Photoplethysmogram signal extraction from facial video signal using Deep Neural Network

ZAHID HASAN, University of Maryland, Baltimore County

Remote cardiovascular signal measurements provide the Photoplethysmogram (PPG) measurement from closely recorded skin videos. In this work, we propose a deep learning scheme to retrieve the PPG signal from the facial videos. We propose and validate personalized PPG extraction model on three video modalities: Near-infrared (IR), RGB, and RGB-demosaiced. We also explore the transfer learning aspects for PPG retrieval and propose two transfer schemes; between persons and modalities; for effective learning for deep neural network to estimate PPG from the facial videos. We analyze the behaviors of the learned convolutional filters (CNN) of the network for the video dataset. Our results of the personalized PPG model on a large publicly available dataset demonstrate the promise of the proposed personalized networks. Our experiments and results transfer learning across persons and modalities resonate the importance of initializing the network with pre-trained weights. Our results on CNN filters visualization provide the evidence for learning appropriate information and opens the prospect of backbone feature extraction network on facial videos.

CCS Concepts: • **Remote PPG → Deep Learning**; • **Neural Network** → *Representation Learning*; • **Transfer learning**;

Additional Key Words and Phrases: Neural networks, Transfer Learning, Video PPG monitor

ACM Reference Format:
Zahid Hasan. 2020. Photoplethysmogram signal extraction from facial video signal using Deep Neural Network. 1, 1 (May 2020), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Remote health monitoring systems enable technology for contactless physiological condition monitoring. This ensures the safety of both healthcare workers and probable patients by avoiding any contact with measurement equipment and person. The technology has been around for different purposes like measuring human body temperature via thermal camera, physiological information like heart rate, breathing rate measurement. These parameters are used for the initial screening for potential diseases like fevers, breathing problems. Recent study has shown the importance of remote health monitor for primary screen in case of contagious pandemics [10], [4]. Besides temperature, a remote sensing system can detect human breathing rate, heart with high accuracy. They are also an important symptom of different diseases and stress [1].

Photoplethysmogram (PPG) signal contains the blood volume change in a given human body section. PPG can provide important information regarding the physiological state [2]. Detecting PPG remotely provides would allow monitoring heartbeat without contact, providing safety in case of contagious patient monitoring. Examining remote PPG enables the system to examine the stress and physical condition of the subject [3]. Videos with enough frame rates and proximity of skin contain blood flow the information in the exposed region [12]. With appropriate signal processing and handcrafted features the face video can be used to extract PPG from the face.

Author's address: Zahid Hasan, zhasan3@umbc.edu, University of Maryland, Baltimore County.

XXXX-XXXX/2020/5-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The first step of getting appropriate skin is hard and error-prone due to the random shift and movement of face/skin in consecutive faces. To get accurate result consistent location of face regions in each frame needs to be extracted precisely. To avoid the handcrafted feature we look into the prospect of the neural network to automatically extract heart rate. Motivated by this prospect of extraction of heart rate from contact-less remote video sensor in this project I have selected a research project based on remote PPG retrieval from the face videos in this course.

Traditional extraction of PPG videos consists of two distinct parts: segmenting skin/face via face detection algorithm and apply signal processing techniques like filtering, PCA along with the video frames of the interested region [12]. The frequency component analysis over the extracted component provides heart rate information. Another line of research in these fields tries to predict heart rate from the video frames directly with the application of deep neural network architectures [19], [9]. There is a research gap in getting the exact PPG signal from the face video. In this project, I plan to extract the raw PPG signal from the raw videos of various imaging scheme using the deep neural network. I also aim to look at the prospect of the transfer learning aspect of neural networks in the context of video physiological information. The information aspect of the whole PPG motivated me further. It would also enable us to better understand the neural network behaviors to obtain the blood volume change signal. Moreover, it would also allow us to measure the signal strength of the PPG signal for controlling the sensor placement.

The hypothesis behind my approach is that there is some relation between face blood volume change and finger blood volume change. The neural network will be able to learn the relation between the face blood volume and finger blood volume based on a particular person in a controlled setting. The network should be able to get blood volume change in the face from the video and transform it into the finger PPG for the corresponding person. The first task (getting face PPG) is universal across the person. The problem here lies in the absence of measurement of face PPG as original ground truth since in most cases only finger PPG measures are available. Most of the works in the remote PPG (rPPG) domain focuses on getting heart rate only. In this work, we plan to particularly extract original PPG from the face video.

In this project, I address the problem of implementing the neural network to recover finger PPG signals on a publicly available dataset. I consider three different video modalities of videos in the prospect of PPG extraction. I also address the issues of the transferability of neural network models between different subjects and modalities to extract the PPG signal. I further dive inside the learned convolution filters and analyze their focus on the input video signal. By researching on the problems I contribute to the rPPG research as following,

- Firstly, I propose neural network architecture to extract raw PPG signal from different facial videos modalities on individual person with high accuracy. I validate the performance of the personalized end-end PPG extraction network by experimenting on multiple subjects.
- Secondly, I experiment and validate the transfer learning aspects of personalized neural network model between different subject and different image modalities. I look into the feature learning aspects and time required for customizing the pre-trained network for new subjects.
- Finally, I look into the neural network's convolution filter to understand and explain their behavior in the learnt model both in source domain and target domain. I seek to find a common backbone networks to extract appropriate features from different video modalities suitable for PPG extraction.

2 PRIOR ARTS

The literature's in rPPG can be broadly categorised in two major research directions. The classical approach of feature extraction for signal processing tools and the deep learning approach. The classical approach tries to find bare skin region and tracks the signal in the consecutive frames. This results a time-series signal from skin intensity in the video frames. Different signal processing techniques especially Green [17], ICA [15], CHROM

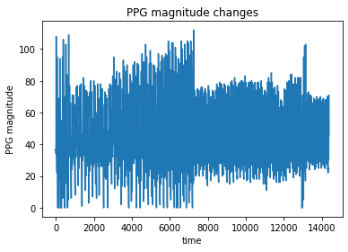
[7], POS [18], singular spectrum analysis [21], spectral peak tracking [20], non-linear mode decomposition [8] showed their methods abilities to estimate PPG and heart rate from time-series data of skin intensity. There are toolbox available for the extraction of PPG signal from facial features [11].

