ABSTRACT

MATHEW, JOSHUA REGI. Remote Blood Oxygen Estimation From Videos Using Neural Networks. (Under the direction of Chau-Wai Wong).

Peripheral blood oxygen saturation (SpO_2) is an essential indicator of respiratory functionality and is receiving increasing attention during the COVID-19 pandemic. Clinical findings show that COVID-19 patients can have significantly low SpO_2 before any obvious symptoms. The prevalence of cameras has motivated researchers to investigate methods for monitoring SpO_2 using videos. Most prior schemes involving smartphones are contact-based: They require using a fingertip to cover the phone's camera and the nearby light source to capture re-emitted light from the illuminated tissue. In this paper, we propose the first convolutional neural network based noncontact SpO_2 estimation scheme using smartphone cameras. The scheme analyzes the videos of a participant's hand for physiological sensing, which is convenient and comfortable and can protect their privacy and allow for keeping face masks on. We design the neural network architectures inspired by the optophysiological models for SpO_2 measurement and demonstrate the explainability by visualizing the weights for channel combination. Our proposed models outperform the state-of-the-art model that is designed for contact-based SpO_2 measurement, showing the potential of our proposed method to contribute to public health. We also analyze the impact of skin type and the side of a hand on SpO_2 estimation performance.

Remote Blood Oxygen Estimation From Videos Using Neural Networks

by
Joshua Regi Mathew

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina
2021

APPROVED BY:

Edgar Lobaton

Tianfu Wu

Chau-Wai Wong
Chair of Advisory Committee

ACKNOWLEDGEMENTS

Thanks to Dr. Wong for your guidance throughout all my research projects. Thanks to Dr. Lobaton and Dr. Wu for being part of my thesis committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

The proliferation of smart devices equipped with various sensors and more powerful hardware has made it commonplace for convenient, at home measurement of biological signals. There has also been development in non-contact video based measurement of various vital signs such as heart rate (De Haan and Jeanne 2013; Li et al. 2014; Tulyakov et al. 2016; Zhu et al. 2017), respiratory rate (Chen et al. 2020; Poh et al. 2010), heart rate variability (Iozzia et al. 2016; Poh et al. 2010; Favilla et al. 2018), and temperature (Zheng et al. 2020). The potential of these contactless measurement methods are especially apparent during outbreaks like the current COVID-19 pandemic, where they can be used to measure vital signs without risk of spreading the disease.

Blood oxygen (SpO_2) is one vital sign that has been shown to have a direct link to COVID infection. One symptom of COVID-19 infection is silent hypoxia, or drop in oxygen saturation; yet many do not exhibit respiratory symptoms (Couzin-Frankel 2020; Starr et al. 2020). This highlights the importance of accurate detection of changes in SpO_2 to facilitate timely management to prevent rapid deterioration. In addition, SpO_2 monitoring is essential for the evaluation of respiratory health and screening for pulmonary diseases (Nitzan et al. 2014). The conventional method for SpO_2 measurement, pulse oximetry, relies on a contact based sensor attached to the finger tip (Severinghaus 2007). Even currently available smart

watch and smart phone based SpO_2 measurement require the device to be in contact with the person (Scully et al. 2011; Lu et al. 2015; Ding et al. 2018). Contact-based methods present the risk of cross-contamination between individuals using the same measurement device. An additional issue with contact-based methods is limb perfusion, especially in the digits. Circulation to the fingers and toes is often impaired in many cardiovascular and pulmonary diseases, which complicates measurement of SpO_2. Also, pulse oximeters may not be widely available in poorer communities and developing countries (Herbert and Wilson 2012).

It has been shown that the circulation of blood through the body induces subtle periodic changes in skin color which can be captured by a camera and extracted from video to accurately measure SpO_2 (Kong et al. 2013; Van Gastel et al. 2016; Shao et al. 2015; Tsai et al. 2016; van Gastel et al. 2019; Bal 2015; Casalino et al. 2020; Tarassenko et al. 2014). In this paper we introduce explainable neural network models for feature extraction and SpO_2 measurement from contactless videos captured using consumer-grade smart phone cameras. To the best of our knowledge, this is the first work using neural networks to remotely monitor SpO_2.

We propose using convolutional neural net-works (CNNs) for contactless SpO_2 monitoring from videos captured by smartphone cameras. Fig. 2.1 gives the overall pipeline for our system. First, we segment the region of interest (ROI), either the palm or back side of the hand, from the video background. Next, the ROI is spatially averaged to produce R, G, and B time series. Finally, the three time series are fed into an CNN architecture inspired by optophysiological models to extract meaningful features and estimate the SpO_2 (Webster 1997; Van Gastel et al. 2016; Scully et al. 2011). Compared to using the face for SpO_2 measurement as in prior art (Bal 2015; Tarassenko et al. 2014), recording hand videos raises less privacy concern to the participants and allows participants to keep their masks on in accordance with the guidelines for human subject research during the COVID-19 pandemic. The contributions of our work are summarized as follows:

- This is the first work to use neural networks to address the challenging problem of contactless SpO_2 sensing using consumer-grade RGB cameras.
- Through a data-driven approach and visualization of the weights for the RGB channel combinations, we demonstrate the explainability of our model and that the choice of the color band learned by the neural network is consistent with the suggested color bands used in the optophysiological methods.

- We analyze the impact of the two sides of the hand and different skin tones on the quality of SpO_2 estimation.
- We achieve more accurate SpO_2 estimation with our optophysiologically inspired neural network structures when compared to the state-of-the-art neural network structure for contact-based SpO_2 prediction.

CHAPTER

2

BACKGROUND AND RELATED WORK

2.1 Blood Oxygen Saturation and the Ratio of Ratios Principle.

Blood Oxygen Saturation is the ratio of the concentration of oxygenated hemoglobin (HbO_2) to the total oxygenated and deoxygenated hemoglobin (Hb) in the blood.

$$SpO_2(\%) = \frac{HbO_2}{Hb + HbO_2}(\%) \tag{2.1}$$

SpO_2 is an important indicator of the ability of the respiratory system to meet metabolic demands. High SpO2 indicates adequate respiratory function with the normal healthy SpO_2 range being between 95% and 100% (Nitzan et al. 2014). Low SpO_2 levels can be an indicator of respiratory illness (Nitzan et al. 2014). The most widespread method for SpO_2 measurement in health care facilites and at home is pulse oximetry (Severinghaus 2007). Pulse oximiters utilize the *principle of ratio of ratios* which takes advantage of the different in light absorption characteristics between oxygenated and deoxygenated hemoglobin at the red and infrared wavelengths as shown in Fig 2.2. Red and infrared light are emitted

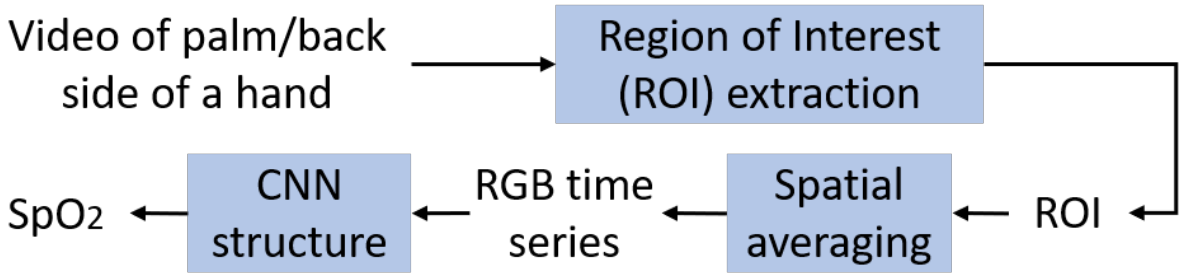


Figure 2.1: Proposed SpO_2 estimation method. Three color time series are extracted from the skin area of a hand video, and are then fed into an optophysiology-inspired neural network for SpO_2 prediction.

through one end of fingertip and an optical sensor at the other end captures the transmitted light, which has been interacted with and attenuated by the blood and tissue. The properties of the transmitted light conveys information about the pulsatile blood volume which is further processed to obtain an SpO_2 estimate.

