Camera-based Recovery of Cardiovascular Signals from Unconstrained Face Videos Using an Attention Network

Yogesh Deshpande

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

A. Lynn Abbott, Chair
Creed F. Jones
Abhijit Sarkar

May 01, 2023
Blacksburg, Virginia

Keywords: Deep Learning, Remote Photoplethysmograph (iPPG), Biometrics

Camera-based Recovery of Cardiovascular Signals from Unconstrained Face Videos Using an Attention Network

Yogesh Deshpande

(ABSTRACT)

This work addresses the problem of recovering the morphology of blood volume pulse (BVP) information from a video of a person's face. Video-based remote plethysmography methods have shown promising results in estimating vital signs such as heart rate and breathing rate. However, recovering the instantaneous pulse rate signals is still a challenge for the community. This is due to the fact that most of the previous methods concentrate on capturing the temporal average of the cardiovascular signals. In contrast, we present an approach in which BVP signals are extracted with a focus on the recovery of the signal morphology as a generalized form for the computation of physiological metrics. We also place emphasis on allowing natural movements by the subject. Furthermore, our system is capable of extracting individual BVP instances with sufficient signal detail to facilitate candidate re-identification. These improvements have resulted in part from the incorporation of a robust skin-detection module into the overall imaging-based photoplethysmography (iPPG) framework. We present extensive experimental results using the challenging UBFC-Phys dataset and the well-known COHFACE dataset. The source code is available at https://github.com/yogeshd21/CVPM-2023-iPPG-Paper.

Camera-based Recovery of Cardiovascular Signals from Unconstrained Face Videos Using an Attention Network

Yogesh Deshpande

(GENERAL AUDIENCE ABSTRACT)

In this work we are trying to study and recover human health related metrics and the physiological signals which are at the core for the derivation of such metrics. A well know form of physiological signals is ECG (Electrocardiogram) signals and for our research we work with BVP (Blood Volume Pulse) signals. With this work we are proposing a Deep Learning based model for non-invasive retrieval of human physiological signals from human face videos. Most of the state of the art models as well as researchers try to recover averaged cardiac pulse based metrics like heart rate, breathing rate, etc. without focusing on the details of the recovered physiological signal. Physiological signals like BVP have details like systolic peak, diastolic peak and dicrotic notch, and these signals also have applications in various domains like human mental health study, emotional stimuli study, etc. Hence with this work we focus on retrieval of the morphology of such physiological signals and present a quantitative as well as qualitative results for the same. An efficient attention based deep learning model is presented and scope of re-identification using the retrieved signals is also explored. Along with significant implementations like skin detection model our proposed architecture also shows better performance than state of the art models for two very challenging datasets UBFC-Phys as well as COHFACE. The source code is available at https://github.com/yogeshd21/CVPM-2023-iPPG-Paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2136915.

Contents

List of Figures vii

List of Tables x

1 Introduction 1

1.1 Motivation 1

1.2 Main Contributions 3

2 Literature Review 5

2.1 Supervised Learning Based Approaches 5

2.2 Unsupervised Learning Based Approaches 7

2.3 Signal Processing 8

2.4 Authentication 8

2.5 Architectural Progressions 9

2.6 Datasets 11

3 Architecture and Approach 14

3.1 Extraction of Areas of Interest 14

3.2 Skin Segmentation Model 16

3.3 Proposed Model 17
3.4 Implementation and Training Details 21
3.5 BVP Annotation and Loss Function 21
3.6 Signal morphology 22

4 Experiments and Results **26**

4.1 Dataset 26
4.2 Data Distribution and Training Details 27
4.3 Signal Morphology Recovery 30
4.4 Standard Cardiac Pulse Metrics 30
4.5 Re-identification 33

5 Conclusion **37**

6 Discussion and Future Work **38**

Appendices **40**

Appendix A Coverage of Edge Cases and Effect of Skin Detection Model **41**

Appendix B Study of Attention Maps and Future Direction for ROI **42**

Bibliography **45**

List of Figures

1.1 Examples of head movement and occlusion of the face in the dataset UBFC-Phys [34], highlighting our inclusion of natural movements. Green boxes indicate face regions that were detected using MTCNN [55]. 2

3.1 Example outputs of some of the important cases using our ROI pipeline. The system extracts the face region and detects skin while excluding major facial hair and divergent factors such as reflections from eyeglasses. 15

3.2 Our skin segmentation model is an FCN-based encoder-decoder network. This subsystem generates a face mask in which skin pixels have been detected. . . 16

3.3 The complete proposed architecture. Face extraction is performed by the MTCNN module, and skin detection is performed on these areas of interest to recover the required ROI's. From a single frame difference, the system generates a single signal value either for a BVP signal or for a temporal derivative of the BVP signal. 18

4.1 Examples of the illumination variation covered in the COHFACE dataset. The figure also represents the data sample variation in terms of age, gender and skin tone variation. Here we also try to present how the illumination varies accross the face and especially in the cases with low illumination how there could be cases with almost partial illumination. This form of illumination variation when combined with other discrepancies mentioned previously makes the overall dataset tough to train on. 28

4.2 The two plots (a) and (b) are for two different individuals from the UBFC-Phys dataset, representing the morphology recovery from our model (top) and from DeepPhys [5] (bottom) in the respective images. This is a qualitative representation of how well our model retrieves signal morphology and its comparison with signal recovery from a state-of-the-art model that focuses on averaged pulse values. 32

4.3 The figure represents the model's ground truth signal (left half (a), (c)) and its corresponding output from the model (right half (b), (d)) for the same candidate. The top half (a), (b) represents an aggregated signal whereas the bottom half (c), (d) represents the corresponding signal from the top half with its maximum deviation, minimum deviation and mean. The figure shows how the over all shape of the signal is retained as well as how both the systolic and diastolic peaks are recovered by our model. Though signal amplitude is a variable factor and is not as important as the morphology, the intention here is to check if the overall aggregated signal is not having large deviations for a constant amplification factor of the signal. 33

4.4 The rank-wise distribution for re-identification is presented here, where the graph in blue represents the re-identification results for the model trained using BVP signals, and the graph in red represents the re-identification results for the model trained using first derivative BVP signals. The variation in the rank improvement for the BVP as well as first derivative BVP based model is explained more in 4.5. 35

B.1 Attention maps in cases with variation with hair color, facial hair and cases like partial occlusion. We can see how the model focuses on the correct ROI's as expected irrespective of the physical variations. 43

B.2 Attention maps in cases with variation with skin color, glair reflection, physique variation and cases like off-frame area of interest (face in our case). We can see how the model incorporates all the variations and gives consistent attention maps even with variation in skin color or off-frame cases. 44

List of Tables

2.1 This table is a compilation of the different available datasets curated by multiple research groups. 13

4.1 Domain-wise morphology metrics outcomes for our model pipelines and for DeepPhys without any post-processing of the output signals. 30

4.2 Domain-wise morphology metrics outcomes for our BVP first derivative-based model pipeline and DeepPhys after integrating output signals to get them in original BVP signal format. 31

4.3 Performance of our architecture pipelines on UBFC-Phys and COHFACE dataset, in terms of heart rate measurements in beats per minute (HR bpm) [9, 11, 15, 40]. Comparisons have been made with literature using available metrics that are used in our study. 31

4.4 Performance (HR BPM-MAE) of our technique in comparison with previously published state-of-the-art models on the UBFC-Phys and the COHFACE dataset [9, 11, 15, 40]. 34

A.1 Performance of our architecture using averaged cardiac pulse metrics with and without the use of skin detection model. The values help us prove the effectiveness our architectural changes on the UBFC-Phys dataset with as well as without the inclusion of our skin detection model. 41

List of Abbreviations

$\overline{SNR}$ Signal to Noise Ratio

BVP Blood Volume Pulse

ECG Electrocardiogram

FCN Fully Convolutional Network

FFT Fast Fourier Transform

HR Heart Rate

HRV Heart Rate Variability

iPPG imaging-based photoplethysmography

MSE Mean Square Error

MTCNN Multi-Task Cascaded Convolutional Neural Networks

PPG photoplethysmography

psd Power Spectral Density

r Pearson's Correlation Coefficient

RMSE Root Mean Square Error

ROI Region of Interest

rPPG remote photoplethysmography

SGDM Stochastic Gradient Decent With Momentum

smm morphology Metrics

Chapter 1

Introduction

1.1 Motivation

Devices that perform convenient measurements of physiological signals have grown in popularity in recent years. For example, wearable devices by Fitbit [8], Apple [20], AliveCor [1], and others are capable of monitoring heart rate and other vital metrics. In addition to wearable devices, researchers have also considered the use of camera-based monitoring of physiological signals (e.g., [25, 35, 50, 52]). Circumstances such as the novel coronavirus pandemic have also increased awareness of benefits that can be obtained from convenient, noninvasive devices [26]. Unlike systems that require contact with the body, camera-based systems have the potential to be less intrusive in many situations like patient monitoring, driver monitoring [24, 36, 44, 46]. Chen et al. [5] have developed imaging-based systems that benefit from deep-learning techniques. However, more work is needed in sensing instantaneous (instance-level) physiological metrics [22].

This work is concerned with monitoring the cardiovascular system through the analysis of image sequences from standard RGB video cameras. Sample frames are shown in Figure 1.1. The approach is based on the principle that each beat of the heart causes blood volume pulses (BVP) to travel through the body; these pulses cause slight changes in reflectance near the skin that are captured by the camera. The resulting intensity changes are very faint and are not noticeable with the unaided eye. The general framework is known as

Figure 1.1: Examples of head movement and occlusion of the face in the dataset UBFC-Phys [34], highlighting our inclusion of natural movements. Green boxes indicate face regions that were detected using MTCNN [55].

remote photoplethysmography (iPPG), which refers to the use of light to perform remote measurements of volumetric changes.

In recent years, there has been a significant effort toward studying cardiovascular signals using camera-based iPPG. However, most previous approaches have focused on scenarios where the subject's head remains stationary during the measurement process. In contrast, our work emphasizes the need to accommodate relatively large head movements. These movements pose significant challenges in detecting skin regions and measuring intensity changes accurately and reliably. Moreover, several confounding factors complicate the problem further, such as facial hair, eyeglasses, occlusion of the face, and variations in skin tone across subjects. Our work proposes a novel approach to address these challenges and improve the accuracy and reliability of camera-based recovery of cardiovascular signals in scenarios with significant head movements.

