, no GRU	63.3 ± 13.1	11.4 ± 4.0	$\mathbf{89.7 \pm 15.4}$	5.0 ± 5.0
Ours, SL	61.7 ± 12.4	13.8 ± 4.2	85.8 ± 18.9	7.3 ± 8.0

tween the ground-truth and estimated HR, is measured in bpm for each 10-second window and averaged over the test sequence.

Comparison methods: We compare with three recent monochromatic iPPG methods based on hand-crafted models: AutoSparsePPG [1], SparsePPG [18], and DistancePPG [17]. We also compare with the end-to-end deep learning method PhysNet-ST [5]. We implemented and modified PhysNet-ST to handle single-channel (monochromatic) frames, and we tested it with different objective functions. The best performing version used the Spectral Loss (SL) objective (see Eqn. (3)), so we call it PhysNet-ST-SL-NIR.

Main Results: Table 1 compares the quantitative performance of all of the methods. Our method outperforms the others in most cases, often by a substantial margin.

PhysNet-ST-SL-NIR performs slightly worse than our method in the Garage condition (TURNIP has 0.9% higher PTE6 and 1.7% lower RMSE), but significantly worse than ours in the Driving condition (TURNIP is better by 11.9% in PTE6). This behavior can be attributed to the first module of our pipeline. Unlike PhysNet, we inject domain knowledge into our method, performing explicit face tracking and selecting face regions that are known to provide a strong PPG signal. In contrast, PhysNet relies solely on spatio-temporal deep convolutional features to implicitly track the face, find the ROI, and extract the PPG signal simultaneously. This works pretty well when there is little head motion (Garage), but for data with moderately large head motion (such as Driving), PhysNet is less successful.

In the Garage condition, TURNIP considerably outperforms AutoSparsePPG (the best hand-crafted model-based method and best published result on the dataset), with a 7.8% PTE6 increase and a 0.5% RMSE decrease. In Driving, the performance is 4.1% better in terms of PTE6 but equivalent in RMSE. As AutoSparsePPG employs a similar video feature extraction module to ours, the main factor of variation is the PPG recovery method.

Ablation study: In Table 2, we compare performance quantitatively when removing different parts of our framework. We can see that data augmentation (DA) is an important part of TURNIP, yielding considerable improvement in PTE6 (3.2% in Driving, 7.8% in Garage), while RMSE is similar. Data augmentation is especially useful for those subjects with out-of-distribution heart rates. It is desirable to train TURNIP with as many examples as possible for a given frequency range. Without data augmentation, the network shows poor performance for subjects with heart rates that are not present in the training set. Table 2 shows that the standard deviation for PTE6 is considerably higher without data augmentation, indicating a high variability across subjects. Figure 2 illustrates why this happens for the Garage data. Subjects 10 and 12 have the lowest and highest resting heart rates in the dataset, ∼40 and ∼100 bpm respectively. Thus when testing on either of those subjects, the training set contains no subjects with similar heart rates. Without data augmentation, the model fails completely for those subjects. With data augmentation, it is much more accurate.

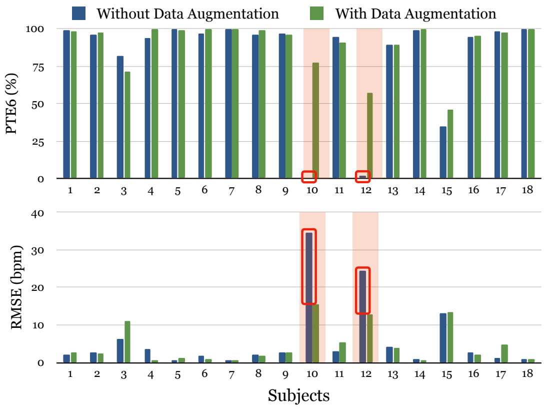

Fig. 2. Impact of data augmentation on tested subjects. Highlighted in red, we show the poor performance of the method without data augmentation for two subjects with out-of-distribution heart rates.

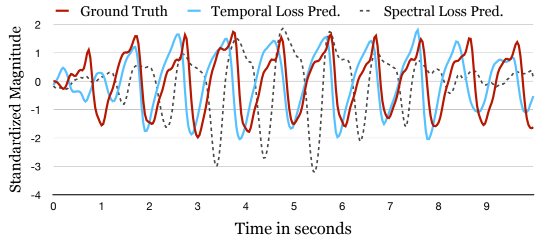

Fig. 3. Comparison of PPG signal estimated by temporal vs. spectral loss for a test subject. Clearly, temporal loss makes better estimates.

Second, we analyze the impact of the GRU cell in our skip connection module. The GRUs process the feature maps sequentially at multiple time resolutions. Thus, they extract features beyond the local receptive field of the convolutional kernels. Addition of this cell improves performance, as shown in the table.

Finally, we compare the two loss functions (TL vs. SL) for training TURNIP, and we see a clear performance drop with SL (Eqn. (3)). However, SL still achieves good performance, outperforming PhysNet-ST-SL-NIR in the Driving condition. Figure 3 compares SL vs. TL for the estimated PPG signals for 10 seconds of a test subject. As shown in the figure, the model trained with TL generates a much better estimate of the ground-truth PPG signal. While the recovered signal with SL has a similar frequency, it often does not match the peaks and distorts the signal amplitude or shape. That is, the spectrum of the recovered signal and the heart rate are similar in both cases, but not the temporal variations.

5. CONCLUSION

In this paper, we proposed a modular framework for estimating the PPG signal from NIR videos. The time series feature extraction module operates on different facial regions and is designed to account for motion variations. The neural network module, TURNIP, maps the extracted multi-dimensional time series to the desired PPG signal, from which the heart rate is determined. TURNIP is an adaptation of U-Nets to time series data, with the addition of GRUs in the passthrough connections. Results show that the proposed framework outperforms existing methods for this application. Investigating the use of other deep architectures for time series processing, such as transformers, presents a promising avenue for future research.

6. REFERENCES

[1] Ewa M. Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan, "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[2] Eugene Lee, E. Chen, and C. Lee, "Meta-rPPG: Remote heart rate estimation using a transductive meta-learner," 2020.

[3] Z. Yu, Wei Peng, Xiao-Bai Li, Xiaopeng Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," *International Conference on Computer Vision (ICCV)*, 2019.

[4] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," *European Conference on Computer Vision (ECCV)*, 2018.

[5] Z. Yu, X. Li, and G. Zhao, "Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks," *The Computing Research Repository (CoRR)*, 2019.

[6] Z. Yu, Xiao-Bai Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," 2019.

[7] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," 2020.

[8] D. McDuff, "Deep super resolution recovering physiological information from videos," *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[9] G. Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 2878–2886, 2013.

[10] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[11] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan, "Algorithmic principles of remote PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017.

[12] D. G. Haan and van Aj Arno Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, pp. 1913–1926, 2014.

[13] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu, "Learning deep models for face anti-spoofing: binary or auxiliary supervision," 2018.

[14] Mark van Gastel, Sander Stuijk, and Gerard de Haan, "Motion robust remote-PPG in infrared," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 5, pp. 1425–1433, 2015.

[15] W. Verkruysse, L. Svaasand, and J. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16 26, pp. 21434–45, 2008.

[16] Xiaobai Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6 5, pp. 1565–88, 2015.

[18] E. Nowara, T. Marks, Hassan Mansour, and A. Veeraraghavan, "SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared," *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[19] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Openface: an open source facial behavior analysis toolkit," 2016.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *The Computing Research Repository (CoRR)*, 2015.