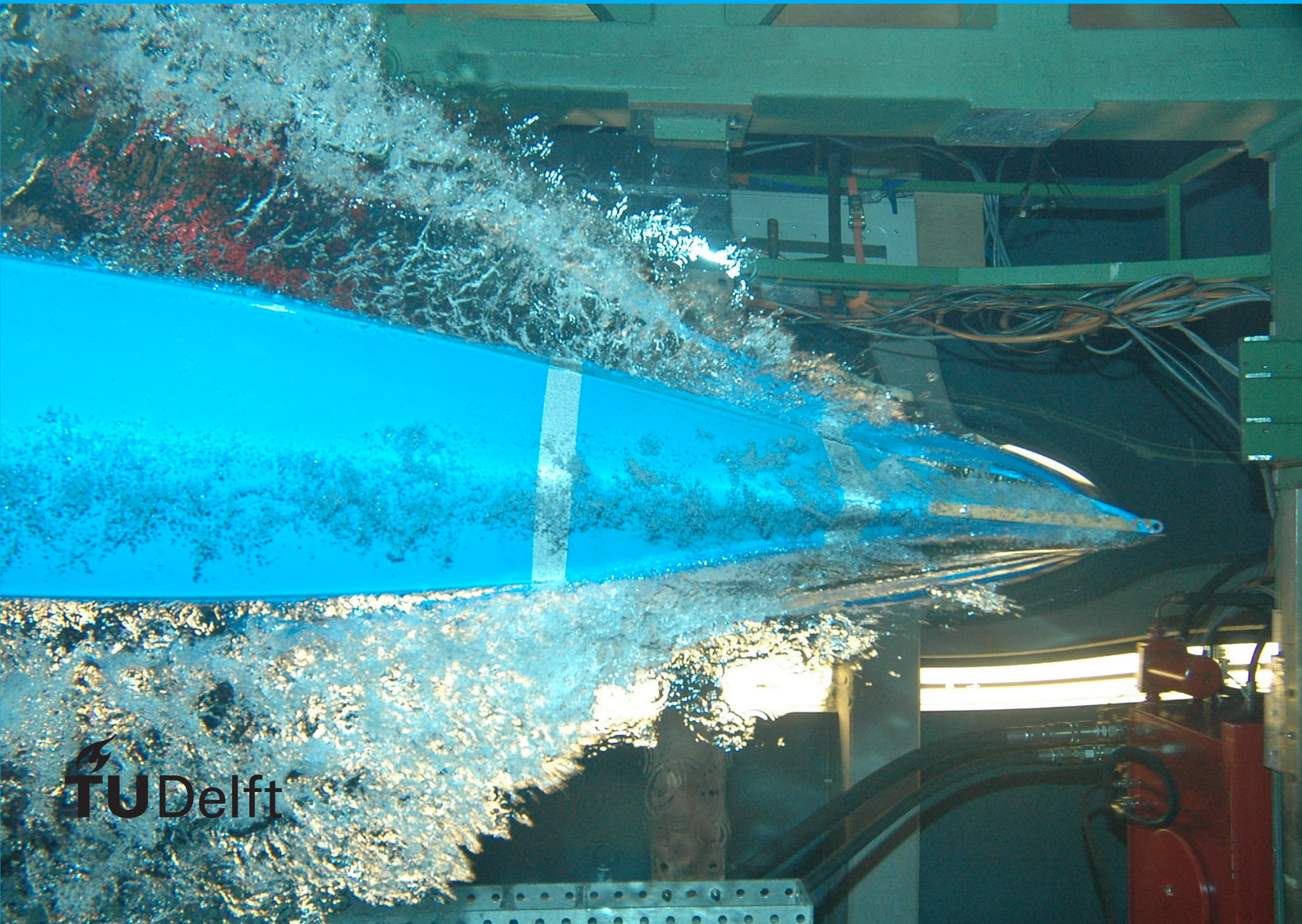
Neural networks for non-contact oxygen saturation estimation from the face

Jim Kok



Neural networks for non-contact oxygen saturation estimation from the face

by

Jim Kok

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Monday June 13, 2022 at 10:00 AM.

Student number:	4679970	
Project duration:	November 8, 2022 – June 13, 2022	
Thesis committee:	Dr. J. C. van Gemert,	TU Delft, Supervisor & Chair
	Dr. K. Liang,	TU Delft, Core member
	M. Bittner,	TU Delft, Daily supervisor & PhD Student

Preface

The work represented in this report is about measuring humans. To be more specific, we have investigated non-contact oxygen saturation estimation from the face by making use of a camera. The study is performed, as part of the Master Thesis graduation procedure, within the Computer Vision Lab at the TU Delft.

In short, the first main finding of this research is that neural networks are able to cope with the presented challenges (e.g. varying lighting conditions) of facial oxygen saturation. Secondly, neural networks show promising results in selecting facial regions that contain oxygen-related information, and we suggest that they should replace traditional facial region selectors. Lastly, we indicate that even single frame differences are a potential source for obtaining oxygen saturation related information.

Finally, I would like to thank Dr J.C. van Gemert for his supervision, M. Bittner for his daily advice and K. Liang for evaluating my work as a core member from a different field.

Jim Kok
Delft, June 2022

Contents

1

Scientific article

Neural networks for non-contact oxygen saturation estimation from the face

Jim Kok
Delft, University of Technology
J.J.M.Kok@student.tudelft.nl

June 3, 2022

Abstract

COVID-19 drastically raised the importance of non-contact based healthcare methods. Low blood oxygen levels of a person, which can be unnoticeable, are potentially a precursor of COVID-19. Contact based methods for measuring blood oxygen saturation could spread the contagious disease. Therefore, this paper investigates non-contact RGB camera-based peripheral oxygen saturation estimation by remote photoplethysmography (rPPG) methods. The novel aspects of non-contact oxygen saturation that we are looking into are: (1) Applying SpO_2 predictor neural networks to rPPG signals obtained from facial regions, instead of the less practical hand based skin regions. To be more specific, we show in a facial based setting that in the relatively uncontrolled environment the traditional Ratio-of-Ratios pulse oximetry principles fail. In the leave-one-participant-out experiments, the RoR method achieved a correlation of -0.05, whereas neural networks showed the capability of dealing with the inherent challenges of the PURE dataset by achieving a superior correlation of 0.64. These challenges are lighting variation due to subtle head motion and clouds alternatively blocking the sun. (2) The first end-to-end neural networks for SpO_2 estimation are introduced by replacing traditional hard pixel region-of-interest selectors, which assign equal weight to each selected pixel, with convolutional soft-attention masks. (3) By using an adapted version of a recent heart and breathing rate estimator network, called DeepPhys, we indicate that the current state-of-the-art is far from optimal. This is done by comparing the window-based constructed end-to-end neural networks with Adapted DeepPhys, which is based on single frame differences. Finally, our research[1] shows that non-contact facial based SpO_2 estimation by RGB camera remains a difficult task. However, as our results indicate, more sophisticated deep learning model might become a viable diagnostic tool for this task in the future.

Index Terms– Neural networks, rPPG, Oxygen saturation, SpO_2.

1 Introduction

Skin colour changes unnoticeable to the human eye contain physiological information, that amongst other things can be used to determine the heart rate [9, 33], oxygen saturation [4, 7, 20, 29] and blood pressure [13, 42]. This paper focuses on blood oxygen saturation. Specifically, this paper focuses on non-contact oxygen saturation estimation. Non-contact based methods could avoid the spread of transferable diseases by taking preemptive measures when low blood oxygen saturation is observed in a person. This is because low blood oxygen saturation may be an early indicator of various diseases (e.g. COVID-19) [6].

Blood oxygen saturation in the blood stream is the ratio of concentrations of oxygenated haemoglobin and total haemoglobin [20]. Arterial blood oxygen saturation (SaO_2) can be measured invasively by a gas chromatograph and is highly correlated with peripheral blood oxygen saturation (SpO_2), which is measured at the skin [41].

[1] Code available on `https://github.com/jimkok9/oxygenSaturation`

Invasive methods require patients' blood samples and do therefore not allow for continuous monitoring. Conventionally, blood oxygen saturation is measured non-invasively with contact-based devices (e.g the pulse oximeter [19]). The SpO_2 measured by pulse oximeters may show a discrepancy with SaO_2 in some cases (e.g. with increased carboxyhaemoglobin concentrations)[3]. The downside of contact-based devices is that they could cause discomfort, especially when someone needs to be monitored constantly. In some cases, for instance, in patients with burn injuries, it would not even be possible to use contact devices at all. In healthcare, especially when dealing with transferable diseases, it is not always possible to treat patients in person. Furthermore, contact based methods could spread contagious diseases when they are applied to multiple patients [6]. Non-contact based methods have the potential to solve these problems [21].

Remote photoplethysmography (rPPG) is a non-contact based method that estimates physiological related signals with a camera, based on reflection variation of the skin [40]. Light reaching the camera is comprised of specular lighting, which is light reflected by the skin, and diffuse lighting, which is light scattered by skin tissue. Specular lighting is not affected by blood volume changes, whereas diffuse lighting penetrates the skin and changes colour depending on the blood volume [10]. As can be seen from figure 1, the skin penetration depth of light is dependent on the wavelength. In visible light, the blue and green wavelengths reach arterioles at the upper dermal layers, whereas the red wavelength reaches subdermal layers [24].

In the literature, various types of cameras have been used to determine oxygen saturation. The types of cameras include charge-coupled device [16], CMOS [32] and RGB cameras [6, 11]. Non-contact based methods, for estimating blood oxygen saturation, rely on the extinction coefficients at different wavelengths of oxygenated and deoxygenated haemoglobin, as can be seen in Figure 2. For example, RGB camera-based SpO_2 estimation makes use of the fact that the absorption coefficient of oxygenated haemoglobin is higher at the blue wavelength and lower at the red wavelength, with respect to deoxygenated haemoglobin. The ratio-of-ratios (RoR) method [41] makes use of this property and is defined as the ratio of absorbance of light at two different wavelengths.

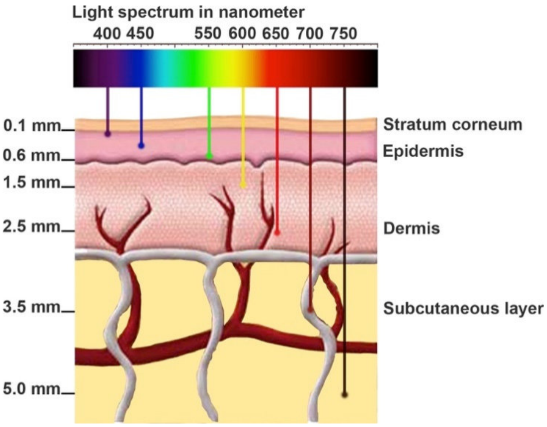


Figure 1: Skin penetration depth of various wavelengths. Figure is copied from [39].

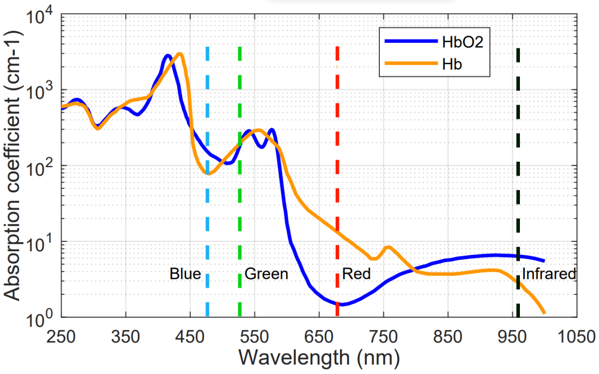


Figure 2: Absorption coefficients of oxygenated and deoxygenated haemoglobin at different wavelengths. Figure is copied from [20].

Several challenges are inherent to rPPG, one of which is changing lighting conditions. Natural lighting changes over time (e.g. clouds blocking the sun) can severely impact the ratio of absorbances [21]. Secondly, head motion introduces changes in pixels' colour intensities which are not related to the absorbance of light [7]. Furthermore, physiological factors, like temperature, change the light scattering properties of the skin and affect the rPPG based SpO_2 estimation [36]. The varying penetration depths of light in the skin may lead to inaccuracies [24]. A linear relation between SpO_2 and RoR is assumed which can be an oversimplified representation due to the previously mentioned factors.

Neural networks have shown to be able to learn complex tasks, and can potentially mitigate the aforementioned challenges. To the best of our knowledge, neural networks have not been used to determine facial oxygen saturation. This leads to the research question of whether neural networks are able to estimate SpO_2 in a relatively uncontrolled facial based setting. Neural networks have previously been used to predict hand based SpO_2 measurements [20, 36, 38], in highly controlled environments, where most of these challenges do not apply to. In contrast to hand based settings, facial based rPPG is more practical at present (e.g. through selfie cameras and webcams). Therefore, in this paper, the explainable neural networks introduced by Methew et al. [20], which have been used to determine oxygen saturation from the hands, are examined by applying them to the face. The contributions of this work are the following:

- We show that SpO_2 facial based neural networks are capable of dealing with the induced challenges, in a relatively low controlled environment with amongst other things changing natural lighting conditions.

- We constructed the first end-to-end neural networks for SpO_2 estimation by adding a convolutional soft-attention mask, for skin pixel averaging, to existing SpO_2 estimator networks. The achieved results of the constructed models are promising for predicting facial rPPG based SpO_2 estimations.

- We adapted DeepPhys [9] to predict oxygen saturation, instead of heart rate (HR), and show that single frame differences are a potential source of SpO_2 relevant information. The results suggest that the current state-of-the-art is not optimal yet.

2 Related work

This section first describes the work done with regard to the traditional ratio-of-ratios methods. Finally, the deep learning methods applied in this field are described.

2.1 Ratio-of-ratios

Traditional photoplethysmography [2] (PPG) methods, which are contact based, make use of the ratio-of-ratios (RoR) principle [25]. The ratio-of-ratios is defined as the ratio of absorption at two different wavelengths, for instance red ($\lambda = 660nm$) and infrared ($\lambda = 940nm$) for pulse oximeters [41].

For computing the RGB camera's RoR, there exists controversy about whether to use the "red and blue channels" or the "red and green channels". Most work in this field uses the blue and red channels, since the absorption coefficient of HbO_2 is higher than Hb at the blue wavelength and for the red wavelength, Hb absorbs more light than HbO_2. Instead of using the red and blue channels averaged over facial regions, the authors in [24] showed that the red and green channels can be used to compute the normalized ratio-of-ratios in controlled settings, with only artificial light. Al-Naji et al. [1], similarly, used the red and green channels to estimate oxygen saturation. The setting is less controlled, in terms that the participants sit still with artificial- and sunlight. The authors concluded that their RoR methods showed similar readings with the pulse oximeter, however, their results showed relatively low correlations. Both authors state that the absorption coefficient of HbO_2 and Hb should be significantly different at one wavelength, which is the case for the red wavelength, and equal at the other wavelength, which makes the green wavelength applicable. These findings lead to the use of the red, green and blue channels, averaged over the facial regions, as input to our neural networks.

Currently three types of digital cameras are used in contactless SpO_2 estimation, which are the RGB, monochrome Complementary Metal Oxide Semiconductor (CMOS) and monochrome Charge-Coupled Device (CCD) camera. In contrast to CMOS cameras, CCD cameras are relatively more sensitive to light and create higher quality low noise images [22]. The authors in [16] used two monochrome CCD cameras to determine facial oxygen saturation for stationary participants under ambient lighting. They empirically showed that the pulse oximeter and camera SpO_2 measurements are in perfect agreement with each other. Shao et al. [32] used the respectively lower light-sensitive monochrome CMOS cameras, with a light source alternating at two different wavelengths ($\lambda = 611nm$ and $\lambda = 880nm$). They performed breath-holding experiments in a setting with no light sources that could act as noise. The results were consistent with those obtained from the pulse oximeter over a wide SpO_2 range. Monochrome CMOS and CCD are, unlike RGB cameras (e.g. smartphones and webcams), generally more expensive and not ubiquitous nowadays. This emphasizes the importance of research with regard to RGB cameras, on which this paper is based.

Fingertip contact-based SpO_2 estimation by RGB camera is relatively less prone to subtle motions and illumination changes, with respect to non-contact based methods. Therefore, fingertip contact-based SpO_2 estimation can be considered less challenging than non-contact based SpO_2 estimation. The authors in [17] used the averaged red and green channels of the participant's fingertip. This is measured by a smartphone touching the fingertip with the flashlight turned on. This work is followed up by Mishra et al. [23] who performed breath-holding experiments on the participants' fingertips. Their setup makes use of polarization techniques to filter components of reflected light. Ding et al. [11] used convolutional neural networks to predict oxygen saturation of videos of the fingertip made by smartphones. However, all these methods are contact-based and require the flashlight and camera of the smartphone to be in touch with the person's fingertip. Contact-based methods can cause inconvenience by among other things the heat of the flashlight. These limitations emphasize the importance of SpO_2 research on non-contact based methods.

In contrast to fingertip contact-based SpO_2 estimation, Sun et al. [36] performed SpO_2 estimation experiments on videos of the hands captured by a smartphone. Lighting from any other source than the smartphone's built-in flashlight was minimized. The authors adapted the RoR method and showed improved performance over the conventional RoR method. Blood oxygen saturation was manually regulated between 90 % and 100% by a blood pressure cuff. Instead of manually regulating oxygen saturation, Mathew et al. [20] were able to accurately predict SpO_2 values in a less controlled environment where the participants breathe normally and hold their breath in cycles. Tian et al. [38] followed this work up and proposed a multi-channel ratio-of-ratios method for non-contact hand based SpO_2 estimation using a smartphone. Their method combines the RGB channels to extract the rPPG wave, after which the heart rate (HR) can be determined. Based on the HR the R, G and B channels are bandpass filtered. A similar approach is used in our paper where we filter the channels based on the ground truth HR, since their method showed promising results. By using this method the raw signals (i.e. unfiltered), which are prone to contain noise (e.g. motion [27] and lighting artefacts), can possibly be filtered more accurately. Their results are based on videos of the hands, which contain relatively few motion artefacts since the participant's hands are lying still in front of a table with a black background. Furthermore, the black background and the stationary nature of the experiments allow for more accurate and simplified region-of-interest (RoI) extraction (e.g. by applying a threshold [38]). Facial RoI extraction is more prone to motion and lighting artefacts, like blinking and shadowing effects [13]. However, nowadays, devices are designed for capturing facial videos (e.g. webcams and selfie-cameras), which makes these hand based methods less practical and realistic.

