Towards Video-based Physiology Estimation

Ambareesh Revanur

CMU-RI-TR-22-71

December 2022

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
László A. Jeni, *Chair*
Artur W. Dubrawski
Conrad S. Tucker
Mononito Goswami
Carnegie Mellon University

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Robotics.

Keywords: RGB video, physiology, heart rate, transformer, blood pressure, vital signs

Abstract

RGB-video based human physiology estimation has a wide range of practical applications in telehealth, sports, and deepfake detection. Therefore, researchers in the community have collected several video datasets and have advanced new methods over the years. In this dissertation, I study these methods extensively and aim to address the following limitations. First, several methods have exclusively focused on addressing the problem of heart rate estimation; but other vital signs such as blood pressure have received little attention. Second, the datasets typically contain few subjects and record physiological data in resting conditions, which does not sufficiently represent patients with cardiovascular issues. Third, the evaluation protocol utilized by these methods does not account for fine-grained variations in physiological signals.

In this work, I introduce three new datasets each collected for a specific purpose. The *Vision-4-Vitals (V4V) dataset* was curated for heart rate estimation and respiration rate estimation using an emotion elicitation protocol to introduce variations in the physiological state under laboratory settings. We are currently collecting another large-scale dataset with over 2000 subjects at India and Sierra Leone sites for *multiple physiological* signal estimation including blood pressure and oxygen saturation. Furthermore, we are also in the process of collecting a new *dual-camera based blood pressure* dataset with a baseline measurement and modified blood pressure measurements through breathing activities. Finally, we introduce the *continuous evaluation protocol* and a new *Transformer-based approach* that outperforms existing heart rate and respiration rate estimation methods.

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. László A. Jeni for his guidance on this challenging research topic. This research would not have been possible without his constant support. I would also like to thank him for creating a strong research environment where new ideas and constructive criticisms are encouraged. I am also extremely grateful to Prof. Conrad Tucker from Carnegie Mellon University and Prof. Lijun Yin from Binghamton University for their valuable inputs on physiological data collection during the course of this research. I would also like to express my sincere gratitude to my friends and collaborators - Ananyananda Dasari, Xiaotian Li, Umur Ciftci, Zhihua Li, and CUBE lab members - for the several brainstorming sessions and discussions we had. I thank my thesis committee members, Prof. Artur Dubrawski and Mononito Goswami, for their input which has helped in shaping this dissertation. Finally, I would like to thank my parents, my brother, and my relatives for their encouragement and support through the years.

Disclosure: This research was in part supported through grants from Bill & Melinda Gates Foundation (BMGF) and National Science Foundation (NSF).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

According to the Centers for Disease Control and Prevention (CDC), heart diseases are the leading cause of death in the United States [121]. In fact, one person dies every 34 seconds in the United States due to cardiovascular disease. CDC estimates that heart diseases cost about $229 billion dollar each year in the US alone. Further, according to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death even globally with an estimated 17.9 million deaths each year which constitutes about 32% of all deaths worldwide [1]. In response to this, several researchers and clinicians are earnestly investigating the intricate composite nature of the human cardiovascular system and are pioneering innovative solutions to these urgent problems.

In the last decade, healthcare systems have started to hugely benefit from the progress of technological developments in related research fields. Miniaturization of sensors has allowed the development of smartwatches that are able to track vital signs such as heart rate, skin temperature, and oxygen saturation among others [70]. These technologies have been very promising and have also brought new wearable technologies [43] such as thoracic pressure sensors, multi-point temperature sensors, and shoe-embedded force sensors along with them [55]. Personalized healthcare is gaining popularity as a means of preventing the occurrence of diseases rather than reacting to the disease after it has occurred. Some recent work in this direction has explored how gene expression could be used to assess the potential occurrence of coronary artery disease and arrhythmia [109].

Owing to the large quantities of collected healthcare in-patient and out-patient data, Artificial Intelligence (AI) has started to take center-stage in healthcare and is now further fueled by the recent outbreak of the COVID-19 pandemic, which has already claimed more than 6 million lives at the time of this writing [11]. AI has a wide and deep impact in several areas of healthcare research including pandemic forecasting [93], precision medicine [128], genome analysis [16, 87, 92] and several others. More recently, AI has been used to understand protein structures [46, 47], discover novel drugs [37, 41], detect diabetes [98] and prevent cardiovascular diseases [103]. There is no doubt that AI methods have been long studied in the medical context. However, a particularly interesting sub-area of AI that has been gaining increasingly wider adoption in

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

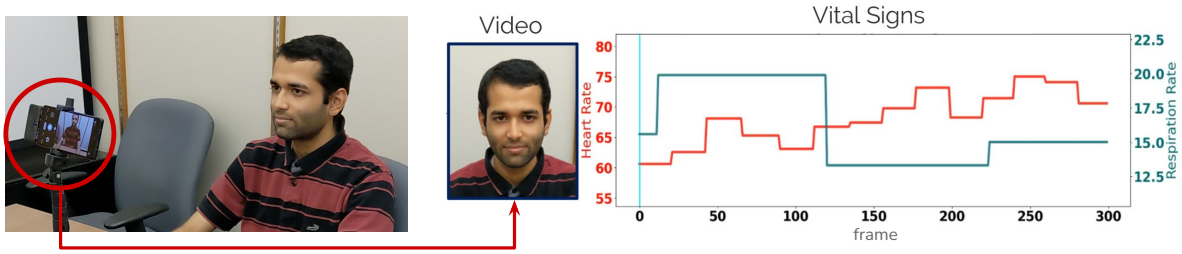

Figure 1.1: The goal of video-based human physiology estimation is to extract human physiological signals such as heart rate and respiration rate from an RGB video. As depicted in the image, the subject is seated in front of a smartphone device that is recording facial video using the rear camera. This video is used to estimate the vital signs of the subject.

healthcare is computer vision.

Computer vision has been studied for several decades. However, with the advent of deep learning, rapid advancements are now being made in several research areas such as self-supervised learning [9, 34], Transformer architecture [18, 136] , and NeRFs [75, 127]. These advancements have led to the development of many novel approaches in medical imaging and surgical robotics [129]. Several computer-vision based methods have been also proposed in the areas of chest radiography [51, 94, 119], mammography [3, 100], cardiac MRI scans [72], lesion detection [22, 25, 122], diabetic retinopathy [33] and breast cancer detection [71].

