Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks

Zitong Yu¹ Zitong.Yu@oulu.fi Xiaobai Li1 Xiaobai.Li@oulu.fi Guoying Zhao*2,1

- ¹ Center for Machine Vision and Signal **Analysis** University of Oulu FI-90014, Finland
- ² School of Information and Technology Northwest University 710069, China

Abstract

Recent studies facial videos based many medical app erage HR is not su variability (HRV) which is the first v rPPG signals from ground truth pulse pulse peaks. Compresults demonstrat HRV levels compa of using reconstruct Recent studies demonstrated that the average heart rate (HR) can be measured from facial videos based on non-contact remote photoplethysmography (rPPG). However for many medical applications (e.g., atrial fibrillation (AF) detection) knowing only the average HR is not sufficient, and measuring precise rPPG signals from face for heart rate variability (HRV) analysis is needed. Here we propose an rPPG measurement method, which is the first work to use deep spatio-temporal networks for reconstructing precise rPPG signals from raw facial videos. With the constraint of trend-consistency with ground truth pulse curves, our method is able to recover rPPG signals with accurate pulse peaks. Comprehensive experiments are conducted on two benchmark datasets, and results demonstrate that our method can achieve superior performance on both HR and HRV levels comparing to the state-of-the-art methods. We also achieve promising results of using reconstructed rPPG signals for AF detection and emotion recognition.

Introduction

The heart pulse is an important vital sign that needs to be measured in many circumstances, especially for healthcare or medical purposes. Traditionally, the Electrocardiography (ECG) and Photoplethysmograph (PPG) are the two most common ways for measuring heart activities. From ECG or PPG signals, doctors can get not only the basic average heart rate (HR), but also more detailed information as the inter-beat-interval (IBI) for heart rate variability (HRV) analysis and supporting their diagnosis. However, both ECG and PPG sensors need to be attached to body parts which may cause discomfort and are inconvenient for long-term monitoring. To counter for this issue, remote photoplethysmography (rPPG) is developing fast in recent years, which targets to measure heart activity remotely without any contact.

In earlier studies of rPPG, most methods [II, II, III, III, III, III, III, III] can be seen as a two-stage pipeline, which first detects or tracks the face to extract the rPPG signals, and then estimates the corresponding average HR from frequency analysis. However, there are two disadvantages of these methods that worth concern. First, each of them works with

^{© 2019.} The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

^{*} indicates the corresponding author.

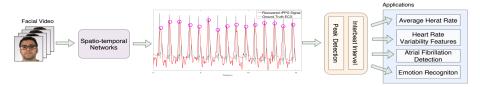


Figure 1: Proposed rPPG signal measurement framework using spatio-temporal networks.

self-defined facial regions that are based on pure empirical knowledge but are not necessary the most effective regions, which should vary across data. Second, the methods involves handcrafted features or filters, which may not generalize well and could lose important information related to heart beat.

There were also several attempts to estimate HR remotely using deep learning [3, 6, 12], [13] approaches. But these studies have at least one of the following drawbacks: 1) The HR estimation task was treated as a one-stage regression problem with one simple output of the average HR, while the individual pulse peak information were lost which limits their usage in demanding medical applications. 2) The approach is not an end-to-end system, which still requires pre-processing or post-processing steps involving handcrafted features. 3) The approach is based on 2D spatial neural network without considering the temporal context features, which are essential for the rPPG measurement problem.

In this paper, spatio-temporal modeling is conducted on facial videos aiming to locate each individual heartbeat peak accurately. Figure 1 shows the framework of the proposed rPPG signal measurement method, and steps for related applications. The heartbeat peaks (red circles) of the measured rPPG signal locate precisely at the corresponding R peaks of the ground truth ECG signal, which allows us to achieve not only the average HR, but also detailed IBIs information and HRV analysis for AF detection and emotion recognition.

The main contributions of this work include: 1) We propose the first end-to-end spatio-temporal network (PhysNet) for rPPG signal measurement from raw facial videos. It takes temporal context into account which was ignored in previous works. 2) Multiple commonly used spatio-temporal modeling methods are explored and compared, which can serve as a foundation for future network optimization specially for rPPG measurement task. 3) Compared with state-of-the-art methods, the proposed PhysNet achieves superior performance for measuring not only the average HRs, but also the HRV features, which were further demonstrated to be effective for AF disease detection and emotion recognition. 4) The proposed PhysNet also has good generalized ability on new data as shown in a cross data test.