One of the key distinguishing features of our work is the focus on extracting individual BVP instances with good approximations of the underlying volumetric signal shape. Unlike previous systems that estimate average heart rate (HR), our approach has the potential to provide information related to inter-beat intervals and heart-rate variability (HRV). Another potential benefit of a pulse-level signal information is the ability to distinguish one person from another. This work also considers the problem of re-identification based on iPPG signals. We present a model that can make such re-identification possible.

1.2 Main Contributions

In summary, the main contributions of this work are as follows,

1) *Improved Attention Pipeline:* The method relies on a deep network that incorporates a novel attention branch with a refined region of interest that emphasizes skin detection and also handles cases with facial hair and specular reflection, over a wide range of skin tones.

2) *morphology Recovery:* The new method emphasizes the recovery of morphology of physiological signals solely from RGB videos of the human face, with emphasis on handling large head movements and partial occlusion.

3) *Improved Standard Cardiovascular Metrics:* We present quantitative experimental results that demonstrate improved estimates of heart rate, as compared to previous state-of-the-art methods.

4) *morphology Metrics:* We present recovered time-variant physiological signal-based metrics and propose a standardized approach that could be followed by future researchers.

5) *Re-identification:* We introduce a candidate approach to subject re-identification, based on the recovered signals from the proposed model rather than averaged cardiovascular metrics.

Our work therefore has the potential for use in biometric authentication tasks.

6) *Generalized ROI With Significant Variations:* Rather than handpicked regions of interest (ROI), our model targets generalization even in extreme cases including partial candidate visibility, occlusion cases, specular reflections from the skin as well as cases such as facial hair; all of them are addressed by our model.

Chapter 2

Literature Review

The research on remote photoplethysmography (iPPG) is based on the biological background where when the visible light goes between 4 to 5mm below the skin surface and the light absorbing components in the skin called chromophores like hemoglobin, change their content. These changes are seen in every pump cycle of the heart which bring up variation in the skin color but is not visible to the human eye. This same phenomenon could be observed using RGB sensors like camera where photoplethysmography comes into play. The approaches considered to work with this idea are generally signal processing based, supervised learning based or unsupervised approaches and their combinations.

2.1 Supervised Learning Based Approaches

One such deep learning based approach is presented in the paper 'DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks' by Weixuan Chen and Daniel McDuff [5]. In this paper the authors are trying to suggest an architecture which allows visualization of spatial-temporal distributions of physiological signals using attention network-based architecture which could also be used to retrieve heart rate and breathing rate. The method is a video based implementation which also considers motion representation based on a skin reflection model and the devised attention mechanism.

Similarly, in the paper 'RhythmNet: End-to-end Heart Rate Estimation from Face via

Spatial-temporal Representation' by Xuesong Niu *et. al.* [28] the authors propose a method for heart rate estimation with an approach based on the generation of a spatial-temporal mask and further using them in a convolution network in order to estimate the heart rate values. They also come-up with a new dataset VIPL-HR specifically created with an intension of working in the area of remote heart rate measurement.

But on the other hand, considering simple CNN based implementations we have the paper 'Visual Heart Rate Estimation with Convolutional Neural Network' by Radim Špetlík *et. al.* [40] in which the authors present a two-step straightforward convolutional neural network to estimate the heart rate from facial images. They also address the lack of variability in the datasets available at that time and have developed their own dataset with essential illumination variation and motion variation among the subjects.

Addressing more on architectures which work towards nullifying the affect on iPPG measurements due to external variations there is this paper 'Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling' by Xuesong Niu *et. al.* [29] through which the authors propose a unique deep learning based implementation where they are trying to address the issue generated due to the overlap of physiological as well as non-physiological parameters in the iPPG based contactless measurements. The disturbances caused due to motion, variation in light, etc. are considered here as the non-physiological parameters and using the cross-verified disentangling strategy they propose a method to retrieve the physiological parameters. They also propose a method to retrieve multi-scale spatial-temporal maps from input frames which are useful for highlighting the physiological information in the face videos.

On the same lines using deep learning techniques but addressing advanced issues like video compression there is this paper titled 'Remote Heart Rate Measurement from Highly Compressed Facial Videos: an End-to-end Deep Learning Solution with Video Enhancement' by

Zitong Yu *et. al.* [53] in which the authors propose an architecture which address the video compression loss as well as measures iPPG signals from highly compressed videos with its help. The method is divided into two parts where one is the input video enhancement based on a unique network and the other part is an attention based network which is used for retrieving the iPPG signals. The authors have named the first block as STVEN (Spatio-Temporal Video Enhancement Networks) and the next one as rPPGNet.

2.2 Unsupervised Learning Based Approaches

Further addressing the overlap of signal processing and unsupervised learning based methods for iPPG based implementations we have the paper 'An Open Framework for Remote-PPG Methods and Their Assessment' by Giuseppe Boccignone *et. al.* [3] where authors have developed an open source iPPG package named pyVHR which stands for Python tool for Virtual Heart Rate. They introduce a logical pipeline useful for addressing the heart rate, heart rate variability and BVP signals in general. They also highlight some important points in the paper where they have mentioned the importance of standardization required in iPPG related research. As the variation in preprocessing, usage of data, testing on across datasets as well as different types of postprocessing methods applied by the various implementations makes the comparison of the different techniques invalid or difficult. They suggest a need of standardization in terms of preprocessing or in general steps as such which have been followed in many implementations and benefit the iPPG related results and implementation.

Similarly in the paper 'Real-Time Webcam Heart-Rate and Variability Estimation with Clean Ground Truth for Evaluation' by Amogh Gudi *et. al.* [11] the authors suggest an approach to recover the iPPG signals in real time such that it calculates the heart rate and gives out the pulse waveform to time heart beats and heart rate variability. They also focus

on performing this as an unsupervised approach which brings a completely unique aspect, as the major data-based training procedure is removed. They have also introduced a new dataset named VicarPPG2 which is specifically useful for heart rate and heart rate variability measurements.

2.3 Signal Processing

Using signal-processing techniques,the variation in average brightness of skin pixels is tracked over time [47]. This variation is too subtle to be noticed by human eyes without digital magnification [35]. Wu et al. [49] proposed a method to amplify such subtle changes. The method, commonly referred to in the literature as VidMag, takes a video sequence as input followed by temporal (band-pass) filtering of frames. The resulting signal after amplification was used to reveal hidden signals. Similarly, Garbey et al. [10] made use of sensitive thermal cameras to acquire the signals from major superficial vessels of face and neck regions. Fourier methods were used to measure cardiac pulse amplitudes. The main problem with these methods is the stability of the face in videos. The observed face is expected to remain stationary, and even small movements cause significant noise during PPG signal recovery. Later, researchers began utilizing a combination of signal-processing techniques and facial tissue trackers to tackle this problem [57].

2.4 Authentication

Every individual possesses a heart and associated vascular system that are inherently unique. When sensing physiological signals such as PPG and ECG, differences between individuals can lead to distinctive characteristics that can be leveraged for the purpose of biometric

authentication [14, 37, 38]. Various research works have shown promising efforts in the field of authentication using contact sensor-based PPG [18]. However, the development of biometric authentication systems using camera-based PPG (rPPG) signals has been a challenging task due to various factors. One of the primary challenges is the presence of noise in iPPG signals, which can lead to inaccurate results. Additionally, the recovery of morphology from iPPG signals has been a difficult task for researchers [23, 30].

2.5 Architectural Progressions

As far as architectures are concerned for the study of deep learning based iPPG methods, they have evolved a lot over the years and have included many different components/modules addressing the problem at a deeper level. Initially the process included face detection, ROI segmentation and further signal processing based on the retrieved ROI's. This approach was too naïve for generalization and did not serve the demanded outcomes. Then with all the different approaches mentioned previously, different techniques and architectural approaches were made available to the research community. Starting with the architecture suggested in the paper DeepPhys which was divided into two parts, a motion model and an appearance model. The motion model was more of a VGG-style CNN model which was useful for identifying the physiological signals based on the motion representations and the appearance model was the attention network devised in order to focus on the skin segmentation such that the predictions are derived based on the expected portions of the face. Similarly in case of the paper Rhytmnet the architecture starts with face detection where they used the SeetaFace face detector and thereby identify 81 facial landmarks. Once the faces are detected and with the corresponding landmarks skin segmentation is done. These processed frames are further converted to YUV color channel and from n face ROI blocks, from which

the respective YUV signals are retrieved. These retrieved signals thus generate the spatial-temporal map which is further used with the convolution network to generate the heart rate signals. The format of the convolution layers is equivalent to the ResNet-18 architecture and the color space transformation from RGB to YUV is computed using,

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \tag{2.1}$$

When addressing advanced issues like video compression the paper 'Remote Heart Rate Measurement from Highly Compressed Facial Videos: an End-to-end Deep Learning Solution with Video Enhancement' by Zitong Yu *et. al.* [53] similar approaches are used with additions of physiological and noise encoder.

Moving on to the Unsupervised and Signal Processing based architectures even they tend to start similar with face area detection and attempts to focus on ROI's but further they adopt different signal processing techniques in order to achieve the desired results. Like in the paper 'An Open Framework for Remote-PPG Methods and Their Assessment' by Giuseppe Boccignone *et. al.* [3], the suggested architecture starts with face extraction using different methods like MTCNN, Dlib and Kalman based face extraction from the input video dataset. Further skin detection is implemented on these extracted faces as the next pre-processing step. They also provide two different options for skin detection one in which standard ROI is considered as in forehead, nose, cheeks, etc. in terms of rectangular regions and the other options includes actual skin segmentation which is achieved by converting the frame in HSV spectrum and then thresholding it based on the possible skin pixel values. Once the ROI's are obtained though it be from a rectangular patch $R(t)$ or skin patch $S(t)$, for each frame an average over all the selected pixels is computed denoted as $q(t)$ and is given as,

$$q(t) = \begin{cases} Patch(R(t)) \\ Skin(S(t)), \end{cases} \tag{2.2}$$

where

$$Patch(R(t)) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|R^{(i)}(t)|}\sum_{(x,y)\in R^{(i)}(t)} R^{(i)}_{x,y}(t) \tag{2.3}$$

$$Skin(S(t)) = \frac{1}{|S(t)|}\sum_{(x,y)\in S(t)} S_{x,y}(t) \tag{2.4}$$

Thus, once the skin detection is performed, from the resultant frame RGB signals are extracted and in order to suppress noise and outliers band-pass filtering is used to get acceptable signals in order to perform BVP signal extraction. They provide different methods like FIR filters, Butterworth IIR filters, Moving Average Filters and methods like detrending so as to achieve the required state of the signals. After this the developed pipeline offers various iPPG algorithms for computing the BVP signal which includes methods like ICA (Independent Component Analysis), PCA (Principal Component Analysis), GREEN, CHROM, POS (Plane Orthogonal to Skin), SSR, LGI (Local Group Invariance), and PBV (Pulse Blood Volume).