Motivated by certain advancements in computer vision, researchers are tackling the intriguing problem of human physiology estimation using RGB videos alone [10, 95, 96, 113, 118]. More specifically, the goal of these works is to extract human physiological signals such as heart rate, respiration rate, oxygen saturation, and other vital signs such as blood pressure from an RGB video of a person. At a high level, physiological extraction from RGB videos is achieved by analyzing subtle color and motion variations in a facial video. Although these subtle changes are not obvious to the naked eye, computer vision techniques can assist in amplifying and analyzing these subtle signals by exploiting the intrinsic periodic nature of physiological activities. In fact, the idea of analyzing color variations from volumetric blood changes in arterial sites such as fingers for physiological extraction has been well studied in Photophlythesmography (PPG). The key principle of PPG is that the light emitted from a source is reflected back to the receiver after it interacts with blood-rich skin tissues containing valuable vital-signs information. Therefore video-based physiology estimation methods are sometimes referred to as remote PPG (rPPG) or imaging PPG (iPPG). Figure 1.1 demonstrates a simple experimental setup for video-based physiology estimation where the subject is seated directly opposite a smartphone camera. We present closely related work in greater detail in Chapter 2, where we explain how computer vision based techniques reliably extract physiological information from the facial video.

In this chapter, I introduce the motivation for our work in Section 1.1. After that, I briefly present some of the challenges in this research area that I aim to address in Section 1.2. I also briefly summarize the contributions of this research in Section 1.3. Finally, I delineate the outline of this report in Section 1.4.

1.1 Motivation

Extraction of human physiological signals from RGB video has several attractive benefits and applications as listed below.

- **Ubiquitous sensors**

  Several individuals readily have access to an RGB camera owing to the ubiquity of affordable smartphones. Therefore, video-based physiological estimation methods can be deployed widely to enable large-scale disease screening and triaging.

- **Telehealth**

  It is difficult to triage the disease conditions of patients in a Telehealth setting [123] since healthcare professionals do not have access to high-quality physiology information that would otherwise be easy to measure in traditional in-person healthcare settings. Therefore, video-based physiological extraction could prove to be a great diagnostic tool in Telehealth. In particular, this technology would aid healthcare systems in developing nations and remote areas where the existing healthcare facilities are limited.

- **Deepfake detection**

  Deepfake videos are typically generated using Generative Adversarial Networks (GANs) [32]. Although GANs generate realistic videos that are hard to distinguish from real-world videos, these videos lack the subtle signature of physiological activities that are found in real-world human videos [12]. Hence video-based physiology extraction could be used in deepfake detection.

- **Affective computing**

  The elicitation of emotions is known to positively correlate with human physiological activity. For example, a study has shown found that emotions associated with negative valence are also associated with higher breathing rate [116]. Therefore, video-based physiology estimation methods could potentially improve affective computing.

- **Other benefits**

  Beyond the above applications, human physiological extraction could be used in passive sports and can assist with continuous monitoring of patients in a *non-contact* fashion.

It is also important to note that the field of video-based physiological estimation is closely related to several computer vision research areas (e.g., head pose estimation, network architectures, and object detection). Therefore, video-based physiological estimation methods directly benefit from new developments and improvements in computer vision.

1.2 Challenges

In this section, I list some of the challenges that exist in the field of video-based human physiology estimation and hope to address them.

- **Datasets** The datasets that have been proposed for the task of RGB-human physiological extraction have often been limited to the subjects with stable physiological experimental settings and have a limited number of subjects (e.g. PFF [39]). Additionally, some of the datasets also lack skin tone diversity [79]. Further, there have been no datasets accepted by the community for benchmarking new methods on the task of video-based blood pressure estimation.
- **Benchmarking and evaluation**

  Existing evaluation protocols only consider summary statistics of the predicted waveform. In other words, these works aim to predict a single heart rate or respiration rate value for a long input video clip (e.g. 60s). This cannot give a complete picture of the quality of the predicted waveform. Further, this evaluation protocol is unable to account for fine-grained variations in physiological signals which are valuable in detecting cardiovascular abnormalities such as Arrhythmia.
- **Limited study on certain vital signs**

  Although the research community has made significant progress in the extraction of heart rate and respiration rate, other vital signs such as blood pressure and oxygen saturation which are clinically valuable have remained to be an open challenge.

1.3 Contributions

With an aim to address the challenges listed in Section 1.2, I introduce the following contributions:

- **Vision-4-Vitals (V4V) dataset**: The V4V dataset is curated with an aim to collect a relatively larger and more diverse dataset with higher physiological variations artificially induced with emotion elicitation. This dataset is publically released and includes vital signs that are valuable for heart rate estimation and respiration rate estimation.
- **Exploiting temporal dimension with Transformers**: Many of the previous methods rely on feature extractors with limited effective input temporal window. We hypothesize that exploiting a larger effective input window for feature extraction is valuable in reducing noise in the extraction of the rPPG signal and model fine-grained variations.
- **Blood pressure estimation**: In this work, I present a dataset collection setup for the extraction of Systolic and Diastolic blood pressure values. This is designed to exploit cameras mounted on a typical smartphone for recording videos at two arterial sites.
- **Large-scale India-Sierra Leone dataset**: As an ongoing work, we are in the process of collecting a large-scale dataset that contains over 2000 subjects including vital signs such as heart rate, respiration rate, blood pressure, oxygen saturation, and skin temperature.

1.4 Outline

This report is organized as follows:

- Chapter 2: Background. This chapter reviews the major principles and ideas of rPPG. I review certain ideas for the extraction of heart rate, respiration rate, and blood pressure.
- Chapter 3: The first V4V dataset and challenge. This chapter presents the proposed dataset and an evaluation protocol that was employed in the first V4V challenge.
- Chapter 4: Transformer-based physiological estimation. This chapter presents our proposed transformer-based approach for the task of heart rate and respiration rate estimation.
- Chapter 5: Novel vital sign data collection. In this chapter, we present a smartphone-based data collection setup for novel vital signs such as blood pressure, oxygen saturation, and skin temperature.

This work is based on the following articles:

- Revanur et. al. "The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision workshops (2021).
- Revanur et. al. "Instantaneous physiological estimation using video transformers" Springer Studies in Computational Intelligence, (2022).

Chapter 2

Background

In this chapter, we explore the principles and methods that have been proposed for the task of video-based estimation of heart rate and respiration rate. Some of these principles are also utilized for video-based blood pressure estimation. The earliest works in the area of physiological extraction studied the absorption spectrum of Haemoglobin (Hg) and showed that a typical RGB video can be used to approximate the underlying physiological signals [113]. Later in 2012, the Eulerian magnification of facial video demonstrated amplification of the underlying physiological blood volume signal [124] that can be perceived visually. These works have inspired several physiological estimation works which led to traditional and deep learning based methods.

The gold standard for measuring heart rate is Electrocardiogram (ECG). It records the electric potentials of heart muscle fibers during each beat cycle. Recently, ECG has started to become more widely accessible due to the miniaturization of ECG sensors on smartwatches [7]. A more popular way of measuring the heart rate is by using a finger PPG device. As described previously, the PPG device analyzes the reflection or transmission of light rays [89] that pass through underlying blood-rich skin tissues. For respiration rate measurement, healthcare professionals typically use a chest belt. The belt measures the contractions and expansions of the chest resulting from the respiratory process.