2 Related Work

Previous methods for remote photoplethysmography measurement, background of HRV measurement, and spatio-temporal networks are briefly reviewed in three subsections.

2.1 Remote Photoplethysmography (rPPG) Measurement

In past few years, several traditional methods explored rPPG measurement from videos by analyzing subtle color changes on facial regions of interest (ROI), including blind source separation [13], [24], least mean square [110], majority voting [32] and self-adaptive matrix completion [252]. There are other traditional methods which utilized all skin pixels for rPPG measurement, e.g., chrominance-based rPPG (CHROM) [32], projection plane orthogonal to the skin tone (POS) [353], and spatial subspace rotation [353]. These methods require complex prior knowledge for ROI selection/skin-pixels detection and handcrafted signal processing

steps, which are hard to deploy and do not necessarily generalize well to new data. Besides, the majority of these works only worked on getting the average HR, but did not consider the accuracy of locating each individual pulse peak, which is a more challenging task.

Recently, a few deep learning based methods were proposed for average HR estimation. In [5], Hsu et al. employed the short-time Fourier transform to build spectrogram and utilized convolutional neural network (CNN) to estimate average HRs. Niu et al. [5] constructed a spatial-temporal map for CNN to measure average HRs. Radim et al. [5] proposed the HR-CNN, which used the aligned face images to predict HRs. In [6], Chen and McDuff exploited normalized frame difference for CNN to predict the pulse signal. These methods are not end-to-end frameworks, as they still rely on handcrafted features or aligned face images as inputs. Besides, they were all based on 2D CNN, which lacks the ability to learn the temporal context features of facial sequences which are essential for rPPG signal measurement.

2.2 Heart Rate Variability Measurement

Most of the mentioned studies were focusing on average HR measurement. The HR counts the total number of heartbeats in a given time period, which is a very coarse way of describing the cardiac activity. On the other side, HRV features describe heart activity on a much finer scale, which are computed from the IBIs of pulse signals. Most common HRV features include low frequency (LF), high frequency (HF), and their ratio LF/HF, which are widely used in many medical applications. Besides, the respiratory frequency (RF) can also be estimated by analyzing the frequency power of IBI, as in [and [] and [] Apparently, compared with the task of estimating the average HR (only one number), measuring HRV features is more challenging, which requires accurate measure of the time location of each individual pulse peak. For the needs of most healthcare applications, average HR is far from enough. We need to step forward to develop methods that can measure heart activity on HRV level.

2.3 Spatio-temporal Networks

Spatio-temporal network plays a crucial role in many video-based tasks (e.g., action detection and recognition [23]) because of the excellent performance. There are two mainstreams of spatio-temporal frameworks. The first category includes 3D convolutional neural networks (3DCNN) such as Convolutional 3D [21], Pseudo 3D [23], Inflated 3D [2] and Separable 3D [24], which are widely used for video understanding as they can capture spatial and temporal context simultaneously. The second category includes recurrent neural network (RNN) based frameworks, such as long short term memory (LSTM) [3] and Convolutioal LSTM [32], which can also capture the temporal context among the CNN spatial features.

Existing spatio-temporal networks are mostly designed for analyzing large scale motions, and it is still unknown whether they are suitable for the rPPG signal measurement task as the temporal skin color variation is extremely subtle. In this paper various spatio-temporal modeling methods and loss functions are evaluated, which will serve as a foundation for future works about network optimization specialized for the rPPG measurement task.

3 Methodology

In this section, the proposed method with several alternative spatio-temporal models is demonstrated. We also describe our customized loss function and rationalize the design.

3.1 Network Architecture

According to [4], [4], there are two important procedures in order to achieve pulse information from facial videos. First is to project RGB into color subspace with stronger representation capacity. After that, the color subspace needs to be reprojected in order to get rid

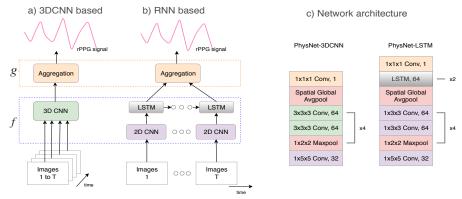


Figure 2: The framework of saptio-temporal networks for rPPG signal recovery. a) 3DCNN based PhysNet; b) RNN based PhysNet; c) Their corresponding network architectures. "3x3x3 Conv, 64" donotes using convolution filter with kernel $3 \times 3 \times 3$ and output channel number is 64 while other operations are analogous.

of irrelevant info (e.g., noise caused by illumination or motion) and achieve the target signal space. Here, we propose an end-to-end spatio-temporal network (denoted as PhysNet), which is able to merge these two steps and achieve the final rPPG signals in one step.