2.6 Datasets

While there are a plethora of suggested architectures and approaches there is a consistent talk about the lack of availability of right datasets which could help in iPPG kind of implementation, to an extent where a custom dataset is developed by many of the groups. The

dataset VIPL-HR is introduced in [28] which is combination of visible light videos (VIS) and near-infrared (NIR) videos including variations like head movements, illumination variations, and acquisition device changes. These samples are collected in a less-constrained scenario thus trying to produce natural environment for Heart rate estimation which is one of the major requirements to avoid overfitting or lack of critical learning scenarios. Similarly, VicarPPG 2 dataset is introduced by [11] which is majorly with an aim of evaluating the heart rate and heart rate variation estimations for iPPG based methods. But as we mention more about such application specific datasets there is also data collection seen like in the case of [34] where the data include videos with natural head motions and ground truth as blood volume pulse (BVP) signals as well as electrodermal activity. This dataset called the UBFC-Phys dataset was original created with an application intension for the study of social stress but even such a dataset is useful in the application of iPPG if used in the right way.

There are different levels of research challenges in this domain of work, and they are highlighted along with some potential solution in the paper 'Camera-Based Physiological Sensing: Challenges and Future Directions' by Xin Liu *et. al.* [22]. In this paper the authors mention the importance of clinical graded equipment, measurements using them and how generally, coarse measurements like pulse rate and breathing rate are submitted as results in iPPG work rather than pulse transit time, oxygen saturation and other clinically meaningful metrics. Similarly, most of the datasets available do not include vital cases like atrial fibrillation, other forms of arrhythmia, low oxygen saturation levels (below 85%), high blood pressure, etc. which could be focused eventually as the aim lies in developing a model which considers the full spectrum of possible cases.

Table 2.1: This table is a compilation of the different available datasets curated by multiple research groups.

Dataset Name	Public (Y/N)	No of participants	No. of Videos	Total duration	Variation	Ground Truth
PURE [41]	N	10 (8 male, 2 female)	60	60 min (1 min each)	Facial Expressions Head Movements	PPG
MMSE-HR [56]	N	40 (58 male, 82 female)	102	-	Facial Expressions Head Movements Age and Ethinicity	HR and BP
VIPL-HR [27]	N	107 (79 male, 28 female)	2378 (RGB) 752 (NIR)	30 sec. each	Facial Expressions Head Movements Illumination	HR and BVP
ECG-Fitness [40]	N	17 (14 male, 3 female)	207	1 min each	Facial Expressions Head Movements	ECG
MAHANOB-HCI [39]	N	30 (13 male, 17 female)	527	-	Facial Expressions	ECG, EEG
Vicar PPG2 [42]	N	10	20	90 sec. each	Facial Expressions Head Movements Illumination Camera Types Occlusion	PPG
DEAP [16]	N	22	874	-	Facial Expressions Occlusion	PPG
COHFACE [12]	N	40 (28 male, 12 female)	160	1 min each	Illumination	PPG
LGI [32]	Y (not all)	25 (20 male, 5 female)	100	1 min each ergometer session 5 min	Facial Expressions Head Movements Illumination	PPG
UBFC-rPPG [2]	N	42	42	1 min each	Facial Expressions Head Movements	PPG
UBFC-Phys [34]	Y	56 (10 male, 46 female)	168	3 min each	Facial Expressions Head Movements Occlusion Facial Hair	BVP and EDA

Chapter 3

Architecture and Approach

The direction in which we address this problem starts with a generalized approach for the inclusion of the right data, which also is inclusive of extreme but acceptable cases. Since we are working with the theoretical concept of Shafer's Dichromatic Reflection model which assumes every pixel considered is a skin pixel, to address this assumption we further work with a skin segmentation model. Having this baseline, we finally work with our convolution attention model which is a modification inspired by the work in the paper DeepPhys, to retrieve the target BVP signals.

The data for addressing the problem related to BVP signal retrieval comes with a lot of discrepancies, especially when it includes cases with extreme head and body movements and to avoid the situation of garbage-in and garbage-out it needs significant preprocessing. As a part of preprocessing we made sure we align the ground truth BVP signal samples as per the number of frame samples for the respective video, but did not perform any extra smoothing or realigning on the signal. This was followed by manually removing the garbage cases from the data which would affect learning.

3.1 Extraction of Areas of Interest

The primary objective of this study is to incorporate natural variations in video data during the training process. To achieve this, extreme cases including occlusion, spatial motion, par-

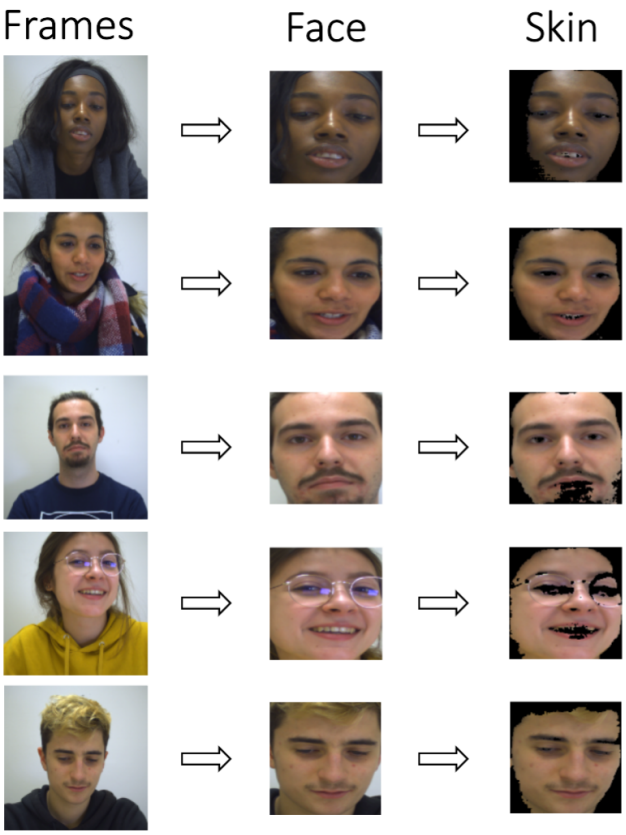

Figure 3.1: Example outputs of some of the important cases using our ROI pipeline. The system extracts the face region and detects skin while excluding major facial hair and divergent factors such as reflections from eyeglasses.

tial frame visibility, and more than just angular movement, must be preserved. Traditional face detection methods, such as center-crop, Haar cascade, and Seetaface, are insufficient for identifying faces with racial variations or handling complex scenarios involving occlusion and spatial movement. Hence, we used MTCNN based face tracking and detection method, which as illustrated in Figure 1.1, helped in both addressing the spatial motion-based variation as well as detection of faces in different angles and natural scenarios in general.

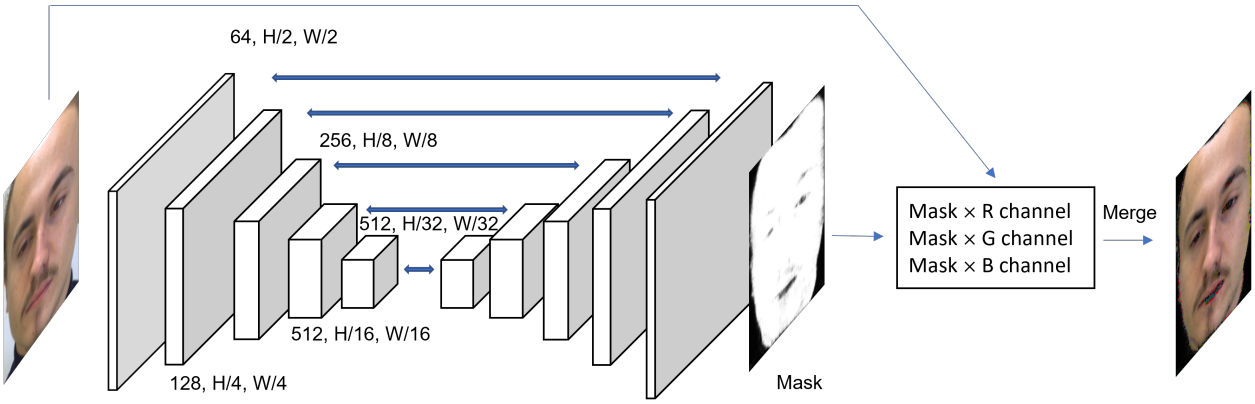

Figure 3.2: Our skin segmentation model is an FCN-based encoder-decoder network. This subsystem generates a face mask in which skin pixels have been detected.

3.2 Skin Segmentation Model

Skin segmentation is a crucial step in training models for remote photoplethysmography (iPPG), as it directs learning toward the important areas of the face from where signals could be recovered. While previous research has emphasized the importance of skin pixels and proper skin detection, there has been less consideration given to the need for having a skin detection algorithm that would avoid facial hair, heavy skin illumination based reflection, glare, etc. Most existing ROI detectors do not have the capability to exclude these regions during learning. Additionally, when employing attention-based networks, it is often assumed that the regions being focused on are exclusively comprised of skin.