The more practical and realistic facial based SpO_2 estimation is examined by various literary works. The authors in [37] tracked changes in oxygen saturation in a non-controlled clinical environment during dialyses, which involves significant oxygen saturation fluctuations, under artificial- and sunlight. They made use of the facial based normalized ratio-of-ratios method and showed robust performance in a setting where participants are lying still. Bal et al. [4] showed that the facial based RoR method is able to track the changes in oxygen saturation, in a setting where participants sit still or lie in a bed, with either fluorescent light or indirect sunlight as the only light source. In contrast to lying in a bed, sitting still is more inherent to contain noise, both due to changing background and motion artefacts. Tarassenko et al. [14] showed that it is possible to determine oxygen saturation changes by using a RGB camera aimed at the face. The experiments took place in an oxygen controlled environment, where the participants sat 1.5 meters in front of the camera. They showed that the accuracy of the results was comparable to that of a commercial pulse oximeter. However, this experiment is performed in a highly controlled non-practical lab environment. Instead, Rose et al. [30] simulated oxygen saturation fluctuation by performing breath-holding experiments with a RGB camera under ambient lighting. To show that the signal is present the authors used Eulerian video magnification to amplify facial skin colour changes. The authors in [5] performed similar facial based experiments, however without trying to achieve fluctuations in oxygen saturation. A relatively constant oxygen saturation makes it harder to tune the ratio-of-ratios, and therefore introduces an additional challenge. Casalino et al. [6, 7] extended this work by determining the facial SpO_2 saturation for stationary participants with changing sunlight as the only source of illumination. The use of sunlight, with changing light intensity over time, introduced additional challenges. Nevertheless, all these papers show that the face could be a potential source for neural networks to show improvement in non-contact oxygen saturation estimation, with respect to the linear RoR method.

2.2 Deep learning applied to rPPG signals

Deep learning based methods have shown promising results with regard to estimating physiological signals from RGB videos. Niu et al. [26] have shown that Rythmnet achieves promising results with regard to heart rate estimation. This was done under less-constrained conditions, such as movements and lighting changes. Chen et al. [9] proposed the first end-to-end convolutional neural network, named DeepPhys, for heart rate (HR) and breathing rate (BR) predictions. DeepPhys combines convolutional soft-attention masks with subsequent frame differences. They showed that their model generalizes to people with different skin types, which led us to adjust and examine it with regard to SpO_2 estimation, which can be considered more difficult than heart and breathing rate estimation. HR and BR estimation requires determining the timestamps of the peaks of the rPPG signal, whereas for SpO_2 estimation both the timestamps and amplitude of the peaks need to be accurately determined.

With regard to oxygen saturation, Ding et al. [11] have shown that convolutional neural networks lead to better results than the ratio-of-ratios principle for videos of the fingertip made by smartphones. However, the method requires the flashlight and camera of the smartphone to be in touch with the person's fingertip, which could cause inconvenience by the heat of the flashlight. Furthermore, in the current COVID-19 situation, contact-based methods should be avoided as much as possible with regard to the possible risk of spreading diseases. The authors in [28, 34] proposed convolutional neural networks for constructing the rPPG signal from facial videos, and thereby replacing the traditional RoI selection procedures. Finally, Mathew et al. [20] proposed the first 3 models to estimate SpO_2 saturation from the hands in a non-contact manner. They were able to accurately predict SpO_2 values in a controlled environment where the participants breathe normally and hold their breath in cycles. Their results are based on videos of the hands with a black background. However, videos of the hands are much less common with current digital devices (e.g. when holding your phone your face is in the camera range). Therefore, this method is further examined, in this paper, with regard to performance on the face, where more challenges are involved (e.g. motion artefacts and changing backgrounds).

3 Method

This section first explains the required background theory. Secondly, the skin pixel extraction methods, by which the rPPG signals are obtained, are described. Thirdly, the 3 proposed models for non-contact SpO_2 estimation by camera, that take as input the rPPG signals, are elaborated. Finally, the hyper-parameter and model structure selection procedure will be explained.

3.1 Theory

Blood oxygen saturation (SO$_2$) is defined as the percentage of oxygenated haemoglobin over the total amount of haemoglobin in the blood, as defined in Equation 1. The total amount of haemoglobin is defined as the concentration of oxygenated haemoglobin (HbO$_2$) and deoxygenated haemoglobin (Hb).

$$SO_2 = \frac{HbO_2}{Hb + HbO_2} \cdot 100\% \tag{1}$$

During systole and diastole the heart transports blood through the body, which causes variation in light absorbed by the skin. Under non-varying lighting conditions, the light perceived by a sensor consists of a direct current (DC) and alternating current (AC). The DC is composed of venous blood, a constant amount of arterial blood and non-pulsatile components like skin pigmentation, while the AC involves the pulsatile components [36].

When the amount of blood passing through the arteries reaches a maximum, the amount of light absorbed is also maximal. Likewise, the amount of light is minimum when the amount of flowing blood is at its minimum. According to the Beer-Lambert law, this phenomenon could be expressed as in Equation 2 and 3, where I_H and I_L are respectively the highest and lowest light intensity, I is the initial light intensity, A_t equals the AC absorbance at time t and A_{DC} is the constant light absorbance. The light absorbance at time t is defined as $A_t = \epsilon_{Hb}(\lambda)c_{Hb}d_{Hb}+\epsilon_{HbO_2}(\lambda)c_{HbO2}d_{HbO_2}$, where λ is the wavelength, ϵ is the extinction coefficient, c is the blood concentration and d is the optical pathlength [36].

$$I_H = Ie^{-A_{DC}}e^{-A_{H,t}} \tag{2}$$

$$I_L = Ie^{-A_{DC}}e^{-A_{L,t}} \tag{3}$$

The ratio-of-ratios is defined as the ratio of absorbance of light at two different wavelengths, as defined in Equation 4.

$$R = \frac{A_{t,\lambda_1}}{A_{t,\lambda_2}} = \frac{\ln(\frac{I_{L,\lambda_1}}{I_{H,\lambda_1}})}{\ln(\frac{I_{L,\lambda_2}}{I_{H,\lambda_2}})} \tag{4}$$

Then, by combining Equation 1, 2, 3 and 4, we can infer a linear relation between oxygen saturation and the RoR, as shown in equation 5, where slope a and bias b are found through calibration. In this paper, least squares regression is used to find the best fitting line.

$$SO_2 = b + a \cdot R \tag{5}$$

Correspondingly, Scully et al. [31] defined the ratio-of-ratios for RGB video sequences as in Equation 6. The AC component is computed as the standard deviation and the DC component as the mean of the corresponding channel over a specified time period.

$$S_pO_2 = A - B\frac{AC_{red}/DC_{red}}{AC_{blue}/DC_{blue}} \tag{6}$$

3.2 Skin colour extraction

The inputs to the neural network are the 3-dimensional RGB times series extracted from the participant's face. To be more specific, the input signals are defined as the spatially averaged RGB time series $X \in \mathbb{R}^{3\times(t\cdot fps)}$, where t is the window length in seconds and fps is the number of frames per second. The selected areas of the face are spatially averaged for each channel. When using traditional RoI selectors each selected pixel is assigned an equal weight, whereas for the convolutional soft-attention masks a weighted average is applied.

3.2.1 Appearance model

The Appearance model, shown in Figure 3, is used in our research to produce the convolutional soft-attention masks. The input of the first convolutional layer is the 32 × 32 pixel cropped face in RGB[2] (i.e. 3 input channels), which is converted to 32 channels. Layer 2 applies 'same' padded convolution to the result of layer 1 and keeps the number of feature maps equal. The convolution in layers 1 and 2 are each followed by batch normalization and the hyperbolic tangent activation function, in this order. The third 1 × 1 convolutional layer combines the resulting features maps into 1 channel.

[2] Examples of cropped faces are shown in Figure 8b and 8d.

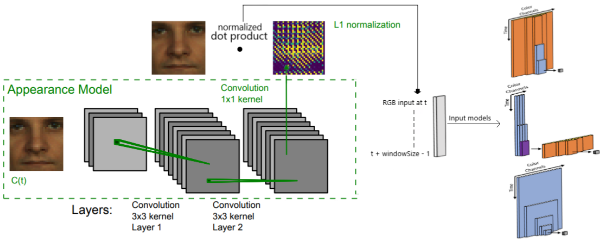


Figure 3: Pipeline of our constructed end-to-end neural networks. The first part is the skin colour extraction performed by the Appearance model. The second part is the SpO_2 prediction from the spatially averaged RGB time series, by the proposed models for SpO_2 extraction. The image is adjusted from [9].

The final soft-convolutional attention mask q is obtained by L1 normalization as shown in Equation 7, where σ is the sigmoid function, H is the height and W is the width of the cropped face. The convolutional weight $\mathbf{w} \in \mathbb{R}^{32x1x1}$ and bias b are applied to resulting feature maps $x \in \mathbb{R}^{32x32x32}$ of the second convolutional layer. Finally, the convolutional soft-attention mask is element-wise multiplied with the input image to extract the raw (i.e. unfiltered) RGB traces.

$$q = \frac{H \cdot W \cdot \sigma(\mathbf{w}x + b)}{2||\sigma(\mathbf{w}x + b)||_1} \tag{7}$$

3.3 Proposed SpO_2 prediction models

The authors in [20] have proposed 3 models for non-contact SpO_2 extraction from the hands, shown in Figure 4. The models, that will be described in this sub-section, are explainable and designed based on domain-specific knowledge [20]. Specifically, the first model structure resembles heart and breathing rate estimation methods [20], where only the timestamps of the peaks of the rPPG signal need to be determined. Generally, first the pulse signal is extracted by spatial combination of the colour channels (e.g. POS [40] and CHROM [10]). Secondly, the pulse signal is temporally combined to get the psychological signal of interest. The structure of model 2 is most in accordance with traditional SpO_2 estimation methods (e.g. the RoR method) where, in contrast to heart and breathing rate estimation, both the timestamps and amplitude of the peaks of the rPPG signal need to be accurately determined. Determining the lowest I_L and highest I_H light intensity of the cardiac cycles is first performed by temporal combination. Secondly, the peak values are spatially combined to determine the SpO_2 saturation.

3.3.1 Model 1

Model 1 performs channel mixing followed by feature extraction. The channel mixing part consists of 3 linear layers, where each layer is followed by the Rectified Linear Unit (ReLU) activation function. Each linear layer combines the values spatially and does not involve temporal combination. The second step is to perform feature extraction twice (i.e. in this order convolution, batch normalization, dropout and temporal sub-sampling are applied). The number of filters of first convolutional layers is determined by the parameter search Hyperband algorithm. The number of filters is halved for each consecutive convolutional layer. The inputs of the convolutional layers are 'same' padded, resulting in equal output width and height. The convolutional layers combine the inputs both spatially and temporally. Sub-sampling decreases the temporal dimension by a factor of two by using the max-pooling operation. To obtain the SpO_2 saturation, the feature extraction result is flattened and linearly combined. To the output, a value of 95 is added for numerical stability.

3.3.2 Model 2

In contrast to model 1, model 2 first performs channel-wise feature extraction followed by channel mixing. Channel-wise feature extraction implies 1d 'same' padded convolution followed by batch normalization, dropout and temporal sub-sampling applied twice to each RGB channel. The 1-d convolution only combines the inputs temporally. The number of filters of the first convolutional layer is equal for each colour channel and determined by the Hyperband algorithm. The second 1-d convolutional layer reduces the initial number of filters by a factor of 2. The channel-wise feature extraction outputs are concatenated after which 3 linear layers are applied, each with ReLU as activation function. After flattening and linearly combining the channel mixing result, 95 is added for numerical stability.

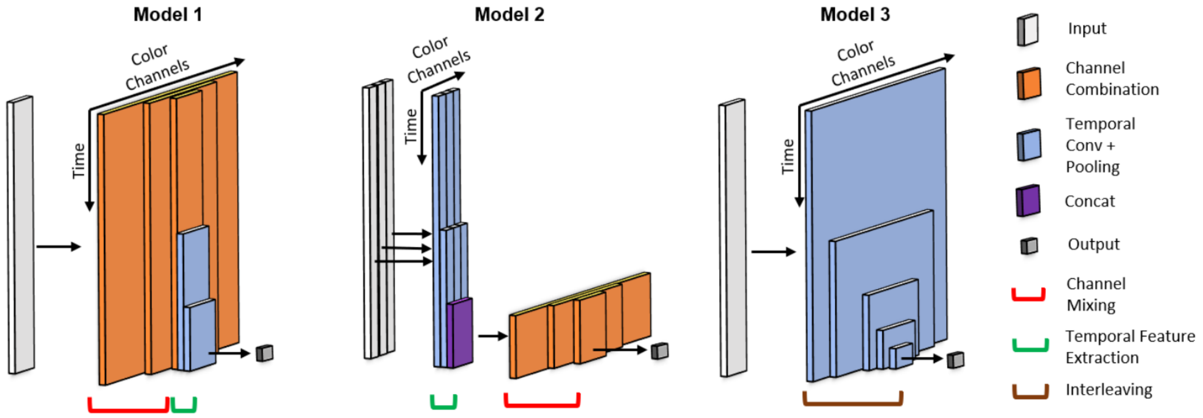


Figure 4: Model structures for predicting SpO_2 saturation, proposed by Mathew et al. [20]. Model 1 first spatially combines the input, after which it performs spatial and temporal combining by performing convolution and pooling. Model 2, in contrast to model 1, performs channel-wise convolution and pooling first. Model 3 interleaves channel combining and convolution. Figure is copied from [20].

3.3.3 Model 3

The third model interleaves channel combination and channel mixing. First, a linear layer which combines the input spatially is applied to the input, followed by the ReLU activation function. Then 2-d convolution is applied to the result of the linear layer, followed by batch normalization, dropout and sub-sampling. The 2×2 max-pooling operation is applied as sub-sampling technique, which decreases the input both temporally and spatially by a factor of 2. The number of filters of the first interleaving layer is reduced by a factor of 2 for each interleaving layer and is determined by the Hyperband algorithm. The result of 4 interleaving layers is flattened and linearly combined to a single value, to which 95 is added to obtain the SpO_2 saturation.

3.4 Hyperband parameter optimization

The hyper-parameters and model structure are chosen by the bandit-based parameter optimization process called Hyperband [18]. The parameters that are tuned include the learning rate, the initial number of filters, kernel size, dropout, batch normalization, and the number of nodes in each layer. Hyperband is an elimination based parameter selection procedure where the best performing parameters, in terms of validation error, are kept for more iterations. This process drastically speeds up the parameter search and trades-off between exploration (i.e. the number of configurations examined) and exploitation (i.e. the extensivity to which the configurations are examined). Hyperband requires 2 inputs, which are the maximum number of iterations per configuration R and the proportion of configurations that is discarded each round η. Finally, the selected parameter configurations for each experiment can be found on our Github[3]

4 Experiments

This section first describes the dataset used. Secondly, the performed experiments are elaborated upon. Finally, the results are presented.

4.1 PURE dataset

The PURE dataset [35] consists of 10 participants each performing 6 tasks. These tasks are the following: 1) Sitting still as much as possible and looking into the camera. 2) Talking while avoiding additional head motion. 3) Head movements parallel to the camera plane. 4) The same as the previous task, however, the speed of the head motion doubled. 5) Orienting the head towards targets placed around the camera, to simulate small rotations. 6) Same as the previous task, however

[3] Configurations available in the /trained_models folder on https://github.com/jimkok9/oxygenSaturation.

with larger head rotations that are on average 35°. For our experiments, only task 1[4] is examined since the other tasks are related to head motion, which is not part of this study. The participants sit in front of the camera at approximately 1.1 meters. The light comes through a large frontal window, where clouds result in illumination changes over time. The videos were captured with the RGB eco274CVGE camera with a resolution of 640x480 at 30 frames per second. The oxygen saturation is measured with a pulse oximeter (CMS50E) at a sampling rate of $60Hz$.

4.2 Leave-One-Participant-Out

To examine how accurately the models are able to predict a relatively constant SpO_2 value, experiments are performed on participants who sit still in front of the camera. This type of task can be considered the easiest task in the datasets, whereas the other tasks introduce additional deliberate motion artefacts caused by head movements. Each of the 10 participants is used for testing once while the rest is used for cross-validation. For cross-validation, we use $n-1$ folds, where n is the total number of participants. The Hyperband parameter R, the maximum number of iterations per configuration, is set to 81 and η, the proportion of configurations that is discarded each round, is set to 3. This setting is empirically determined, based on the size of the PURE dataset, to obtain a trade-off between training time and exploration. Since the hyper-parameters, weights and structure are tuned and trained to $n-2$ participants, they are selected at the last fold of the Hyperband iteration with the lowest summed cross-validation loss.

4.3 Neural networks and RoR performance

This experiment aims to give insight into how the models are able to generalize to different persons. The performance of the neural networks are compared to the traditional RoR method, in a facial based environment with changing natural lighting conditions. The questions it tries to answer are: (1) Are the models able to predict a relatively constant SpO_2 value of participants that the model has not seen. (2) Are the models able to improve upon the traditional ratio-of-ratios methods. For this experiment, the parameters of the spatially averaged RGB time-series inputs $X \in \mathbb{R}^{3\times(t\cdot fps)}$ are set to $t = 10s$ with a sliding window of 0.2 seconds (i.e. 6 frames). According to the authors of the models [20], the neural network input segments should be an order of magnitude longer than 1 heartbeat cycle to add resilience against sensing noise. We have chosen Gudi et al.'s [15] relatively large area RoI selector, since it provides smoother rPPG signals in comparison to Casalino et al.'s [6] relatively dense area RoI selector.