The another research direction focuses on application of deep learning methodologies to extract PPG from facial/ skin videos. The work in VitaMon [9] shows the neural networks ability to detect PPG peaks from raw frames. Works in DeepPhys [6] showed the neural networks ability to learn the facial features to decide on blood volume pressure (BVP). Two-Stream CNN [19] and 3d NN [5] approximated heart rate directly from video input. Neural network has also been used for skin detection in rPPG [16]. Neural network have been also used on extracted spatio-temporal video features to get the heart rate [13].

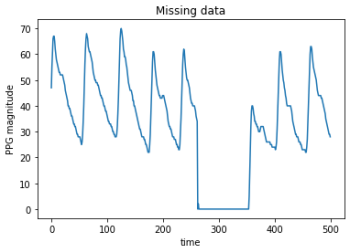
3 DATASET DESCRIPTION

For this project, I have used public datasets for this experiment [14]. The dataset contains 8 person's data. For each person, the dataset provides three modalities; RGB, RGB-demosaiced, and Near Infrared (IR) video along with the ground truth of the finger PPG. The database is of the size of 100GB. Each person contains on average 16GB of video data. The corresponding finger PPG data is also available as the ground truth. The dataset contains data of male and female participants. The male participants have beard variables in data. The details of data collection procedure are available in the source [14]

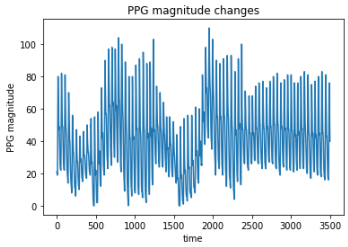
Since this project, I aim to extract the original finger PPG signal. The dataset offers great data in the rPPG region for heart rate estimation. The target output finger PPG is well collected in the dataset. But due to sensor placement variability and stochastic movement of the subject, the output sometimes got corrupted by various artifacts. The dataset poses some problems towards the direction of exact PPG extraction. The PPG for each subject is not aligned. This causes a huge problem. Sometimes PPG is missing in some frames. The PPG location of each person is varied which significantly changes the property of the target PPG signal. Another future direction of hand-crafted target for PPG from face removing finger PPG dependency. Face to face makes more sense. Neural Network to remove the face dependency. The shift in the output causes a huge error. It is really hard to remove data bias. For this shift and scaling issues of data output, it's hard to get a global network to extract the PPG signal across all the datasets. But, the dataset is sufficient to get a personalized network proposal on steady condition. In some cases, the magnitude range for PPG changes suddenly due to sensor movement or placement. Figure 1 shows some of the problems with the ground truth.



(a) PPG magnitude variability across dataset



(b) Missing PPG signal in the ground truth



(c) PPG signal changes by the presence of slow random trending artifact

Fig. 1. Different issues with the target ground truth PPG data

4 DATA PREPARATION

To create the training instances for training the neural network model, we consider the green channel of each of the video frames. The green channel consists of the most PPG signal as has higher absorption [17]. We first resize the image by $100x100$ and stack green channels of the 40 consecutive frames together which provides us with an input instance size of $100x100x40$. The 40 frames mark the 1.33 seconds video data as the original frame rate is 30 fps for the collected videos. We mark the starting position of the 40 video frames. As the assumption is that the PPG and video are synchronous together, we take the 1.33 seconds of the PPG signal starting from the same time of the video frame for the particular person. In this dataset, PPG is collected at a 60 Hz rate. So, to match the video we take 80 points of PPG from the same timestamp of the video frame starting. This becomes one training instance. By taking different starting points in the video we create the dataset. For each person, we consider 5000 different starting points, which returns us 5000 training instances per person for each modality. Both the input and output data has been also normalized between 0 to 1 for the network training. The summarized data preprocessing pipeline is as follow,

- Resize the video frames to 100x100
- From a random starting points in video stack 40 consecutive frame's green channel as input and take PPG signal from the same starting timestamp as the ground truth signal for the input.
- Select of 5000 different starting point and clip video frames and corresponding PPG signal to get 5000 training instances and normalization of both input and output between 0 to 1.

5 NETWORK DESIGN

The network design has been motivated by the work of [9]. I implement three layers of CNN followed by inception layers and two fully connected layers. Most of the parameters are at the CNN flatten to fully connected layer (FC layers) layers. In the initial layers, batch normalization and ReLU activation layers are following the CNN layers. The further information are available in the figure 2. I have used python 3.7 with Tensorflow 2.1 API for model design and implementation. For initial data preparation, the OpenCV library has been employed.

```
In [122]: neural_net.summary(line_length =80)
Model: "conv_net_2"
_________________________________________________________________________________
Layer (type)                    Output Shape                  Param #
=================================================================================
conv_bn_relu_18 (ConvBNRelu)    multiple                      11680
_________________________________________________________________________________
conv_bn_relu_19 (ConvBNRelu)    multiple                      18752
_________________________________________________________________________________
max_pooling2d_6 (MaxPooling2D)  multiple                      0
_________________________________________________________________________________
conv_bn_relu_20 (ConvBNRelu)    multiple                      37184
_________________________________________________________________________________
conv_bn_relu_21 (ConvBNRelu)    multiple                      55776
_________________________________________________________________________________
max_pooling2d_7 (MaxPooling2D)  multiple                      0
_________________________________________________________________________________
incept_mod_2 (InceptMod)        multiple                      58000
_________________________________________________________________________________
average_pooling2d_2 (AveragePooling multiple                  0
_________________________________________________________________________________
flatten_2 (Flatten)             multiple                      0
_________________________________________________________________________________
dense_6 (Dense)                 multiple                      4719104
_________________________________________________________________________________
dense_7 (Dense)                 multiple                      262656
_________________________________________________________________________________
dense_8 (Dense)                 multiple                      41040
=================================================================================
Total params: 5,204,192
Trainable params: 5,203,520
Non-trainable params: 672
_________________________________________________________________________________
```

Fig. 2. Network architecture and no of parameters in each layer. From CNN to FC layers have the most number of parameters.