To address these problems, we have used a skin segmentation model as shown in Figure 3.2. This is a fully convolutional network (FCN), with an encoder-decoder architecture. The model was trained using the benchmark ECU dataset [31] using binary cross entropy loss function and SGD optimizer. where the loss function could be represented as follows,

$$L_{Skin} = -1/N \sum_{i=1}^{N} y_i \log(f(y_i) + (1 - y_i) \log(1 - f(y_i)) \tag{3.1}$$

where N represents the number of classes (here skin pixels and non-skin pixels), y_i represents labels, and $f(y_i)$ represents predicted probability.

This trained model gives a skin probability mask as the output which is used in our main architecture to gain the output skin frame. To retrieve the skin frame, a thresholding operation is performed on the mask generated from the FCN model such that 0 represented non-skin pixels and 1 represented skin pixels. This computed mask is then multiplied with the RGB channels of the original face frame thus giving us the required skin ROI from the respective input video frame. While training the model along with the standard data augmentation techniques variations like image color variation are also included to consider skin segmentation during uncertain or extreme light effects.

The primary objective of this implementation is to focus on skin regions and eliminate areas that include significant facial hair, heavy illumination points, and reflections from glasses or skin in general. The architecture of the model is presented in Figure 3.2. Additionally, Figure 3.1 illustrates some critical test cases from the UBFC-Phys dataset along with their corresponding skin segmentation results, providing an intuition of how disregarding such scenarios can lead to a lack of generality and natural scenario consideration in iPPG-based model training.

3.3 Proposed Model

Our proposed architecture is based on Shafer's dichromatic reflection model (DRM) and the mathematical analogy introduced by DeepPhys [5]. We can represent the time-varying function of the RGB values of the k^{th} skin pixel in an image sequence as follows,

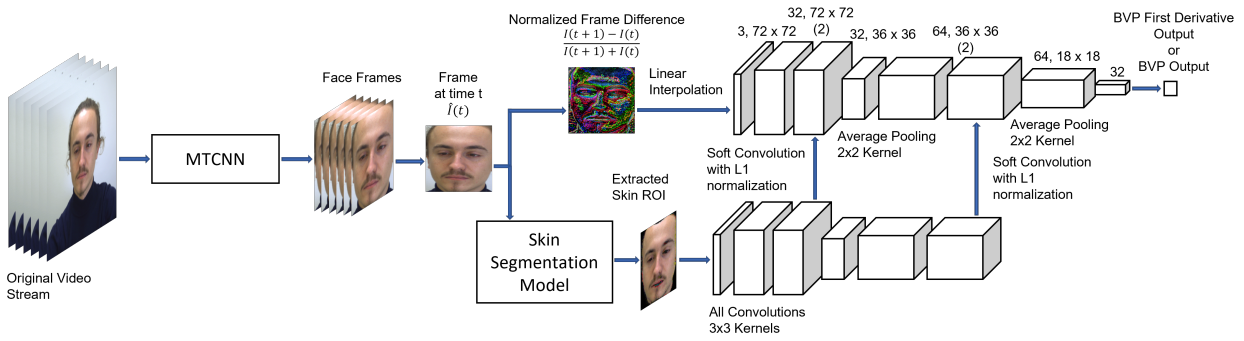

Figure 3.3: The complete proposed architecture. Face extraction is performed by the MTCNN module, and skin detection is performed on these areas of interest to recover the required ROI's. From a single frame difference, the system generates a single signal value either for a BVP signal or for a temporal derivative of the BVP signal.

$$C_k(t) = I(t) \cdot (v_s(t) + v_d(t)) + v_n(t) \tag{3.2}$$

where $C_k(t)$ represents a vector of the RGB values, $I(t)$ represents luminance intensity level, $v_s(t)$ represents specular reflection, $v_d(t)$ represents diffuse reflection and $v_n(t)$ represents camera quantization noise. Here to reduce the effects due to camera quantization error, every frame is down-sampled to a size preferred by the model (72×72 in our case). Then bilinear interpolation for downsampling is used. This is in contrary to using conventional bicubic interpolation, as it helps in avoiding excessive smoothing effects and influences better learning from the face features for the BVP signal retrieval. So equation 3.3 is represented as:

$$C_l(t) = I(t) \cdot (v_s(t) + v_d(t)) \tag{3.3}$$

where $C_l(t)$ represents a vector of the RGB values for the l^{th} skin pixel from the resized frames. In previous modeling approaches, the l^{th} pixel in an image sequence is naively assumed to be a skin pixel. On contrary, we added an additional attention branch in the

training pipeline (see Figure 3.3) that can improve the automatic selection of regions of interest to skin areas.

As a result, we can update 3.3 as,

$$C_l(t) = \begin{cases} 0, & \text{if } Skin(l) < \delta, \\ I(t) \cdot (v_s(t) + v_d(t)), & \text{if } Skin(l) \geq \delta \end{cases} \tag{3.4}$$

where $Skin(l)$ represents the outcome of our skin detection model for the l^{th} pixel and δ represents the threshold for the binary cross entropy probabilities (0.5 in our case).

We use an MTCNN [55] based face tracking and detection to extract target face regions from video frames. This helps spatial motion cancellation and our overall model focus on the horizontal angular motion as well as yaw rotation as briefly highlighted in Figure 1.1 which shows the natural motion of the candidates within a camera frame, that is addressed by our model. Considering previous work and the need of addressing motion changes we use a normalized face frame difference of two consecutive face frames as the input to the main branch of our convolution attention network, whose maxima are clipped to the third standard deviation above the mean. The input normalized face frame difference can be represented as,

$$D_{ip}(t) = min(D(t), D_{ipmax}) \tag{3.5}$$

and,

$$D(t) = \frac{C_l(t+1) - C_l(t)}{C_l(t+1) + C_l(t)} \tag{3.6}$$

$$D_{ipmax}(t) = \mu(D(t)) + (3 \times \sigma(D(t))) \tag{3.7}$$

where $C_l(t)$ is the vector of the RGB values for the resized face frame, $D_{ip}(t)$ is the input (clipped normalized face frame difference), $D(t)$ is the normalized face frame difference without clipping, $D_{ipmax}(t)$ is the maxima of the threshold for clipping, μ represents mean and σ represents standard deviation.

The attention branch of our model which includes the skin segmentation model, helps us retrieve the skin regions from the face which are then batch standardized and passed on to the network further. As depicted in Figure 3.3 the architecture after the skin segmentation model, in the attention branch, is the same as the one we have in the main branch of the CAN network. In case of our model we use 72×72 input frame size which is a signification modification to the DeepPhys model which uses 36×36 input frame size, as it changes the complete feature size consideration for all the layer of the model. We use dropout layers, with dropout rates of 0.5, before every average pooling layer and also before the last fully connected layer which is followed by tanh activation function. The mask generated from our attention branch uses sigmoid activation over the respective branch outcome which is multiplied by the height and width of the respective layer prior to pooling layers and then this outcome is divided by twice the L_1 normalization on the output of the sigmoid activation. Finally, for our feature extracting dense layers we consider 32 feature parameters in the output layer, which has a tanh activation function to keep the outcomes bounded. We trained our attention model using SGDM optimizer, a momentum of 0.9, a batch size of 128, and a learning rate of 10^{-4}.

3.4 Implementation and Training Details

The implementation for our complete model after the required data cleaning and preprocessing is divided into three parts. First we trained the skin segmentation model using the ECU dataset and kept the trained weights ready for skin segmentation on the UBFC Phys dataset. Next we implemented the face tracking and detection using MTCNN in order to get the input video frames ready along with their corresponding BVP signal values. At our final step we use the devised face frames and the pre-trained skin segmentation model to compute the required face frame-difference normalization and skin ROI generation, which was then passed to the main branch and attention branch respectively of our convolution attention network.

In order to reduce the computation time for each run we also implemented GPU based parallel processing using NCCL backend based technique, under the Distributed Data Parallel module available with PyTorch. The required human understandable metrics as well as the metric required for authentication where then calculated based on the output BVP signals retrieved from our model.

3.5 BVP Annotation and Loss Function

During training, the ground truth blood volume pulse (BVP) values are first resampled to match the sampling rate of the video frames. The first derivative of these BVP signals is computed and batch standardized to be used as the ground truth in one training pipeline. We also use the original batch-standardized BVP signal as the ground truth, thus computing outcomes on both the first derivative as well as the original BVP signals in two different training pipelines. To compute the loss during training, we utilize the mean square error

between the model outcome and the standardized ground truth. Hence, in the case of the first derivative BVP signal retrieval, mathematically it could be represented as,

$$b_{der}(t) = b(t+1) - b(t) \tag{3.8}$$

$$b_{gt}(t) = \frac{b_{der}(t) - \mu(b_{der}(t))}{\sigma(b_{der}(t))} \tag{3.9}$$

$$Loss_{CAN} = \frac{1}{N}\sum_{i=1}^{N}(b_{gt}(t) - b_{pred}(t))^2 \tag{3.10}$$

where $b(t)$ is the BVP value collected from the sensor at time t, $b_{der}(t)$ is the first derivative signal, $b_{gt}(t)$ is the standardized first derivative BVP signal that we use as the ground truth and $b_{pred}(t)$ is the predicted model outcome.

3.6 Signal morphology

The field of remote photoplethysmography (iPPG) has mainly focused on extracting average cardiac pulse-based metrics. However, as physical sensor-based technology advances, the potential for generating instantaneous physiological data also increases, highlighting the need for more research in this area [22]. Though recovery of signal morphology is not specifically focused by the previous research, but there has been significant discussion around it that has led to notable contributions [6, 13, 17, 21]. Based on those studies we understood using first derivative BVP signals was important but it was not evident enough on not using original BVP signal based models. This led us to include model pipeline having both the approaches one with original BVP ground truth and another with first derivative BVP ground truth.

There on the objective was to have a qualitative proof of signal recovery as well as quantative metric based signal recovery understanding based on which this section develops further. This work is one of the first to tackle the challenges of detailed shape morphological features and represent a morphology metrics. In this section, we present a set of metrics that can be used to study conformity of the recovered signals.

For morphology-based metrics, we compute the mean of the normalized cross-correlation between the model output signals and the ground truth BVP signals for every candidate in the dataset. These metrics are computed for the respective signals in the time domain, frequency domain as well as power domain, which are reported further giving us a complete idea of how well the model could retrieve correct signal morphology.