4.3.1 The proposed models' capability

To answer question (1), the models are first trained on the raw (i.e. unprocessed) signals, of which the results are shown in Table 1a, 1b and Figure 5a. The training mean is defined as the mean of the pulse oximeter's measurements of the training set. As can be seen from Table 1b, all results by the models have a relatively high negative correlation with the training mean. Whereas, Table 1a shows that the correlation with the ground truth is positive yet relatively low. The negative train correlation and positive test correlation empirically demonstrate that the models do not simply minimize the error by outputting the mean of the training set. The results of the models trained on the raw signal show relatively low test correlations. Therefore, we hypothesise that the raw signals contain noise that the models are not able to extract and significantly affects the SpO_2 predictions.

To verify the hypothesis that the raw signals contain noise that the models are not able to extract, the models are trained on the band-pass ($0.7-4Hz$) filtered signals, instead of the raw signals. The results are shown in Table 2a, 2b and Figure 5b. The correlation with regard to the ground truth improved significantly with respect to training on the raw signal, by comparing the results in Table 1a and 2a. On top of that, models 2 and 3 show improvement in both mean absolute error (MAE) and root mean square error (RMSE). The improvement in test correlation and error metrics indicates that predictions are more accurate. Despite model 1

[4] Examples of this task are shown in Figure 8a and 8c.

achieving a higher correlation, the MAE and RMSE remained relatively stable. The results of this sub-experiment strengthen the hypothesis that the raw signals contain noise, that the models are not able to extract.

The hypothesis is further examined by narrowing the band-pass filter based on the participant's heart rate. In this sub-experiment, the models are trained on heart rate band-pass filtered signals (i.e. frequency HR$\pm 0.2 Hz$), where the heart rate is obtained from the pulse oximeter at the timestamp of the centre of the window. Since the heart rate can change over time, the centre of the window is chosen to obtain a more representative value. Results are shown in Table 3a, 3b and Figure 5c. For all models, the test correlation decreases and the error metrics increase, with respect to the broad band-pass filtered results. On top of that, all the models' test MAE and RMSE increase in comparison to the raw signal and $0.7 - 4Hz$ band-pass filtered signal results. The decrease in test correlation and highest error rates suggest that valuable oxygen saturation information is being filtered out. Referring back to question (1), the models show the potential to predict oxygen saturation in the given setting. The broad static band-pass filtering improves the models' results, whereas the narrow dynamic heart rate band-pass filter seems to respectively degrade performance. In conclusion, we empirically showed that the models are capable of predicting oxygen saturation but require carefully selected band-pass signal filtering.

4.3.2 Proposed models vs RoR

For question (2) "Are the models able to improve upon the traditional ratio-of-ratios methods in a facial based environment.", the RoR method is examined and compared against the models in a similar setup as in question (1). First, the comparison is made between the trained models and the RoR methods tuned to the same unfiltered training set, of which the results can be seen in Figure 5a. Table 1a shows that the correlation for the RoR methods is lower than all the models and the MAE and RMSE are higher than all the models, which indicates that the RoR method is to a lesser extend able to learn the SpO$_2$ saturation. Furthermore, Table 1b shows that the train correlation of the RoR method is significantly higher and train error metrics are significantly lower than the models. This indicates that the RoR method is biased towards the average of the training set. In this setting, therefore, the models are outperforming the RoR method.

	Correlation	MAE	RMSE
Model 1	0.31	1.34	1.61
Model 2	**0.36**	**1.13**	**1.44**
Model 3	0.29	1.45	1.62
RoR	-0.66	1.49	1.85

(a) Results with respect to the GT.

	Correlation	MAE	RMSE
Model 1	-0.49	0.86	1.25
Model 2	-0.49	0.61	0.82
Model 3	-0.33	0.91	1.25
RoR	0.61	0.27	0.34

(b) Result with respect to the mean of the training set.

Table 1: Models trained on raw signal (i.e. unfiltered) and RoR tuned to raw signal comparison, for steady experiment. The ground truth and training mean for each test participant is shown in Figure 5a.

Table 2a shows the results of the $0.7-4Hz$ band-pass filtered signal. This table shows that the RoR test correlation is significantly lower than all the models. The RoR method error metrics are inferior to the models', except for the MAE of model 3. Table 2b shows that the RoR is, likewise as for the raw signal tuning, biased towards the training mean. As in the raw signal scenario, the models outperform the RoR method by a large margin.

Table 3a shows the comparison with regard to the HR filtered signal and indicates that the RoR method in terms of test error rates performs relatively similar to the models. However, for the RoR method the error metrics are less meaningful, since the test correlation is negative. Furthermore, Table 3b shows that the RoR is the most correlated with the training mean.

With regard to question (2) "Are the models able to improve upon the traditional ratio-of-ratios methods in a facial based environment.", we have shown that all models exhibit superior results in training on the raw and broad static band-pass filtered signals. The HR band-pass filtered results showed that the RoR method in terms of error metrics performs similar to the models. However, the negative RoR ground truth correlation makes these lower errors less meaningful.

	Correlation	MAE	RMSE
Model 1	0.41	1.36	1.57
Model 2	**0.64**	**0.98**	**1.21**
Model 3	0.51	1.32	1.44
RoR	-0.05	1.28	1.66

(a) Results with respect to the GT.

	Correlation	MAE	RMSE
Model 1	-0.17	1.09	1.41
Model 2	-0.41	0.73	0.92
Model 3	-0.28	1.08	1.42
RoR	0.20	0.45	0.55

(b) Results with respect to the mean of the training set.

Table 2: Models trained on band-pass (i.e. $0.7 - 4Hz$) filtered signal and RoR tuned to band-pass filtered signal comparison, for steady experiment. The ground truth and training mean for each test participant is shown in Figure 5b.

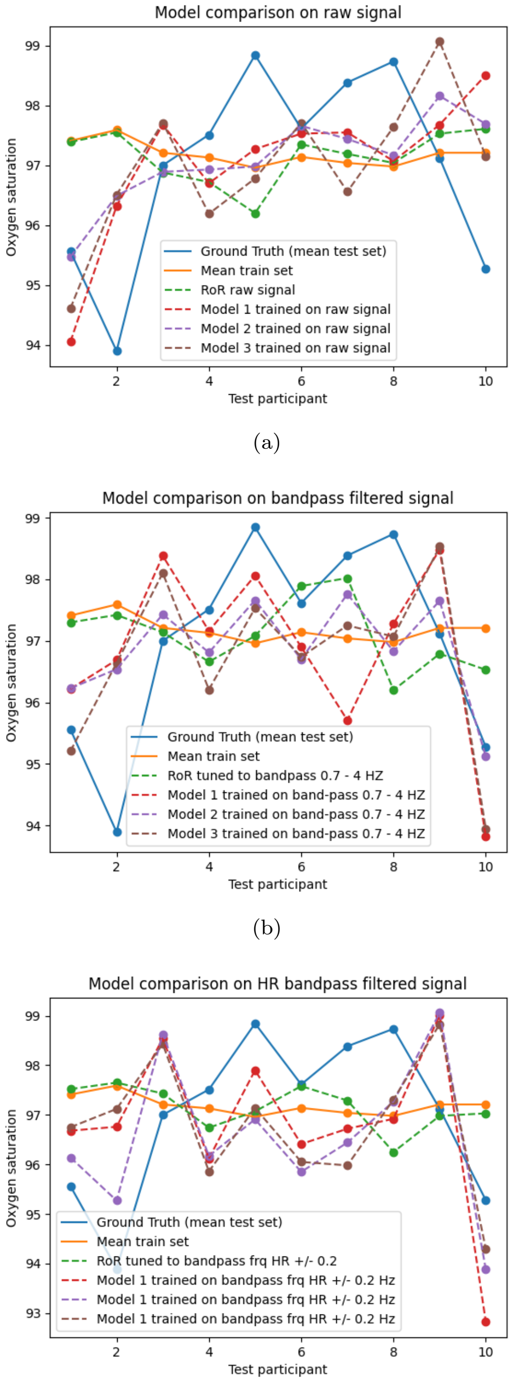
	Correlation	MAE	RMSE
Model 1	0.36	1.69	1.78
Model 2	**0.48**	1.54	**1.59**
Model 3	0.16	1.73	1.83
RoR	-0.53	**1.42**	1.80

(a) Results with respect to the GT.

	Correlation	MAE	RMSE
Model 1	-0.08	1.22	1.68
Model 2	-0.27	1.34	1.63
Model 3	0.07	1.08	1.32
RoR	0.62	1.42	1.80

(b) Results with respect to the mean of the training set.

Table 3: Models trained on heart rate band-pass (i.e. frq HR$\pm 0.2Hz$) filtered signal and RoR tuned to heart rate band-pass filtered signal comparison, for steady experiment. The ground truth and training mean for each test participant is shown in Figure 5c



(a)

(b)

(c)

Figure 5: Model comparison based on training on raw signal (a), $0.7 - 4Hz$ band-pass signal (b) and HR band-pass filtered signal (c). The ground truth is the mean of the pulse oximeter measurements of a test participant.

4.4 End-to-end neural networks

This sub-experiment makes use of the same 3 models as in the previous experiment, however, the RGB skin averaging procedure is replaced by the Appearance model convolutional neural network. For this experiment, the parameters of the spatially averaged raw RGB time-serie inputs $X \in \mathbb{R}^{3\times(t\cdot fps)}$ are set to $t = 2s$ with a sliding window of 2 seconds (i.e. non-overlapping windows). We have chosen these parameters since Casalino et al. [7] obtained a relatively high correlation, on the raw signal obtained with their RoI selector. They used the RoR method in the same settings on the PURE dataset. The questions we aim to answer are: (1) Are convolutional neural networks able to learn facial features that contain SpO_2 related information. (2) Are convolutional neural networks able to improve upon traditional RoI selector methods.

4.4.1 Learning SpO_2 related facial features

To answer question (1), all 3 previously used models are trained in combination with the Appearance model, which outputs a weighted facial skin pixel average. The learned masks of the 300^{th} frame are shown in Figure 9, for each model and participant. The 300^{th} frame is chosen to reduce head motion distortion at the start of the video. The forehead in the selected frames is excluded since some participants' forehead is covered by hair and we are interested in general facial features. The learned masks definitely show their capability to exclude certain facial features (e.g. Figure 9ad, 9ah and 9aj the eyes and Figure 9g, 9o and 9al parts of the mouth). We have shown the capability of the Appearance model to exclude certain facial features, however, is it also capable of including certain facial features that contain SpO_2 related information. According to the literature [8, 12], the cheeks, forehead and nasal area contain the strongest rPPG signal and thus most SpO_2 related information. Although sub-optimal solutions are obtained by selecting large parts of the image (e.g. Figure 9j and 9n), some masks show promising results in selecting the areas of interest (e.g. Figure 9b, 9an and 9ai the cheek and nose and Figure 9m the nose). Therefore, question (1) "Are convolutional neural networks able to learn facial features that contain SpO_2 related information." can be answered by that the constructed end-to-end neural networks show promising results in including and excluding SpO_2 related facial features.

4.4.2 Traditional RoI selection vs convolutional soft-attention masks

Now, that we have shown the capability of the Appearance model to learn SpO_2 related facial features, we are interested in question (2) "Are convolutional neural networks able to improve upon traditional RoI selector methods". To answer this question, all 3 models are trained on the RGB traces obtained by the RoI selectors by Casalino et al. [6], Gudi et al. [15] and in combination with the first DeepPhys soft-attention mask (i.e. the Appearance model).

Figure 6a shows the result of model 1 trained in combination with the Appearance model and both traditional RoI selectors. The corresponding Table 4 shows that the Appearance model achieves the lowest RMSE, however, performs worst in terms of correlation. As can be seen in Figure 6a, the Appearance model achieves the lowest absolute error for test participants 1, 5, 6, 8 and 9. This shows the potential improvements that can be achieved by CNNs over traditional RoI selector methods. Referring back to the examination in question (1), the masks of these participants show relatively high skin pixel weights in the SpO_2 related facial regions.

	Correlation	MAE	RMSE
Appearance model	-0.19	1.62	**1.95**
Casalino et al.	-0.05	**1.55**	1.98
Gudi et al.	**0.01**	2.38	2.68

Table 4: Model 1 RoI selector comparison in terms of ground truth correlation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

Figure 6b and Table 5 show results of the RoI selector methods in combination with model 2. Casalino et al.'s method achieves the highest performance in terms of correlation and error metrics. The Appearance model achieves the lowest absolute error for test participants 1, 4, 5, 6 and 8. This again shows a correlation with the relatively high quality learned masks. Furthermore, the result of the Appearance model are impacted by outlier test participant 10, where the mask includes features of the mouth as shown in Figure 9am. By removing the 10^{th} test participant, the correlation significantly increases from -0.10 to 0.29.

	Correlation	MAE	RMSE
Appearance model	-0.10	2.02	2.29
Casalino et al.	**0.33**	**1.88**	**2.11**
Gudi et al.	0.20	1.93	2.21

Table 5: Model 2 RoI selector comparison in terms of ground truth correlation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

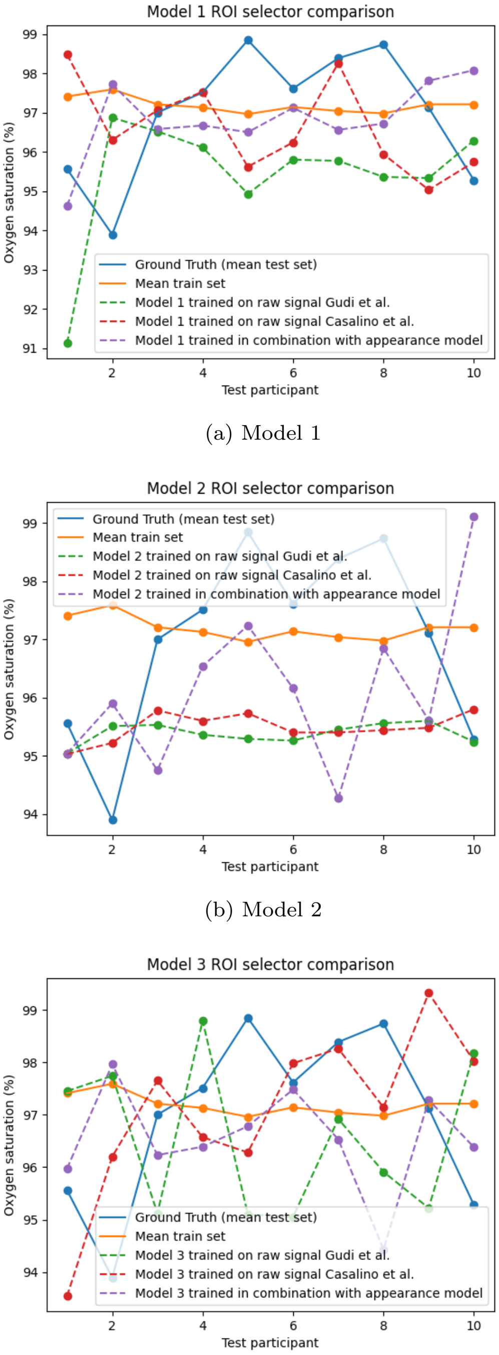
Figure 6c and Table 6 show the results with regard to model 3. Like the results of model 2, Casalino et al.'s method achieves the best results for all 3 metrics. However, the lowest absolute error is obtained by the Appearance model for half of the total number of test participants. To be more specific, it obtains the lowest absolute error for participants 1, 5, 6, 9 and 10. With regard to question (2), the CNN weighted RoI selector is capable of improving over traditional RoI selectors, however, is dependent on the quality of the learned masks.

	Correlation	MAE	RMSE
Appearance model	-0.42	1.60	2.15
Casalino et al.	**-0.05**	**1.55**	**1.80**
Gudi et al.	-0.52	2.43	2.58

Table 6: Model 3 RoI selector comparison in terms of ground truth correlation, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

4.4.3 Weight visualisation

In Figure 12 the first linear layer of model 2, which combines the RGB channels, is examined for the test participants where significant performance difference is achieved by the masks and traditional RoI selectors. The first 2 examined test participants are 4 and 5, for which the Appearance model achieved superior performance. From Figure 12a and 12c can be seen that the absolute red and blue channel weight correlation of the Appearance model for participant 4 and 5 is the highest. Furthermore, Figure 12b and 12d show that all RoI selector methods



(a) Model 1

(b) Model 2

(c) Model 3

Figure 6: Model 1 (a), 2 (b) and 3 (c) RoI selector comparison. RoI selectors used are described in Gudi et al. [15] and Casalino et al. [6]. The RGB traces of the Appearance model are obtained by multiplying DeepPhys' first soft-attention with the input image.

have relatively low absolute red and green channel weight correlation for these participants. The Appearance model showed inferior performance for participant 7, which is accompanied by relatively low absolute red and blue channel weight correlation (Figure 12e) and absolute lowest red and green channel weight correlation (Figure 12f). For participant 8, the Appearance model outperforms and shows a relatively high correlation in both the absolute red and green channel weights (Figure 12h) and the red and blue channel weights (Figure 12g). Whereas, the traditional RoI selectors show either relatively high absolute correlation for the red and green channel weights or the red and blue channel weights. Lastly, for participant 10, Gudi et al.'s RoI selector achieved superior performance with an absolute error of 0.02, whereas the prediction of the Appearance model can be considered as an outlier. Gudi et al.'s absolute red and blue weight correlation is with a value of 0.38 the highest, whereas the absolute red and green channel weight correlation of the Appearance model is the lowest.

These observations are in accordance with the optophysiological properties of oxygenated and deoxygenated haemoglobin, shown in Figure 2. To be more precise, the relatively large difference in extinction coefficients of the blue wavelength with respect to the green wavelength are in accordance with the perceived performance. With this weight visualisation sub-experiment, we have shown a correlation between performance and RGB weight combination, which is in turn influenced by the RoI selector method used.