The overall architecture of PhysNet is shown in Figure 2. The input of the network is T-frame face images with RGB channels. After forwarding several convolution and pooling operations, multi-channel manifolds are formed to represent the spatio-temporal features. Finally, the latent manifolds are projected into signal space using channel-wise convolution operation with $1 \times 1 \times 1$ kernel to generate the predicted rPPG signal with length T. The whole procedure can be formulated as

$$[y_1, y_2, ..., y_T] = g(f([x_1, x_2, ..., x_T]; \theta); w), \tag{1}$$

where $[x_1, x_2, ..., x_T]$ are the input frames, and $[y_1, y_2, ..., y_T]$ is the output signal of the network. As shown in Figure 2(a)(b), f is the spatio-temporal model for subspace projection, θ is a concatenation of all convolutional filter parameters of this model, g is the channel aggregation for final signal projection, and w is a set of its parameters. For the spatio-temporal model f there are two mainstream models, and here we explore and compare both, as 3DCNN based and RNN based PhysNet.

3DCNN based PhysNet Denoted as "PhysNet-3DCNN" and shown in Figure 2(a)(c), 3DCNN is imposed as the spatio-temporal model f, which adopts $3 \times 3 \times 3$ convolutions to extract the semantic rPPG features in both spatial and temporal domain simultaneously. It helps to learn more robust context features and recover the rPPG signals with less temporal fluctuation. Inspired by successful leveraging encoder-decoder in action segmentation task [\square], we also attempt a temporal encoder-decoder (ED) structure for rPPG task, denoted as "PhysNet-3DCNN-ED", which intends to exploit more effective temporal context and reduce temporal redundancy and noise.

RNN based PhysNet In Figure 2(b)(c), 2DCNN is deployed to extract the spatial features firstly and then RNN based module is exploited for propagating the spatial features in temporal domain, which may improve the temporal context features via forward/backward information flows. LSTM and ConvLSTM can be formulated as

$$\begin{aligned}
i_{t} &= \delta(W_{i}^{X} * X_{t} + W_{i}^{H} * H_{t-1}), \\
f_{t} &= \delta(W_{f}^{X} * X_{t} + W_{f}^{H} * H_{t-1}), \\
o_{t} &= \delta(W_{o}^{X} * X_{t} + W_{o}^{H} * H_{t-1}), \\
c_{t} &= f_{t} \circ c_{t-1} + i_{t} \circ tanh(W_{c}^{X} * X_{t} + W_{c}^{H} * H_{t-1}), \\
H_{t} &= o_{t} \circ tanh(c_{t}),
\end{aligned} \tag{2}$$

where \ast denotes the multiplication and convolution operator for LSTM and ConvLSTM respectively, and \circ denotes the Hadamard product. Here, bias terms are omitted. All the gates i, f, o, memory cell c, hidden state H and the learnable weights W are 3D tensors. In later sections, "PhysNet-LSTM", "PhysNet-BiLSTM" and "PhysNet-ConvLSTM" represent the networks with LSTM, bi-directional LSTM and Convolutional LSTM respectively. For PhysNet-ConvLSTM, global average pooling is deployed after temporal propagation.

About implementation details, all the convolution operations use $1 \times 1 \times 1$ stride, and are cascaded with batch normalization and nonlinear activation function ReLU except the last convolution with a $1 \times 1 \times 1$ kernel. The strides of all the "Maxpool" operations are set as $1 \times 2 \times 2$ except "PhysNet-3DCNN-ED". For "PhysNet-3DCNN-ED", both strides and kernel sizes of the second and third "Maxpool" are $2 \times 2 \times 2$ in encoder while there are two deconvolution layers before "Spatial Global Avgpool" in decoder to reach back to the original temporal length. In addition, spatial and temporal padding for all convolutions are needed to keep a consistent size. The number of the stacked LSTM layers are set as 2. As our method is designed as a fully convolutional framework so theoretically facial sequences with arbitrary spatial and temporal size are feasible as inputs.