The normalized cross-correlation is computed as:

$$ncr(x_{gt}(n), x_{op}(n)) = \frac{\sum_{i=1}^{N} x_{gt}(n_i) x_{op}(n_i)}{\sqrt{\sum_{i=1}^{N} x_{gt}(n_i)^2} \sqrt{\sum_{i=1}^{N} x_{op}(n_i)^2}} \tag{3.11}$$

where $x_{gt}(n)$ is the ground truth signal, $x_{op}(n)$ is the model output signal and N is the number of signal samples.

The signal in the time domain is represented as $x_{gt}(t)$ and $x_{op}(t)$ where $x_{gt}(t)$ is ground truth signal in time domain and $x_{op}(t)$ is model output signal in time domain.

Thus, the same signals in the frequency domain could be represented as follows,

$$x_{gt}(f) = FFT_{mag}(x_{gt}(t)) \tag{3.12}$$

$$x_{op}(f) = FFT_{mag}(x_{op}(t)) \tag{3.13}$$

where FFT_{mag} is the magnitude of the Fast Fourier Transform for a signal in the time domain.

Similarly, the signals in the power domain will be as follows,

$$psd(x(n), f_s) = \lim_{x\to\infty} \frac{1}{T} \left| \sum_{n=1}^{N} x_n e^{-i2\pi fn} \right|^2 \tag{3.14}$$

$$x_{gt}(p) = psd(x_{gt}(t), f_s) \tag{3.15}$$

$$x_{op}(p) = psd(x_{op}(t), f_s) \tag{3.16}$$

where psd is power spectral density, f_s is sampling frequency and N is the number of signal samples.

Thus based on this developed baseline we further compute our morphology metrics in the time, frequency, and power domain denoted as smm_t, smm_f and smm_p respectively, which could be given as follows,

$$smm_t = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(t)_{gt}, x_i(t)_{op}) \tag{3.17}$$

$$smm_f = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(f)_{gt}, x_i(f)_{op}) \tag{3.18}$$

$$smm_p = \frac{1}{C} \sum_{i=1}^{C} ncr(x_i(p)_{gt}, x_i(p)_{op}) \tag{3.19}$$

It is important to follow the normalized cross-correlation based approach in this case as it helps us learn better the morphology of the signal. Generally cross-correlation is used to address this kind of metrics computation but since we have to focus on systolic peak, diastolic peak as well as dicrotic notch with that kind of approach a zero value in the signal

is not taken into consideration, both metrics need to have a similar amplitude which takes away the focus from morphology and it becomes difficult to understand the scoring value because of the lack of a normalizing factor.

Through all our efforts the focus is to retrieve as much of morphology based details as possible. We make it a point to address the systolic peak, diastolic peak as well as dicrotic notch because later these same signals are use to check the scope of human re-identification. It is important to understand that since morphology is not focused-on traditionally, we have to make sure we are not producing signals similar to just a sine wave addressing all the systolic peaks, such that the averaged cardiac pulse based values are intact as per the expectation. Later when we perform re-identification (covered in subsection 4.5) we use Pearson correlation coefficient which helps us in keeping in track the rises and falls in the corresponding signals thus performing the re-identification process correctly as well as addressing all the required details in the signal shape.

Chapter 4

Experiments and Results

Since our implementation addresses the process of BVP signal recovery from face videos as well as re-identification based on those recovered signals, we divided our experiments into three parts, starting with the signal recovery forefront, followed by computation of standard human understandable metrics and then computing re-identification procedure. For better evaluation, we have two different pipelines covering both the recovery from ground truth BVP signals as well as first derivative BVP signals.

4.1 Dataset

We evaluated our work on two different datasets UBFC-Phys as well as COHFACE. We can see glimpse of the variation covered in the UBFC-Phys dataset in Figure 1.1 and because of the kind of nuances introduced through the data it becomes tough to handle and hence not used much in different research papers. The dataset covers all the different possible variations including occlusion, facial hair, cases with glasses, as well as skin tone variation all of which is added with significant body movements which makes it very challenging. Similarly in case of COHFACE data samples presented in Figure 4.1 we can see significant illumination changes as well as a good variation in terms of candidate age, gender and skin tone variation. The more details on both the dataset are covered below explaining how challenging they are and encouraging more research using such datasets, as well as motivating new dataset collection

including all such variations together.

UBFC-Phys [34]: This dataset includes 56 candidates with a distribution of 10 males and 46 females. The video frame rate is 35 FPS and has a resolution of 1024×1024, BVP signals are also provided in this dataset which are collected at a rate of 64 Hz using the E4 wristband. Each candidate is subjected to 3 tasks, which are rest, speech, and arithmetic tasks, respectively, and a 3-minute RGB video is collected for each task. All the videos were captured in a lab environment, with natural movements incorporated into all tasks. Additionally, the dataset features natural translational and rotational movements, along with various attributes such as facial hair, glasses, skin color, and occlusion. The annotated labels in the UBFC-Phys dataset for tasks two and three are not suitable for training, and hence we have only considered the data from the first task, which still encompasses all the necessary variations in the video data.

COHFACE [12]: This dataset includes 40 candidates with a distribution of 28 males and 12 females. The videos were captured at a frame rate of 20 FPS and have a resolution of 640×480. The dataset also includes corresponding BVP signals collected at a rate of 256 Hz. Each candidate in the dataset was recorded for a duration of one minute, under two different illumination scenarios, i.e., good lighting and low light conditions. All the videos were collected in a lab environment. Though there are no intentional movements, the challenge in this dataset is brought by the low illumination samples. The data also include a considerable variation of skin tones which makes it even more useful.

4.2 Data Distribution and Training Details

Previously we have covered everything with respect to model details, training details and dataset in sections 3.3, 3.4 and 4.1 respectively. In this section we bring it all together and

Figure 4.1: Examples of the illumination variation covered in the COHFACE dataset. The figure also represents the data sample variation in terms of age, gender and skin tone variation. Here we also try to present how the illumination varies accross the face and especially in the cases with low illumination how there could be cases with almost partial illumination. This form of illumination variation when combined with other discrepancies mentioned previously makes the overall dataset tough to train on.

highlight how our exact data distribution strategy as well training is carried out. It is very important to understand that this is a time series domain based application area and hence needs consistent data handling is much needed. Also, we need to make sure that the testing

data is a representation of the whole dataset and shouldn't be based on a few candidates from the dataset. Hence generally approaches like cross validation are taken into consideration when testing such applications.

As the size of the dataset increases the processing time for approaches like cross validation also increases exponentially and hence is not considered the best approach to perform testing. Hence to address this problem with large datasets as well as to have a all inclusive test data we distribute our dataset such that for every 3 minutes candidate video 2 minutes of data is used for training and the rest of the 30 seconds each from the remaining 1 minute, are used for validation and testing respectively. In case of out model initially we tried using 36×36 as the input frame size for our model with bicubic interpolation, but that introduced a lot of blurring effect in the frames and insufficient features to perform better signal morphology learning. Hence we work with 72×72 as the input frame size along with linear interpolation which introduces the model to more exploitative face features and hence better learning.

Additionally to address the changes that we have introduced in particular to the original DeepPhys model, we start by more feature inputs by increasing the frame size, all the following layers in the architecture are change because of that. We also used batch standardization on the input frames which gave us better results, the number of output features is kept to just 32 features, we have also tuned the dropout throughout the model to a value of 0.5 and we used SGDM optimizer with momentum of 0.9, batch size of 128, and a learning rate of 10^{-4}. We have trained two model pipelines one with BVP as the ground truth and another with first derivative BVP as the ground truth as mentioned and explained earlier both of these models were trained and test on the UBFC Phys as well as COHFACE dataset.

4.3 Signal Morphology Recovery

This section reports results on the recovery of signal morphology. Here, we demonstrate our model's performance on the UBFC-Phys dataset quantatively using (3.17)-(3.19) as well as using a visual depiction as shown in Figure 4.2. Table 4.1 presents the morphology metric outcomes for our models with ground truth as the original BVP signal and first derivative BVP signal, along with recovered signals from our implementation of the DeepPhys model all without any form of the post-processing involved. Similarly, in Table 4.2, we present the morphology metrics on the integrated signals after post-processing for the first derivative ground truth BVP signal-based models. Based on the recovered signals we also show an aggregated representation of same in Figure 4.3, as a visual comparison of the expected signal outcome with the corresponding output signal from the model, along with its maximum and minimum deviation.

Table 4.1: Domain-wise morphology metrics outcomes for our model pipelines and for DeepPhys without any post-processing of the output signals.

Metrics	**Our Model (BVP)**	**Our Model (First Der. BVP)**	**DeepPhys**
Time↑	**0.088**	0.077	0.06
Frequency↑	0.443	**0.511**	0.359
Power↑	0.45	**0.47**	0.339

4.4 Standard Cardiac Pulse Metrics

The computation of the heart rate in beats per minute, is often performed by post-processing the model's output BVP signals using a bandpass filter. We have used a cutoff frequency of 0.75 Hz and 2.5 Hz (since the expected range for heart rate is 45 beats/min to 150

Table 4.2: Domain-wise morphology metrics outcomes for our BVP first derivative-based model pipeline and DeepPhys after integrating output signals to get them in original BVP signal format.

Metrics	**Our Model (First Der. BVP)**	**DeepPhys**
Time↑	**0.120**	0.118
Frequency↑	**0.670**	0.645
Power↑	**0.594**	0.472

beats/min). We next compute the power spectral representation of the band-passed signal where the highest peak is considered as the estimated HR. We report root mean square error (RMSE), mean absolute error (MSE), and Pearson correlation coefficient between the heart rate for the ground truth BVP signal and the estimated BVP signal. The results are shown in Table 4.3 for both UBFC-Phys and COHFACE.

Further, in Table 4.4 we present a comparison of the MAE-based outcomes for the average cardiac pulse-based measurements from our models with respect to the state-of-the-art models previously published.

Table 4.3: Performance of our architecture pipelines on UBFC-Phys and COHFACE dataset, in terms of heart rate measurements in beats per minute (HR bpm) [9, 11, 15, 40]. Comparisons have been made with literature using available metrics that are used in our study.