The gold standard for measuring blood pressure is the Arterial catheter which is generally used in perioperative and emergency medicine. The catheter is generally inserted into the peripheral artery to measure the mechanical motion of blood. However, cuff-based instruments developed on the principles of oscillometry are more widely used to measure blood pressure outside perioperative care.

2.1 Heart Rate and Respiration Rate

We describe the estimation of video-based heart rate and respiration rate estimation in this section. We begin with traditional methods that were developed for these tasks and then describe more recent deep-learning methods.

2.1.1 Green channel

In this work, the authors empirically show that the green wavelength [114] contains the highest relative strength when compared with red and blue wavelengths for the task of physiology extraction. This is the result of two different observations. On one hand, the molar extinction coefficient of oxygenated Haemoglobin decreases as the wavelength increases (from $B \rightarrow G \rightarrow R$). On the other hand, the penetration strength of the light through the skin segment increases as the wavelength increases (in the range of visible light). Due to this trade-off, wavelengths corresponding to the green color have the highest relative strength. Therefore, the authors extract the signal from a predetermined patch of skin by spatially averaging the pixel values within the green channel of RGB video. Once the signal is extracted, it can be detrended and filtered using a bandpass filter to remove the frequencies outside the range of heart rate (0.7Hz - 4Hz; which translates to 42bpm-240bpm). The power spectrum analysis of the extracted signal reveals peaks that correspond to the heart rate and respiration rate (and their harmonics).

However, it should be noted that the red and blue channels of RGB video have physiological information, albeit at a lower relative strength. This observation has led to additional works that study how these channels can be linearly combined to extract physiological signals. In particular, [114], demonstrates that other channels - blue and red channels have a higher strength of respiratory signal information.

2.1.2 ICA

Independent Component Analysis (ICA) has been well studied in the context of cocktail problem [108] to separate out the speech signals from an audio track consisting of multiple speakers. This idea of blind source separation has been adapted to RGB video-based physiological extraction [85] in a similar setup, where the goal is to extract the physiological signal from three signals of the recorded RGB video - red, green, and blue channels. Unlike the green method described above, ICA-based physiological extraction has been empirically shown to be robust to motion. In this study, the authors conducted the experiments on a video dataset consisting of 12 subjects recorded in predetermined lighting conditions. The subject pool included male and female genders, and also had a sufficient representation of different skin tones.

2.1.3 POS

Plane Orthogonal to Skin (POS) [118] is a principled heart rate estimation method that was carefully designed based on the skin reflection model which explains how light interacts with the skin. One of the well-known models utilized for the same is Shafer's dichromatic reflection model. According to this model, the light incident on the skin surface can be decomposed into three main components - specular reflection, transmission, and diffuse reflection. The specular reflection is a light reflection that bounces back to the camera and does not contain color-variation information for physiological signals. Diffuse reflection is the component of light that is reflected back to the camera after passing through blood-rich layers of skin. We will again revisit the Skin

reflection model in Chapters 3 and 4. The benchmark dataset used for this study contains 60 videos recorded at 20fps; with ECG as the ground-truth signal.

2.1.4 PhysNet

With the advent of deep learning, several computer vision researchers have focused on improving heart rate estimation further by using deep learning architectures and loss formulations. In PhysNet [130], the authors propose a 3D convolution neural network that operates on 64 frames as input and predicts the rPPG signal directly. They use a loss formulation that maximizes the correlation of the ground truth waveform with the predicted waveform. Experimental results suggest that PhysNet is able to outperform traditional methods across multiple evaluation metrics for the task of heart rate estimation.

2.1.5 DeepPhys and MTTS-CAN

DeepPhys [10] was trained to predict the first derivative signal of the underlying physiological signal. The model is trained on normalized frame differences and input frames using the motion branch and appearance branch respectively. The goal of the motion branch is to extract the physiological signal while the appearance branch aims to predict an attention map that assists in the determination of the region of the image that would likely contain the physiological signal.

A mean squared loss between the ground-truth first-derivative signal and the predicted first derivative signal is employed to train the model. During the inference phase, these are the steps followed by DeepPhys to estimate the heart rate. First, the predicted signal is integrated (cumulative sum) to obtain the rPPG signal. Next, a detrending algorithm is applied to detrend the signal. Next, a bandpass filter is applied to rPPG signal to eliminate frequencies outside the frequency range of heart rate or respiration rate. Finally, the frequency corresponding to the peak value from power spectral density is predicted as the heart rate or respiration rate by DeepPhys.

MTTS-CAN [64] is similar in architecture and loss formulation to DeepPhys. However, there are two main differences. First, the MTTS-CAN model is trained to simultaneously predict heart rate and respiration rate using the same model in a multi-task learning setup. Second, the MTTS-CAN model embeds a temporal learning module called Temporal Shift Module (TSM) [59] which helps with learning over a short temporal window without adding additional trainable parameters to the network. Both DeepPhys and MTTS-CAN were able to achieve lower mean prediction errors than traditional heart rate and respiration rate estimation methods. At the time of publication, MTTS-CAN held a state-of-the-art performance on multiple datasets.

2.2 Blood Pressure estimation

In this section, we present some of the works that are closely related to cuffless blood pressure extraction. In Table 2.1, we summarize the related works and list **(1)** the input signals used by

Table 2.1: Summary of recent cuffless BP estimation methods sorted chronologically. Refer to Table 6.1 for notations. Here PTT stands for "Pulse Transit Time" and PPGm stands for "Photoplethysmography waveform morphology".

Paper	Input signal	Methodology	Model	Metric
Yang et al. [126] (2021)	Finger-PPG, ECG	PPGm, PTT	LiR, RF, ANN	ME, SD, MAD
Hsu et al. [40] (2020)	PPG	PPGm	ANN	MAE, RMSE
Lin et al. [63] (2020)	PPG	PPGm	n/a	p-value
Schoettker et al. [101] (2020)	Finger-PPG	PPGm	LiM	ME, STD
Luo et al. [68] (2019)	Face-video	PPGm	ANN	ME, SD, r
Lee et al. [54] (2019)	ECG, PPG	PTT	LiM	r, ME, SD, MAD
Lin et al. [62] (2018)	ECG, PPG	PPGm, PTT	LiR	MAE, STD
Wang et al. [117] (2018)	Finger-video SCG, APG	PTT	LiM	r, RMSE
Dey et al. [17] (2018)	PPG	PPGm	LasR	MAE
Kachuee et al. [48] (2017)	ECG, PPG	PPGm, PTT	LiR, DT, SVM	STD, MAE, r.
Watanabe et al. [120] (2017)	PPG	PPGm	n/r	MAD, r
Gao et al. [28] (2016)	Finger-video	PPGm	SVM	MAE, STD
Gaurav et al. [29] (2016)	PPG	PPGm	ANN	ME, STD, MAE
Datta et al. [14] (2016)	PPG	PPGm	Latent parameterization	MAE, STD, MAPD, r
Jeong et al. [45] (2016)	Face Video, Finger-video	PTT	LiM	r
Peng et al. [83] (2015)	Heart-sound	Spectral analysis	SVM	MAE, ME, STD, r
Banerjee et al. [4] (2015)	Finger-video	PPGm	ANN	n/r
Visvanathan et al. [115] (2013)	Finger-video	PPGm	LiR, SVM	Acc
Poon et al. [86] (2006)	ECG, PPG	PTT	Inverse-square	MAD, STD