3.2 Loss Function

Besides designing the network architecture, we also need an appropriate loss function to guide the networks. One major step-forward of the current study is that we aim to recover rPPG signals which have matching trend and accurately estimated pulse peak time locations that match with ground truth signals, which are essential for detailed HRV analysis. In order to maximize the trend similarity and minimize peak location errors, negative Pearson correlation is utilized as the loss function

$$Loss = 1 - \frac{T \sum_{1}^{T} xy - \sum_{1}^{T} x \sum_{1}^{T} y}{\sqrt{(T \sum_{1}^{T} x^{2} - (\sum_{1}^{T} x)^{2})(T \sum_{1}^{T} y^{2} - (\sum_{1}^{T} y)^{2})}},$$
(3)

where T is the length of the signals, x is the predicted rPPG signals, and y indicates the ground truth PPG signals.

PPG signals are used as the ground truth for training our network instead of ECG because PPG measured from fingers resembles more to the rPPG measured from faces, as they both measure the peripheral blood volume changes, while ECG measures electrical activities thus contains extra components that do not present in rPPG. In the testing stage, ECG is adopted as the ground truth following previous works like [4, 12], [13] for fair comparison.

4 Experiments

Two datasets are employed in our experiments. First, we train the proposed PhysNet on the OBF dataset [III]. OBF has large number of facial videos and with corresponding PPG signals, which are suited for our training need. The trained PhysNet is first tested on OBF for evaluation of HR and HRV measurement accuracy, and then demonstrated for an extended application of AF detection. At last the trained PhysNet is also crossly tested on the MAHNOB-HCI [III] dataset and another application of emotion recognition is explored.

4.1 Data and Experimental Settings

OBF dataset [121] is used for both training and testing. OBF contains totally 212 videos recorded from 100 healthy adults and six patients with atrial fibrillation (AF). Each subject were recorded for two five-minute sessions, in which facial videos and the corresponding physiological signals (two video cameras, ECG and breathing signal measured from chest, and PPG from finger) were recorded simultaneously. In the current study we use the RGB videos, which were recorded at 60 fps with resolution of 1920x2080.

MAHNOB-HCI dataset [□] is used for cross testing the generalization of our model. It includes 527 videos from 27 subjects, with recording speed of 61 fps and resolution of 780x580. We use the EXG2 signals as the ground truth ECG signal in evaluation. In order to make fair comparison with previous works [□, □], we follow the same routine as their works and use 30 seconds clip (frames 306 to 2135) of each video.

Training Settings. Facial videos and corresponding PPG signals are synchronized before training. For each video clip, we use the Viola-Jones face detector [24] to crop the face area at the first frame and fix the region through the following frames. Then the face images are normalized to 128x128. We set the length of training clips as $T = \{32,64,128,256\}$ and both videos and ground truth signals are downsampled to 30 fps and 30 Hz respectively. The proposed method is trained on Nvidia P100 with PyTorch. The Adam optimizer is used and the learning rate is set as 1e-4. We train all the models for 15 epochs.

Testing Settings and Performance Metrics. As PhysNet is designed as fully-convolutional structure, it is easily to do inference in arbitrary long video clips. In the testing stage, both the recovered rPPGs and their corresponding ground truth ECG signals go through the same process of filtering, normalization, and peak detection to obtain the inter-beat-intervals, from which the average HR and HRV features are calculated. For HRV level evaluation, we followe paper [12] to calculate three commonly used HRV features (in normalized units, n.u.) and the RF (in Hz). Details about the features are referred to [12], [24]. Performance metrics for evaluating both the average HR and HRV features include: the standard deviation (SD), the root mean square error (RMSE), the Pearson's correlation coefficient (R), and the mean absolute error (MAE). For atrial fibrillation detection and emotion recognition evaluation, the accuary (ACC) and specificity (SP) are used as the validation metrics.

4.2 Experiments on OBF

We evaluate several aspects of the proposed method separately, i.e., the loss function, spatiotemporal networks and clip length, and report performance on both HR and HRV levels. Subject-independent 10-fold cross validation is adopted here. We also report the AF detection accuracy of using the measured HRV features as an application scenario.