Methods	UBFC-Phys				COHFACE			
	MAE↓	RMSE↓	r↑	$\overline{SNR}$(dB)↑	MAE↓	RMSE↓	r ↑	$\overline{SNR}$(dB)↑
ICA [33]	6.71	-	-	-	12.24	15.67	0.24	-4.43
CHROM [7]	4.39	-	-	-	7.80	12.45	0.26	-
POS [47]	5.98	-	-	-	13.43	17.05	0.24	-4.43
HR-CNN [40]	-	-	-	-	8.10	10.78	0.29	-
DeepPhys [5]	11.78	17.848	0.174	-7.676	6.60	10.788	0.524	-6.425
Our Model (BVP)	5.02	10.673	0.701	-1.792	4.02	6.799	0.80	-6.317
Our Model (1^{st} Der. BVP)	**4.05**	**8.438**	**0.828**	**-0.78**	**2.92**	**6.128**	**0.86**	**-2.685**

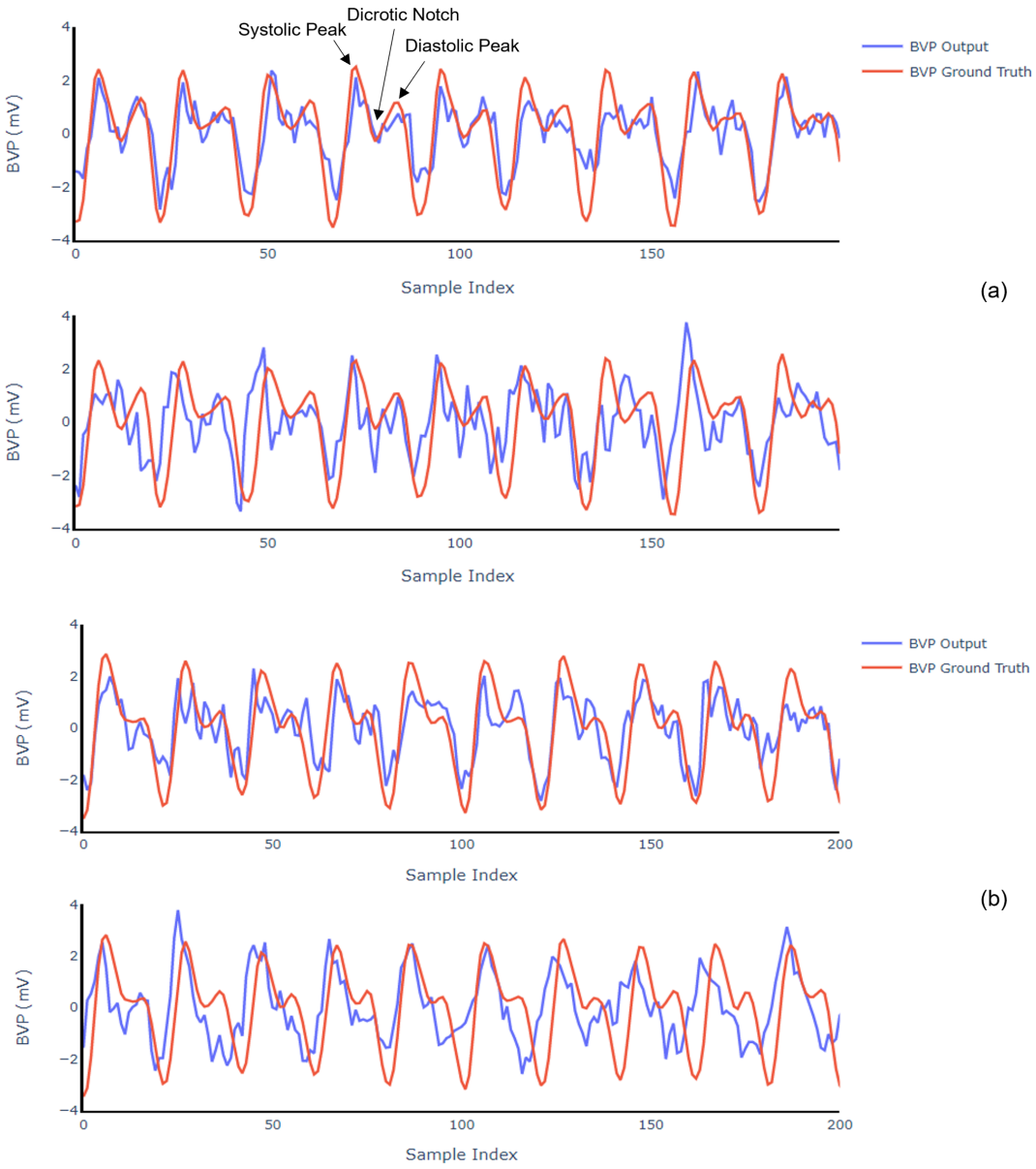

Figure 4.2: The two plots (a) and (b) are for two different individuals from the UBFC-Phys dataset, representing the morphology recovery from our model (top) and from DeepPhys [5] (bottom) in the respective images. This is a qualitative representation of how well our model retrieves signal morphology and its comparison with signal recovery from a state-of-the-art model that focuses on averaged pulse values.

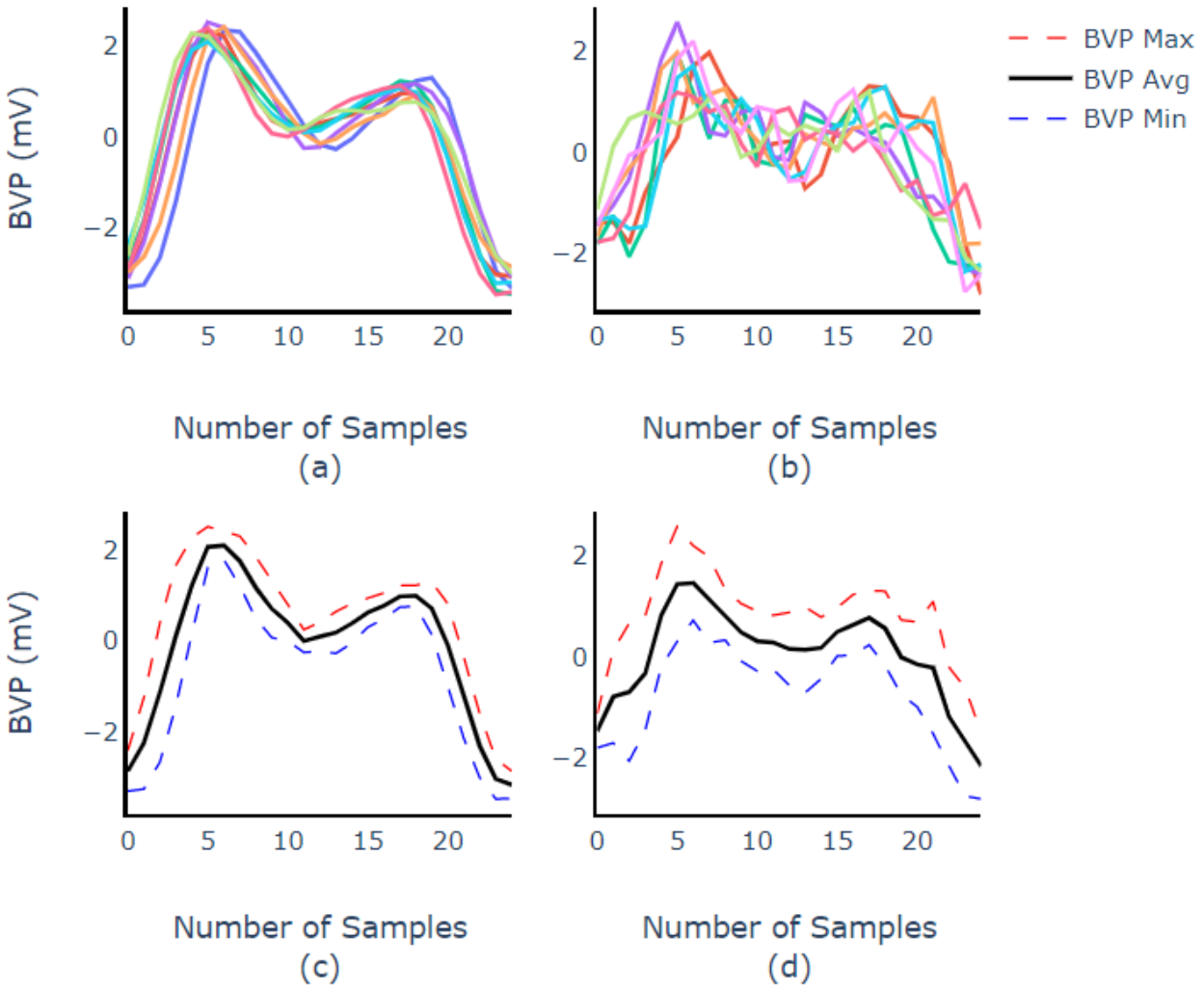

Figure 4.3: The figure represents the model's ground truth signal (left half (a), (c)) and its corresponding output from the model (right half (b), (d)) for the same candidate. The top half (a), (b) represents an aggregated signal whereas the bottom half (c), (d) represents the corresponding signal from the top half with its maximum deviation, minimum deviation and mean. The figure shows how the over all shape of the signal is retained as well as how both the systolic and diastolic peaks are recovered by our model. Though signal amplitude is a variable factor and is not as important as the morphology, the intention here is to check if the overall aggregated signal is not having large deviations for a constant amplification factor of the signal.

4.5 Re-identification

As a part of our experiments, our aim was to evaluate the possible scope and extent of re-identification using our devised architecture. Since we were focusing primarily on good

Table 4.4: Performance (HR BPM-MAE) of our technique in comparison with previously published state-of-the-art models on the UBFC-Phys and the COHFACE dataset [9, 11, 15, 40].