the model for estimating the blood pressure, **(2)** the principle or the methodology adopted (such as Pulse Transit Time or Photophlethysmography waveform morphology) which we describe in this section, the **(3)** mathematical model utilized by their method and **(4)** the metric they used to benchmark the performance of their approach. While early methods have utilized simpler models such as linear regression and SVMs for blood pressure estimation using few handcrafted features, more recent methods have developed extensive feature sets and also make use of deep neural networks to predict blood pressure. Unfortunately, these methods have been tested on private datasets collected by respective authors. Therefore, in order to enable comparisons among different approaches, we propose a data collection setup in Chapter 5.

Traditionally blood pressure is measured using a cuff-based sphygmomanometer which is developed based on the principles of oscillometry [26]. Recently, many works have explored the estimation of blood pressure without using a cuff thereby allowing for a more comfortable experience [77]. Although there have been other works that have demonstrated utilization of ultrasound [132] and volume control approaches [27], we primarily limit our focus on these three groups of work: Pulse Transit Time (PTT) based, Photoplethysmography waveform (PPG) based and video-based blood pressure estimation. However, it should be video-based blood pressure estimation methods are directly drawn from the principles of PTT-based methods and PPG-based methods.

2.2.1 Pulse Transit Time

Pulse Transit Time (PTT) is a strong indicator of blood pressure. The mechanical model of blood vessels clearly shows that there is an inverse relationship between PTT and blood pressure [40, 42, 110]. Based on this observation, several methods have empirically verified the existence of such a relationship between PTT and BP. In [69, 73], the author utilized a simple linear model of the form $P = a \cdot \frac{1}{PTT} + b$ and showed that the model is able to estimate blood pressure with a correlation coefficient of 0.89 with systolic blood pressure. On similar lines, [90] have utilized an inverse-logarithmic relationship between PTT and blood pressure of the form $P = -a \cdot \ln(PTT) + b$ which had a correlation coefficient of 0.82 with systolic blood pressure. As tabulated in Table 2.1, methods have relied on a variety of signals such as finger PPG, ear PPG, or video-based rPPG signal for computing Pulse Transit Time (PTT) [5, 54, 117, 126]. In [45], the authors extract a video-based PPG signal (rPPG) at facial and palm locations to compute PTT using a high-frequency camera. Another related study showed that the PTT values are typically as low as 30ms and a high-frequency camera (120 fps) could mitigate error due to signal discretization [104].

2.2.2 PPG morphology

Several works have demonstrated that the signal morphology is directly influenced by blood pressure [20, 38]. Therefore, in principle, it should be possible to extract blood pressure from the PPG signal alone (without any PTT-related information). Inspired by this idea, several methods have studied features of PPG morphology that are influenced by blood pressure such as temporal

features [1, 40, 48, 58], the area features [19, 40, 58], signal amplitude features [48, 62, 91] and derivative features [40]. The authors of these methods empirically test their models and have demonstrated high accuracy on different datasets. However, these methods generally utilize a PPG device and have access to a higher-quality signal in comparison to the signal extracted from a video. As listed in Table 2.1, video-based methods tend to use the morphological information present in the PPG signal to compute blood pressure. In [68], authors recorded a large dataset consisting of over 1326 people and extracted morphological features such as PPG, HR, and wave energy using the facial video alone. Unlike this dataset which consists mostly consists of normotensive videos, in Chapter 5 we present our dataset collection which is aimed to record facial videos of subjects with BP interventions to artificially modify blood pressure. In [102], the authors used the PPG signals obtained from the face and used it to fine-tune an LSTM-based model. This method also demonstrated the high generalizability of the model to new unseen subjects. However, while deep learning based blood pressure estimation methods have started to gain popularity [21, 23, 106], these methods are compute-hungry and often cannot operate on smartphones due to limited computing budget.

Some works that estimate blood pressure through PPG morphological features also benefit from PTT correlation [48, 63]. In [68], the authors utilized PTT as one of the features and trained models such as Linear Regression (LR) and Artificial neural network (ANN). However this work computes PTT between two confined subregions of the face which is prone to higher approximation error.

Chapter 3

Vision-4-Vitals (V4V) dataset and challenge

Telehealth has the potential to offset the high demand for help during public health emergencies, such as the COVID-19 pandemic. Remote Photoplethysmography (rPPG) - the problem of non-invasively estimating blood volume variations in the microvascular tissue from video - would be well suited for these situations. Over the past few years a number of research groups have made rapid advances in remote PPG methods for estimating heart rate from digital video and obtained impressive results. How these various methods compare in naturalistic conditions, where spontaneous behavior, facial expressions, and illumination changes are present, is relatively unknown.

To enable comparisons among alternative methods, the 1st Vision for Vitals Challenge (V4V) presented a novel dataset containing high-resolution videos time-locked with varied physiological signals from a diverse population. In this chapter, we outline the evaluation protocol, the data used, and the results. V4V challenge has been held in conjunction with the 2021 International Conference on Computer Vision. The dataset can be downloaded from here [1].

3.1 Introduction

There has been a tremendous interest in the extraction of human physiological signals using just facial videos. Computer vision based physiological extraction has been gaining momentum steadily because this technology has significant benefits over traditional contact-based measurements. Firstly, these methods allow reliable estimation of Heart Rate (HR) and Respiration Rate (RR) in absence of specialized equipment such as the electrocardiogram (ECG). These methods depend only on the video feed recorded from a general RGB camera readily available in a commodity smartphone. Secondly, these methods operate without any contact with the subject. Hence, video-based physiology estimation promotes social distancing and is more patient-friendly than contact-based devices. Thirdly, these methods aid in the remote diagnosis of patients located in remote areas where quality healthcare facilities are limited. These methods have a wide range of applications including telehealth, deep fake detection, affective computing, human behavior understanding, and sports.