In this work, as I am focusing on point-wise signal reconstruction. To penalize the reconstruction error the minimum squared error for each point is a potential candidate. The final layer output of the proposed network calculated the mean squared error (MSE) with the PPG of the corresponding input video frames. The Adam optimizer has been used to update the network weight to minimize the MSE of the network output and the target PPG signal.

$$MSE = \frac{\sum_{i=1}^{N} (y_{pred,i} - y_{true,i})^2}{N} \tag{1}$$

The other hyperparameters for the neural networks are the learning rate schedule. I begin the experiment with a learning rate of 0.001 for the optimizer and reduce by 1 percent after each epoch. In this experiment, the batch size was fixed to 16 for all the experiments. Finally, I controlled the randomness parameters of the model and data preparation for the reproducible result.

6 EXPERIMENTS

To meet the contribution criterion, I have experimented in mainly in three parts. Firstly to train on individual persons different modality and test on the left out validation set. In this experiment we train individual model on each person and each modality separately. We hold out in prior some video frames of the considered video data to validate the model performance. This provides us with three models for each person; on IR, RGB and RGB-demosaiced. We hypothesis that, each model learns the persons characteristic and the modality property to extract the raw PPG.

Secondly, I experimented on the transfer learning aspect of the neural network across the subjects of the same modalities and the modalities of the same subjects. This falls in the domain of inductive transfer learning since the target domain and the source domain have similar characteristics. Both the source and target domain have their labels available for our dataset. I further look at the representation learning of the model. I investigate this by observing the learned CNN filters and comparing their states after training.

Thirdly, I experimented with the overall learning capabilities of our proposed network across all the data. In this experiment, I have combined multiple subjected dataset and trained the model over all the dataset. During this experiment, I left one person's data out to conduct leave-one-out cross-validation. By leaving one person, we hypothesis, the model can only perform better on test data by learning the mechanism of getting raw PPG from video, not by memorizing the person's pattern. Throughout this experiment, I will use only IR modalities of the data for better explaining the model performance.

Throughout the experiments, the neural network architecture has not been modified. But some hyperparameters like learning rate, training epochs are fine-tuned for better convergence of the proposed deep learning model. For the sake of unbiased evaluation of model performance, there has been no post-processing. As the result section depicts the post-processing scheme like smoothing filtering or envelop detection can provide better visual results and reduced MSE on the test case. So the overall experiments I have conducted experiments to claim the contributions using the single neural network model are as follow,

- Train and validation on same person, single modality.
- Transfer learning prospect across person on same modality and same modality different persons.
- Training on multiple person single modality and test on different person on same video modality.

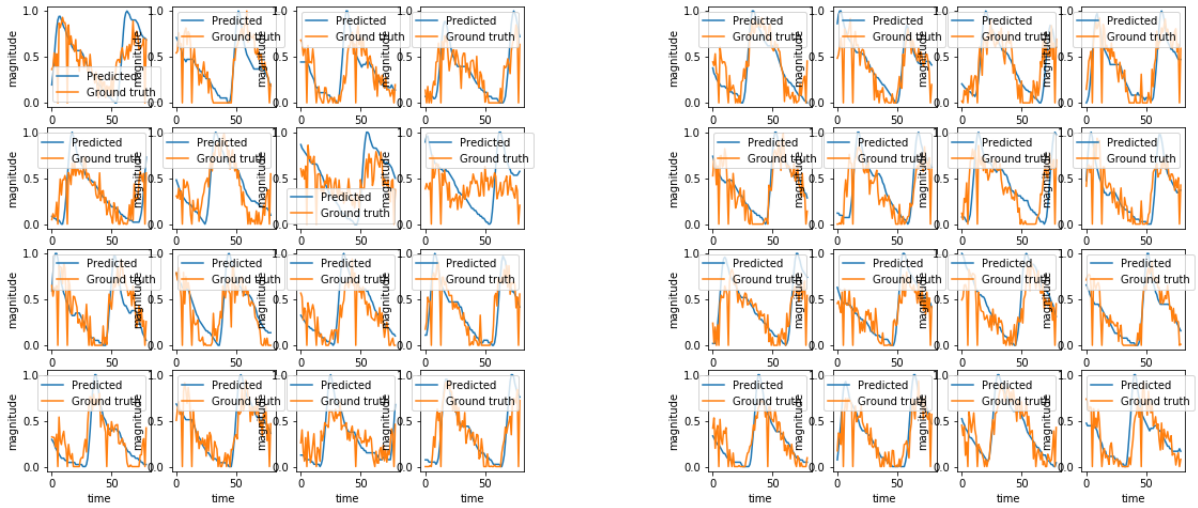
7 RESULTS AND DISCUSSION

Since we have trained our network to reduce the MSE between the ground truth and estimated signal, we will use the MSE between prediction and actual result for describing model performance. We will also visualize the model predicted signal alone with the original signal. We will look into the transfer learning aspect of the network in regard to the video data by considering two lines of the transfer prospects: Across persons and across modalities. Further, we plan to understand and explain learned neural network filters. We hope to find that the neural network focuses on bare skin to alone the axis to get the PPG signal from the face.

I have conducted three different experiments to show the contributions described in section 1. This section will cover the result and discussion from each of the experiments. Due to the combinatorial option available for

these experiments like person 1 IR, person 2 IR, person 1 RGB, or RGB to IR, IR to RGB for each person, person 1 RGB to person 3 RGB, person 1 IR to person 2 IR there is large number of possible test case scenario. For concise representation, the focus will be on subject 1 IR, transfer between subject 1 IR to RGB, and subject 1 IR to subject 2 IR. For the generalized model, I will show the model performance by training on subjects 1,2 and 3 IR and test on subject 4 IR. The other possible results are somewhat comparable with the result using IR modality.

Firstly, I focus on the result of personalized models. The validation results on the same subject and modality are promising. For this particular experiment, we first separate 15 seconds of person data to test the performances of the network. We train our network by the rest of the time frames using the discussed training scheme and data preparation. We train our model for 30 epochs on the training instances. Figure 3 shows the model effectiveness on extracting PPG on the two random test batches. The result shows that our neural network model successfully adopts the person's trait also while learning to extract PPG synchronous with the finger PPG data.