Methods	**UBFC Phys**	**COHFACE**
GREEN [45]	14.17	-
ICA [33]	6.71	12.24
CHROM [7]	4.39	7.8
POS [47]	5.98	13.43
1D-CNN [40]	5.41	-
LSTM-rPPG [4]	6.48	-
SQA-rPPG [9]	6.01	-
2SR [48]	-	20.98
LiCVPR [19]	-	19.98
HR-CNN [40]	-	8.10
SAMC [43]	-	6.23
DeepPhys [5]	11.78	6.6
Our BVP	5.02	4.02
Our BVP Der.	**4.05**	**2.92**

BVP signal shape retrieval and thereby performing computations based on the retrieved BVP signals, hence instead of evaluating authentication based on an Inter Beat Interval (IBI) or similar averaged cardiac pulse-based metrics which is the standard approach to go about authentication/re-identification, we computed the Pearson correlation coefficient between the ground truth and the output BVP signals for each candidate. So, the outcome for every candidate was compared with the annotated BVP signals for every other candidate in the test set and the one with the maximum correlation was acknowledged as the identified candidate for the respective output signal. Considering rank 5, we could re-identify 14 candidates from a pool of 20 candidates from a diverse dataset such as UBFC-Phys. The

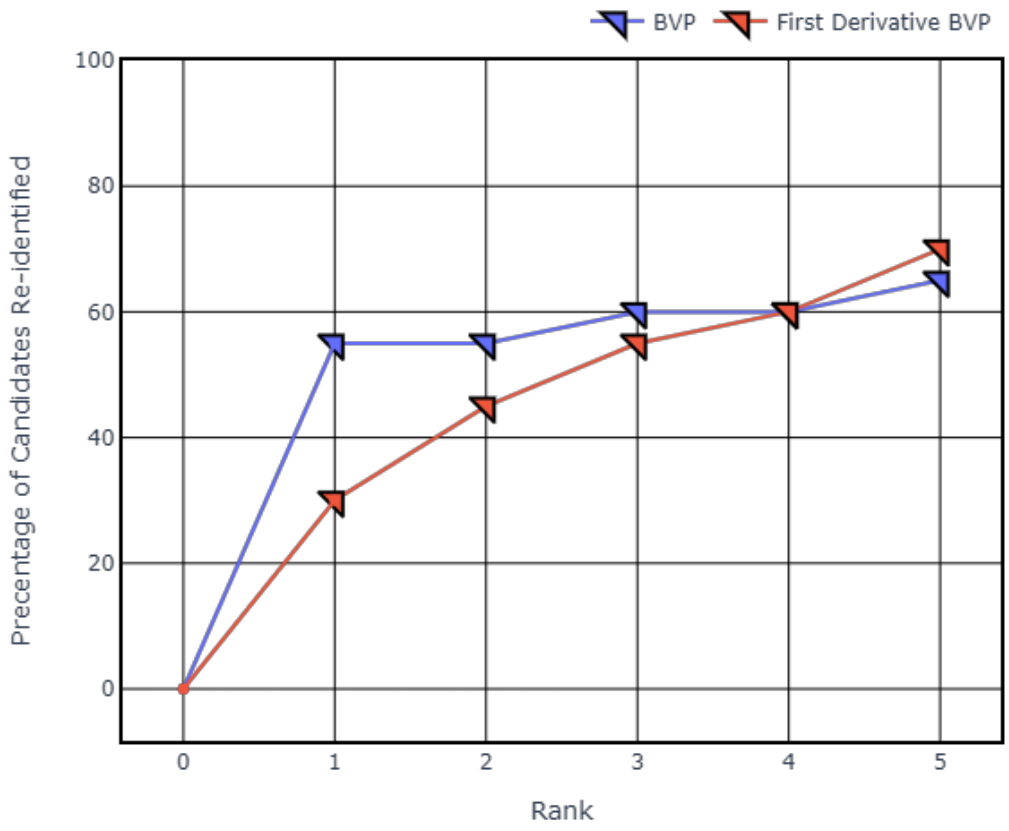

Figure 4.4: The rank-wise distribution for re-identification is presented here, where the graph in blue represents the re-identification results for the model trained using BVP signals, and the graph in red represents the re-identification results for the model trained using first derivative BVP signals. The variation in the rank improvement for the BVP as well as first derivative BVP based model is explained more in 4.5.

rank-5 accuracy was therefore 70%, which demonstrates the potential of this approach. The ratio of comparison for every candidate 1:20 and gives us an accuracy of 70%, thereby supporting our idea of considering the scope of re-identification just by using raw BVP signal outcomes for a small group of people. The rank-wise distribution is presented in Figure 4.4 where we show the rank-wise re-identification for both of our models including first derivative BVP signal outcomes as well as integrated BVP signal outcomes. As seen the Figure 4.4 the re-identification accuracy is quite high initially for the raw BVP annotated model as compared to the first derivative BVP annotated model, which is majorly because of the

similarity between raw BVP signal and an output sine wave kind of signal where the actual morphology in terms of systolic and diastolic peaks takes a step back and also because of the noise introduced through the first derivative signal computation. Hence as we move on with the ranks there is less significant improvement using the raw BVP based model and a significant growth is seen with the first derivative BVP based model.

Chapter 5

Conclusion

This thesis has presented an iPPG method that can extract BVP signals from standard RGB video of a person's face. The primary emphasis has been to recover the shape (morphology) of the BVP signal. We have shown that recovery of systolic and diastolic peaks is possible through camera-based iPPG. Using large-scale benchmark datasets and a series of metrics, we have demonstrated that our method performs better than previous state of the art methods to extract the BVP signal.

A better understanding of BVP will help iPPG research in many ways. First, there is no longer a need to place so much emphasis on recovering average heart rate only. Direct BVP signal recovery will help in studying inter-beat intervals with greater accuracy than is now possible. In turn, this work opens up the potential of iPPG in performing measurements related to heart rate variability. We have shown that better recovery of BVP signals significantly reduces error associated with other HR metrics. Finally, we also demonstrated that extracted BVP signals can be used for person reidentification.

Chapter 6

Discussion and Future Work

Based on our experiments we conclude that to retrieve better BVP signal shape and to have a better visibility of the systolic peak, diastolic peak and dicrotic notch it is import to train the model using the first derivative BVP signals which helps in better signal shape metric learning. There is a lot of value in focusing on the retrieval of physiological signal shape as it is the core to so many different averaged cardiac pulse metrics. As far as ROI selection for this domain of work is concerned, it is very important to consider a generalized approach and work on getting the models focus on the appropriate face regions, instead of naively choosing general regions of interest. Finally, its a high time that more work on physiological signal recovery be done instead of recovering just averaged cardiac pulse based metrics. At the same time, it is important to start considering cases like dense facial hair, reflection from glasses, extreme age as well as skin variation, heady-body movements and similarly, more study with respect to the effect of health factors on ones physiological signals needs to be done.

This thesis work could be further progressively extended by working with architectures like vision transformer. The current presented work has exhausted the maximum potential of this deep learning based model and has also included all the considerations for generalized variation in the dataset. Now with the future work the focus should be more on incorporating NLP and Computer Vision crossover ideas to progress the work in this domain. Since it is more of a time series problem, so just like any logical sentence formation problem where

every next word has a dependability on the words that occurred in the past, even here it has large dependability on the physiological signal status that was in the near past. Thus using architectures like vision transformer would rather help in better recovery of the physiological signals where the model would consider the data from previous frames as well as the time based variations in the annotated BVP signals. There is already some work in this direction covered in [51, 54] but with the advancements in recent Generative AI based models this could be a very good potential direction for future work as well recovery of good physiological signals.

Appendices

Appendix A

Coverage of Edge Cases and Effect of Skin Detection Model

In this section we dive in deep to study the effect on averaged cardiac pulse metric based results due to the use of skin detection model. We implement two different architecture pipelines, one with the inclusion of skin detection model and another without the inclusion of the same. This is a good approach to conduct this ablation study since it also helps us in proving the claims with respect to handling the different edge cases like facial hair, glair from glasses, etc. as well as help us understand the improvements brought in just by using the architectural additions without the skin detection model, as well as using the added improvements with the use of our skin detection architecture.

Table A.1: Performance of our architecture using averaged cardiac pulse metrics with and without the use of skin detection model. The values help us prove the effectiveness our architectural changes on the UBFC-Phys dataset with as well as without the inclusion of our skin detection model.

Metrics	With Skin Model		Without Skin Model	
	Our Model (BVP)	**Our Model (1^{st} Der. BVP)**	Our Model (BVP)	Our Model (1^{st} Der. BVP)
MAE (BPM)↓	5.02	**4.05**	5.2	4.297
RMSE (BPM)↓	10.673	**8.438**	10.624	8.623
r ↑	0.701	**0.828**	0.813	0.671
$\overline{SNR}$(dB)↑	-1.792	**-0.78**	-0.885	-1.623

Appendix B

Study of Attention Maps and Future Direction for ROI

In this part of the ablation study we have computed the attention maps based on our model for the UBFC-Phys dataset. The purpose of this effort is to study what are the regions of interest chosen by our model and further have our own conclusions for future ROI selection based study.

Based on the attention map outcomes seen in Figure B.1 and Figure B.2 it is clear that the model gives consistent outcomes and the additions as well as changes incorporated in the proposed architecture help in handling the vast set of variations encountered in datasets like UBFC Phys. As covered in literature skin area around cheek bone as well as forehead are good regions for physiological signal recovery. This is also verified through our attention map based outcomes but at the same time we can also conclude that for many cases the upper as well as lower cartilage area on the nose is also a very good source for physiological signal retrieval.

Based on this ablation study we can make sure that in future work if we consider face region wise ROI selection we can avoid considering areas below eyelids, lip region, as well as extreme edges on the face. Also as far as forehead is concerned central forehead as well as the area just above eyebrows seems like a really active zone for physiological signal recovery.