4.5 Adapted DeepPhys

For this sub-experiment, an adapted version of DeepPhys [9] is used to investigate whether single frame differences contain SpO_2 related information. Instead of heart and breathing rate, Adapted DeepPhys outputs SpO_2, which can be considered a more difficult task. This is because for SpO_2 estimation, in contrast to heart and breathing rate estimation, both the timestamps and amplitude of the peaks of the rPPG signal need to be accurately determined. Furthermore, Adapted DeepPhys does not normalize the frame differences to keep the SpO_2 related RGB information intact. In this experiment Adapted DeepPhys is trained sequentially and non-sequentially on the face extracted frames. The sequential training inputs consist of 2-second non-overlapping windows, whereas the non-sequential training inputs consist of 128 batch sizes of frame differences randomized over time and participants. The 2-second non-overlapping windows are chosen to keep the settings of the previous sub-experiment the same. The 128 batch size is chosen since this is the initial DeepPhys setting for HR based training. The question this sub-experiment aims to answer is: (1) Do single frame differences contain SpO_2 related information.

With regard to question (1), Figure 7 and Table 7a show that the way Adapted DeepPhys is trained has a significant influence on the performance. Sequentially trained Adapted DeepPhys achieves higher performance than non-sequential training, with respect to the mean of the test set, in terms of correlation and MAE. Table 7b shows that the non-sequentially trained Adapted DeepPhys, with respect to the mean of the training set, has lower MAE, RMSE and higher correlation than sequentially trained Adapted DeepPhys. This indicates that non-sequential Adapted DeepPhys is more biased towards the training means. Furthermore, by comparing Tables 4, 5 and 6 with Table 7a, sequentially trained Adapted DeepPhys achieves superior performance in terms of correlation, MAE and RMSE with respect to the end-to-end neural networks investigated in Section 4.4, regardless of which RoI selector method is used.

	Correlation	MAE	RMSE
Non-sequentially	0.11	1.45	**1.62**
sequentially	**0.38**	**1.32**	1.69

(a) Results with respect to the GT.

	Correlation	MAE	RMSE
Non-sequentially	-0.24	0.67	0.82
sequentially	-0.53	1.40	1.60

(b) Results with respect to the mean of the training set.

Table 7: The results of Adapted DeepPhys, which is trained non-sequentially and sequentially. Non-sequential inputs consist of single frame differences of 128 batch sizes. Sequential inputs consists of 2 sequential second non-overlapping windows. The ground truth and training mean for each test participant is shown in Figure 7.

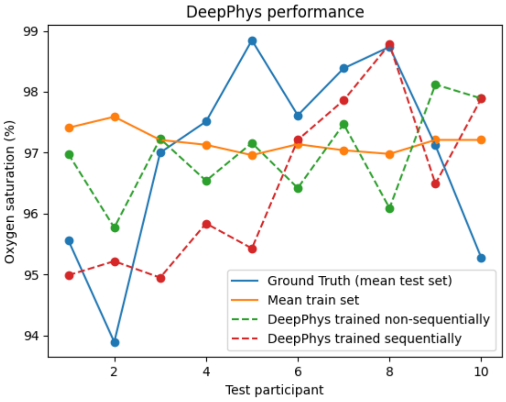


Figure 7: The results of Adapted DeepPhys trained sequentially and non-sequentially. Non-sequential inputs consist of single frame differences of 128 batch sizes. Sequential inputs consists of 2 sequential second non-overlapping windows.

To further investigate whether single frame differences contain SpO_2 related information, the learned convolutional soft-attentions masks are examined. Figure 10 and 11 show respectively the sequentially and non-sequentially learned masks of each participant. The fact that the non-sequential trained model is relatively more biased towards the training mean shows coherence with the learned masks. To be more specific, non-sequential training shows respectively less attention to SpO_2 related areas (e.g. Figure 10c shows that sequential trained Adapted DeepPhys learns the cheek areas, whereas non-sequential trained Adapted DeepPhys includes parts of the right eye 11f). Furthermore, Figure 11f, 11u and 11ad show that for participant 2, 7 and 10 non-sequential training gives out of proportional weights to dense areas.

In conclusion, the way Adapted DeepPhys is trained has significant influence on the performance. Sequentially trained Adapted DeepPhys achieved promising results and outperformed the window-based end-to-end neural networks. Furthermore, the learned convolutional soft-attention masks showed coherence with the performance achieved by the training methods. These findings shows that single subsequent frames difference potentially contain subtle skin colour changes which are related to SpO_2 estimation.

5 Discussion

The first leave-one-participant-out sub-experiment, in which the raw (i.e. unprocessed) signals are input to the neural networks, empirically showed that the 3 networks are able to pick up a proportion of the signal of interest. In this setting, model 2 achieved the highest correlation with a value of 0.36. To enhance the learning process of the models, the signals are band-pass filtered from 0.7 to $4Hz$. This corresponds to a range of 42 to 240 heartbeats per second. The highest correlation is again achieved by model 2 and significantly increased from 0.36 to 0.64, with respect to raw signal training. Remarkably by further narrowing down the band-pass filter range to the heart rate frequency, degraded performance is perceived. This indicates that important information is contained outside the HR filtered ranges (i.e. frq HR$\pm 0.2Hz$). Model 2 achieved the highest correlation of 0.48 for the HR based filter, which is nevertheless higher than the highest correlation of 0.36 obtained from raw signal training. The superior performance of model 2 on the raw signal, broad band-pass and dynamic band-pass filter signal training is in accordance with the explainability of the model structure, which is based upon traditional SpO_2 estimation methods. Furthermore, the correlation of model 2 increased with respect to raw signal training by 77.78 %, for the broad band-pass filtered signal training, and by 33.33 %, for dynamic HR band-pass filtered signal training. This suggests that future work should carefully select the band-pass signal filtering range.

Next, we compared the performance of the neural networks against the performance of the RoR method in the same settings. The results suggest that the RoR oversimplifies the problem by predicting values near the mean of the training set, and thus fails in a facial based setting with changing natural lighting conditions. The superior results achieved by the models indicate that we need a more complex procedure to estimate non-contact SpO_2 measurements from the face in a relatively uncontrolled environment. In conclusion, the RoR based method fails in this type of setting regardless of the signal preprocessing method used, whereas the models seem to be able to mitigate the induced facial challenges.

The second sub-experiment investigated different RoI selection methods. To be more precise, two traditional hard pixel assignment RoI selectors, with relatively large and small areas, and the soft-attention masks learned by a convolutional neural network are examined. This sub-experiment showed that the used RoI method has a significant influence on the SpO_2 estimation. The Appearance model in combination with the SpO_2 predictor networks showed promising results on a large subset of the test participants, however, produced outliers due to wrongly learned masks. The fact that we trained the end-to-end neural networks on the relatively small PURE dataset could be the reason for degraded performance. Nevertheless, the Appearance model showed the capability of learning facial features that inherently contain SpO_2 related information. On top of that, it showed to be able to exclude facial features where the rPPG signal is less prominent. The rPPG related signals can differ per person, therefore one of the strengths of the convolutional soft-attention masks is the provided flexibility. Nevertheless, only the first soft-attention mask of DeepPhys' Appearance model is used in our end-to-end neural network, which results in less sophisticated facial features. This led to the investigation of Adapted DeepPhys, which consists of two soft-attention masks and is based on frame differences. Finally, the weakness of this current study that should be addressed in future work is the fact that we use cropped facial regions, where the forehead is not included. The term end-to-end neural network is not completely appropriate, because of the fact that we first track and crop the face. Future work could include examination of the performance of applying the networks on the non-cropped images, on possibly larger datasets.

In the final sub-experiment, an adapted version of DeepPhys was trained sequentially and non-sequentially to investigate whether single frame differences contain SpO_2 related information. Sequentially trained Adapted DeepPhys outperforms non-sequentially trained Adapted DeepPhys significantly, emphasizing the importance of the way the network is trained. Participant specific skin characteristics (e.g. skin colour and temperature) could potentially be the cause of the diminished performance of non-sequential training. Nevertheless, the sequentially trained Adapted DeepPhys model showed superior performance with respect to the SpO_2 predictor models combined with the RoI detector methods. This indicates that single frame differences are a potential source of SpO_2 related information. Adapted DeepPhys, which is based on single frame differences, achieved a higher correlation than the window-based models trained on the raw signal. Although Adapted DeepPhys achieved superior correlation, we would not recommend using it with regard to SpO_2 estimation. The reason for this is that window-based networks take into account time sequences that minimally contain one heartbeat cycle. This makes it possible for the window-based networks to determine the minimum I_L and maximum I_H light intensity of a cardiac cycle, where the principles of SpO_2 estimation are based upon. Finally, DeepPhys' relatively large network size and the fact that it uses 2 soft-attention masks could be the reason for the improved performance. Therefore, with regard to future work, more complex and sophisticated window-based end-to-end neural networks (e.g. extending Adapted DeepPhys to a window-based network by applying 3D convolutional neural networks) should be examined to improve the current state-of-the-art.

In conclusion, we showed that the RoR method fails in a facial based relatively uncontrolled environment. Neural networks in combination with traditional RoI selector methods were able to mitigate the induced lighting challenges and outperformed the RoR method. Furthermore, our constructed end-to-end neural networks showed promising results by replacing hard skin pixel RoI selection with learned convolutional soft-attention masks. Finally, our Adapted DeepPhys model showed that even single frame differences are a potential source for obtaining SpO_2 related information. Sequentially trained Adapted DeepPhys obtained superior performance over our constructed window-based end-to-end neural networks, and thereby emphasizes the performance gap that needs to be bridged in this field.

References

[1] Ali Al-Naji, Ghaidaa A Khalid, Jinan F Mahdi, and Javaan Chahl. Non-contact spo2 prediction system based on a digital camera. *Applied Sciences*, 11(9):4255, 2021.

[2] Aymen A Alian and Kirk H Shelley. Photoplethysmography. *Best Practice & Research Clinical Anaesthesiology*, 28(4):395–406, 2014.

[3] Mona Ascha, Anirban Bhattacharyya, Jose A Ramos, and Adriano R Tonelli. Pulse oximetry and arterial oxygen saturation during cardiopulmonary exercise testing. *Medicine and science in sports and exercise*, 50(10):1992, 2018.

[4] Ufuk Bal. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical optics express*, 6(1):86–97, 2015.

[5] Amitabha Bhattacharjee and Md Salah Uddin Yusuf. A facial video based framework to estimate physiological parameters using remote photoplethysmography. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–7. IEEE, 2021.

[6] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. A mhealth solution for contact-less self-monitoring of blood oxygen saturation. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7. IEEE, 2020.

[7] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. Evaluating the robustness of a contact-less mhealth solution for personal and remote monitoring of blood oxygen saturation. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2022.

[8] AVJ Challoner and CA Ramsay. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Physics in Medicine & Biology*, 19(3):317, 1974.

[9] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

[10] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[11] Xinyi Ding, Damoun Nassehi, and Eric C. Larson. Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2603–2610, 2019.

[12] Qiang Fan and Kaiyang Li. Non-contact remote estimation of cardiovascular parameters. *Biomedical Signal Processing and Control*, 40:192–203, 2018.

[13] Xijian Fan, Qiaolin Ye, Xubing Yang, and Sruti Das Choudhury. Robust blood pressure estimation using an rgb camera. *Journal of Ambient Intelligence and Humanized Computing*, 11(11):4329–4336, 2020.

[14] Alessandro R Guazzi, Mauricio Villarroel, Joao Jorge, Jonathan Daly, Matthew C Frise, Peter A Robbins, and Lionel Tarassenko. Non-contact measurement of oxygen saturation with an rgb camera. *Biomedical optics express*, 6(9):3320–3338, 2015.

[15] Amogh Gudi, Marian Bittner, and Jan van Gemert. Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation. *Applied Sciences*, 10(23):8630, 2020.

[16] Lingqin Kong, Yuejin Zhao, Liquan Dong, Yiyun Jian, Xiaoli Jin, Bing Li, Yun Feng, Ming Liu, Xiaohua Liu, and Hong Wu. Non-contact detection of oxygen saturation based on visible light imaging device using ambient light. *Optics express*, 21(15):17464–17471, 2013.

[17] Francesco Lamonaca, Domenico Luca Carnì, Domenico Grimaldi, Alfonso Nastro, Maria Riccio, and Vitaliano Spagnolo. Blood oxygen saturation measurement by smartphone camera. In *2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA) Proceedings*, pages 359–364. IEEE, 2015.

[18] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hy-

perband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[19] Santiago Lopez and RTAC Americas. Pulse oximeter fundamentals and design. *Free scale semiconductor*, page 23, 2012.

[20] Joshua Mathew, Xin Tian, Min Wu, and Chau-Wai Wong. Remote blood oxygen estimation from videos using neural networks. *arXiv preprint arXiv:2107.05087*, 2021.

[21] Daniel McDuff. Camera measurement of physiological vital signs. *arXiv preprint arXiv:2111.11547*, 2021.

[22] Sanket Mehta, Arpita Patel, and Jagrat Mehta. Ccd or cmos image sensor for photography. In *2015 International conference on communications and signal processing (ICCSP)*, pages 0291–0294. IEEE, 2015.

[23] Deepak Mishra, Neha Priyadarshini, Supriya Chakraborty, and Mukul Sarkar. Blood oxygen saturation measurement using polarization-dependent optical sectioning. *IEEE Sensors Journal*, 17(12):3900–3908, 2017.

[24] Andreia Moço and Wim Verkruysse. Pulse oximetry based on photoplethysmography imaging with red and green light. *Journal of Clinical Monitoring and Computing*, 35(1):123–133, 2021.

[25] P Madhan Mohan, A Annie Nisha, V Nagarajan, and E Smiley Jeya Jothi. Measurement of arterial oxygen saturation (spo 2) using ppg optical sensor. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 1136–1140. IEEE, 2016.

[26] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018.

[27] Fulai Peng, Zhengbo Zhang, Xiaoming Gou, Hongyun Liu, and Weidong Wang. Motion artifact removal from photoplethysmographic signals by combining temporally constrained independent component analysis and adaptive filter. *Biomedical engineering online*, 13(1):1–14, 2014.

[28] Abdul Qayyum, MKA Ahamed Khan, Moona Mazher, M Suresh, D Najumnissa Jamal, and John Tran Duc Chung. Convolutional neural network approach for estimating physiological states involving face analytics. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 68–72. IEEE, 2019.

[29] Abdul Qayyum, Moona Mazher, Aliyu Nuhu, Abdesslam Benzinou, Aamir Saeed Malik, and Imran Razzak. Assessment of physiological states from contactless face video: a sparse representation approach. *Computing*, pages 1–21, 2022.

[30] Alessandra de Fátima Galvão Rosa and Roberto Cesar Betini. Noncontact spo 2 measurement using eulerian video magnification. *IEEE Transactions on Instrumentation and Measurement*, 69(5):2120–2130, 2019.

[31] Christopher G Scully, Jinseok Lee, Joseph Meyer, Alexander M Gorbach, Domhnull Granquist-Fraser, Yitzhak Mendelson, and Ki H Chon. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering*, 59(2):303–306, 2011.

[32] Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Transactions on Biomedical Engineering*, 63(6):1091–1098, 2015.

[33] Rencheng Song, Senle Zhang, Chang Li, Yunfei Zhang, Juan Cheng, and Xun Chen. Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks. *IEEE Transactions on Instrumentation and Measurement*, 69(10):7411–7421, 2020.

[34] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin W Bowyer, Diane Wright, and Patrick Flynn. Deception detection and remote phys-

iological monitoring: A dataset and baseline experimental results. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[35] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062. IEEE, 2014.

[36] Zhiyuan Sun, Qinghua He, Yuandong Li, Wendy Wang, and Ruikang K Wang. Robust non-contact peripheral oxygenation saturation measurement using smartphone-enabled imaging photoplethysmography. *Biomedical Optics Express*, 12(3):1746–1760, 2021.

[37] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological measurement*, 35(5):807, 2014.

[38] Xin Tian, Chau-Wai Wong, Sushant M Ranadive, and Min Wu. A multi-channel ratio-of-ratios method for noncontact hand video based spo _2 monitoring using smartphone cameras. *arXiv preprint arXiv:2107.08528*, 2021.

[39] Junqing Wang, Gang Liu, Ken Cham-Fai Leung, Romaric Loffroy, Pu-Xuan Lu, Xiang J Wang, et al. Opportunities and challenges of fluorescent carbon dots in translational optical imaging. *Current pharmaceutical design*, 21(37):5401–5416, 2015.