2.2 Video Based SpO_2 Measurement

It has been shown that skin color changes induced by blood circulation is captured in video and can be used to estimate many vitals signs. There has been recent work done exploring contact-based measurement of SpO_2 from video. Contact based methods involves pushing the fingertips against a smartphone camera and light source and analyzing the diffusely reflected light captured by the camera (Scully et al. 2011; Lu et al. 2015; Ding et al. 2018). These works use a modified ratio of ratios method utilizing some combination of the red, blue, or green color channels from the captured RGB video instead of red and infrared light as is done in pulse oximetry Nemcova *et al.* and Lamonaca *et al.* use a modified version of the ratio of ratios using the red and green channels to estimate SpO_2 from video (Nemcova et al. 2020; Lamonaca et al. 2015).

There are two groups of noncontact SpO_2 estimation methods from video. The first category utilizes monochromatic sensing similar to pulse oximetry. They use either special monochromatic cameras with selected optical filters or controlled monochromatic light sources (Kong et al. 2013; Van Gastel et al. 2016; Shao et al. 2015; Tsai et al. 2016). The other category uses consumer-grade RGB cameras (Tarassenko et al. 2014; Bal 2015; Casalino et al. 2020; Al-Naji et al. 2021). All the noncontact SpO2 estimation methods rely on opto-

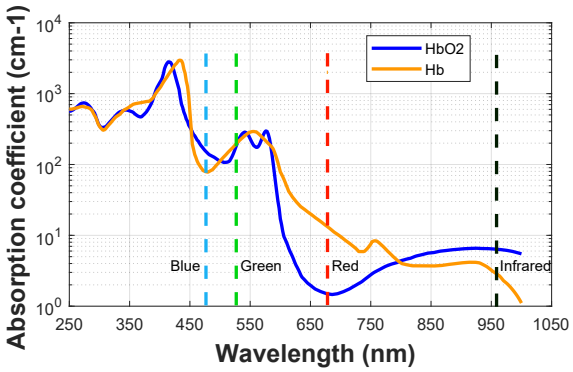


Figure 2.2: Extinction coefficient curves of hemoglobin. The curves were plotted based on (Ding et al. 2018; hbc 2020). The difference between oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) at the red and blue wavelengths means that these color channels contain useful information for SpO_2 prediction by means of optophysiological principles.

physiological modelling or some form of the ratio of ratios principle. The monochromatic light sources and sensors are selected to have accurate control of the absorption effect of hemoglobin, while the consumer-grade digital cameras, including webcams and smartphone cameras, have a wider sensing band and are more challenging for SpO2 sensing.

2.3 Deep Learning Aided Camera-based Physiological Monitoring.

There have been many developments in the application of deep learning for physiological monitoring in recent years. Neural networks have been used to estimate heart rate and breathing rate (Niu et al. 2019; Chen and McDuff 2018; Špetlík et al. 2018). Convolutional neural networks have been utilized to blood volume pulse signals from videos which are manually analyzed to obtain heart rate and breathing rate (Chen and McDuff 2018). CNNs have also been used to infer the heart rate directly from video (Niu et al. 2019). Mobile applications have been developed which utilize CNNs to measure body temperature form facial images (Zheng et al. 2020).

There has been little research into the use of deep learning for the estimation of SpO_2. Ding *et al.* investigated the use of CNNs for SpO_2 estimation from contact-based smart phone videos (Ding et al. 2018). They have demonstrated that their model is capable of outperforming the ratio of ratios method. Their work has shown the potential for neural

network based SpO_2 estimation from video but is limited to scenarios where the finger must be in contact with the camera. This motivates us to investigate neural network models that can estimate SpO_2 from contactless videos taken at a distance from the hand which would avoid issues with sanitation and patient isolation.

CHAPTER

3

PROPOSED METHOD FOR SPO_2 FROM VIDEOS

We aim to estimate SpO_2 levels using a hand video by leveraging the fact that the color of the skin changes subtly when red cells in the blood carry/release oxygen. In our proposed method, we extract three color time series from the skin area of the hand video. We feed the extracted time series to optophysiology-inspired neural networks designed to achieve better and more explainable SpO_2 predictions.

3.1 Extraction of Skin Color Signals

The physiological information related to SpO_2 is embedded in the color of the reflected/reemitted light from a person's skin. Hence, a preprocessing step that precisely extracts the color information from the skin area is crucial to the design of an effective SpO_2 estimation method. For each participant's video, we aim to extract the R, G, and B time series and refer to these 1-D time series as *skin color signals*. We first need to locate the ROI of the skin pixels from the video. We found that it is most effective to discriminate the skin pixels from the

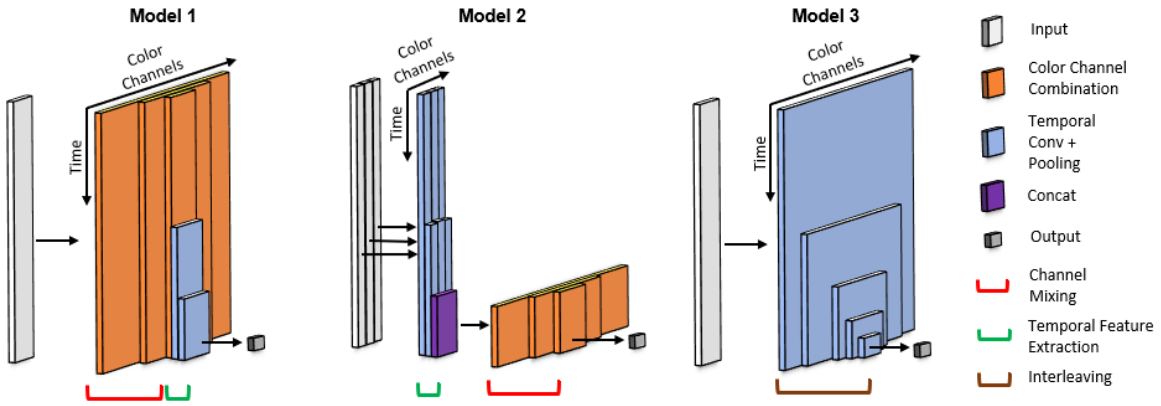


Figure 3.1: Proposed network structures for predicting an SpO_2 level from a fixed-length segment of skin color signals. We highlight the differences among the three model configurations instead of showing the exact model structures. Model 1 combines the RGB channels before temporal feature extraction. Model 2 extracts the temporal features from each channel separately and fuses them toward the end. Model 3 interleaves color channel mixing and temporal feature extraction.

background along the Cr color direction of the YCbCr color space (Burger and Burge 2008).

We use Otsu's method (Otsu 1979) to determine a threshold that best separates the skin pixels from the background by minimizing the variance within the skin and non-skin classes. Once the ROI corresponding to the hand is located, the R, G, and B time series are generated by spatially averaging over the values of skin pixels for each frame of the video.