Loss Function. In order to demonstrate the advantages of our proposed negative Pearson (NegPea) loss function, we compare it with the mean square error (MSE), which is used in [3]. Both experiments employed 3DCNN based PhysNet with the training clip length fixed to 64 (as PhysNet64-3DCNN), and the results are listed in Table 1. Results show that NegPea performs better than MSE on both HV and HRV levels, which support the efficacy of the proposed loss function as we explained in Sec 3.2. Note that as the amplitude of peaks are irrelevant with our task (i.e., to measure accurate time location of heartbeats), the MSE loss may direct wrongly and introduces extra noises.

Table 1: Performance comparison of two loss functions with negative Pearson and MSE. Smaller RMSE and bigger R values indicate better performance.

	HR(bpm)	RF(Hz)	LF(u.n)	HF(u.n)	LF/HF	
Method	RMSE R	RMSE R	RMSE R	RMSE R	RMSE R	
PhysNet64-3DCNN-MSE PhysNet64-3DCNN-NegPea	4.012 0.95 2.143 0.98	5 0.069 0.435 5 0.067 0.494			0.721 0.659 0.647 0.72	

Table 2: Performance comparison of spatio-temporal networks.

	HR(bpm)		RF(Hz)		LF(u.n)		HF(u.n)		LF/HF	
Method	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R
PhysNet64-2DCNN	10.237	0.928	0.092	0.104	0.247	0.321	0.247	0.321	0.962	0.318
PhysNet64-3DCNN PhysNet64-3DCNN-ED	2.143 2.048	0.985 0.989	0.067 <u>0.066</u>	0.494 0.501	0.15 <u>0.146</u>	0.749 0.772	0.15 <u>0.146</u>	0.749 0.772	0.647 <u>0.624</u>	0.72 0.748
PhysNet64-LSTM PhysNet64-BiLSTM PhysNet64-ConvLSTM	3.139 4.595 2.937	0.975 0.945 0.977	0.084 0.085 0.083	0.189 0.183 0.191	0.226 0.231 0.22	0.478 0.421 0.485	0.226 0.231 0.22	0.478 0.421 0.485	0.928 0.956 0.896	0.404 0.396 0.44

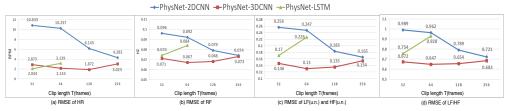


Figure 3: Performance comparison of various training clip lengths T = 32,64,128,256. RMSE is used as the evaluation metric, lower RMSE value indicates better performance.

Spatio-temporal Networks. In these experiments we evaluate the effectiveness of spatiotemporal networks, and we fixed the training clip length as T = 64 and NegPea as the loss function. First, we evaluated a 2DCNN based PhysNet (PhysNet-2DCNN) and reported its result as the baseline (Table 2 top). Second, we evaluate 3DCNN based models either with (PhysNet64-3DCNN-ED) or without (PhysNet64-3DCNN) the ED component(Table 2 middle). It is clear that 1) compared with 2DCNN, temporal convolutions boost performance on both HR and HRV levels, and 2) benefited by the temporal encoder-decoder structure, "PhysNet64-3DCNN-ED" achieved better performance than "PhysNet64-3DCNN". It can be explained that semantic features with less temporal redundancy can be extracted in such squeeze-stretch-like encoding and decoding process. Third, we also evaluate how RNN based models perform, i.e., using LSTM (PhysNet64-LSTM), Bidirectional LSTM (PhysNet64-BiLSTM) and Convolutional LSTM (PhysNet64-ConvLSTM). Results (Table 2 bottom) show that 1) "PhysNet64-LSTM" achieved better performance than the baseline "PhysNet64-2DCNN" but not as well as 3DCNN, which implies that LSTM is able to improve the performance but not so effective as 3DCNN for long-term temporal context aggregation; and 2) LSTM and ConvLSTM are about the same level while BiLSTM is the worst, which indicates the backward information of the highest-level features seems to be not necessary.