[1] `https://vision4vitals.github.io/`

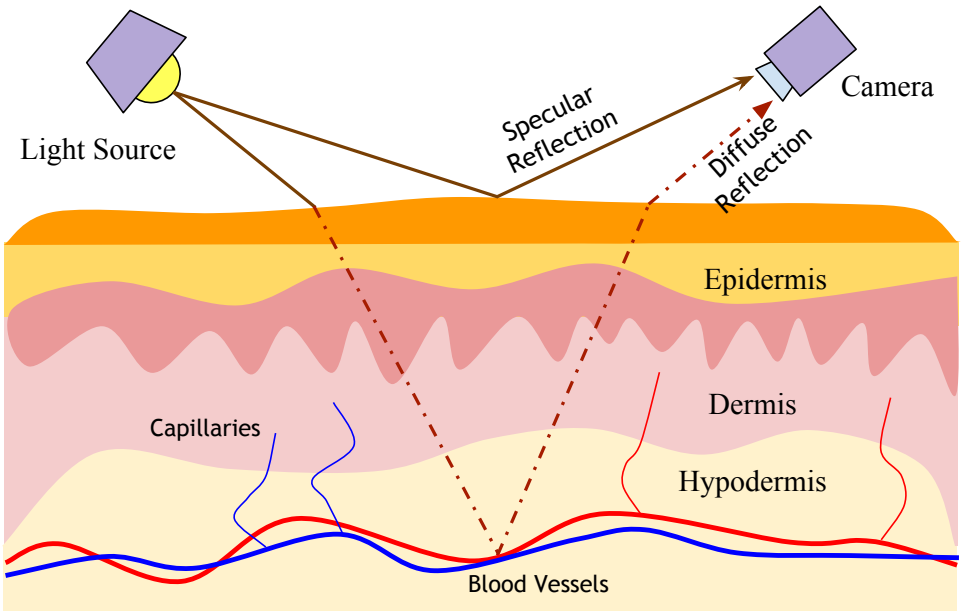

Figure 3.1: Skin reflection model [118]. The light reflected back to the camera consists of "specular reflection" and "diffuse reflection".

Clinicians use FDA-approved devices such as an electrocardiogram (ECG), a chest belt, and a photoplethysmography (PPG) device for extracting human physiology signals such as HR and RR. Since PPG is closely related to the V4V challenge, we describe the functioning of a PPG device. It is a contact-based device capable of extracting the subtle imperceptible color changes induced as a result of periodic changes in the volume of blood flowing in the underlying skin tissues. In a simplified setup, it consists of a light emitter and receiver. While the emitter is used to focus the light beam on the skin tissue, the receiver is used to record the intensity of light transmitted back to the PPG device. It is known that the absorption spectrum of (oxy-) hemoglobin lies in the color band corresponding to green [113]. Accordingly, the emitter and receiver are designed to capture the periodic color variations in the frequency range of heart rate. Studies have also shown that the respiration rate can be extracted either through motion analysis [44] or using the PPG signals [111]. However, a PPG device is a contact-based method and does not offer the attractive benefits offered by a non-contact-based method. To this end, several video-based non-contact remote physiology estimation methods have been advanced.

The video-based physiology estimation methods [15, 85, 113, 118] exploit the reflectance properties of the skin (typically facial region) with an aim to extract the human physiological signals. Often, the skin tissue is modeled under Shafers Dichromatic Reflection Model (DRM) [118] that provides a way to model the behavior of the light energy incident on surfaces. As shown in Fig. 3.1, the light incident on the skin tissue reflects back to the camera as two components - specular and diffuse reflectance. A fraction of the incident light energy that is reflected right off the skin surface is the specular component. This appears as a glossy/shiny reflection on the image captured using the camera. Diffuse reflectance is the light component that passes through the blood-rich tissues under the skin and is then transmitted out. Therefore, the diffuse component contains the signature of physiological signals, while the specular component does

not. Similar to PPG, a careful analysis of the variations in the diffuse component of the reflected light shows a pulsatile signal in the frequency range of heart rate. Therefore, video-based physiological measurement techniques are often referred to as Remote-PPG (rPPG) methods.

Several methods have been advanced for the extraction of rPPG signals [10, 15, 52, 65, 67, 80, 85, 113, 118]. Owing to the advancements in computer vision and deep learning, remarkable results have been achieved on the task of human physiology estimation. However, there are two drawbacks of existing methods. First, it is not clear how these various methods compare in naturalistic conditions, where spontaneous movements, facial expressions, and illumination changes are present. Second, most previous benchmarking efforts focused on posed situations. No commonly accepted evaluation protocol exists for estimating vital signs in spontaneous behavior with which to compare them. Therefore, in the 1st Vision-For-Vitals 2021 challenge, we introduce a new dataset called *Vision-For-Vitals (V4V) dataset* which includes challenging elements such as spontaneous behavior. We also contribute a new evaluation metric for stronger benchmarking on the *V4V dataset*.

3.2 Related works

Table 3.1 summarizes the related challenges and most recent methods in the area of remote video-based human physiological estimation.

Related challenges. In conjunction with CVPR'20, the 1st Remote Physiological Signal Sensing (RePSS) challenge was organized for the estimation of heart rate using RGB videos. The challenge consisted of a training dataset that was drawn from VIPL-HR-v2 [79] and a test set that was drawn from VIPL-HR-v2 [79] and the OBF [56] datasets. There were about 2500 samples, each 10 seconds long, in the training dataset. In order to test the efficacy of different methods employed in the challenge, the organizers used a segment-level evaluation protocol, i.e., one heart rate prediction for a 10s segment.

The 2nd Remote Physiological Signal Sensing (RePSS) challenge was held recently[2], and it included samples drawn from VIPL-HR-V2 and OBF. In this challenge, the organizers introduced an inter-beat-interval (IBI) based metric for measuring the performance of the participants' methods. Unlike the segment-level metric used in the 1st RePSS, the IBI-based metric is more granular as it penalizes missed/extra heartbeat predictions. In the 2nd RePSS challenge, the organizers also included respiration rate as a separate challenge track. Our V4V challenge is similar in spirit to the 2nd RePSS as we used granular evaluation metrics and have both tracks (HR and RR). Further, we introduce a newly curated dataset that contains challenging elements such as spontaneous behavior and varied physiological signals as part of our V4V challenge.

Methods. Traditional rPPG extraction methods typically involve two stages. The signal is first extracted based on the rPPG principles, and then the signal processing methods are used to compute the HR and RR. In [113], the pixels in the green channel are used to extract the physiological signals since it contains a strong signature of the pulsating rPPG signal. Some methods such as

[2]in conjunction with ICCV'21

Table 3.1: Summary of previous challenges and recent methods in the area of video-based non-contact physiology estimation. Many of the previous evaluation protocols use non-overlapping segments. In V4V 2021, we use continuous frame-level error measurement metrics.