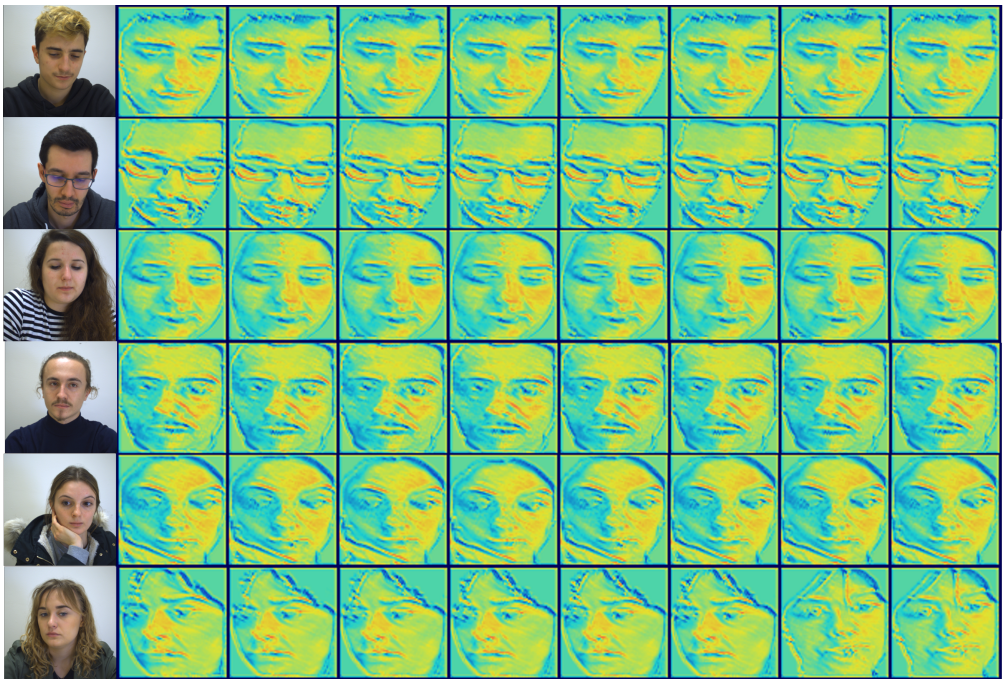

Figure B.1: Attention maps in cases with variation with hair color, facial hair and cases like partial occlusion. We can see how the model focuses on the correct ROI's as expected irrespective of the physical variations.

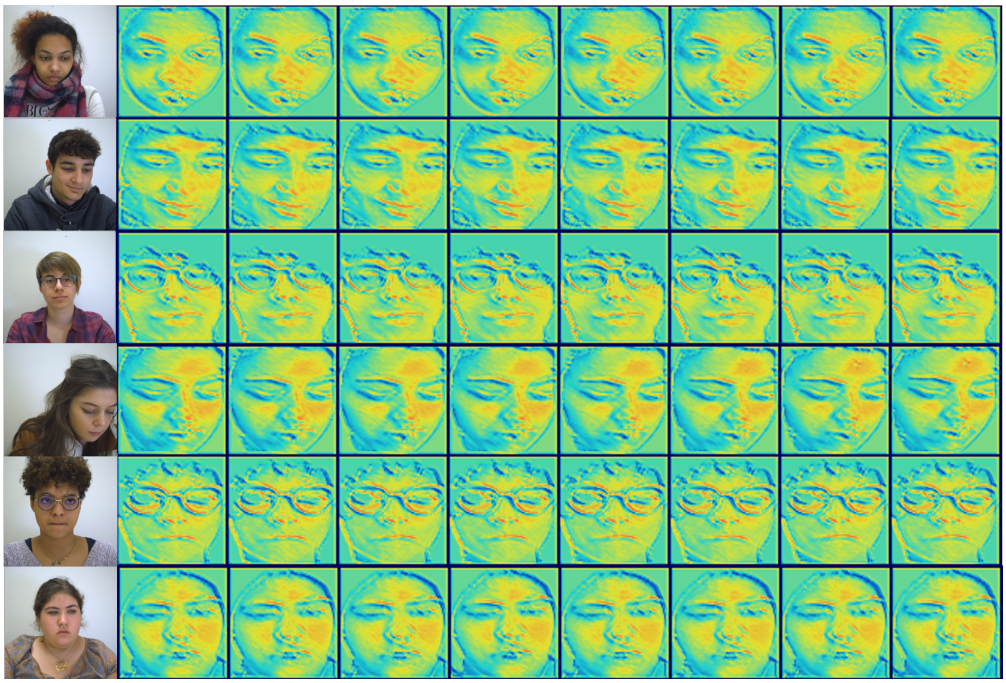

Figure B.2: Attention maps in cases with variation with skin color, glair reflection, physique variation and cases like off-frame area of interest (face in our case). We can see how the model incorporates all the variations and gives consistent attention maps even with variation in skin color or off-frame cases.

Bibliography

[1] AliveCor, Inc. *AliveCor for Physiological Measurements.* Web page: https://www.alivecor.com. Accessed January 2023.

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

[3] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro d'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. An open framework for remote-PPG methods and their assessment. *IEEE Access*, 8:216083–216103, 2020.

[4] Deivid Botina-Monsalve, Yannick Benezeth, Richard Macwan, Paul Pierrart, Federico Parra, Keisuke Nakamura, Randy Gomez, and Johel Miteran. Long short-term memory deep-filter in remote photoplethysmography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[5] 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.

[6] Joaquim Comas, Adria Ruiz, and Federico Sukno. Efficient remote photoplethysmography with temporal derivative modules and time-shift invariant loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2182–2191, 2022.

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

[8] Fitbit, Inc. *Fitbit Watch for Physiological Measurements*. Web page: https://www.fitbit.com. Accessed January 2023.

[9] Haoyuan Gao, Xiaopei Wu, Jidong Geng, and Yang Lv. Remote heart rate estimation by signal quality attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2122–2129, 2022.

[10] Marc Garbey, Nanfei Sun, Arcangelo Merla, and Ioannis Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Biomedical Engineering*, 54(8):1418–1426, 2007.

[11] 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.

[12] Guillaume Heusch, André Anjos, and Sébastien Marcel. A reproducible study on remote heart rate measurement. *arXiv preprint arXiv:1709.00962*, 2017.

[13] Brian L Hill, Xin Liu, and Daniel McDuff. Learning higher-order dynamics in video-based cardiac measurement. *arXiv preprint arXiv:2110.03690*, 2021.

[14] Rudi Hoekema, Gérard J.H. Uijen, and Adriaan Van Oosterom. Geometrical aspects of the interindividual variability of multilead ECG recordings. *IEEE Transactions on Biomedical Engineering*, 48(5):551–559, 2001.

[15] Min Hu, Dong Guo, Xiaohua Wang, Peng Ge, and Qian Chu. A novel spatial-temporal convolutional neural network for remote photoplethysmography. In *12th International*

Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–6. IEEE, 2019.

[16] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011.

[17] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 405–410. IEEE, 2011.

[18] Lin Li, Chao Chen, Lei Pan, Jun Zhang, and Yang Xiang. SoK: an overview of PPG's application in authentication. *arXiv preprint arXiv:2201.11291*, 2022.

[19] Xiaobai Li, Jie Chen, Guoying Zhao, and Matti Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4264–4271, 2014.

[20] G. Lin, T. Nakajima, P. Rahul, and A. Hodge. *Seamlessly embedded heart rate monitor*. U.S. Patent 8,615,290.

[21] Xin Liu, Brian L Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement. *arXiv preprint arXiv:2110.04447*, 2021.

[22] Xin Liu, Shwetak Patel, and Daniel McDuff. Camera-based physiological sensing: Challenges and future directions. *arXiv preprint arXiv:2110.13362*, 2021.

[23] Giulio Lovisotto, Henry Turner, Simon Eberz, and Ivan Martinovic. Seeing red: PPG biometrics using smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[24] Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9):1–40, 2023.

[25] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2957–2960, 2014.

[26] Seyedfakhreddin Nabavi and Sharmistha Bhadra. Design and development of a wristband for continuous vital signs monitoring of COVID-19 patients. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6845–6850. IEEE, 2021.

[27] 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.

[28] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019.

[29] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. Video-based remote physiological measurement via cross-verified feature disentangling. In *European Conference on Computer Vision*, pages 295–310. Springer, 2020.

[30] Omkar R. Patil, Wei Wang, Yang Gao, Wenyao Xu, and Zhanpeng Jin. A non-contact

PPG biometric system based on deep neural network. In *IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2018.

[31] Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005.

[32] Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1254–1262, 2018.

[33] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

[34] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. UBFC-Phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 2021.

[35] Abhijit Sarkar. *Cardiac signals: remote measurement and applications*. PhD thesis, Virginia Tech, 2017.

[36] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Assessment of psychophysiological characteristics using heart rate from naturalistic face video data. In *IEEE International Joint Conference on Biometrics*, pages 1–6, 2014.

[37] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. ECG biometric authentication using a dynamical model. In *IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015.

[38] Abhijit Sarkar, A. Lynn Abbott, and Zachary Doerzaph. Biometric authentication using photoplethysmography signals. In *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2016.

[39] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2011.

[40] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network. In *Proceedings of the British Machine Vision Conference, Newcastle, U.K.*, pages 3–6, 2018.

[41] 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.

[42] H Emrah Tasli, Amogh Gudi, and Marten Den Uyl. Remote PPG based vital sign measurement using adaptive facial regions. In *2014 IEEE international conference on image processing (ICIP)*, pages 1410–1414. IEEE, 2014.

[43] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2396–2404, 2016.

[44] Rik van Esch, Kambez Ebrahimkheil, Iris Cramer, Wenjin Wang, A. T. M. Kaandorp, Carla Kloeze, Cindy Verstappen, Tomas van't Veer, Marcel, Federica Sammali, and Dierick van Daele. Remote PPG for heart rate monitoring: lighting conditions and

camera shutter time. In *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2021.

[45] Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.

[46] Wenjin Wang and Albertus C. den Brinker. Camera-based respiration monitoring: Motion and PPG-based measurement. In *Contactless Vital Signs Monitoring*, pages 79–97. Elsevier, 2022.

[47] 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.

[48] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE Transactions on Biomedical Engineering*, 63(9):1974–1984, 2015.

[49] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):1–8, 2012.

[50] Umang Yadav, Sherif N. Abbas, and Dimitrios Hatzinakos. Evaluation of PPG biometrics for authentication in different states. In *2018 International Conference on Biometrics (ICB)*, pages 277–282, 2018.

[51] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 28:1290–1294, 2021.

[52] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38(6):50–58, 2021.

[53] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 151–160, 2019.

[54] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4186–4196, 2022.

[55] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[56] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.

[57] Yan Zhou, Panagiotis Tsiamyrtzis, Peggy Lindner, Ilya Timofeyev, and Ioannis Pavlidis. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5):1280–1289, 2012.