The skin color signals are split up into 10-second segments using a sliding window with a step size/stride of 0.2 seconds to serve as the inputs for neural networks. From an opto-physiological perspective, the reflected/reemitted light from the skin for the duration of one cycle of heartbeat, i.e., 0.5–1 seconds for a heart rate of 60–120 bpm, should contain almost the complete information necessary to estimate the instantaneous SpO_2 (Severinghaus 2007). In our system design, we use longer segments to add resilience against sensing noise. Since the segment length is one order of magnitude longer than the minimally required length to contain the SpO_2 information, we can use a fully-connected or convolutional structure to adequately capture the temporal dependencies without resorting to a recurrent neural network structure.

3.2 Neural Network Architectures

The previous neural network work for SpO_2 prediction mainly explored prediction, but not the model explainability (Ding et al. 2018). Explainability/interpretability is highly desirable in many applications yet often not sufficiently addressed, partly due to the black box nature of neural networks. From a healthcare standpoint, explainability is a key factor that should be taken into account at the beginning of the design of a system. To extract features from the skin color signals and estimate SpO_2, we propose three physiologically motivated neural network structures. These structures are inspired by domain knowledge-driven physiological sensing methods and designed to be physically explainable. For heart rate sensing (Zhu et al. 2017; Niu et al. 2019) and respiratory rate sensing (Nam et al. 2015; Sohn et al. 2017), the RGB skin color signals are often combined first followed by temporal feature extraction, as is done in the plane-orthogonal-to-skin (POS) algorithm (Wang et al. 2016). In contrast, for conventional SpO_2 sensing methods such as the ratio-of-ratios (Webster 1997), the temporal features are extracted first and the color components are combined at the end. Our proposed neural network structures explore different arrangements of channel combination and temporal feature extraction. We want to systematically compare the performance of our explainable model structures.

Channel Mixing Followed by Feature Extraction. In Model 1, shown as the leftmost structure depicted in Fig. 3.1, we combine the color channels first using several channel combination layers and then extract temporal features using temporal convolution and max pooling. A channel combination layer first linearly combines the C_{in} input channels/vectors into C_{out} activation vectors and then applies a rectified linear unit (ReLU) activation function to obtain the output channels/vectors. Mathematically, the channel combination layer is described as follows:

$$\mathbf{V} = \sigma(\mathbf{W}\mathbf{U} + \mathbf{b}\mathbb{1}^T), \tag{3.1}$$

where $\mathbf{U} \in \mathbb{R}^{C_{\text{in}} \times L}$ is the input comprised of C_{in} time series/vectors of length L. The initial channel combination layer has an input of three channels with 300 points along the time axis. $\mathbf{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ is a weight matrix, where each of the C_{out} rows of the matrix is a different linear combination for the input channels. A bias vector $\mathbf{b} \in \mathbb{R}^{C_{\text{out}}}$ contains the bias terms for each of the C_{out} output channels, which ensures that each data point in the artificially created segment of length L has the same intercept. $\mathbb{1}^T \in \mathbb{R}^{1 \times L}$ is a row vector of all ones. The nonlinear ReLU function $\sigma(x) = \max(0, x)$ is applied elementwise to the

activation map/matrix. The output of the channel combination layer $\mathbf{V} \in \mathbb{R}^{C_{\text{out}} \times L}$ contains C_{out} channels of nonlinearly combined input channels.

The channel mixing section concatenates multiple channel combination layers with decreasing channel counts to provide significant nonlinearity. The output of the last channel combination layer has seven channels. After the channel mixing, for temporal feature extraction, we utilize multiple convolutional and max pooling layers with a downsampling factor of two to extract the temporal features of the channel-mixed signals. When there are multiple filters in the convolutional layer, then there will also be some additional channel combining with each filter outputting a channel-mixed signal. Finally, a single node is used to represent the predicted SpO_2 level. This model has three channel combination layers, three feature extraction layers, and 34K trainable parameters.

Feature Extraction Followed by Channel Mixing. In Model 2, the middle structure depicted in Fig. 3.1, we reverse the order of channel mixing and temporal feature extraction from that in Model 1. The three color channels are separately fed for temporal feature extraction. The convolutional layers learn different features unique to each channel. At the output of the temporal feature extraction section, each color channel has been downsampled to retain only the important temporal information. The color channels are then mixed together in the same way as described for Model 1 before outputting the SpO_2 value. This model has three channel combination layers, 2 feature extraction layers, and 12K parameters.

Interleaving Feature Extraction and Channel Mixing. In our third model, we explore the possibility of interleaving the color channel mixing and temporal feature extraction steps. As illustrated by the rightmost structure depicted in Fig. 3.1, the input is first put through a convolutional layer with many filters and then passed to max pooling layers, resulting in feature extraction along the time as well channel combinations through each filter. The number of filters is reduced with each successive convolutional layer, gradually decreasing the number of combined channels and downsampling the signal in the time domain. This model has 4 layers and 307K parameters.

Loss Function and Parameter Tuning. We use the root-mean-squared-error (RMSE) as the loss function for all models. During training, we save the model instance at the epoch with the lowest validation loss. The neural network inputs are scaled to have zero mean and unit variance to improve the numerical stability of the learning. The parameters and hyperparameters of each model structure were tuned using the HyperBand algorithm (Li et al. 2018) which allows for faster and more efficient search over a large parameter space

than grid search or random search. It does this by running random parameter configurations on a specific schedule of iterations per configuration, and uses earlier results to select candidates for longer runs. The parameters that were tuned include the learning rate, the number of filters and kernel size for convolutional layers, the number of nodes, the dropout probability, and whether to do batch normalization after each convolutional layer.

CHAPTER

4

EXPERIMENTAL RESULTS

4.1 Dataset and Capturing Conditions

Our proposed models were evaluated on a self-collected dataset. The dataset consisted of hand video recordings and SpO_2 data from 14 participants, of which there were six males and eight females between the ages of 21 and 30. Participants were asked to categorize their skin tone based on the Fitzpatrick skin types (ski 2020). The distribution of the participants' skin types is as follows: Two participants of type II, eight participants of type III, one participant of type IV, and three participants of type V. This research was using protocol #1376735-2 approved by the University of Maryland Institutional Review Board (IRB).

Our dataset consists of four recordings per participant for a total of 56 recordings. Each participant was asked to place his/her hands still on a table to avoid hand motion. Their palm of the left hand and the back of the right hand are facing the camera, as illustrated in Fig. 4.1. We refer to these two hand-video capturing positions as *palm up (PU)* and *palm down (PD)*, respectively. Each participant was asked to follow the breathing protocol outlined in Fig. 4.2a. The participant breathes normally for 30–40 seconds and then holds his/her breath for 30–40 seconds, and this process is repeated three times for each recording.

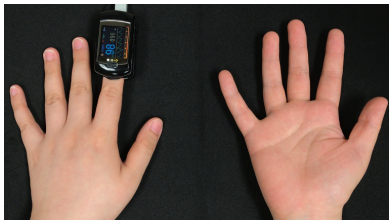
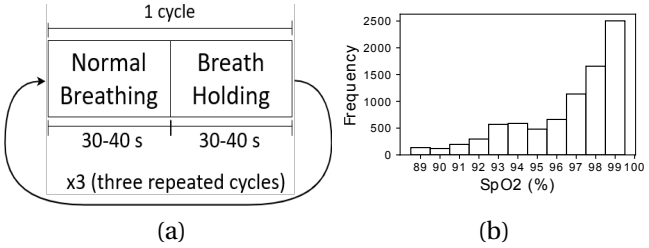


Figure 4.1: Illustration of two hand-video capturing positions. The hand on the left is in the palm down (PD) position and the hand on the right is in the palm up (PU) position.