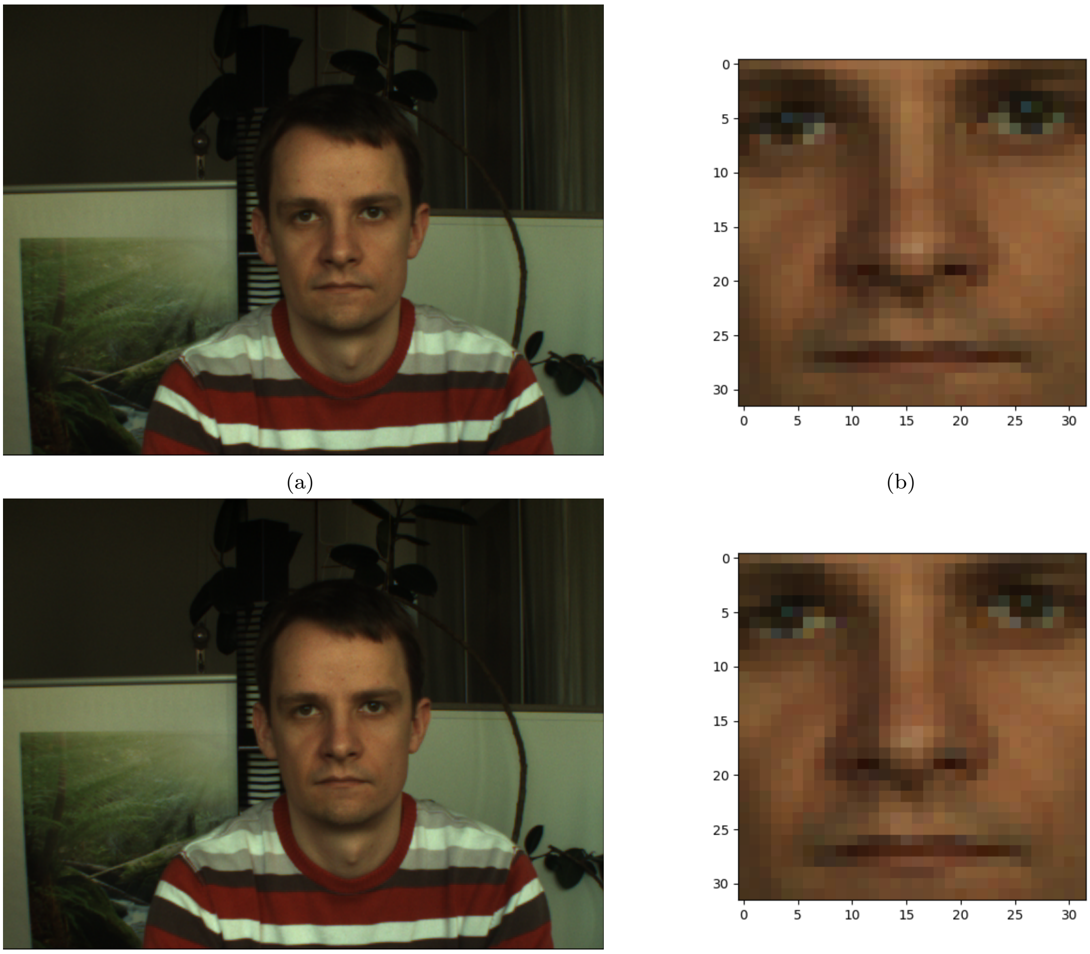
[40] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

[41] John G Webster. *Design of pulse oximeters.* CRC Press, 1997.

[42] Jialiang Zhuang, Bin Li, Yun Zhang, and Xiujuan Zheng. Insightnet: non-contact blood pressure measuring network based on face video. *arXiv preprint arXiv:2203.03634*, 2022.

Appendices

A PURE dataset



(a)

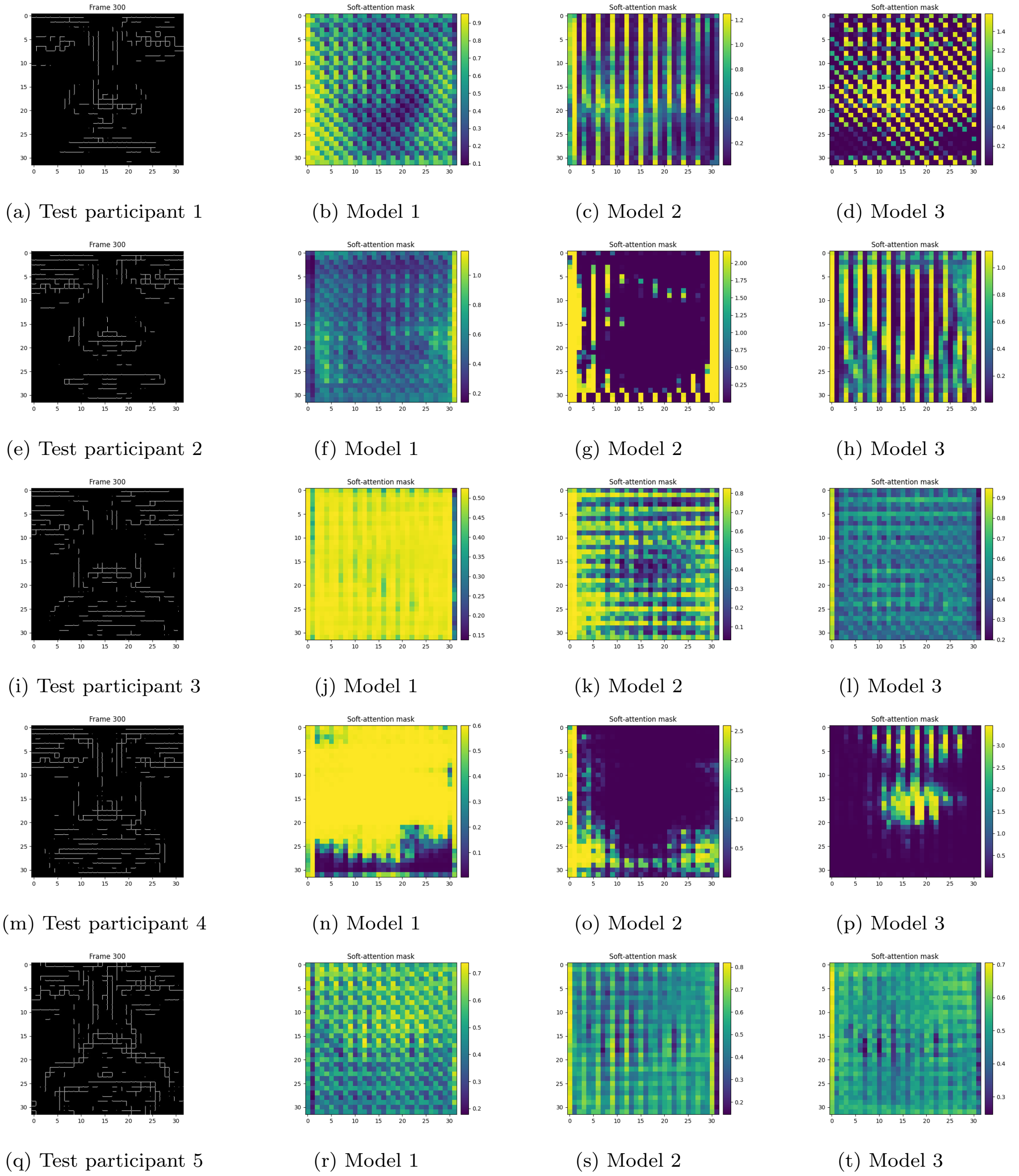
(b)

(c)

(d)

Figure 8: Examples of frames of the PURE dataset (left column), with their corresponding cropped image (right column). The images are extracted from the video sequence of participant 1 for the steady experiment.

B Appearance model's soft-attention mask 1



(a) Test participant 1 (b) Model 1 (c) Model 2 (d) Model 3

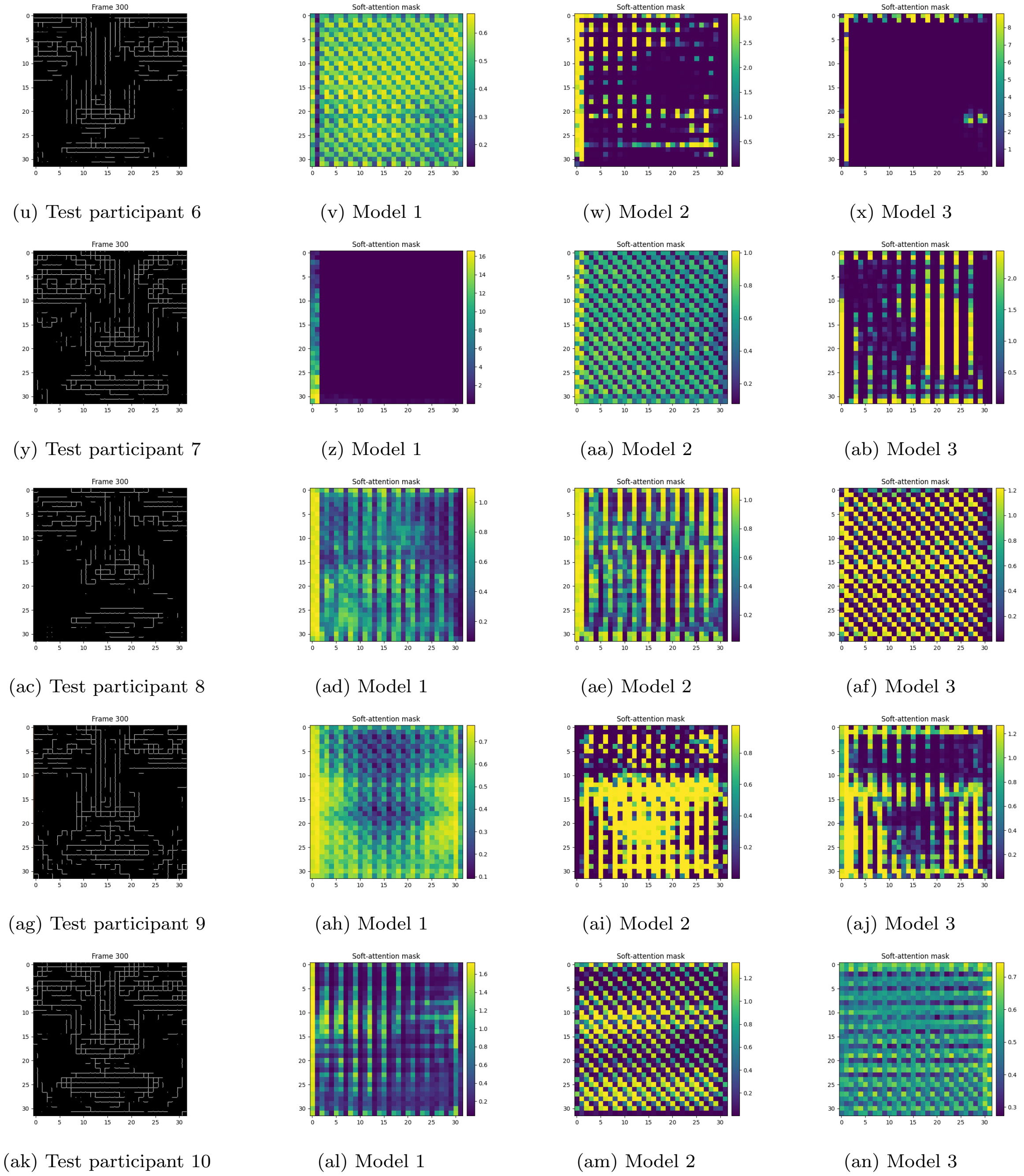
(e) Test participant 2 (f) Model 1 (g) Model 2 (h) Model 3

(i) Test participant 3 (j) Model 1 (k) Model 2 (l) Model 3

(m) Test participant 4 (n) Model 1 (o) Model 2 (p) Model 3

(q) Test participant 5 (r) Model 1 (s) Model 2 (t) Model 3

Figure 9: Learned masks of the cropped 300^{th} frame for each participant (first column) of the Appearance model trained in combination with model 1 (second column), 2 (third column) and 3 (fourth column). For privacy reasons the persons are made unrecognizable.



(u) Test participant 6 (v) Model 1 (w) Model 2 (x) Model 3

(y) Test participant 7 (z) Model 1 (aa) Model 2 (ab) Model 3

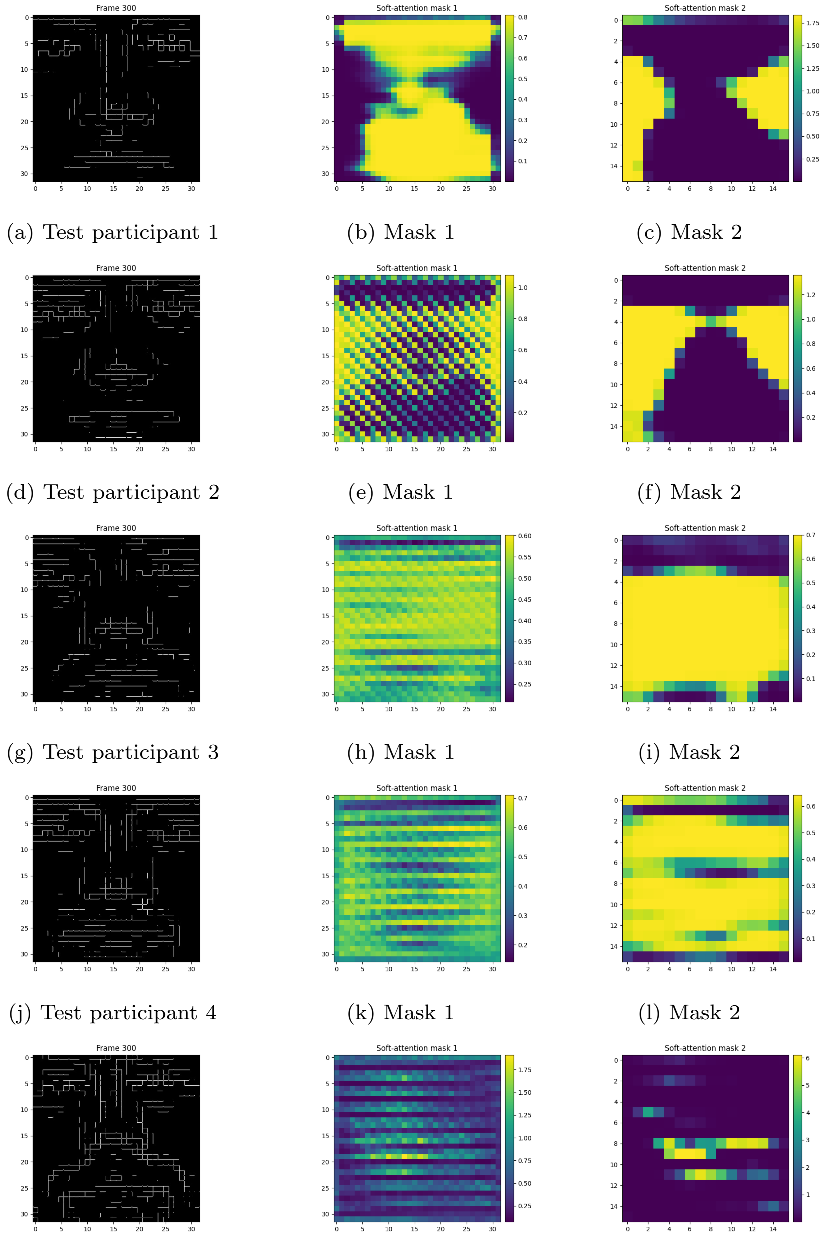
(ac) Test participant 8 (ad) Model 1 (ae) Model 2 (af) Model 3

(ag) Test participant 9 (ah) Model 1 (ai) Model 2 (aj) Model 3

(ak) Test participant 10 (al) Model 1 (am) Model 2 (an) Model 3

Figure 9: Learned masks of the cropped 300^{th} frame for each participant (first column) of the Appearance model trained in combination with model 1 (second column), 2 (third column) and 3 (fourth column). For privacy reasons the persons are made unrecognizable.

C Adapted DeepPhys trained sequentially



(a) Test participant 1 (b) Mask 1 (c) Mask 2

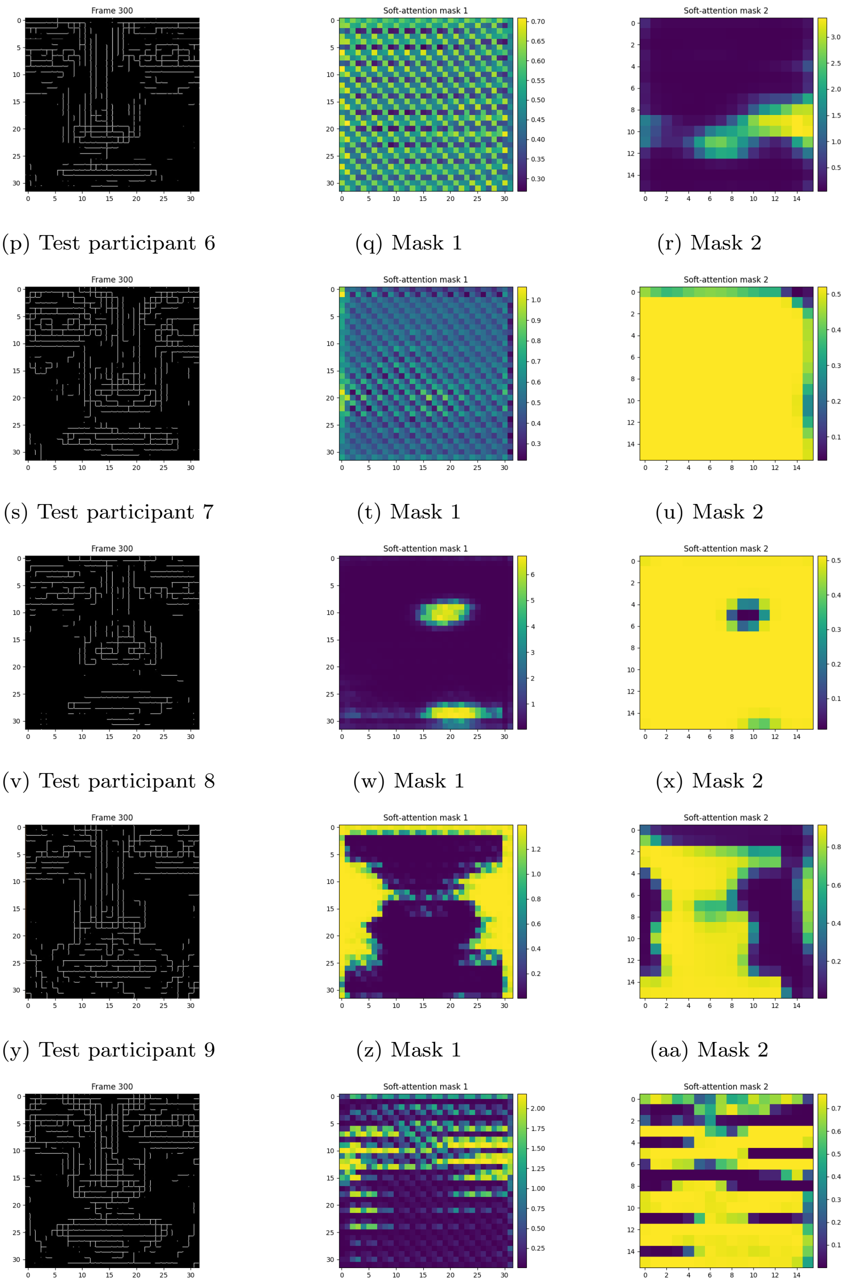
(d) Test participant 2 (e) Mask 1 (f) Mask 2

(g) Test participant 3 (h) Mask 1 (i) Mask 2

(j) Test participant 4 (k) Mask 1 (l) Mask 2

(m) Test participant 5 (n) Mask 1 (o) Mask 2

Figure 10: Results for Adapted DeepPhys trained sequentially. The first column represents the cropped 300^{th} frame of the video sequence of each participant. The second column displays the learned soft-attention masks 1 and the third column displays the learned soft-attention masks 2. For privacy reasons the persons are made unrecognizable.