Clip Length T. In training stage, the video length may impact each network differently, and we evaluate $T = \{32,64,128,256\}$ here. Results are shown in Figure 3. For "PhysNet-2DCNN" it is clear that longer inputs lead to better performance. "PhysNet-3DCNN" has more stable performance over clip length. Note that with temporal convolution layers, "PhysNet32-3DCNN" outperformed "PhysNet-2DCNN" with much shorter inputs, as the

	HR(bpm)		RF(Hz)		LF(u.n)		HF(ı	ı.n)	LF/HF	
Method	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R
ROI_green [2.162 2.733 1.906	0.99 0.98 0.991	0.084 0.081 0.07	0.321 0.224 0.44	0.24 0.206 0.158	0.573 0.524 0.727	0.24 0.206 0.158	0.573 0.524 0.727	0.832 0.863 0.679	0.571 0.459 0.687
PhysNet128-3DCNN-ED	1.812	0.992	0.066	0.507	0.148	0.766	0.148	0.766	0.631	0.739

Table 3: Performance comparison between previous methods and our proposed method.

Table 4	4:	Resul	ts o	T A	Atrial	Fit	orıl	Hati	lon	D	eteci	tion	on	UI	SF.

Metric	ROI_green [CHROM [1]	POS [26]	PhysNet128-3DCNN-ED (ours)
ACC	77.23%	70.61%	76.5%	80.22%
SP	75.61%	74.18%	79.37%	81.71%

temporal convolution filters can provide extra help in learning temporal representations. For "PhysNet-LSTM" we only compared $T = \{32,64\}$ because of its limited long-term temporal propagation ability, and T = 32 achieved better performance. For "PhysNet-3DCNN", the best HR and HRV performance were achieved at T = 128 and T = 64 respectively.

Comparison with Previous Methods. We replicate three previous methods, i.e., [] as "ROI_green", [] as "CHROM", and [] as "POS", and compare with our method in Table 3. Our best performance was achieved with "PhysNet128-3DCNN-ED" (marked in bold), which outperforms all compared methods on both HR and HRV levels indicating efficacy and robustness of the proposed method.

Atrial Fibrillation Detection. Followed the protocol in [III], we extracted ten dimensional HRV features from the recovered rPPG signals for detecting AF cases against healthy ones. As seen in Table 4, results show that PhysNet achieves better performance than previous methods. Note that the pre-trained parameters of classifers were fixed and kept same, so the improvement was purely based on the more accurately measured HRV features by PhysNet.

4.3 Evaluation on MAHNOB

As "PhysNet128-3DCNN-ED" achieved the best performance on OBF, we use the model trained on OBF to cross test on MAHNOB-HCI to validate its generalization ability. Average HR results of our method are compared with previous methods in Table 5 (as previous works only reported performance on average HR level but not on HRV level). The first four [1, 11, 12] are earlier methods not involving neural network. Although performance of [11] and [12] are good, the approaches are not trained-ready but require computational costly processing steps for each input, which are limited for real-time usage. On the other side

Table 5: Results of average HR measurement on MAHNOB-HCI

Table 3. Results of average TR measurement on MATHOD-TICI.									
Method	HR _{SD} (bpm)	HR _{MAE} (bpm)	HR _{RMSE} (bpm)	HR_R					
Poh2011 [12]	13.5	-	13.6	0.36					
CHROM [₫]	-	13.49	22.36	0.21					
Li2014 [III]	6.88	-	7.62	0.81					
SAMC [🔼]	5.81	-	6.23	0.83					
SynRhythm [□]	10.88	-	11.08	-					
HR-CNN [██]	-	7.25	9.24	0.51					
DeepPhys [1]	-	4.57	-	-					
PhysNet128-3DCNN-ED (ours)	7.84	5.96	7.88	0.76					

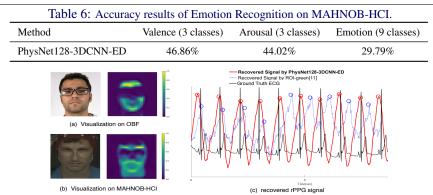


Figure 4: Visualization of original faces, learned neural features and recovered rPPG signals.

the proposed method is a pre-trained end-to-end system which runs very fast on new test samples (discussed later in Sec 4.4). The later three [1, 12] are all neural network based methods, but note that their test protocols vary (e.g., [12])(Table 4) trained on MAHNOB-HCI, [13] trained on a lot more (2302) videos from both the whole MAHNOB and another dataset PURE [13], and [3] trained on other self-collected data), so the results are compared on a general level. Another fact is that all three approaches need pre-processing steps but ours does not thus easy and efficient to deploy. As being cross tested, our method achieved top level performance which indicates its generalization ability for HR measurement.