(a) Result from same person validation data

(b) Result on another random validation set

Fig. 3. Sample test result of the personalized trained model. The training and test data are from the same person and modality. The test set are from different video regions from training. (*blue lines are ground truth)

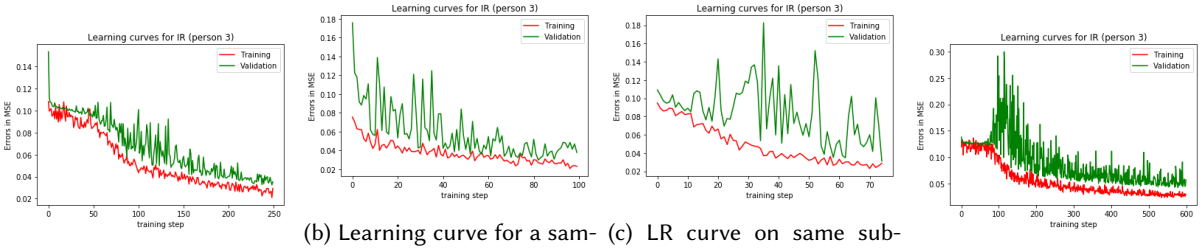
The following table summaries the training MSE and validation MSE for the models trained from scratch on individual subjects after 30 epochs on 5000 data instances. The subject 2 MSE is higher for some frames missing PPG. Subject 2's PPG property also changes during the collection procedure due to different stochastic factors like finger sensor location, movement, or physical condition. This has been discussed in the data description section. For the other modalities, the errors are comparable with the following results after the model trained properly.

I closely monitor the learning curves in these experiments to check the learning behaviors of the neural network. The sample learning curve from person 1 IR data shows effective learning and convergence of personalized model using a single dataset as depicted in figure 4a. It took about 30 epochs to reach the minimum error of 0.03 in training single person.

Next, I present and discuss the transfer learning results. We have two schemes for transfer learning; between subjects and between imaging modalities. At first, I discuss the transfer between modalities for the same person. In the dataset, the label for the source domain (training subject/modality) and the target domain (different subject/modality) have target data available. Since both input domains are controlled facial videos with background,

Subject	Training MSE Error	Validation MSE Error
Subject 1 (male, facial hair)	0.024	0.031
Subject 2 (female, no facial hair)	0.031	0.039
Subject 3 (male, facial hair)	0.028	0.034
Subject 4 (Male, no facial hair)	0.027	0.031
Subject 5 (male, no facial hair)	0.029	0.034
subject 7 (female, no facial hair)	0.03	0.035
Subject 8 (male, facial hair)	0.028	0.031

Table 1. Training and validation MSE for the personalized PPG extraction models



(a) Learning curve for a sample subject with random initialization

(b) Learning curve for a sample subject with pre-trained weight initialization (transfer across person)

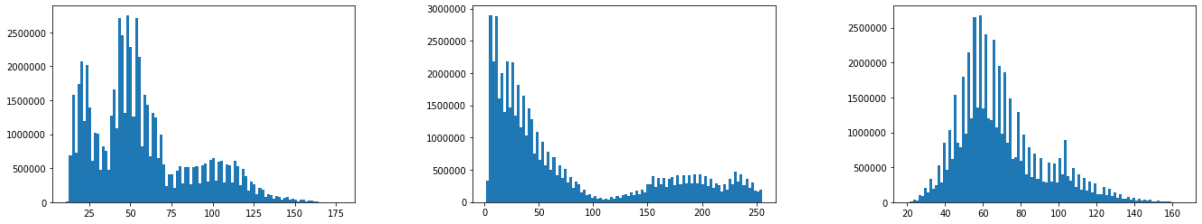
(c) LR curve on same subject's RGB modality with pre-trained weight on IR (transfer across modalities)

(d) Learning curve for the dataset combined from three subjects together

Fig. 4. Sample Learning curves of different models and training schemes.

they have a very similar distribution shape. But the pixel values for the face are different across modalities. I analyze their distribution properties. The result below shows input domain distributions are of similar shape with different means and variances. The face pixels values of IR and RGB images are in different ranges.

In figure 5 we see the histograms for IR and RGB side by side. It seems that the pixel values around 100 in IR has been shifted to the right for the RGB case. This distribution similarity and presence of label in target domain allows us to apply inductive transfer learning. The new task for the neural network to learn to focus on new pixel value groups and but similar spatial features in the target domain.



(a) Histogram of IR video: subject 1

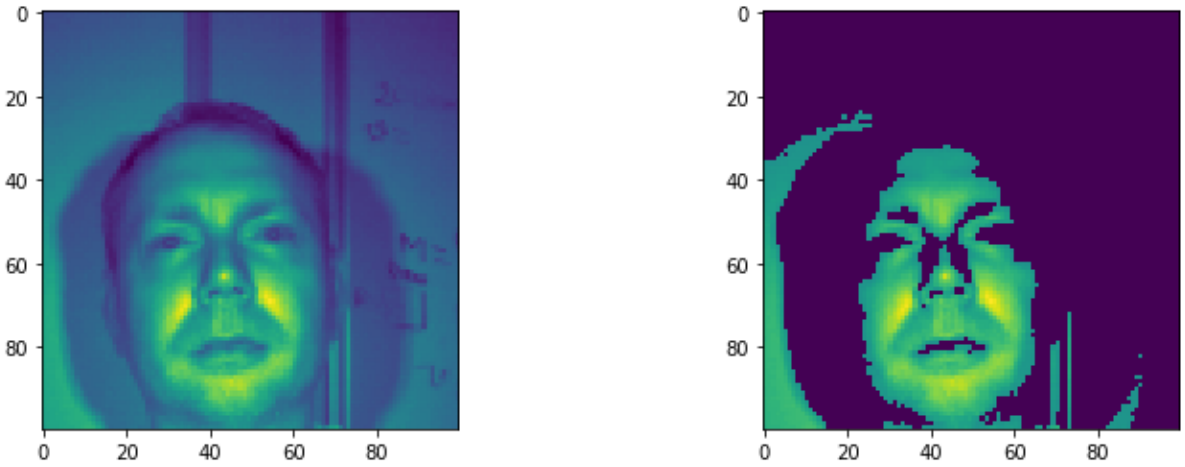
(b) Histogram of RGB video: subject 1

(c) Histogram of IR video: subject 4

Fig. 5. Comparison histogram for IR and RGB video of the same subject and IR of different object.