(p) Test participant 6 (q) Mask 1 (r) Mask 2

(s) Test participant 7 (t) Mask 1 (u) Mask 2

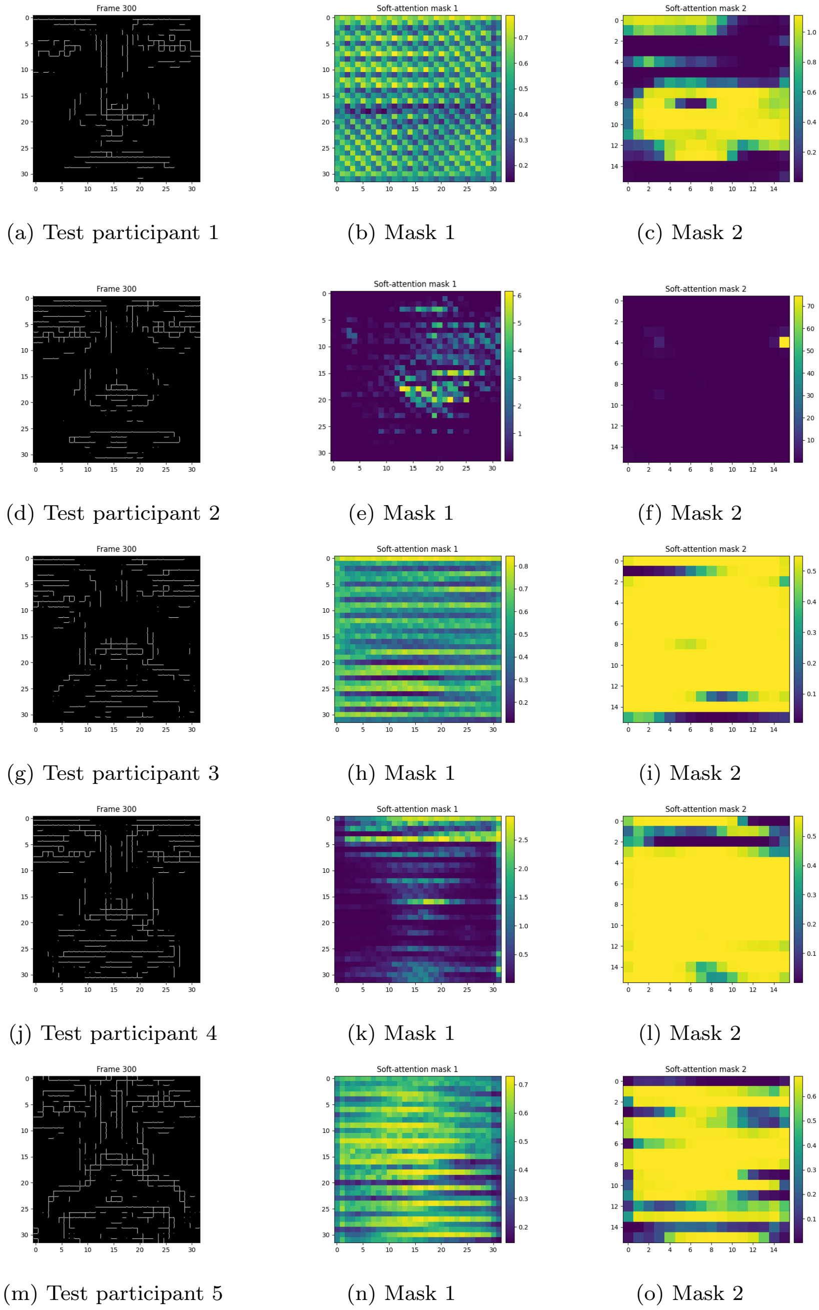
(v) Test participant 8 (w) Mask 1 (x) Mask 2

(y) Test participant 9 (z) Mask 1 (aa) Mask 2

(ab) Test participant 10 (ac) Mask 1 (ad) Mask 2

Figure 10: Results for Adapted DeepPhys trained sequentially. The first column represents the cropped 300^{th} frame of the video sequence of each participant. The second column displays the learned soft-attention masks 1 and the third column displays the learned soft-attention masks 2. For privacy reasons the persons are made unrecognizable.

D Adapted DeepPhys trained non-sequentially



(a) Test participant 1 (b) Mask 1 (c) Mask 2

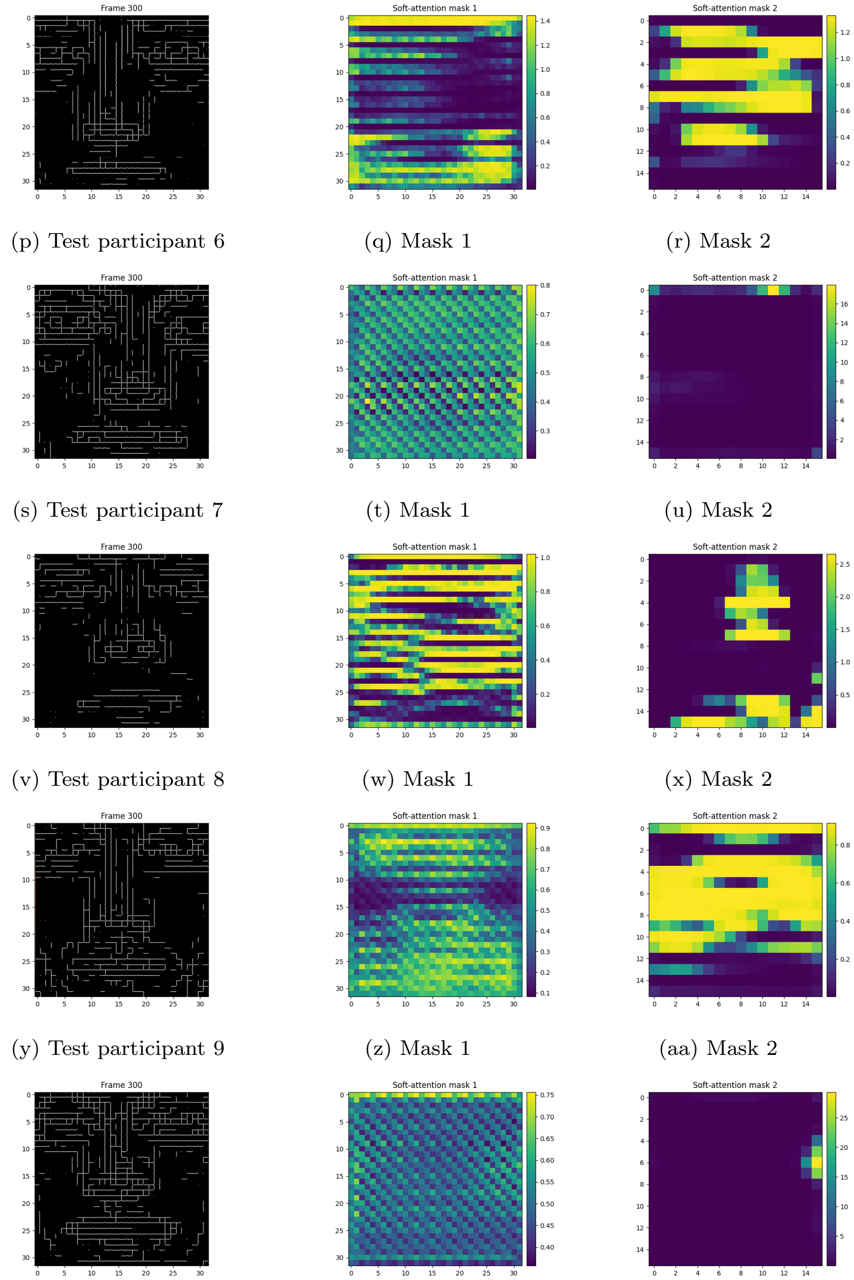
(d) Test participant 2 (e) Mask 1 (f) Mask 2

(g) Test participant 3 (h) Mask 1 (i) Mask 2

(j) Test participant 4 (k) Mask 1 (l) Mask 2

(m) Test participant 5 (n) Mask 1 (o) Mask 2

Figure 11: Results for Adapted DeepPhys trained non-sequentially. The first column represents the cropped 300^{th} frame of the video sequence of each participant. The second column displays the learned soft-attention masks 1 and the third column displays the learned soft-attention masks 2. For privacy reasons the persons are made unrecognizable.



(p) Test participant 6 (q) Mask 1 (r) Mask 2

(s) Test participant 7 (t) Mask 1 (u) Mask 2

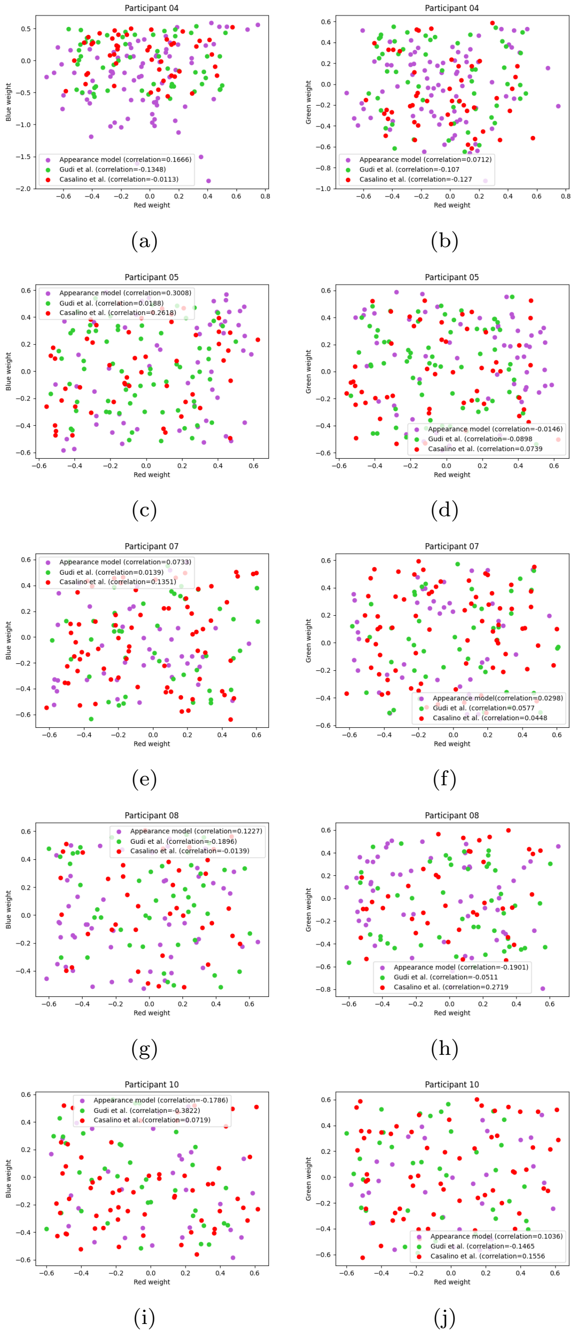
(v) Test participant 8 (w) Mask 1 (x) Mask 2

(y) Test participant 9 (z) Mask 1 (aa) Mask 2

(ab) Test participant 10 (ac) Mask 1 (ad) Mask 2

Figure 11: Results for Adapted DeepPhys trained non-sequentially. The first column represents the cropped 300^{th} frame of the video sequence of each participant. The second column displays the learned soft-attention masks 1 and the third column displays the learned soft-attention masks 2. For privacy reasons the persons are made unrecognizable.

E Model 2 weight visualization



(a) (b)

(c) (d)

(e) (f)

(g) (h)

(i) (j)

Figure 12: The comparison of the RGB channel weights of the first linear layer of model 2 trained in combination with traditional RoI selectors (i.e. Gudi et al. [15] and Casalino et al. [6]) and DeepPhys' Appearance model, as described in Section 4.4. The first column displays the weights of the red and blue channel, and the second column displays the weights of the red and green channel.

2

Basic Deep Learning concepts

In this chapter, the basic Deep Learning concepts used in our research are explained and are intended as refreshments. The readers who are already familiar with the concepts can feel free to skip over this Section.

2.1. Linear Layers

This sub-section describes the basics of linear layers using Figure 2.1. A linear layer has an input and output layer consisting of an arbitrary number of nodes. The value of an output layer's node is computed by multiplying each node of the input layer by the corresponding weights of the node of the output layer. A constant value, called the bias, is added to the results of the multiplication. In other words, the nodes of the output layer are a linear combination of the nodes of the input layer.

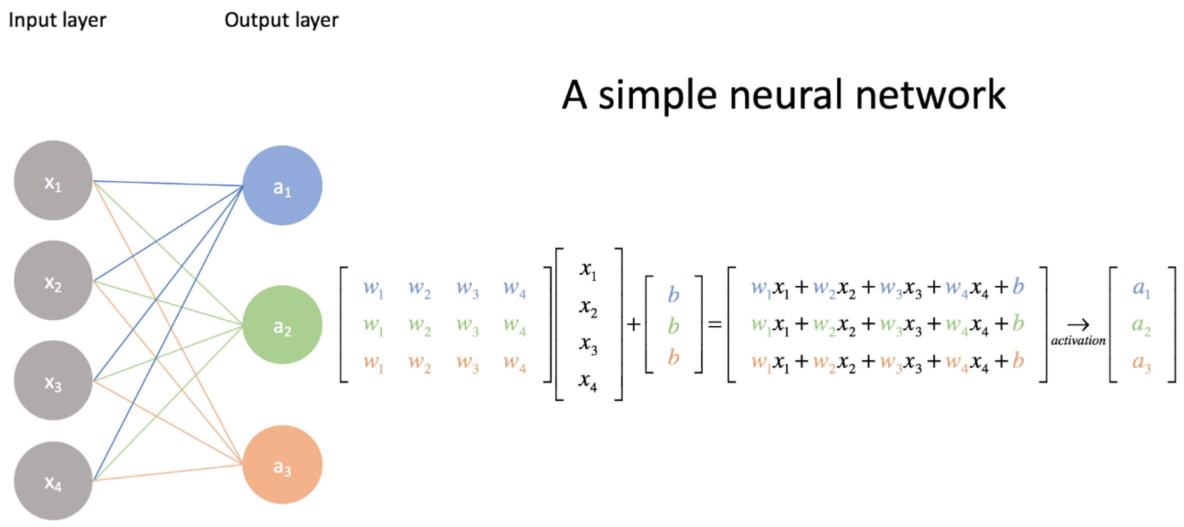

Input layer
Output layer
A simple neural network
x1
x2
x3
x4
a1
a2
a3
$\begin{bmatrix} w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \\ w_1 & w_2 & w_3 & w_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow[activation]{} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$

Figure 2.1: Example of a 3-node linear layer. Image obtained from https://ashwinhprasad.medium.com/pytorch-for-deep-learning-nn-linear-and-nn-relu-explained-77f3e1007dbb

2.2. Convolutional layers

Convolutional neural networks (CNNs) [14] consist of convolutional layers and are inspired by the natural visual perception mechanism of living creatures [7]. CNNs have shown their adequacy in a wide variety of image processing tasks (e.g. image classification [16, 24] and object detection [3, 6]).

Figure 2.2 shows an example of a convolutional layer. The output is obtained by sliding a kernel of arbitrary size, which is in this example 3×3, over the input image. The centre pixel of the kernel corresponds to the location of the output value, which is computed by multiplying the kernel weights with their corresponding input value.

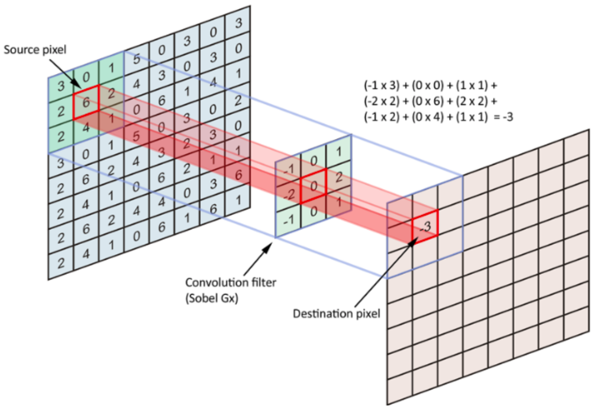

Source pixel
(-1 x 3) + (0 x 0) + (1 x 1) +
(-2 x 2) + (0 x 6) + (2 x 2) +
(-1 x 2) + (0 x 4) + (1 x 1) = -3
Convolution filter
(Sobel Gx)
Destination pixel

Figure 2.2: Example of a convolutional layer with a 3×3 kernel. Image obtained from https://medium.com/ai-salon/understanding-deep-self-attention-mechanism-in-convolution-neural-networks-e8f9c01cb251.

2.2.1. Padding

Determining the border pixels of the image is in the given example not possible since this would result in the kernel exceeding the boundaries of the image. This is so-called 'valid' padding and we, therefore, lose the boundary values of the image. For our research 'same' padding is used, which is also known as zero padding, where the values outside the image are set to 0. This allows the kernel to determine the boundary values, which results in the same input and output size.

2.3. Pooling

The pooling operation downsamples the input image, depending on the kernel size. Figure 2.3 shows an example of the max and average pooling operation. In this example, a 2×2 kernel is slid over the image in a non-overlapping way, as shown by the colours. The stride is defined as the number of pixels the kernel moves horizontally and vertically. To make the 2×2 kernel non-overlapping, the stride is set to 2. Finally, in the case of max pooling the maximum value of the region, indicated by a particular colour, is output and in the case of average pooling the average value.