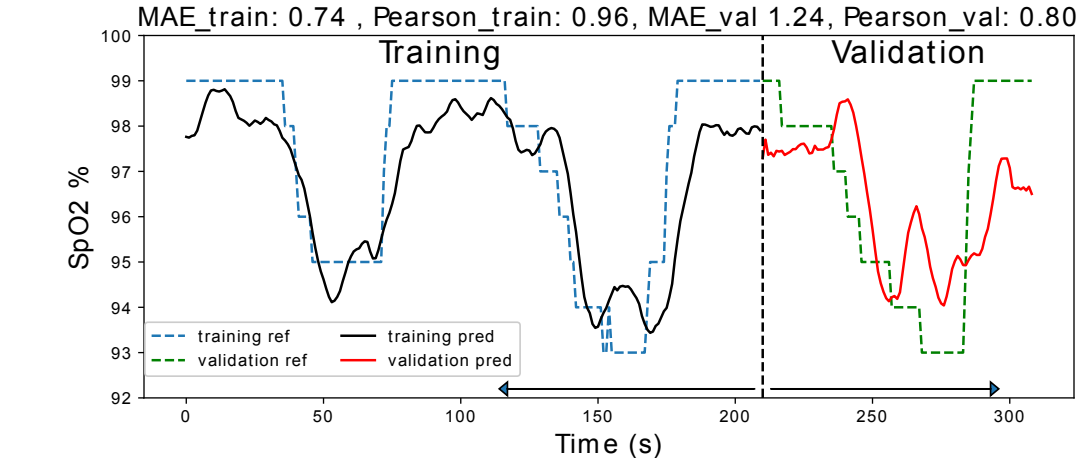


(a) (b)

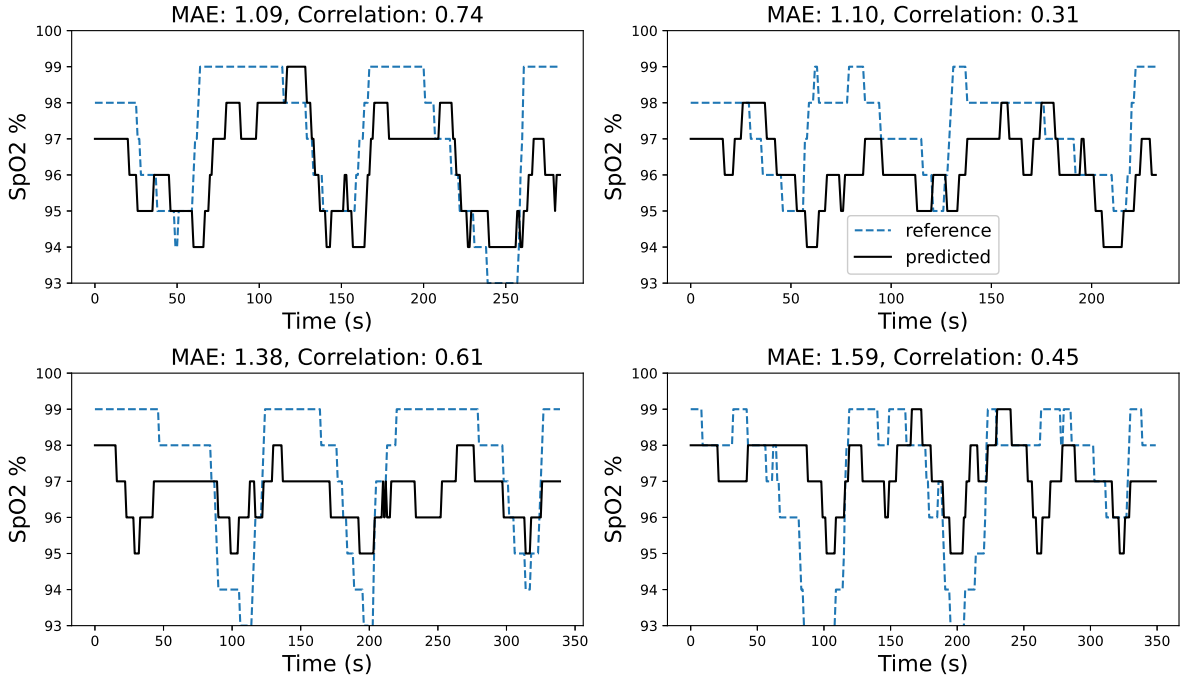
Figure 4.2: (a) Breathing protocol that participants were asked to follow, including 3 cycles of normal breathing and breath holding. (b) Histogram of SpO_2 values in the collected dataset.

All videos were recorded using an iPhone 7 Plus. The participant's SpO_2 was simultaneously measured using a Contec CMS50E pulse oximeter clamped to the left index finger of the hand. We use this pulse oximeter as the reference measurement as it has been validated to be within ±2% of the true SpO_2 level for the range of SpO_2 levels in our dataset. The video frame rate is 30 fps and the sampling rate for the reference SpO_2 measurements is 1 Hz. This data capturing procedure was repeated twice for each participant with at least 15 minutes between sessions.

The reference SpO_2 signal is interpolated to 5 sample points per second to match the segment sampling rate using a smooth spline approximation (Green 1990). Each RGB segment and SpO_2 value pair is fed into our models as a single data point, the models output a single SpO_2 estimate per segment. To evaluate a model on a recording, the model is sequentially fed all RGB segments from the recording to generate a time series of preliminarily predicted SpO_2 values. All predictions greater than 100% SpO_2 are clipped to 100% since they are physiologically impossible. A 10-second long moving average filter is applied to generate a refined time series of predicted SpO_2 values.



(a)



(b)

Figure 4.3: (a) Training vs. validation predictions. (b) Test predictions of varying performance with reference SpO_2. The higher the Pearson correlation, the better the prediction captures the reference SpO_2 trend. The lower the MAE, the better the prediction captures the dips in SpO_2.

4.2 Participant-Specific Results

To investigate how well the proposed models could learn to estimate a specific individual's SpO_2 from his/her own data, we first conducted participant-specific experiments, that is, we learn individualized models for each participant.

Experimental Setting. Two recordings per participant were captured with at least 15 minutes in between. One recording is used for training and validation of the model and the remaining recording is for testing. An example of the training and validation predictions curves is shown in Fig. 4.3a. Each recording contains three breathing cycles, for each training/validation recording, the first two breathing cycles are taken for training and the third cycle is used for validation. Splitting the recordings into cycles instead of randomly sampling the 10-sec overlapping RGB segments ensures that there are no overlapping segments of data between the training and validation set. Example test prediction curves and their correlation and mean-absolute-error (MAE) are shown for reference in Fig. 4.3b. It should be noted that if the correlation is low, e.g., a constant temporal estimate, then the MAE and RMSE metrics are less meaningful. For the participant-specific experiments, due to the small dataset size, we augment the training and validation data by sampling with replacement. This is an example of the bootstrapping data reuse strategy (James et al. 2013, Chapter 5). The oversampling also helps address the imbalance in SpO_2 data values that is shown in Fig. 4.2b.

In each experiment, the model structure and hyperparameters are first tuned using the training and validation data. Once the model has been tuned, we train multiple instances of the model using the best tuned hyperparameters. Between each instance, we vary the random seed used for model weights initialization and random oversampling. Each model instance is evaluated on the training/validation recording, the model instance that achieves the highest validation RMSE is selected for evaluation on the test recording. This model is then evaluated on the test recording to obtain the final test results.