We further show the evidence for the previous section by showing image for pixel groups. Figure 6 shows that in IR modality the pixel value higher than 80 contains the face information. The hypothesis is also a valid cross

person transfer. By observing 5c and 5a we observe that the histogram for subject 1 and subject 4 for IR images are of similar distribution in the facial regions with different mean. The different means for face an IR camera pixel values for the same object can be explained by the exposure difference of persons to the camera thermal sensor and background temperature.



(a) IR image without modification

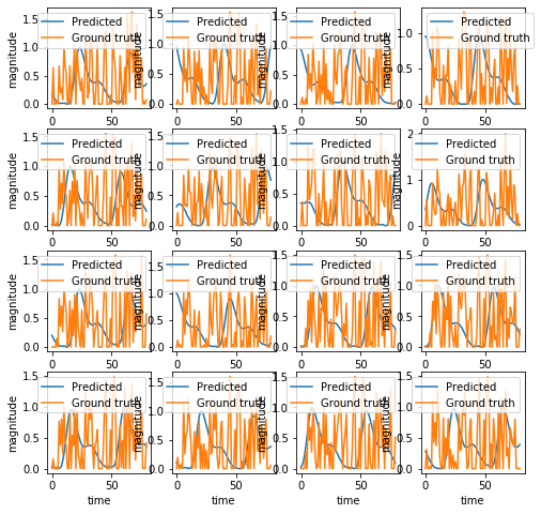
(b) IR image for pixel value of 80 or higher

Fig. 6. IR and RGB image with all the pixel and selected pixel region.

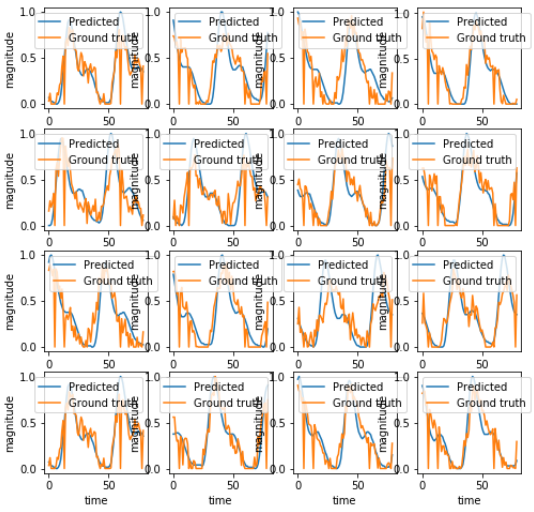
Firstly, I transfer the model between the modalities of image for the same person. As evident from earlier figure 5 we see than the pixel distribution between IR and RGB is different. It is expected to need more information for the neural network model to adapt the PPG extraction for the RGB camera as it has only seen the examples of IR. But the principles are identical both in IR and RGB data. We put forth the result of the pre-trained only model and compare it with the fine-tuning model while using the IR trained model on RGB data of the same person. Fine-tuning for different modalities needs longer training compared to transfer between person. The learning curve in figure 4c after weight initialization from different modality shows faster convergence for the network. The test batches result in figure 7 shows the capacity of networks quick learning in different domains after fine-tuning the network weights only for two epochs.

For personalized model transfer, I use the same neural network architecture by initializing with the previously trained model on different subject. The sample pre-trained model results on the new subjects test case are shown in figure 8a. The result shows that across a person the model fails to track the PPG peaks consistently. This can be explained by dataset problems and neural networks' ability to learn personal traits. But, after fine-tuning in the new domain data the model outputs almost overlap the ground truth PPG. This result is evident from figure 8b. We observe the improvement in test batches. We also observe the faster training for the model after initializing with the previously learned model on different subjects as the comparison is shown in figure 4b.

Thirdly, beyond personalized models, I conducted experiments on building a global model by training on three different subjects together. The learning curves are shown in figure 4d. After training, I tried to measure the performance of the neural network on a different subject in the same modalities. But the results are not convincing 9a. I suspect this is mostly because of the problem of the training dataset, not the neural network itself. As I discussed one of the problems in the data as different shifts in the PPG and the Video alignment. This

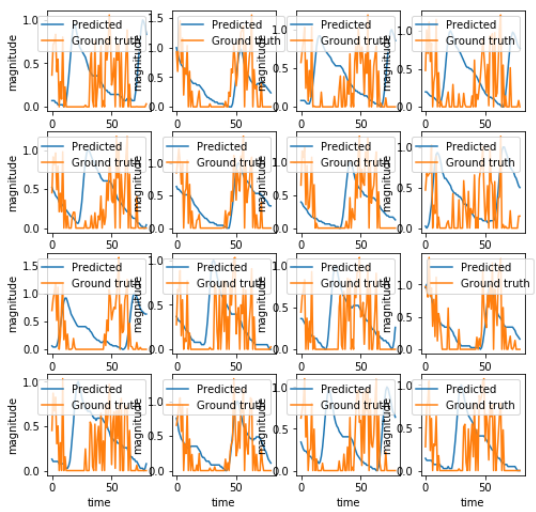


(a) Sample output result on different modality data of same person without any additional training

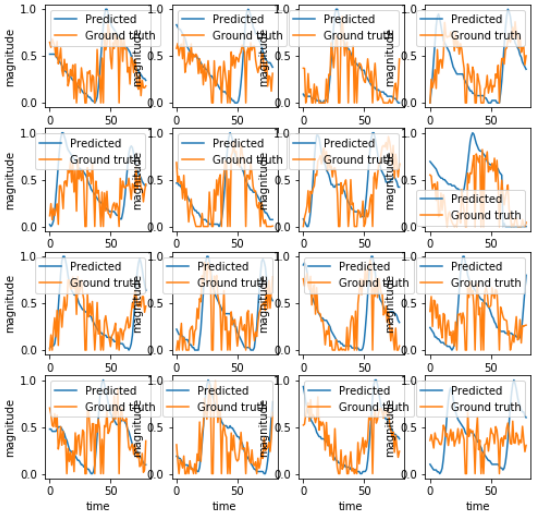


(b) Sample output result by little training on the target data domain for same person