2.4. Activation functions

Neural networks try to approximate an arbitrary function, which is also known as the true function. The loss of the neural network is determined by the loss function, output of the neural network and the corresponding ground truth (i.e. the output of the true function). The true function can be either linear or non-linear, therefore, non-linearity needs to be added to the linear layers. This can be done by making use of activation functions [19] (e.g. the binary step and sigmoid function). In this sub-section, we will dive deeper into two activation functions used in our research, which are the Rectified Linear Unit (ReLU) and the hyperbolic tangent (TanH) function.

The ReLU function $f(u)$ outputs u if u is larger than 0, else it outputs 0, as defined in Equation 2.1 and shown in Figure 2.4. The output range of ReLU is $[0, \infty]$ and the derivative is defined in Equation 2.2, of which the usefulness will be explained in Section 2.5.

$$f(u) = \begin{cases} u, & \text{if } u \geq 0 \\ 0, & \text{otherwise} \end{cases} = \max(0, u) \tag{2.1}$$

$$f'(u) = \begin{cases} 1, & \text{if } u \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

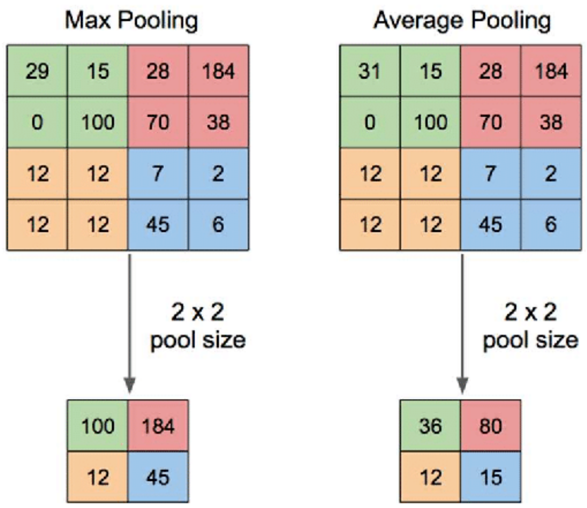


Figure 2.3: Example of a 2×2 kernel max and average pooling operation, with a stride of 2. Image obtained from [25].

The symmetric TanH function, with output range $(-1, 1)$, is shown in Figure 2.5 and defined as in Equation 2.3. The corresponding derivative is defined in Equation 2.4. The property of the TanH function is that relatively large positive and negative input values correspond to derivatives approximately equal to 0. Whereas, the derivatives for input values close to 0 are significantly higher.

$$f(x) = \frac{2}{1 + e^{-2}} - 1 \tag{2.3}$$

$$f'(x) = 1 - f(x)^2 \tag{2.4}$$

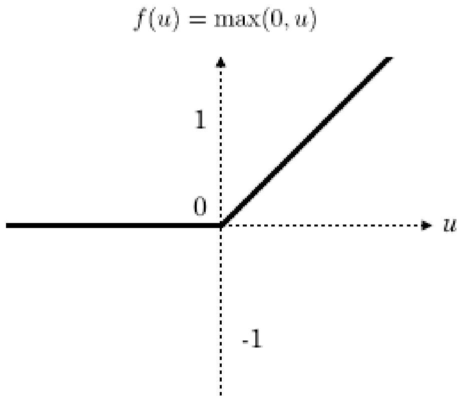


Figure 2.4: The Rectified Linear Unit (ReLU) activation function. Image obtained from [15].

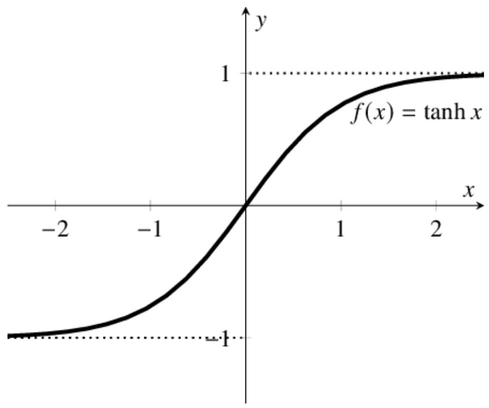


Figure 2.5: The Hyperbolic Tangent (TanH) activation function. Image obtained from [9]

2.5. Training neural networks

In our work, the neural networks are trained by a process called mini-batch gradient descent [17], which updates the weights and biases of the nodes based on a loss criterion computed over a certain number of samples (i.e. the batch size). In contrast to updating the weights over a batch, gradient descent updates the parameters (i.e. weights and biases) over the entire dataset and stochastic gradient descent (SGD) updates the parameters for each sample [17].

2.5.1. Gradient descent

Figure 2.6 shows an example of gradient descent, where the objective of is to minimize the cost function $f(x) = \frac{1}{2}x^2$. The minimum is at $x = 0$ and for values $x > 0$ the derivative $f'(x)$ is larger than 0 and for $x < 0$ the derivative $f'(x)$ is negative. Furthermore, the cost function $f(x)$ is prior unknown but can be explored according to Equation 2.5. This implies that we can minimize the cost by moving small steps in the opposite direction of the gradient as specified in Equation 2.6. By setting the step size α too high, the approximation specified in Equation 2.5 does not hold any more. This can result in overshooting the minimum.

$$f(x + \delta) \approx f(x) + f'(x) \cdot \delta \tag{2.5}$$

$$x_{new} = x_{old} - \alpha f'(x), \text{ where } \alpha \text{ is the step size.} \tag{2.6}$$

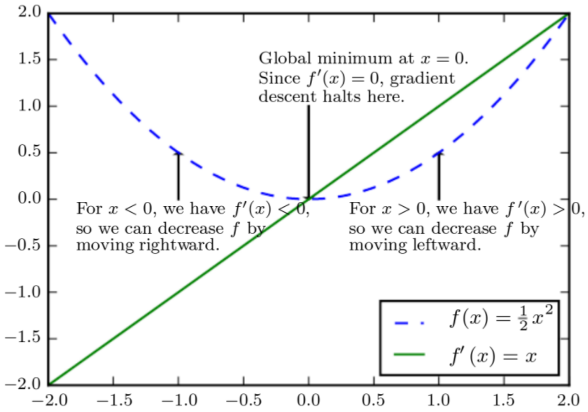


Figure 2.6: Example of gradient descent. Image obtained from the Deep Learning course slides given by J.C. van Gemert at the TU Delft.

2.5.2. Backpropagation

Backpropagation [2] is a way to update the parameters of the network based on gradients. First, the sample(s) are forward passed through the network, which means that the input values are propagated through the network. Secondly, the result of the forward pass (i.e. the output of the network) is used to compute the loss, according to the used loss criterion. Finally, the gradients of the loss are computed over the sample(s) with respect to the weights and biases of the network. After computation of the gradients, one of the variants of gradient descent (i.e. SGD, gradient descent or mini-batch gradient descent) can be performed.

3 Non-contact oxygen saturation estimation

This chapter first describes the principles of oxygen saturation estimation. Secondly, non-contact oxygen saturation estimation by RGB camera and two traditional skin pixel trace extraction methods are described.

3.1. Pulse oximetry

Conventionally, peripheral oxygen saturation (SpO_2) is measured by a pulse oximeter which requires being in contact with the skin. In contrast to arterial oxygen saturation (SaO_2), which is measured by taking blood samples, SpO_2 is measured non-invasively at the skin surface [13]. The LEDs of a pulse oximeter commonly emit light at a wavelength of $660nm$ (i.e. red light) and $940nm$ (i.e. infrared light) [21]. At the infrared wavelength, as can be seen from Figure 3.1, oxygenated haemoglobin (O_2Hb) absorbs more light than deoxygenated haemoglobin (HB). At the red wavelength it is reversed; Hb has a higher extinction coefficient than O_2Hb. By measuring the lighting changes of the skin at 2 wavelengths over time, which depend on the corresponding extinction coefficients of Hb and HbO_2, the peripheral oxygen saturation can be determined.

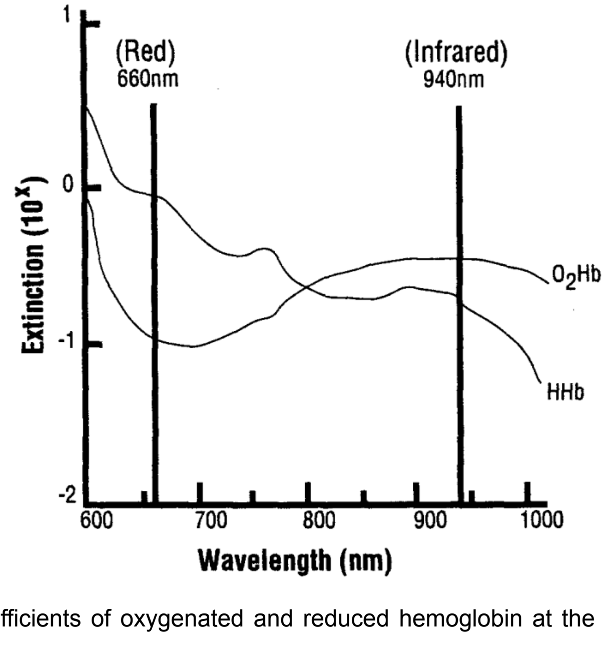


Figure 3.1: Extinction coefficients of oxygenated and reduced hemoglobin at the red to infrared wavelengths. Image obtained from [21].

3.1.1. Ratio-of-ratios

The difference in light absorption of HbO_2 and Hb at 2 different wavelengths is utilized to compute the ratio-of-ratios (RoR). Blood is continuously transported through the body in a pulsating manner by the heart. During 1 heart cycle, the amount of arterial blood fluctuates, which is referred to as the alternating current (AC). The direct current (DC), which does not fluctuate during a heart cycle, consists of tissue, venous blood, and non-pulsatile arterial blood [21]. The AC and DC at the red and infrared wavelength can be used to compute the RoR, as shown in Equation 3.1. The RoR is in turn linearly related to oxygen saturation, of which a derivation can be found in Section 3.1 of our scientific article.

$$RoR = \frac{\frac{AC_{red}}{DC_{red}}}{\frac{AC_{infrared}}{DC_{infrared}}} \tag{3.1}$$

3.2. Non-contact based SpO_2 estimation by RGB camera

Non-contact based oxygen saturation estimation by RGB camera relies on the difference in extinction coefficients of HbO_2 and Hb at the human visual wavelength spectrum, as shown by Figure 3.2. For RGB cameras the blue or green wavelength is used instead of the infrared wavelength, which is used for pulse oximeters. Similar to pulse oximeter contact-based SpO_2 estimation, the authors in [18] defined the ratio-of-ratios for non-contact RGB video sequences as in Equation 3.2, in where AC is computed as the standard deviation and DC as the mean of the corresponding channel over a specified time period.

$$S_pO_2 = A - B\frac{AC_{red}/DC_{red}}{AC_{blue}/DC_{blue}} \tag{3.2}$$

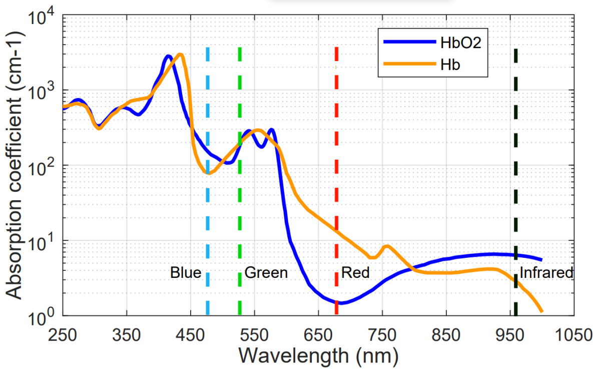
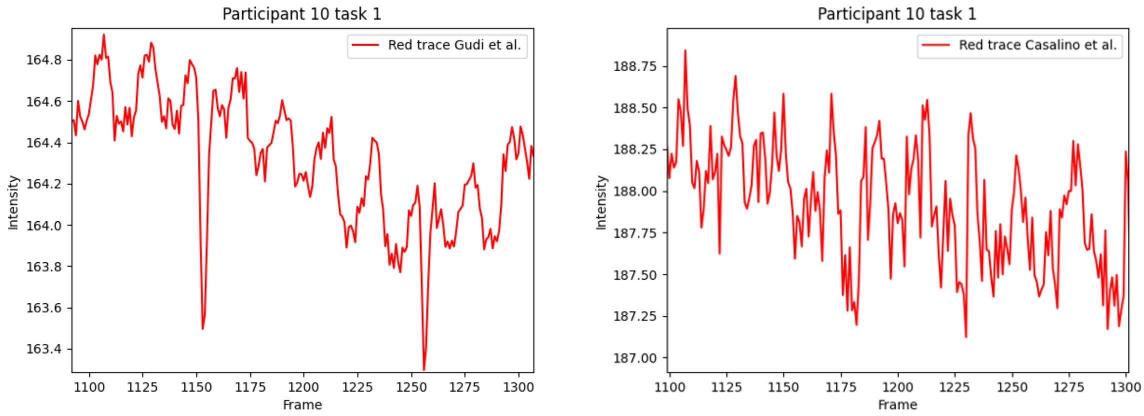


Figure 3.2: Extinction coefficients of oxygenated and reduced hemoglobin at the human visual wavelength spectrum. Image obtained from [12].

The method of obtaining the physiological related signals in a non-contact based manner, by a camera, is called remote photoplethysmograph (rPPG) [5]. To obtain the physiological related signals, from cameras, spatial averaging over SpO_2 related facial regions is performed.

3.2.1. Region of interest selection

In our paper, we use 2 traditional Region-of-Interest (RoI) selectors, introduced by Casalino et al. [1] and Gudi et al. [8]. Casalino et al.'s method is designed for non-contact oxygen saturation measurements, whereas Gudi et al.'s method is initially designed for heart rate (HR) and heart rate variability (HRV) estimation. The selected RoI by Gudi et al.'s method is larger than each of the RoIs of Casalino et al.'s method. This results in Gudi et al.'s method to obtain relatively smoother signals, as shown in Figure 3.3. However, the pulsatile component for Gudi et al.'s method is generally diminished with respect to Casalino et al.'s method. This is because of the fact that we spatially average over a larger region, where most likely not all pixels contribute equally to the constructed rPPG signal.



(a) Trace obtained by Gudi et al.'s RoI selector.

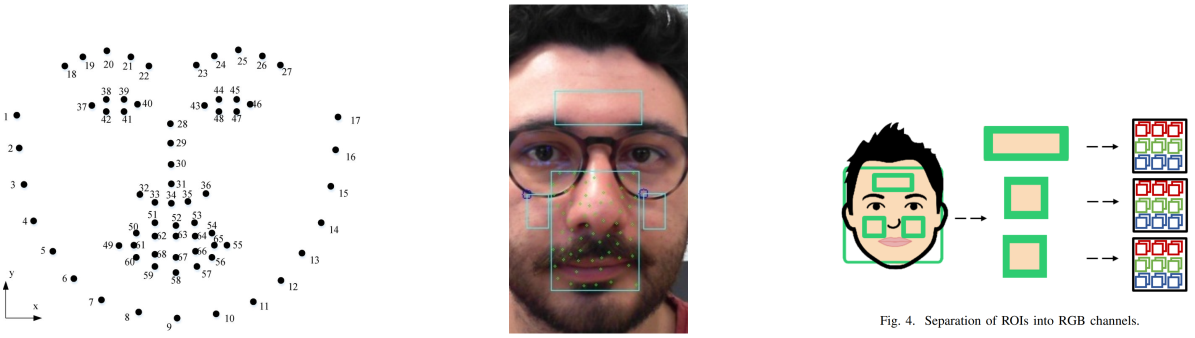
(b) Trace obtained by Casalino et al.'s RoI selector.

Figure 3.3: The red trace obtained by the two traditional RoI selector from test participant 10, for the task of sitting still.

Casalino et al.'s method first uses Dlib's pre-trained face detector to crop the face. Secondly, Dlib is used to obtain the 68 facial landmark points as shown in Figure 3.4a. The 68 landmark points are used to select the corner points of the regions of interest to include as much skin as possible. The RoIs include the forehead, left cheek and right cheek and are shown in Figure 3.4b. The centre box containing parts of the nose and mouth is used for tracking head movements. To be more specific, feature points are selected in this box according to [20]. Then Kanade-Lucas-Tomasi's method [22] is used to track the feature points in each frame. A feature point is selected if both eigenvalues λ_1 and λ_2 of matrix $Z = \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix}$, where g_x and g_y are respectively the gradients in the x and y direction, are above a predefined threshold α. The selected feature points in the image correspond to the most prominent corner points. The number of selected points can be tuned by changing α; the lower alpha the more points are selected and vice versa. Subsequent frames are usually highly correlated with each other. The horizontal ξ and vertical η pixel displacement from the frame at time t to the next frame at time $t + \tau$ can be modelled according to Equation 3.3 [22]. For the equation to hold, the lightning environment is assumed to be static. Changes in lighting conditions due to for instance head motion and sunlight violate this assumption. The displacement **d** is chosen such as to minimize the error taken over a predefined window size. The previous (x_i, y_i) and updated points (x_{i+1}, y_{i+1}) in the central box are used to determine the transformation matrix H, as defined in Equation 3.4. The transformation matrix is in turn used to update the RoI corner points. Then, each RoI is spatially averaged for each frame to obtain the rPPG signals, as shown in Figure 3.4c. The rPPG signal of each RoI is analyzed in the frequency domain. Finally, the rPPG signal of the RoI where the heart rate frequency is most prominent is chosen.

$I(x, y, t + \tau) \approx I(x - \xi, y - \eta, t)$, where $I(x, y, t)$ is the pixel intensity at location (x, y) at time t, τ is the time between 2 subsequent frames and $\mathbf{d} = (\xi, \eta)$ is the displacement vector in the *x* and *y* direction. (3.3)

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \text{ where } H \text{ is the transformation matrix.} \tag{3.4}$$



(a) The 68 facial landmark points. Image obtained from [11].

(b) Example of the 3 RoIs selected by Casalino et al. Image obtained from [1].