	HR	RR	Datasets used	Evaluation Protocol
Related Challenges				
The 1st Remote Physiological Signal Sensing (CVPR'20) [57]	✓	✗	VIPL-HR-v2, OBF	10s non-overlapping segment
The 2nd Remote Physiological Signal Sensing (ICCV'21)	✓	✓	VIPL-HR-v2, OBF	Continuous metric: Inter-beat interval
Recent Methods				
MTTS-CAN (NeurIPS'20) [65]	✓	✓	AFRL, MMSE-HR	30s non-overlapping segment
Feature Disentanglement (ECCV'20) [80]	✓	✓	OBF, VIPL-HR, MMSE-HR	30s non-overlapping segment
1st Vision-For-Vitals Challenge				
Ours (ICCV'21)	✓	✓	*V4V dataset*	Continuous metric: Frame-level HR/RR

[13, 85, 88] use an ICA method to determine the underlying rPPG signal followed by 3rd order Butterworth bandpass filtering to obtain the power spectrogram whose peak corresponds to the heart rate (in a valid range of 0.7Hz - 2.5Hz). Tarassenko [111] et. al propose a pipelined approach that uses face tracking and pole selection mechanism to estimate heart rate, respiration rate, and oxygen saturation. However, these methods are susceptible to the noise, motion, and lighting conditions of the environment.

Advancements in deep learning [52, 53, 65, 66, 81, 99] have made it possible to achieve remarkable performance in the task of video-based human physiological estimation. One of the early popular methods in this direction is DeepPhys [10]. This model is trained on many facial videos in a supervised fashion, where ground truth blood volume pulse was used as the label. The key idea is to use separate branches for modeling motion and appearance. The latter branch aids the former by providing attention to facial pixels. Similarly, a more recent method MTTS-CAN [65] used the attention mechanism in conjunction with Temporal Shift Modules [60] to achieve real-time performance on the task. In [80], the authors demonstrated an effective method to disentangle spatio-temporal representation of the video called MSTmaps into noisy signals and physiological signals. In summary, deep learning methods have demonstrated reliable performance on datasets containing relatively stable physiological signals. In the V4V challenge, we introduce a dataset in which the physiological signals vary significantly due to the elicitation of spontaneous emotions.

3.3 *V4V dataset* curation

As part of the V4V challenge, we curated a database called the *V4V dataset* by carefully selecting subjects from the Multimodal Spontaneous Emotion database (BP4D+) [134] along with a number of new subjects that are collected as BP4D+ extension. In this section, we describe the data collection process, distribution of subjects, and data annotation process used for creating the *V4V dataset*. In this section, we also describe the significance of the dataset for the V4V

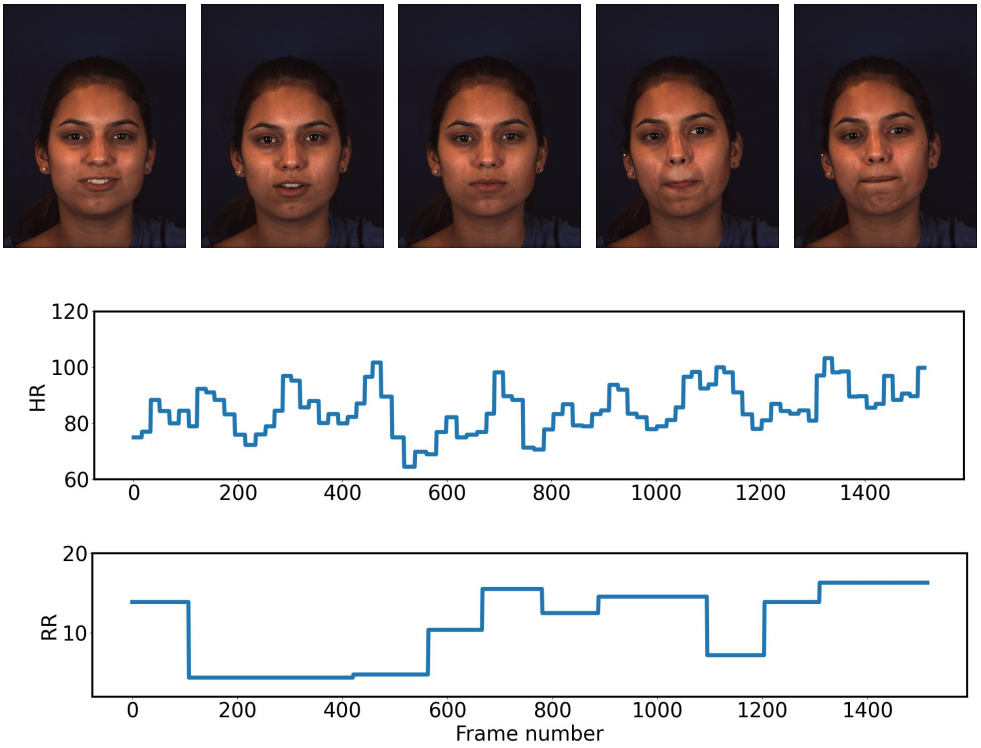

Figure 3.2: Subject participating in a cold presser (T8) task

challenge since the dataset contains challenging variations in the physiological signals induced through emotion elicitation.

3.3.1 Data collection and annotation

The *V4V dataset* was curated with the goal of obtaining a large-scale emotional corpus for human behavioral and physiological analysis. The dataset is collected at Binghamton University and includes subjects of age groups ranging from 18 to 66. It has subjects from diverse ethnicities/racial ancestries - African American, White, Asian (East and Middle-east), Hispanic/Latino, and Native American. There are 179 subjects in total with a maximum of 10 experimental tasks per subject. Each task was specifically designed to induce specific emotions among participants.

For recording the facial videos of each subject, a 3D Dynamic imaging system (Di3D) was used. All the videos were recorded with a resolution of 1040×1392 pixels and a fixed frame rate of 25 FPS in portrait mode. The Di3D system also has a symmetric lighting system that is used as the light source for capturing the videos. A board has been placed in the background while recording the video to limit any background motion and noise.

For collecting human physiological data, the Biopac MP150 system was used. The specification of the device is as follows:

- Blood pressure: For monitoring the blood pressure, the Biopac NIBP100D system with a