Fig. 7. Comparison of across image modality transfer model result with pre-trained model without and with fine tuning (transfer across modalities) *blue lines are ground truth



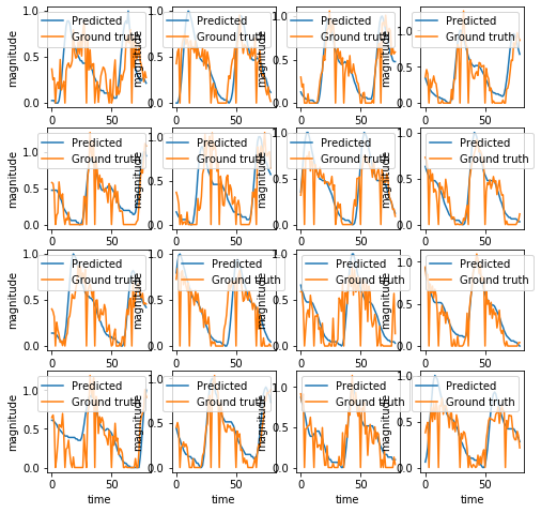
(a) Sample output result on transferred model without any additional training



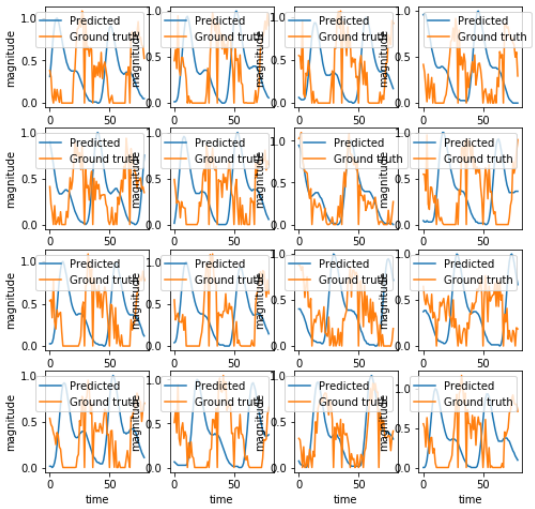
(b) Sample output result by little training on the target data

Fig. 8. Comparison of transfer model result with pre-trained model without and with fine tuning for a different subject's same video modality data (transfer across persons) * blue lines are ground truth

changes the whole learning structure. Since now the neural network has similar input and but multiple output patterns at different training instances. The neural network learns an interesting shortcut. To minimize the error it generates a PPG like structure in the middle. The results are shown in figure 9. In the case of test data from a seen subject the model can somewhat track the PPG but for the unseen subject the network randomly generates PPG shape marking its failure to tract the ground truth.



(a) Performance of model trained on multiple person on the test data of an involved person.



(b) Performance of model trained on multiple person on the test data of an excluded person.

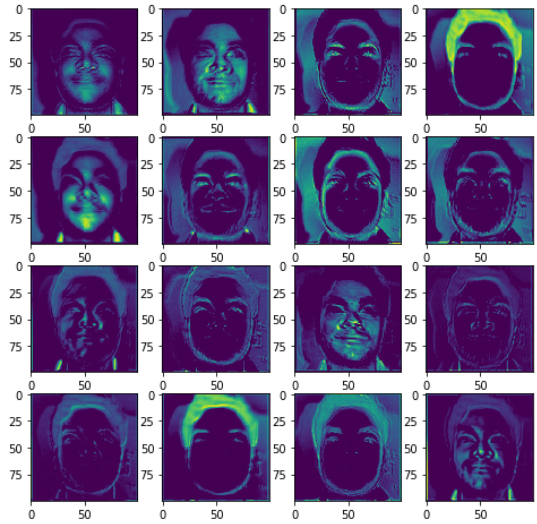
Fig. 9. Comparison of global model on test data from seen subject and unseen subject. *blue lines are ground truth

Lastly, while conducting the previous experiments, I simultaneously looked inside the filter layers to find the activation of the input in different layers of the neural network. For this, I took test images and pass through the trained network and observe the layer after CNN filters and ReLU activation. This provides an interesting prospect in this research domain. The hypothesis is that the correctly learned neural network should focus on the face while getting the PPG signal. Figure 10 shows the activation of the face in the layers as features for PPG extraction. This result ensures the focus of neural networks in estimating PPG from facial videos. Some of the initial filters focus on different unrelated regions like hair, background as seen in figure 10a. But in the deeper network seems to focus entirely on different face regions as evident in figure 10d. These results are promising for further analysis of filter behaviors and finding the most important filters for feature extraction.

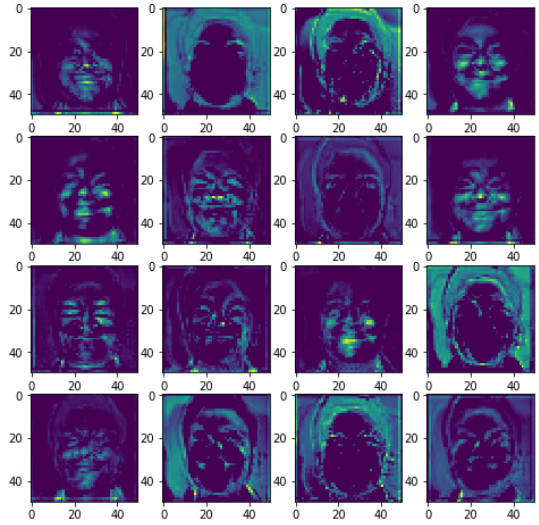
In transferring the model between the modalities of the same person, we find that the CNN layers activation changes and learns to focus more on faces for the new modality after the fine-tuning. The result in figure 11 depicts the neural network learns to activate face more in RGB after retraining with the new modality data as the initial model was trained upon IR images.