(c) Spatial averaging performed over each channel and each RoI. Image obtained from [1]

Figure 3.4: Casalino et al.'s pipeline. The 68 landmarks (a) are used to determine the RoI corner points (b). For each frame the red, green and blue channel intensity is averaged over each RoI to obtain the rPPG signals (c).

Gudi et al.'s [8] region of interest selection method first employs the Viola-Jones algorithm [23] to detect the face and fits an active appearance model [10] to it. The facial landmarks determined by the active appearance model are used to define the RoI, which is shown in Figure 3.5.

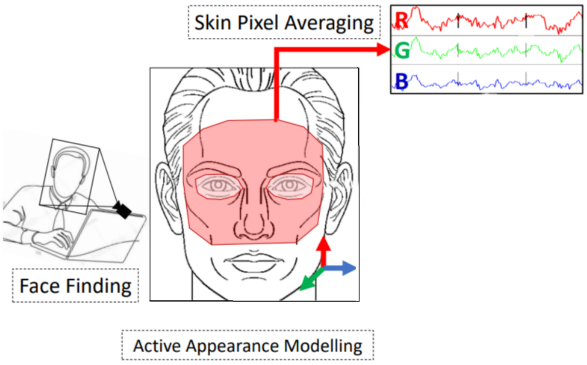


Figure 3.5: The region of interest selection procedure introduced by Gudi et al. The R,G and B traces are obtained by spatial averaging over the RoI for each frame. Image adapted from [8].

4

Proposed deep learning models

This chapter elaborates upon the deep learning models used in our research, which are used to estimate oxygen saturation in a non-contact manner.

4.1. First neural networks for camera based SpO_2 prediction

Mathew et al. [12] introduced the first 3 non-contact video hand based models for SpO_2 estimation, which are shown in Figure 4.1. The neural networks are the current state-of-the-art in non-contact oxygen saturation estimation from the hands. Neural networks have not been applied to the face before and our work examines how these models perform in a facial based setting. This section dives deeper into the rationale of these models.

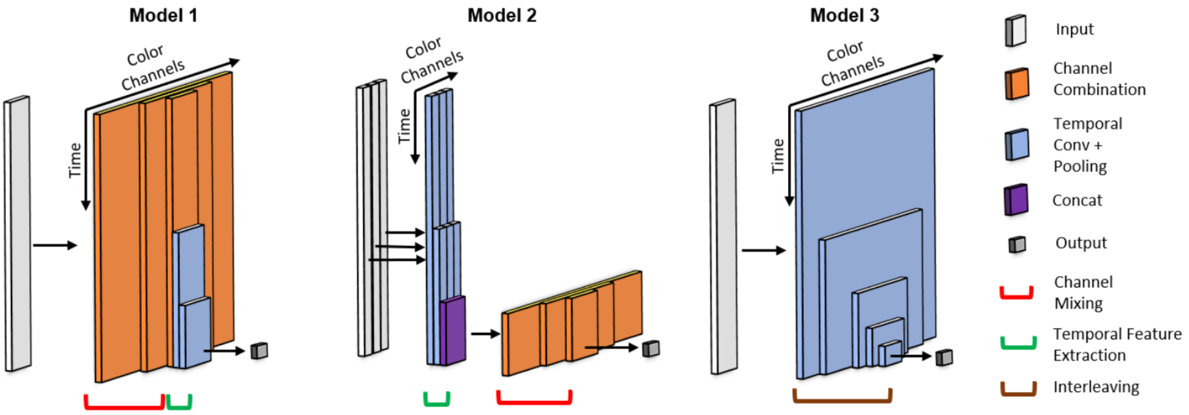


Figure 4.1: The 3 proposed non-contact based SpO_2 estimation neural networks. Image obtained from [12].

4.1.1. Model 1

Model 1 first combines the averaged R, G and B pixel intensities spatially by applying 3 linear layers. After spatial combination, the result is both temporally and spatially combined by applying 2 convolutional layers, which are interleaved with max pooling. Max pooling ensures that maxima and minima in the signal are maintained. This is important for determining the alternating current (AC) component in non-contact SpO_2 estimation. After the convolutional layers, the result is flattened to predict the SpO_2 saturation. Finally, this model is the least complex out of the 3 models in terms of number parameters.

4.1.2. Model 2

Model 2 first combines the averaged R, G and B pixel intensities over time, by applying 1-d convolutional layers in combination with max pooling. After temporally combining the signal, it is spatially combined by applying linear layers. Finally, to predict the output SpO_2 saturation the result of the linear layers is flattened and linearly combined, which is both a spatial and temporal combination. Since the number of feature maps is propagated to the linear layers, this model is more complex than model 1 in terms of number of parameters.

4.1.3. Model 3

Model 3 interleaves channel combination with convolutional layers. The input is first spatially combined by a linear layer. The result after applying the linear layer is combined both spatially and temporally by the convolutional layer, which in turn is combined by the max pooling operator. The result of the 4 interleaving layers is flattened and linearly combined to predict the SpO_2 output. Finally, this model is the most complex out of the 3 models in terms of number parameters.

4.2. DeepPhys

DeepPhys [4], which is shown in Figure 4.2, is the first end-to-end deep convolutional neural network for heart and breathing rate estimation. DeepPhys consists of a Motion model, which takes as input the normalized frame differences, and an Appearance model, which takes as input the current frame. The Appearance model consists of two convolutional soft-attention masks and is used to determine regions that contain the signal of interest. The Motion model uses the change in colour between subsequent frames to determine the output. The purpose of training the Motion model in combination with the Appearance model is to give attention to regions that based on single frame differences contain relevant information with regard to the to be determined physiological output.

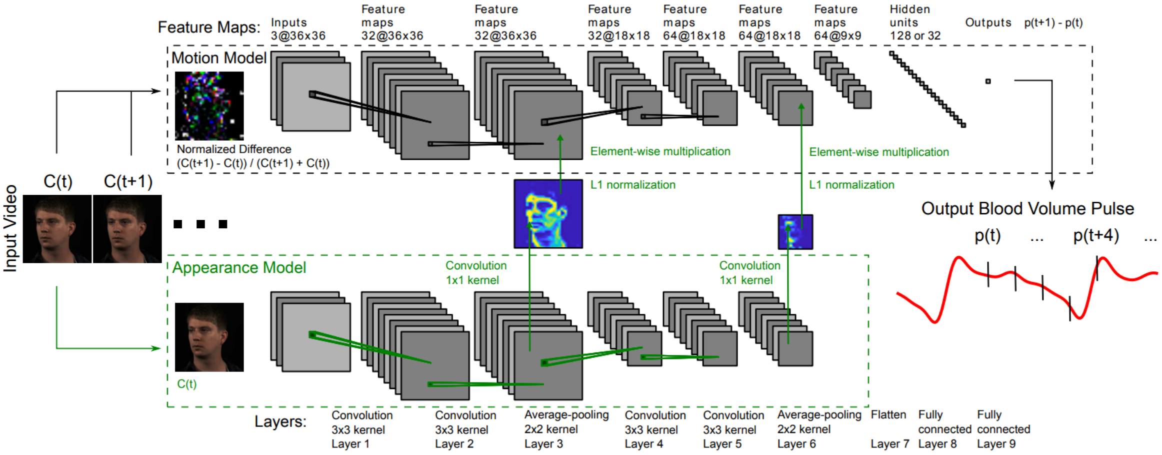


Figure 4.2: DeepPhys; the first end-to-end deep convolutional neural network for heart rate and breating rate estimation. The Appearance model of DeepPhys uses 2 convolutional soft-attention masks to learn weighted facial regions that are related to the physiological signal of interest. Image obtained from [4].

4.2.1. Appearance model

Instead of the traditional RoI selector methods, we use the first part of the Appearance model of DeepPhys to obtain the rPPG signal, as shown in Figure 4.3. The final 1×1 convolutional layer combines the by layer 2 resulting features maps x into a single feature map $\sigma(\mathbf{w}x + b) \in \mathbb{R}^{1x32x32}$. This is done by applying the convolutional weight $\mathbf{w}$ to x, followed by adding the bias b and applying the sigmoid activation function σ. The output of the appearance model is the L1 normalized convolutional soft-attention mask q. This is computed by element-wise multiplying the single feature map $\sigma(\mathbf{w}x + b)$ by the height H and width W of the input image and element-wise dividing the result by the L1-norm of $\sigma(\mathbf{w}x + b)$ times 2, as shown in Equation 4.1. The dot product of the input image (i.e. $C(t)$) and the output of the Appearance model q is normalized to obtain the weighted average R, G and B pixel intensities. This process is repeated for input $C(t)$ up to $C(t + windowSize)$, where $windowSize$ is the number of frames contained in the window. The weighted average R, G and B intensities are concatenated over time, which results in input $I \in \mathbb{R}^{3xwindowSize}$. The input I acts as input to the models designed for SpO$_2$ estimation. Finally, the models are trained separately on the RGB traces.

$$q = \frac{H \cdot W \cdot \sigma(\mathbf{w}x+b)}{2||\sigma(\mathbf{w}x+b)||_1}, \text{ where } \sigma \text{ is the sigmoid function, } H \text{ is the height and } W \text{ is the width.} \tag{4.1}$$
The convolutional weight $\mathbf{w} \in \mathbb{R}^{1x1x32}$ and bias b are applied to feature maps $x \in \mathbb{R}^{32x32x32}$.

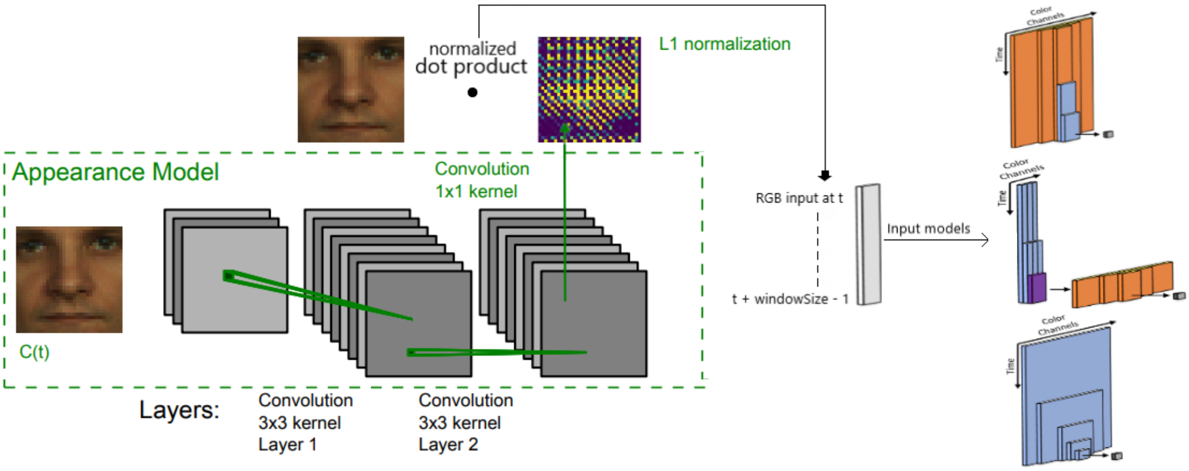


Figure 4.3: Appearance model in combination with SpO$_2$ predictor networks. The convolutional soft-attention mask of the Appearance model is used to compute the weighted average of the input image. The resulting R,G and B intensities are concatenated over time, which act as input the SpO$_2$ models. Image adapted from [4, 12].

4.2.2. Adapted DeepPhys

To compute oxygen saturation, instead of heart and breathing rate, we have adapted DeepPhys. The adapted network can be seen in Figure 4.4. The first part that is adjusted is the cropped input size. For the original DeepPhys network the input images are of size $36x36$, which includes, besides parts of the face, the background. The resolution for Adapted DeepPhys is reduced to $32x32$ by applying bicubic interpolation. Bicubic interpolation is chosen since it reduces noise caused by subtle head motions. By lowering the resolution, more computational efficiency and smoother signals are obtained. Furthermore, the forehead in the cropped areas is not included, since for some participants in the PURE dataset the hair occludes parts of the forehead. By not including the forehead it makes it easier for the Appearance model to learn global features. Secondly, we do not normalize the frame differences, which act as input to the Motion model. Normalizing the frame differences would result in loss of SpO$_2$ related information. This is because we are interested in the change in intensity of the R, G, and B channels with respect to each other. Finally, instead of outputting the difference (i.e. $p(t+1) - p(t)$), we directly output the SpO$_2$ saturation.

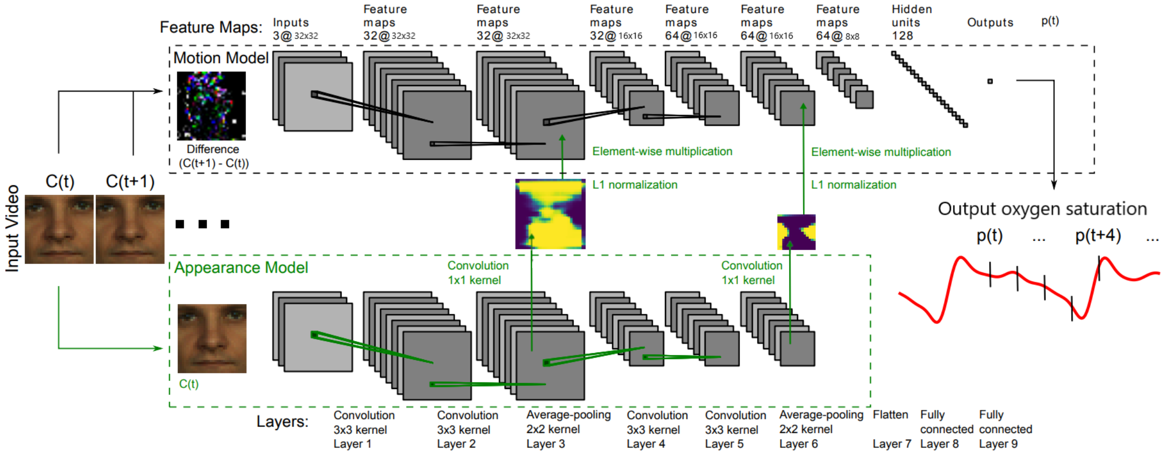


Figure 4.4: Adapted DeepPhys, which consists of the Motion and Appearance model. The Appearance model takes as input the first of the two consecutive frames, whereas the Motion model takes the difference of the two consecutive frames as input. The output of Adapted DeepPhys is the SpO_2 saturation. Image adapted from [4].

Bibliography

[1] Gabriella Casalino, Giovanna Castellano, and Gianluca Zaza. “A mHealth solution for contactless self-monitoring of blood oxygen saturation”. In: *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE. 2020, pp. 1–7.

[2] Yves Chauvin and David E Rumelhart. *Backpropagation: theory, architectures, and applications*. Psychology press, 2013.

[3] Chenyi Chen et al. “R-CNN for small object detection”. In: *Asian conference on computer vision*. Springer. 2016, pp. 214–230.

[4] Weixuan Chen and Daniel McDuff. “Deepphys: Video-based physiological measurement using convolutional attention networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 349–365.

[5] Ananyananda Dasari et al. “Evaluation of biases in remote photoplethysmography methods”. In: *NPJ digital medicine* 4.1 (2021), pp. 1–13.

[6] Spyros Gidaris and Nikos Komodakis. “Object detection via a multi-region and semantic segmentation-aware cnn model”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1134–1142.

[7] Jiuxiang Gu et al. “Recent advances in convolutional neural networks”. In: *Pattern Recognition* 77 (2018), pp. 354–377.

[8] Amogh Gudi, Marian Bittner, and Jan van Gemert. “Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation”. In: *Applied Sciences* 10.23 (2020), p. 8630.

[9] Arthur Selle Jacobs et al. “Artificial neural network model to predict affinity for virtual network functions”. In: *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE. 2018, pp. 1–9.

[10] Hans van Kuilenburg, Marco Wiering, and Marten den Uyl. “A model based method for automatic facial expression recognition”. In: *European conference on machine learning*. Springer. 2005, pp. 194–205.

[11] Dahua Li et al. “Facial expression recognition based on Electroencephalogram and facial landmark localization”. In: *Technology and Health Care* 27.4 (2019), pp. 373–387.

[12] Joshua Mathew et al. “Remote Blood Oxygen Estimation From Videos Using Neural Networks”. In: *arXiv preprint arXiv:2107.05087* (2021).

[13] Meir Nitzan, Ayal Romem, and Robert Koppel. “Pulse oximetry: fundamentals and technology update”. In: *Medical Devices (Auckland, NZ)* 7 (2014), p. 231.

[14] Keiron O’Shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).

[15] Leo Pauly et al. “Deeper networks for pavement crack detection”. In: *Proceedings of the 34th ISARC*. IAARC. 2017, pp. 479–485.

[16] Jiaohua Qin et al. “A biological image classification method based on improved CNN”. In: *Ecological Informatics* 58 (2020), p. 101093.

[17] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[18] Christopher G Scully et al. “Physiological parameter monitoring from optical recordings with a mobile phone”. In: *IEEE Transactions on Biomedical Engineering* 59.2 (2011), pp. 303–306.

[19] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. “Activation functions in neural networks”. In: *towards data science* 6.12 (2017), pp. 310–316.

[20] Jianbo Shi et al. “Tomasi. Good features to track”. In: *Computer Vision and Pattern Recognition*. 1994, pp. 593–600.

[21] James E Sinex. “Pulse oximetry: principles and limitations”. In: *The American journal of emergency medicine* 17.1 (1999), pp. 59–66.

[22] Carlo Tomasi and Takeo Kanade. “Detection and tracking of point”. In: *Int J Comput Vis* 9 (1991), pp. 137–154.

[23] Paul Viola and Michael Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. I–I.

[24] Jiang Wang et al. “Cnn-rnn: A unified framework for multi-label image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2285–2294.

[25] Muhamad Yani et al. “Application of transfer learning using convolutional neural network method for early detection of terry’s nail”. In: *Journal of Physics: Conference Series*. Vol. 1201. 1. IOP Publishing. 2019, p. 012052.