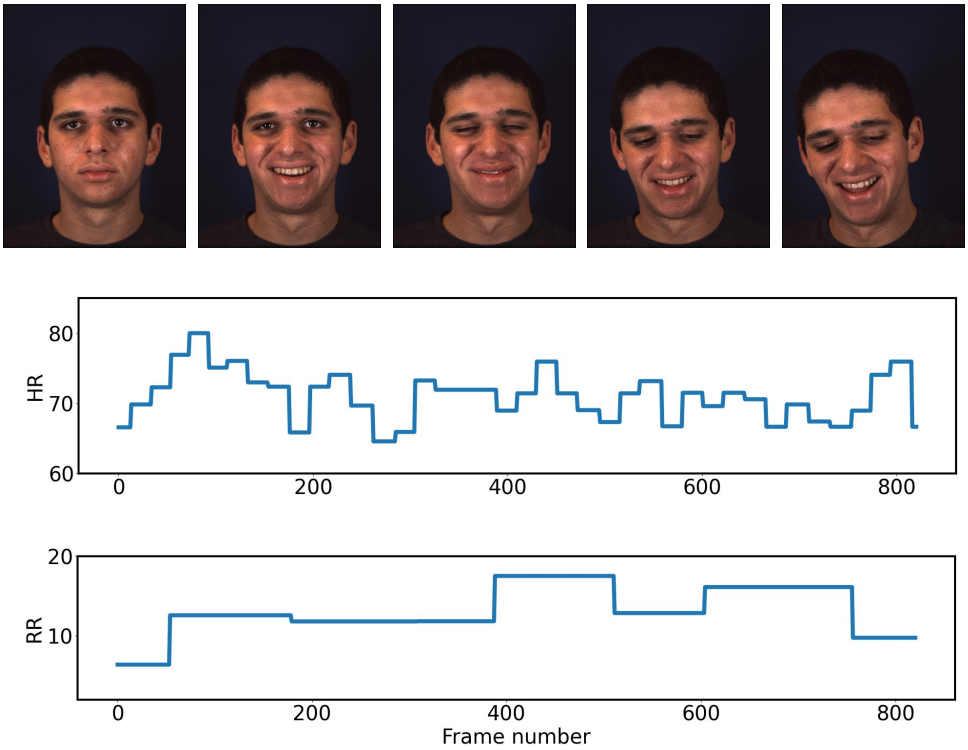

Figure 3.3: Subject participating in a silly song (T6) task

measurement range of (-25mmHg, 300mmHg) was used. It contains a finger unit and an inflatable cuff that can be placed on the arm to measure blood pressure. The device recorded high-quality measurements of systolic and diastolic blood pressure and also recorded the continuous blood pressure waveform at 1000Hz.

- Heart Rate (HR) measurement: We used off-the-shelf software called Biopac AcqKnowledge to derive HR measurement from the continuous blood pressure signal. This was achieved by performing noise removal, followed by peak-to-peak time calculation. The software used an HR range preset to (40, 180) beats-per-min. The HR signal is then downsampled and synchronized with each frame.
- Respiration Rate (RR) measurement: The respiratory signal was captured by using the Biopac Respiration Belt. Similar to the extraction of HR, we used the Biopac Acqknowledge software to extract the respiration rate of the subject by computing peak-to-peak time with an RR range preset to (4, 20) breaths-per-min.

In order to synchronize the camera and physiological devices, a controller was used to trigger the start of video capture and physiology measurement simultaneously.

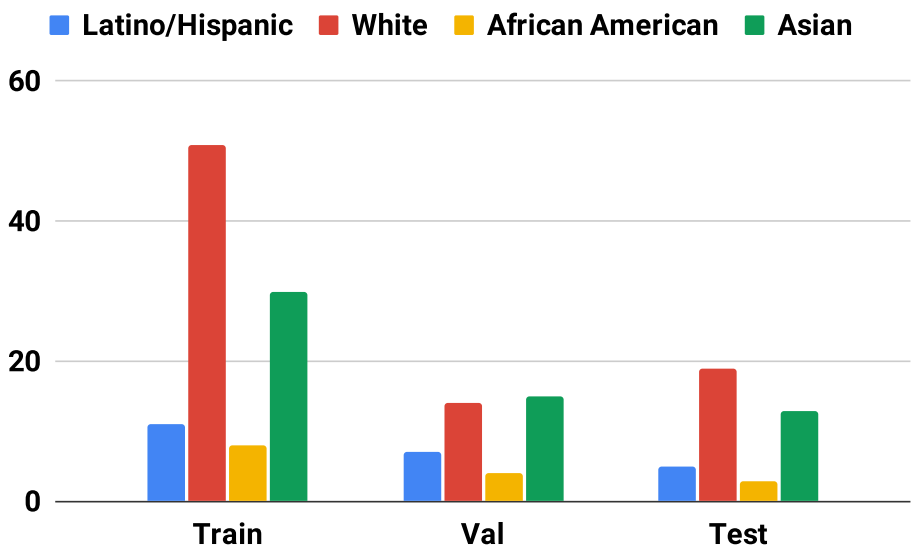

Figure 3.4: Distribution of videos according to the ethnicity of subject for train, validation, and test subjects.

Table 3.2: Ten tasks and the target emotion

Task	Activity performed	Emotion induced
T1	Funny joke	Happiness
T2	Watch 3D avatar of self	Surprise
T3	911 emergency call	Sadness
T4	Sound	Surprise
T5	True / False question	Skepticism
T6	Silly song	Embarrassment
T7	Dart threat	Fear
T8	Cold presser	Pain
T9	Complaining against participant	Angry
T10	Odor experience	Disgust

Table 3.3: *V4V Dataset*: Data split

Data Fold	Number of subjects	Number of Videos	Average video length (in sec)
Train	100	724	44.2
Validation	39	276	42.9
Test	40	358	45.8

3.3.2 Emotion elicitation protocol of the *V4V dataset*

Each subject participated in 10 different tasks that were carefully designed to evoke a specific emotion. It is known that an increased emotional activity often alters the human physiological signals [125]. For example, fear arousal spontaneously increases the heart rate and respiration rate of a subject. To achieve this goal, a skilled interviewer was hired to conduct the tasks.

As shown in Table 3.2, the tasks included interpersonal communication, film watching, cold pressure, and physical activities. There is also a smooth transition in the emotions induced across the 10 tasks. There was a brief pause between any two tasks for self-reporting purposes.

First, the interviewer ensured that the participant felt comfortable and relaxed at the start of data collection by advancing a joke (T1). Then the subject was shown their own 3D avatar to invoke the feeling of surprise (T2). Next, the T3 task required participants to watch a short film of a 911 emergency call to elicit a feeling of sadness. In the T4 task, a loud noise was played to startle and surprise the participant. The interviewer then induced skepticism by advancing a question (T5), followed by arousal of embarrassment when the subject was required to conduct a silly task (T6). In the T7 task, the interviewer invoked fear in the participant by threatening to throw a dart at the subject. In T8, the participant was required to submerge their hand into the ice water which invoked physical pain. The interviewer pretended that the participant demonstrated poor performance in task T8 and complained to the participant to evoke anger (T9). In the final task, T10, the subject experienced a smelly odor to evoke a feeling of disgust. At the end of each task, the subject was asked to report the emotions experienced from a list of choices and also rate the emotional intensity in a 5-point rating style.

Owing to the carefully designed emotion elicitation, the *V4V dataset* contains challenging intra-video physiological variations. Further, due to the nature of the tasks T6-T10, they often are associated with large head movements (> 10 deg) adding an additional element of challenge for physiological estimation. Therefore, *V4V dataset* offers desirable elements for benchmarking approaches effectively.