8 LIMITATION

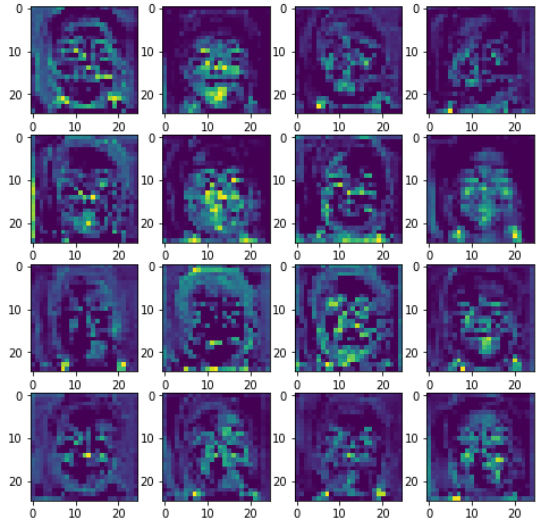
The current limitation of our current architecture is the ReLU activation at the output layer. The network is tasked to estimate time series data but consists of no recurrent blocks in the architecture. This hindered the smooth outputs in the final layers as there is no communication between layer output nodes. Further, if we see



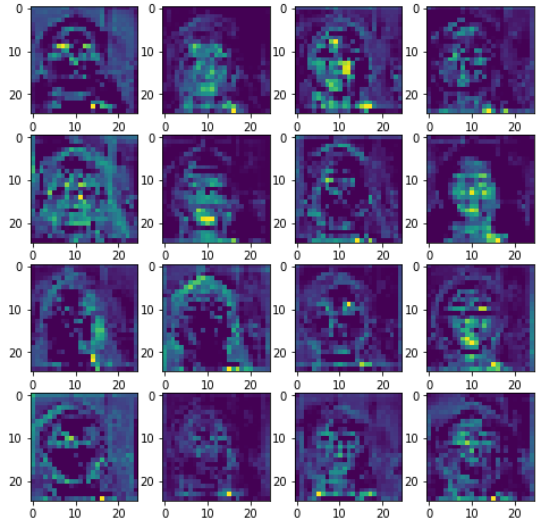
(a) The intermediate filer output of an sample test input



(b) The intermediate filer of layer 4 output of an sample test input



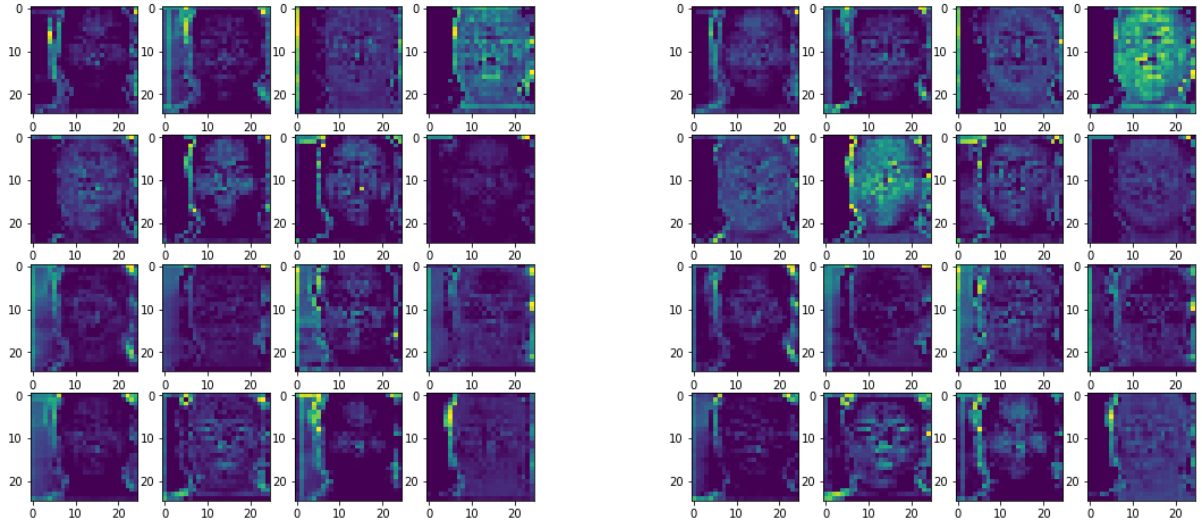
(c) The intermediate filer of layer 6 output of an sample test input



(d) The intermediate filer of layer 6 output of an sample test input on different dataset

Fig. 10. Inside the neural network: In the later layers neural network focus on faces

the distribution of the parameters over the network we find parameter heavy FC layers layers after flatten the CNN outputs. Most of the features are extracted by CNN layers because of their capabilities to learn spatial features but the FC layers has the most number of parameters. Another interesting learning behavior have been



(a) Output the network layer 6 for the RGB image (Pre-trained on IR videos)

(b) Sample output result by little training on the target data

Fig. 11. Comparison of transfer model result with pre-trained model without and with fine tuning for same subject's different video modality data (Transfer from RGB to IR)

observed in NN. In global model it learned to generate random PPG shapes to minimize the errors, which is potentially misleading.

9 FUTURE WORKS

As I discussed some of the problems with the dataset, I seek to get a unbiased data for the PPG extraction task. In the future, I aim to design our system of PPG extraction. To propose and validate such a system, I seek to collect our video heart rate data and expand the horizon of the rPPG research.

We have our preliminary results on the public dataset. I plan to extend the works based on the limitation of the current work. The first problem our network contains is the lack of learning capacity for the time series data. We have used only CNN and FC layers in our deep neural network (NN) model so far. They capture the spatial information quite well. But the network lacks layers to learn the temporal dependency inside the output PPG. The PPG has time-series data and contains certain temporal information. We believe adopting a recurrent neural network (RNN) or long short term memory (LSTM) on top of CNN extract features would enable the network to learn the time characteristic of the output and smoothing the output PPG signal. My plan of extending the work is via applying LSTM in the neural network to better capture the temporal pattern of the output.