Emotion Recognition. Another advantage of the proposed method is that the recovered rPPG signals allow HRV feature analysis for more sophisticated applications, i.e., emotion recognition, while previous methods that estimate only HR values are not feasible. We extracted ten dimensional HRV features (the same as in AF detection task, see [III]) from rPPG signals measured with "PhysNet128-3DCNN-ED" on MAHNOB-HCI clips, and feed them to a support vector machine (with a polynomial kernel) as the classifier for estimating the person's emotion status of each video. Several emotion labels are provided by MAHNOB-HCI, and we follow [II] to estimate "Arousal" and "Valence" on three levels, and "Emotions" in nine categories. As listed in Table 6, the results are very promising especially for the valence recognition. To the best of our knowledge this is the first exploration of using face video measured physiological features for emotion analysis. In next work we will try fusing it with facial expression analysis for multimodal emotion recognition.

4.4 Visualization and Inference Speed

Visualization. Mid-level neural features extracted from both dataset samples by "PhysNet 128-3DCNN-ED" are shown in Fig. 4 (a) and (b). The high light areas show that the network is able to learn and select skin regions with the strongest rPPG infomation (e.g., forehead, cheeks and lower jaw). Besides, in Fig. 4 (c) we also show a sample rPPG signal recovered with the proposed PhysNet (red) comparing with that from a baseline method "ROI_green" [III] (blue) and the ground truth ECG (black). The red curve matches much better to the ground truth than the blue one in terms of peak time locations, which demonstrates the effectiveness of the proposed method.

Inference Speed. As our method does not require any pre-processing step as previous networks [122] did, it works faster and allow real-time rPPG signal recovery. For a test video of 30s, the "PhysNet64-3DCNN-ED" takes only 0.235s (3830 fps) on a Tesla P100 GPU, which suits most real-time applications.

5 Conclusion

In this paper, we proposed an end-to-end framework with spatio-temporal networks which is able to recover rPPG signals from raw facial videos fast and efficiently. We tested on OBF and MAHNOB-HCI datasets, and results showed that the proposed PhysNet can recover rPPG signals with accurate time location of each individual pulse peak, which allows measuring not only the average HRs, but also IBIs information and HRV level features that enable potential applications in e.g., remote AF detection and emotion recognition.

6 Aknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61772419), Tekes Fidipro Program (No. 1849/31/2015), Business Finland Project (No. 3116/31/2017), Academy of Finland, and Infotech Oulu. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

References

- [1] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, 2013.
- [2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.
- [4] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.
- [5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [6] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In 2017 IEEE International Joint Conference on Biometrics (IJCB), pages 383–389. IEEE, 2017.
- [7] Xiaohua Huang, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen, and Matti Pietikäinen. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer Vision and Image Understanding*, 147: 114–124, 2016.
- [8] Antony Lam and Yoshinori Kuno. Robust heart rate measurement from video using select random patches. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3640–3648, 2015.

- [9] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
- [10] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4264–4271, 2014.
- [11] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying 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), pages 242–249. IEEE, 2018.
- [12] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Synrhythm: Learning a deep heart rate estimator from general to specific. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 3580–3585. IEEE, 2018.
- [13] 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*, 18(10):10762–10774, 2010.
- [14] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2011.
- [15] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.
- [16] Jingang Shi, Iman Alikhani, Xiaobai Li, Zitong Yu, Tapio Seppänen, and Guoying Zhao. Atrial fibrillation detection from face videos by fusing subtle variations. *IEEE Transactions on Circuits and Systems for Video Technology*, DOI 10.1109/TCSVT.2019.2926632, 2019.
- [17] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.
- [18] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *The British Machine Vision Conference (BMVC)*. 2018.
- [19] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014.
- [20] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

- [21] Duc Nhan Tran, Hyukzae Lee, and Changick Kim. A robust real time system for remote heart rate measurement via camera. In 2015 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2015.
- [22] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016.
- [23] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.
- [24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2001.
- [25] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2016.
- [26] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.
- [27] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.
- [28] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [29] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. *The British Machine Vision Conference* (BMVC), 2017.