Results. Table 4.1 shows the performance comparison of our proposed models with the prior-art model from Ding *et al.* (Ding et al. 2018). To the best of our knowledge, Ding *et al.*'s model is the only convolutional neural network structure that has been tried for contact-based SpO_2 estimation. Its structure is similar to our Model 3 but with fewer layers. We also compare with the classic ratio-of-ratios method proposed by Scully *et al.* (Scully et al. 2011). The performance is measured in Pearson's Correlation, mean absolute error (MAE), and root mean square error (RMSE) and results of each condition are summarized in the median and interquartile range (IQR). IQR quantifies the spread of an empirical distribution of a set

Table 4.1: Performance comparison of each model structure for participant-specific experiments. Results are given as the test median and IQR of all participants.

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	0.41	0.40	2.12	0.91	2.51	0.78
	PU	0.39	0.37	2.16	1.80	2.70	2.09
Model 2 (Proposed)	PD	**0.46**	0.44	2.09	1.32	2.52	1.63
	PU	**0.41**	0.32	1.96	0.68	2.48	0.89
Model 3 (Proposed)	PD	0.44	0.40	**1.93**	1.11	2.48	1.31
	PU	**0.41**	0.46	**1.81**	1.83	2.43	2.44
Scully *et al.* (Scully et al. 2011)	PD	0.08	0.37	1.94	0.92	**2.22**	0.77
	PU	0.19	0.24	2.01	0.80	**2.36**	0.78
Ding *et al.* (Ding et al. 2018)	PD	0.38	0.39	3.25	2.85	3.83	3.24
	PU	0.34	0.56	3.40	3.16	4.58	3.12

of data points by computing the difference between the first quartile and the third quartile of the distribution.

Table 4.1 reveals that Model 2 achieves the best correlation in both PD and PU cases, whereas Model 3 achieves the best MAE and a comparable correlation with Model 2, suggesting that Model 2 and Model 3 are comparably the best in the individualized learning. Even though the method proposed in Scully *et al.* (Scully et al. 2011) achieves the best (lowest) RMSE, its correlations are the worst (lowest). This suggests that the classic ratio-of-ratios method cannot track the trend of SpO_2 well using the contactless smartphone measurement. All of our model configurations outperform Ding *et al.* (Ding et al. 2018). For example, in the PU case for Model 3, the correlation is improved from 0.34 to 0.41 and the MAE is lowered from 3.40% to 1.81%. It is worth noting that the international standard for clinically acceptable pulse oximeters tolerates an error of 4% (ISO 2011), and our estimation errors are all within this range.

There are two factors, including the skin type and the side of the hand, that might influence the performance of SpO_2 estimation. We therefore analyze the following two questions: (1) Whether the different skin types matter in PU or PD case, and (2) whether the side of hand matters in lighter skin (types II + III) or darker skin (types IV + V). The box plots in Fig. 4.4 shows the distributions of the test correlations from all the three proposed models in PU and PD modes of (a) lighter skin and darker skin participants, and (b) all participants.

To answer question (1), we focus on the left panel of Fig. 4.4a. We note that overall, the medians of darker skin group are larger than those of the lighter skin group. Zooming into the PD case, we can confirm that the darker skin group indeed outperforms the light group

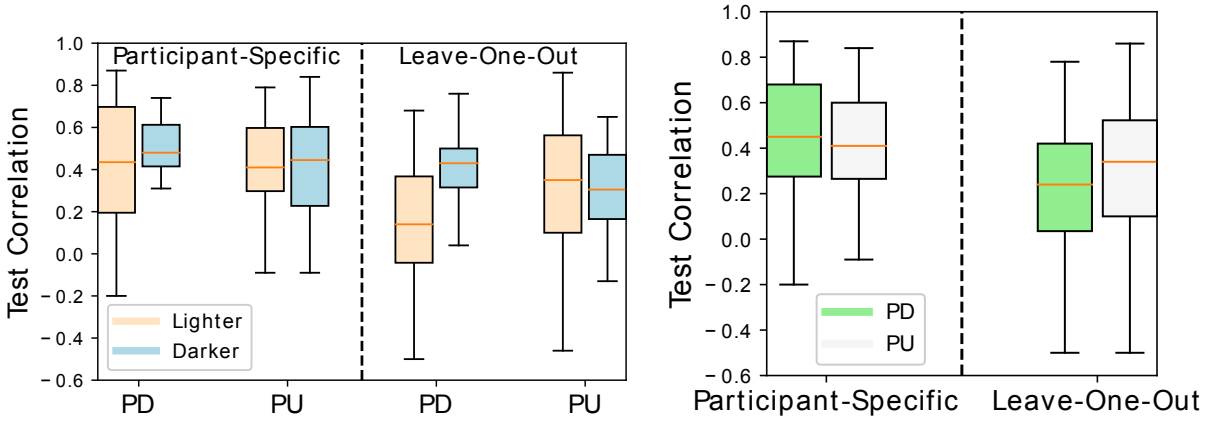


Figure 4.4: Box plots comparing distributions of correlations for (a) lighter vs. darker skin types, and (b) PD vs. PU for all skin types. The PD results are better for darker skin tones in both the participant-specific and leave-one-out cases.

since the former has a smaller interquartile range (IQR). However, for the PU case, the no significant performance difference can be observed, because while the dark skin group is better in a larger median, the light skin group is better in a narrower IQR. To answer question (2), we first focus on the left panel of Fig. 4.4b. We note that no significant performance difference can be observed between PD and PU given one has a better median and the other has a better IQR, when participants of all skin colors are considered together. However, if we zoom into the subset of darker skin group as shown in the left panel of Fig. 4.4a, we observe that PD is better than PU given its higher median and narrower IQR. To summarize, in the participant-specific experiments, (1) darker skin group outperforms the lighter skin group when using the back side of the hand as the ROI for SpO_2 prediction but they are comparable when using the palm of the hand; and (2) the side of the hand has an impact on SpO_2 prediction for the darker skin group but not for the lighter skin group.

4.3 Leave-One-Participant-Out Results

To investigate whether the features learned by the model from other participants are generalizable to new participants whom it has not seen before, we conduct leave-one-participant-out experiments. For each experiment, when testing on a certain participant, we use all the other participant's data for training and leave the test participant's data out. The recordings

Table 4.2: Performance comparison of each model structure in leave-one-participant-out experiments. Results are given as the test median and IQR of all participants.

	Hand Mode	Correlation		MAE (%)		RMSE (%)	
		Median	IQR	Median	IQR	Median	IQR
Model 1 (Proposed)	PD	**0.33**	0.42	2.33	1.07	3.07	1.52
	PU	**0.46**	0.36	**1.97**	0.80	**2.32**	0.87
Model 2 (Proposed)	PD	0.15	0.50	2.43	0.94	3.35	1.11
	PU	0.33	0.39	2.08	0.73	2.41	0.71
Model 3 (Proposed)	PD	0.23	0.38	2.48	1.18	2.98	1.33
	PU	0.27	0.31	2.02	1.03	2.54	1.28
Scully *et al.* (Scully et al. 2011)	PD	0.05	0.43	**2.08**	0.65	**2.44**	1.14
	PU	0.01	0.54	2.08	0.60	2.43	1.20
Ding *et al.* (Ding et al. 2018)	PD	0.11	0.56	3.19	1.61	3.76	1.52
	PU	0.26	0.42	2.43	1.22	2.85	1.51

from all the non-test participants are used for participant-wise cross-validation to select the best model structure and hyperparameters. The selected model is evaluated on the two recordings of the test participant, whose data was never seen by the model during training.