In this work, I prepared the input by taking 40 frames together obviating the necessity for the temporal learning across different inputs. In our set up, the input instance of 40 frames contains the necessary information for extracting PPG. In my case, CNN has been deployed to learn the temporal features in terms of spatial features. Another way to extend the work can be considering LSTM at the input side and dis-entangle the input frames. This would be a sequence to a sequence learning problem. Another prospect in future work can be the implementation of a multitasking model for different subjects. In this direction, I plan to extend the network in two stages. Firstly, a shared backbone network for facial features detection from video frames for a fixed modality. The second stage

is a personalized network for extracting PPG signals from the common features. Training such a network should train the shared backbone network better to find the implicit features. Multitasking can also improve the feature for a particular image modality. As an extension to these, I believe multitasking at the initial layers for different modality followed by the shared network also might resolve the image modality variability. In this case, we need a separate network for each modality to transform video frames in some common latent space. Then followed by a shared network for facial feature detection and finally some personalized layers to capture the characteristics for individual subjects. In the test phase, only the individual network needs to be fine-tuned.

In some cases, facial video to finger PPG makes little sense. In this work, the initial hypothesis is that the video signals are related to face PPG and face PPG are related to fingering PPG. So with an appropriate learning scheme, the learner can learn the transformation from face video to finger PPG. Instead of learning finger PPG, the model to learn face PPG makes more sense. But the problem is the availability of face PPG. The current algorithm allows us to compute face PPG by selecting the specific face region and applying the signal processing technique. The major drawbacks of this approach are the consistent face tracking in video. To create the target face PPG signal, we can use a handcrafted method then train the neural network to learn the inherent actions implicitly to get face PPG from face video. This will be my next direction to work with. So, I will be preparing new ground truth by handcrafted labeling and train the network to learn all the steps end to end to enable complete automation.

In line with the face video to face PPG, object detection algorithms may play a crucial role. We need teach the network to learn to detect skin in consecutive video frames for correct signal extraction. Object detection backbone like feature pyramid networks, region proposal network might localize the region of interest for PPG signal and rest networks can focus of PPG extraction from the segmented video section. This problem can be formulated as another multitasking learning frameworks but the sub-task learning of skin detection will be performed inside the network instead of separate task proposed in earlier sections. This work has potential to generate PPG from each segmented skin portions.

Another important issue I didn't discuss too much in this current work is the loss function, the key driver to train the learner network. Here, I implement vanilla MSE loss between the network output and ground truth. I think this is not enough to appropriately measure model performance. The MSE ignores the correctly detected peak or crest position which is an important criterion for network performance. I plan to look into different shape similarity metrics to reward/penalize the network based on the shape matching with the target. I also look to shift variant loss function for cross person transfer. I want to penalize the network for a shift in the target output. This would teach the network to shift the output from the backbone network in the transfer learning scheme.

Lastly, the neural network architecture was largely motivated by earlier works. Model hyper-parameter tuning may provide better estimation with the current setup. I aim to look into the architecture issues in the future task. The advanced pruning and network compression techniques technique may provide a comparable result with reducing parameter numbers. This will enable to implement the network in various edge and memory constraint device for ubiquitous health monitoring.

10 CONCLUSION

In this project work, We have successfully demonstrated the neural networks' potential to end-end extraction of exact PPG signals from raw facial videos of different imaging modalities. We propose three personalized PPG extractor neural network architecture and validate the performance of the model on publicly available data. Besides, we present the transfer learning scheme between person to person and IR to RGB and vice-versa for the models providing support using the data distribution. The results show the efficacy of initial weight initialization by the pre-trained networks. Finally, we closely monitor the features of learned filters in the source domain and target domain in case of the video heart rate monitoring data. Our current results on the dataset are encouraging

to extend our works for top-tier conference publications. For that, we have put forth the probable future direction we can advance towards.

11 ACKNOWLEDGMENT

I acknowledge the meticulous guidance, suggestions, encouragement and funding of Dr. Nirmalya Roy throughout the project. I also acknowledge the suggestions and effective discussion with Sreenivasan Ramasamy Ramamurthy. I acknowledge MERL-lab and UBFC for generously sharing their dataset.

REFERENCES

[1] 2020. Common Covid 19 Symptomes. https://www.webmd.com/lung/covid-19-symptoms#1
[2] 2020. Photoplethysmogram wikipedia information. https://en.wikipedia.org/wiki/Photoplethysmogram
[3] 2020. Stress impacts heart rate. https://www.webmd.com/balance/stress-management/qa/can-stress-impact-your-heart-rate-and-blood-pressure
[4] 2020. Thermal camera to detect fever. https://www.wired.com/story/can-an-infrared-camera-detect-a-fever/
[5] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 2019. 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video. *Applied Sciences* 9, 20 (2019), 4364.
[6] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 349–365.
[7] Gerard De Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.
[8] Halil Demirezen and Cigdem Eroglu Erdem. 2018. Remote photoplethysmography using nonlinear mode decomposition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1060–1064.
[9] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. 2019. VitaMon: measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 1–14.
[10] Ruby Lane and Merissa Kelley. 2020. How Infrared Cameras Can Help Prevent the Spread of COVID-19. https://infraredcameras.com/in-the-news/infrared-cameras-coronavirus-spread/
[11] Daniel McDuff and Ethan Blackford. 2019. iphys: An open non-contact imaging-based physiological measurement toolbox. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6521–6524.
[12] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. 2014. Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering* 61, 12 (2014), 2948–2954.
[13] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. 2019. RhythmNet: End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation. *IEEE Transactions on Image Processing* (2019).
[14] Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavany. 2018. SparsePPG: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1353–135309.
[15] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express* 18, 10 (2010), 10762–10774.
[16] Chuanxiang Tang, Jiwu Lu, and Jie Liu. 2018. Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1309–1315.
[17] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. 2008. Remote plethysmographic imaging using ambient light. *Optics express* 16, 26 (2008), 21434–21445.
[18] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (2016), 1479–1491.
[19] Zhi-Kuan Wang, Ying Kao, and Chiou-Ting Hsu. 2019. Vision-Based Heart Rate Estimation Via A Two-Stream CNN. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3327–3331.
[20] Bing-Fei Wu, Po-Wei Huang, Chun-Hsien Lin, Meng-Liang Chung, Tsong-Yang Tsou, and Yu-Liang Wu. 2018. Motion resistant image-photoplethysmography based on spectral peak tracking algorithm. *IEEE Access* 6 (2018), 21621–21634.
[21] Changchen Zhao, Chun-Liang Lin, Weihai Chen, and Zhengguo Li. 2018. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1299–1308.