Table 4.2 shows the performance comparison of each model in leave-one-participant-out experiments. Model 1 achieved the best performance in terms of correlation and achieved the best MAE and RMSE for the PU case. Similar to the participant-specific case, the classic ratio-of-ratios method proposed in Scully *et al.* (Scully et al. 2011) achieved better MAE and RMSE results for the PD case but the correlation result was low, suggesting that the model achieved low error by simply predicting a nearly constant SpO_2 near the middle of the SpO_2 range. The best performance of Model 1 in the leave-one-participant-out experiment may imply that the features extracted after combining the color channels at the beginning of the pipeline can be generalized better to unseen participants than the features extracted before channel combination or through interleaving as in Models 2 or 3.

In the participant-specific case, the model is specifically tailored to the test individual, whereas the leave-one-participant-out case is more difficult because the model needs to accommodate for the variation in the population. As expected, in Fig. 4.4, we observe that the overall results from the leave-one-participant-out experiments do not match those from the participant-specific experiments. Because of the modest size of the dataset, the model has not seen as diverse data as a larger and richer dataset would offer. The generalization capability to new participants can be improved when more data is available.

We now revisit the two research questions raised in Section 4.2 under the leave-one-participant-out setup. First, we analyze the impact of skin type given the same side of the hand. From the right panel of Fig. 4.4a, we observe that in the PD case, the darker

skin group outperforms the lighter skin group , whereas in PU case, the performances are comparable. This observation is consistent with the participant-specific experiments that when using the palm as the ROI, the skin color is not a factor to the accuracy of SpO_2 estimation. Second, we analyze the impact of the side of the hand for two skin color groups. The right panel of Fig. 4.4a reveals that for darker skin group, the PD case outperforms the PU case, which is consistent with the results from the participant-specific experiments. However, in contrast to these experiments, the PU outperforms the PD in both lighter skin group as well as the mixed group as illustrated in the right panel of Fig. 4.4b. This different generalization capability in the PU and PD cases may be attributed to skin color difference between the palm and the back of the hand. The color of the back of the hand tends to be darker than the color of the palms, and has larger color variation among participants due to different degrees of sunlight exposure. In contrast, the color variation of the palms is much milder among participants. Furthermore, in the participant-specific experiments, the individualized models learn the traits of the skin type and the side of the hand from each participant, whereas in the leave-one-participant-out experiments, the learned model must capture the general characteristics of the population.

4.4 Ablation Studies

To justify the use of nonlinear channel combinations and convolutional layers for temporal feature extraction in our proposed models, we conduct two ablation studies comparing the performance of these model components to other generic ones. We focus on the PU case to avoid the uncontrolled impact of such factors as skin tone and hair. In the first ablation study, we compare nonlinear to linear channel combination. We create a variant of Model 1 with only a single linear channel combination layer with no activation function and repeat the leave-one-participant-out experiments. In the second study, we compare the performance of using convolutional layers for temporal feature extraction to using fully-connected dense layers. We create this second variant of Model 1 and repeat leave-one-participant-out experiments.

Table 4.3 presents the medians and IQRs specified for numerical comparison of the ablation study. First, we compare the first and the third rows in Table 4.3 for ablation study 1. Our proposed Model 1 achieves a better correlation with a median of 0.46 and IQR of 0.36 and a better RMSE with a median of 2.32 and IQR of 0.87 than its linear channel combination variant. Besides, Model 1 achieves a comparable MAE with a better median of 1.97 but a

Table 4.3: Numerical results of the ablation studies for Model 1 (M1) in the leave-one-participant-out mode. Comparisons among the proposed (nonlinear) M1, modified M1 with only linear channel combinations, and modified M1 with fully connected dense layers instead of convolutional layers are listed. Ablation studies confirm that the nonlinear channel combinations and convolutional layers improve model performance.

Method		ρ	MAE(%)	RMSE(%)
Linear Ch. Comb.	Median	0.46	2.14	2.66
+ Conv. layer for Feat. Extra.	IQR	0.38	0.73	0.93
Nonlinear Ch. Comb.	Median	0.41	2.29	2.66
+ Fully Connec. layer for Feat. Extra.	IQR	0.39	0.63	0.70
Model 1 (Proposed): Nonlinear Ch. Comb.	Median	**0.46**	**1.97**	**2.32**
+ Conv. layer for Feat. Extra.	IQR	0.36	0.80	0.87

wider IQR of 0.80. The overall better performance of Model 1 suggests the necessity of using the nonlinear channel combination method. Second, in ablation study 2, we compare the second and the third rows in Table 4.3. We observe that Model 1 outperforms its second variant with fully connected layers for feature extraction with better medians in terms of correlation (0.46 vs. 0.41), MAE (1.97 vs. 2.29), and RMSE (2.32 vs. 2.66), and narrower IQR of correlation. This suggests that convolutional layers are better than fully connected layers for temporal feature extraction.

CHAPTER

5

DISCUSSION

5.1 Contact-based Dataset Testing

We also test our models on the publicly available dataset gathered by Nemcova *et al.* for their SpO_2 estimation work (Nemcova et al. 2020). This dataset consists of contact-based smartphone video recordings where a participant placed a finger on the smartphone camera and was illuminated by the camera flashlight. Participants were asked to breathe normally without following any sophisticated breathing protocol. Each recording lasts about 10 to 20 seconds. The subject for each recording is not identified, so subject-specific and leave-one-participant-out experiments cannot be conducted. There is a single reference SpO_2 value associated with each recording. We used 14 recordings for training and seven recordings for testing and compared them with the modified ratio-of-ratios method proposed in their paper.

As shown in Table 5.1, Models 1 and 2 outperform the method used by Nemcova *et al.* on both the training and test recordings. Model 3 is not able to generalize well from the training set to the test set, which may be due to the small size of the dataset. It should be noted that because the participants were not asked to follow any sophisticated breathing

Table 5.1: Experimental results on the contact-based video SpO_2 dataset obtained by Nemcova *et al.* (Nemcova et al. 2020). One SpO_2 estimate was output per recording and MAE and RMSE were calculated across all recordings. Models 1 and 2 outperform the method proposed in their work, model 3 was unable to generalize well to the test set.

	MAE (%)		RMSE (%)	
	Training	Test	Training	Test
Model 1	0.86	**1.19**	0.94	**1.36**
Model 2	0.50	1.28	0.59	1.64
Model 3	0.75	3.28	0.99	3.69
Nemcova *et al.* (Nemcova et al. 2020)	2.05	2.18	2.24	2.36

protocol, the dynamic range of SpO_2 values is narrow. These results show that our CNN Models 1 and 2 work well for contact-based video recordings in addition to contactless videos recordings.

5.2 Prediction Correlation

By employing a training-validation-test split in Section 4, we guarantee the generalizability of our models (Shalev-Shwartz and Ben-David 2014, Chapter 11). Because we use this approach and ensure the models can perform well on unseen data, we show that our models are indeed outputting meaningful predictions.

As extra evidence that the correlation values achieved by our models are meaningful, we compare our results to randomly generated SpO_2 predictions. For each reference signal, a random prediction signal was generated by choosing a random SpO_2 value for each time between the minimum and maximum values from the reference signal and applying a moving average in the same way as is applied to the neural network predictions. Fig. 5.1 shows the distribution of the correlations between the predicted and reference SpO_2 signals from the randomly generated predictions and the predictions generated by Model 2. It is clear that the neural network performs much better than random guessing. It has been shown in other applications that even very low correlation coefficients can be meaningful. For example, in photo response non-uniformity (PRNU) work, the device used to take a photo can be predicted with correlation values below 0.1 (Baar et al. 2012).

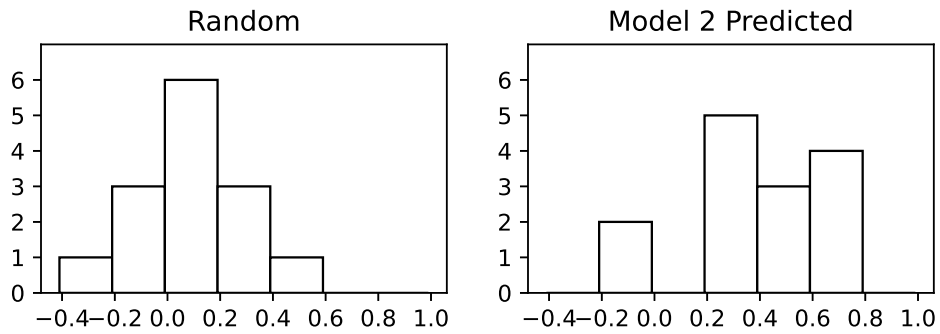


Figure 5.1: Correlation distributions of randomly generated SpO_2 signals vs. SpO_2 predicted by neural network Model 2. It is clear that the correlation distribution for Model 2 is centered much higher than the random signals.

5.3 Visualizations of RGB Combination Weights

To understand and explain what our physiologically inspired models have learned, we conduct a separate investigation to visualize the learned weights for the RGB channels. Our goal is to understand the best way to combine the RGB channels for SpO_2 prediction. Having an explainable model is important for a physiological prediction task like this. Our neural network models can be considered as nonlinear approximations of the hypothetically true function that can extract the physiological features related to SpO_2 buried in the RGB videos. The ratio-of-ratios method, for example, is another such extractor that combines the information from the different color channels at the end of the pipeline. For this experiment, we use the modified version of Model 1 from the ablation studies that has only a single linear channel combination at the beginning. Seeing that using a single linear channel combination did not significantly reduce model performance in the ablation studies, and understanding that the linear component may dominate the Taylor expansion of a nonlinear function, we use only linear combinations for this model to facilitate more interpretable visualizations.

We have trained 100 different instances of the model on the first two cycles from all the recordings and tested on the third cycle from all recordings. The difference between each instance is that the weights are randomly initialized. The weights for each channel learned by the model instances were visualized as points representing the heads of the linear combination vector in RGB space. Each point is colored according to the average test correlation achieved by the model instance. Figs. 5.2a and 5.2b show the projections of these points onto the RB and RG planes. The subfigures reveal that the majority of the

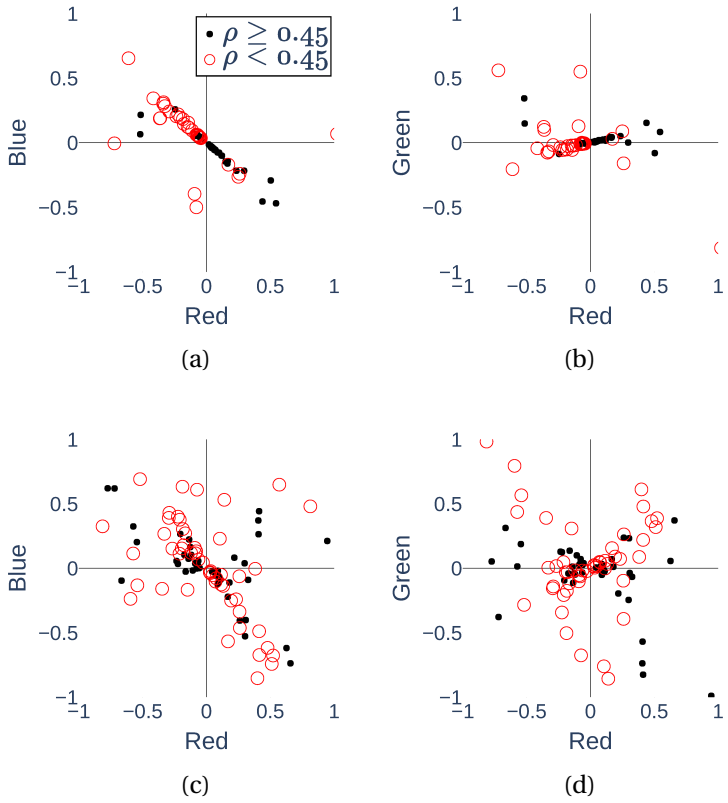


Figure 5.2: Learned RGB channel weights. Plots (a) and (b) are the channel weights learned by different model instances trained on the data of all study participants together, projected onto the RB and RG planes in the RGB space. Plots (c) and (d) are the RB and RG projections of the learned channel weights for model instances trained on random subsets of the participants' data. Each point is color-coded according to the correlation ρ achieved by the instance.

channel weights lay along certain lines in the RGB space. For the weights on the line, the ratio of the blue channel weight to the red channel weight is 0.87, the ratio of the green channel weight to red channel weight is 0.18. It is clear that the red and blue channels are the dominating factors for SpO_2 prediction.

To further verify this result, we repeat this experiment under the following setup: instead of using the data from all participants, for each model instance, we randomly select seven participants and use their data for training and testing. In this case, the difference between each model instance is not only the initialized weights but also the random subset of participants that the model was trained on. Fig. 5.2d reveals that most of the better-performing instances (with $\rho \geq 0.45$) have little contribution from the green channel. In Fig. 5.2c, we again see that most of the points lay on a line in the RB plane, the ratio of the blue channel

weight to the red channel weight for these points is 0.80.

These results are in accordance with the physical understanding of how light is absorbed by hemoglobin in the blood. Recall that Fig. 2.2 reveals a large difference between the extinction coefficients, or the amount of light absorbed, by deoxygenated and oxygenated hemoglobin at the red wavelength. There is a significantly smaller difference at the blue wavelength and almost no difference at green. The amount of light absorbed influences the amount of light reflected which can be measured through the camera. A larger difference in extinction coefficients makes it easier to measure the ratio of light absorbed by oxygenated vs. deoxygenated hemoglobin over time. This ratio indicates the level of blood oxygen saturation. Therefore, from a physiological perspective, it makes sense for the neural networks to give larger weight to the red and then blue channels and give little to the green channel. These visualizations indicate that the models are learning physically meaningful features.

CHAPTER

6

CONCLUSION

In this paper, we have proposed the first CNN-based work to solve the challenging problem of video-based remote SpO_2 estimation. We have designed three optophysiologically inspired neural network architectures. In both participant-specific and leave-one-participant-out experiments, our models are able to achieve better results than the state-of-the-art method. We have also analyzed the effect of skin color and the side of the hand on SpO_2 estimation and have found that in the leave-one-participant-out experiments, the side of the hand plays an important role with better SpO_2 estimation results achieved in the PU case for the lighter skin group. We have also shown the explainability of our designed architectures by visualizing the weights for the RGB channel combinations learned by the neural network, and have confirmed that the choice of the color band learned by the neural network is consistent with the established optophysiological methods.

REFERENCES

(2011). *Particular requirements for basic safety and essential performance of pulse oximeter equipment* . International Organization for Standardization.

(2020). *Fitzpatrick skin phototype*. Australian Radiation Protection and Nuclear Safety Agency.

(2020). Optical Absorption of Hemoglobin. `https://omlc.org/spectra/hemoglobin/`. Accessed: 2021-03-09.

Al-Naji, A., Khalid, G., Mahdi, J., and Chahl, J. (2021). Non-contact spo2 prediction system based on a digital camera. *Applied Sciences*.

Baar, T., van Houten, W., and Geradts, Z. (2012). Camera identification by grouping images from database, based on shared noise patterns.

Bal, U. (2015). Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical Optics Express*.

Burger, W. and Burge, M. J. (2008). *Digital Image Processing - An Algorithmic Introduction using Java*. Springer.

Casalino, G., Castellano, G., and Zaza, G. (2020). A mHealth solution for contact-less self-monitoring of blood oxygen saturation. In *IEEE Symposium on Computers and Communications (ISCC)*.

Chen, M., Zhu, Q., Wu, M., and Wang, Q. (2020). Modulation model of the photoplethysmography signal for vital sign extraction. *IEEE Journal of Biomedical and Health Informatics*.

Chen, W. and McDuff, D. (2018). DeepPhys: Video-based physiological measurement using convolutional attention networks. In *The European Conference on Computer Vision (ECCV)*, pages 349–365.

Couzin-Frankel, J. (2020). The Mystery of The Pandemic's "Happy Hypoxia". *Science*.

De Haan, G. and Jeanne, V. (2013). Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*.

Ding, X., Nassehi, D., and Larson, E. C. (2018). Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*.

Favilla, R., Zuccala, V. C., and Coppini, G. (2018). Heart rate and heart rate variability from single-channel video and ica integration of multiple signals. *IEEE journal of biomedical and health informatics*.

Green, P. J.; Silverman, B. (1990). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall.

Herbert, L. J. and Wilson, I. H. (2012). Pulse oximetry in low-resource settings. *Breathe*, 9(2):90–98.

Iozzia, L., Cerina, L., and Mainardi, L. (2016). Relationships between heart-rate variability and pulse-rate variability obtained from video-ppg signal using zca. *Physiological measurement*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kong, L., Zhao, Y., Dong, L., Jian, Y., Jin, X., Li, B., Feng, Y., Liu, M., Liu, X., and Wu, H. (2013). Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Opt. Exp.*

Lamonaca, F., Carni, D., Grinaldi, D., Nastro, A., Riccio, M., and Spagnolo, V. (2015). Blood oxygen saturation measurement by smartphone camera. *IEEE Intl. Symposium on Medical Measurements and Applications (MeMeA) Proceedings*.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*.

Li, X., Chen, J., Zhao, G., and Pietikainen, M. (2014). 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.

Lu, Z., Chen, X., Dong, Z., Zhao, Z., and Zhang, X. (2015). A prototype of reflection pulse oximeter designed for mobile healthcare. *IEEE Journal of Biomedical and Health Informatics*.

Nam, Y., Reyes, B. A., and Chon, K. H. (2015). Estimation of respiratory rates using the built-in microphone of a smartphone or headset. *IEEE Journal of Biomedical and Health Informatics*.

Nemcova, A., Jordanova, I., Varecka, M., Smiseka, R., Marsanova, L., Smital, L., and Vitek, M. (2020). Monitoring of heart rate, blood oxygen saturation, and blood pressureusing a smartphone. *Biomedical Signal Processing and Control*.

Nitzan, M., Romem, A., and Koppel, R. (2014). Pulse oximetry: Fundamentals and technology update. *Medical Devices (Auckland, NZ)*, 7:231.

Niu, X., Shan, S., Han, H., and Chen, X. (2019). RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Trans. on Image Processing*.

Otsu, N. (1979). A threshold Selection Method from Gray-level Histograms. *IEEE Trans. Syst., Man, and Cybernet.*

Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2010). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*.

Scully, C. G., Lee, J., Meyer, J., Gorbach, A. M., Granquist-Fraser, D., Mendelson, Y., and Chon, K. H. (2011). Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Trans. on Biomedical Eng.*

Severinghaus, J. W. (2007). Takuo Aoyagi: Discovery of pulse oximetry. *Anesthesia & Analgesia*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Shao, D., Liu, C., Tsow, F., Yang, Y., Du, Z., Iriya, R., Yu, H., and Tao, N. (2015). Noncontact Monitoring of Blood Oxygen Saturation Using Camera and Dual-wavelength Imaging System. *IEEE Trans. Biomed. Eng.*

Sohn, K., Merchant, F. M., Sayadi, O., Puppala, D., Doddamani, R., Sahani, A., Singh, J. P., Heist, E. K., Isselbacher, E. M., and Armoundas, A. A. (2017). A novel point-of-care smartphone based system for monitoring the cardiac and respiratory systems. *Scientific Reports*.

Špetlík, R., Franc, V., and Matas, J. (2018). Visual heart rate estimation with convolutional neural network. In *British Machine Vision Conf., Newcastle, UK*.

Starr, N., Rebollo, D., Asemu, Y. M., Akalu, L., Mohammed, H. A., Menchamo, M. W., Melese, E., Bitew, S., Wilson, I., Tadesse, M., et al. (2020). Pulse oximetry in low-resource settings during the COVID-19 pandemic. *The Lancet Global Health*.

Tarassenko, L., Villarroel, M., Guazzi, A., Jorge, J., Clifton, D., and Pugh, C. (2014). Non-contact Video-based Vital Sign Monitoring Using Ambient Light and Auto-regressive Models. *Physiol. Meas.*

Tsai, H.-Y., Huang, K.-C., and Yeh, J. A. (2016). No-contact oxygen saturation measuring technology for skin tissue and its application. *IEEE Instrum. Meas. Magazine*.

Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J. F., and Sebe, N. (2016). 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.

Van Gastel, M., Stuijk, S., and De Haan, G. (2016). New Principle for Measuring Arterial Blood Oxygenation, Enabling Motion-Robust Remote Monitoring. *Scientific Reports*.

van Gastel, M., Verkruysse, W., and de Haan, G. (2019). Data-driven Calibration Estimation for Robust Remote Pulse-oximetry. *Applied Sciences*.

Wang, W., den Brinker, A. C., Stuijk, S., and De Haan, G. (2016). Algorithmic principles of remote PPG. *IEEE Trans. on Biomedical Eng.*

Webster, J. G. (1997). *Design of Pulse Oximeters*. CRC Press.

Zheng, Y., Wang, H., and Hao, Y. (2020). Mobile application for monitoring body temperature from facial images using convolutional neural network and support vector machine. *Mobile Multimedia/Image Processing, Security, and Applications*.

Zhu, Q., Wong, C.-W., Fu, C.-H., and Wu, M. (2017). Fitness heart rate measurement using face videos. In *IEEE Int'l Conf. on Image Proc. (ICIP)*.

APPENDIX

APPENDIX

A

FITZPATRICK SKIN TYPES

Our self-collected dataset consists of hand video recordings and SpO_2 data from fourteen participants, of which there were six males and eight females between the ages of 21 and 30. Participants were asked to categorize their skin tone based on the Fitzpatrick skin types (ski 2020) shown in Fig. A.1. The Fitzpatrick skin types classify the skin by its reaction to exposure to sunlight and pigmentation. From type I to type VI, the skin color becomes darker and less prone to be burned by the sunlight. Among the fourteen participants, two are from type II, eight are from type III, one is from type IV, and three are from type V.

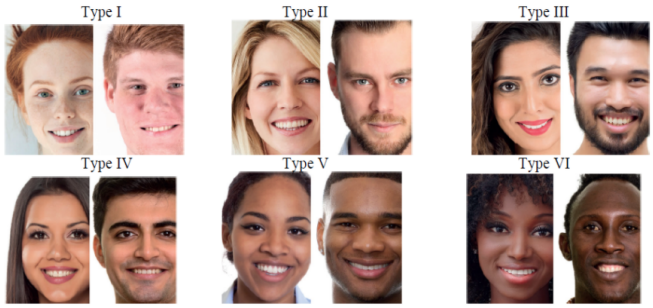


Figure A.1: Fitzpatrick skin types. Reproduced from (ski 2020).