3.3.3 Post-processing

After creating the ground truth, we eyeballed each of the heart rate and respiration rate sequences to discard any video that had noisy readings, e.g. Shaking the contact-based device during emotion elicitation tasks. We ensured that every video had exactly 1 HR and 1 RR reading per frame of the video after aligning the physiological signals. After processing the data as described, we obtained 179 subjects and 1358 videos with heart rate and respiration rate readings.

3.3.4 V4V Challenge phases and dataset split

The 1st V4V challenge was organized in two phases. In the first phase, the participants used the validation set to improve the performance of their methods, and in the second phase, the participants evaluated their method on the test set. A public leaderboard was maintained in both

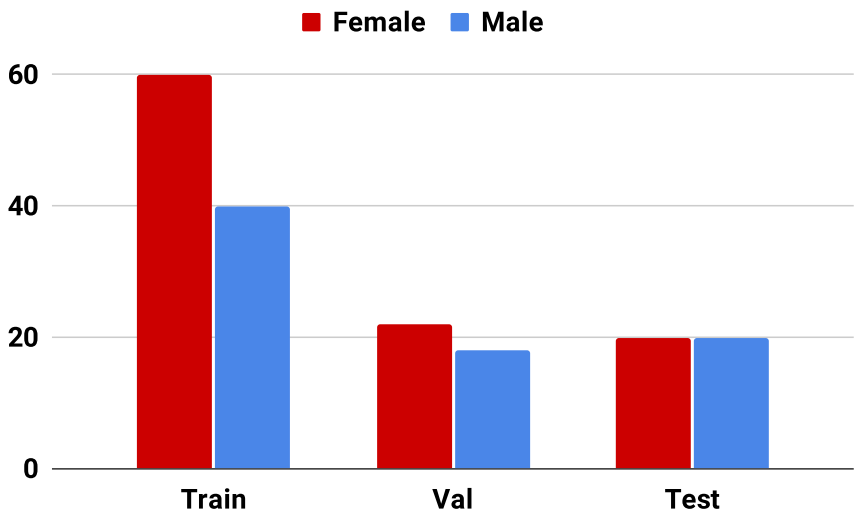

Figure 3.5: Distribution of videos according to subject gender for train, validation, and test splits.

phases indicating scores obtained by the methods on the evaluation metrics described in Sec. 3.4.

As shown in Table 3.3, we used 1000 videos in phase 1 of the challenge where the training set included 724 videos and the validation set included 276 videos. At the start of phase-1, we released the training dataset and validation set videos. The training set contains the compressed videos, heart rate signal, respiration rate signal, and raw 1000 Hz blood pressure waveform. At the start of phase-2, we released the test set videos (a total of 358 videos) and validation-set ground truth as well.

In order to ensure that the methods performed fairly across different population groups, we used gender and ethnicity information self-reported by the subjects to create the data folds. Fig. 3.4 shows the ethnic distribution of subjects of train, validation, and test set. We tried to balance the ethnic distribution keeping them similar across the data splits to avoid the biases brought by skin tone and affective attributes. Fig. 3.5 presents the distribution of videos according to the gender of the subject, which has been adjusted to have near-balanced distribution, with the number of female subjects slightly more than the number of male subjects in the train and validation set.

3.4 Evaluation metrics

In the existing literature, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (R) have been used commonly for evaluating the efficacy of the proposed method. The caveat is that almost all of the existing works construct 30-second non-overlapping segments and predict a single HR/RR value per segment. While this is a good measure of the performance of the method, it has some potential drawbacks.

Table 3.4: Results obtained on the *V4V dataset* (test set) sorted by cMAE (lower is better ↓)

Approach	cMAE (↓)	cRMSE (↓)	cR (↑)
Stent et. al [107]	9.22	14.18	0.47
Hill et. al [36]	9.37	14.59	0.44
Kossack et. al [49]	10.15	15.38	0.44
Ouzar et. al [82]	11.60	14.90	0.20
Baseline (Green [113])	15.45	20.73	0.05

Unlike other physiological datasets, the *V4V dataset* consists of short video clips that have significant intra-video variations in the HR and RR (as seen in Fig. 3.2 and Fig. 3.3) for the duration of the video. Therefore, an accurate method should be able to predict HR and RR at a more fine-grained frame-level rather than at a segment-level. Further, contact-based devices predict the heart rate continuously, instead of one prediction for every 30s. This is especially useful in diagnosing Atrial fibrillation (Afib) [84] which is associated with high Heart Rate Variability. Based on these observations, we propose a metric that computes error by taking into account the frame-level HR/RR.

In the 1st V4V challenge, we employ the following three metrics - MAE, RMSE, and R - at a frame-level rather than at segment-level. We denote these granular evaluation metrics as cMAE, cRMSE, and cR respectively. (where *c* indicates a continuous metric)

$$cMAE = \frac{\Sigma_i |\widehat{HR}_i - HR_i|}{N} \tag{3.1}$$

$$cRMSE = \sqrt{\frac{(\Sigma_i |\widehat{HR}_i - HR_i|^2)}{N}} \tag{3.2}$$

$$cR = \mathbf{PearsonCorrelation}(\widehat{HR}, HR) \tag{3.3}$$

where HR_i is the ground truth heart rate of the frame i in the test set, $\widehat{HR}_i$ is the predicted heart rate for the frame i in the test set and N is the total number of frames in the test set. Similarly, we define equivalent evaluation metrics for Respiration Rate using RR_i and $\widehat{RR}_i$.

3.5 Methods used in the V4V challenge

There were four teams that participated in the V4V challenge. In Table. 3.4 results obtained by different teams and a baseline (Green) method [74, 113] have been listed.

3.5.1 Estimating Heart Rate from Unlabelled Video

Stent et. al [107] use a self-supervised approach based on [31] to overcome challenges that are typically faced by a supervised approach such as imprecise and noisy data collection owing to complex ground truth capturing setup, proprietary hardware computations, etc. Further, by virtue of the self-supervised nature of the proposed method, it can also tackle any domain shift problem that arises during the test phase. First, an image of size 192×128 is extracted by using a face detector [133] and padded with an additional 25% buffer. Then a PPG estimator consisting of 3DCNN [130] is used to extract the rPPG signal.

The core idea of this approach is to formulate a self-supervised framework by utilizing a video resampling module to augment the dataset with new heart rate labels. E.g., the heart rate of the subject can be increased by squeezing the video. Similarly, stretching the video decreases the heart rate. The module is not only used to increase/decrease the heart rate by controlling the speed for video but it is also used to resample the output rPPG signal with corresponding inverse frequency. Based on this idea, positive and negative pairs are constructed for effective contrastive learning [9, 34].

Owing to the self-supervised nature of the proposed method, the method is able to train on the test samples directly. The paper also implements several hand-crafted tricks to further improve performance of the their method. The authors employ specialist models to account for multi-mode lighting conditions in the test set and also propose a confidence model to mitigate errors caused by faulty PPG prediction.

3.5.2 Beat-to-Beat Cardiac Pulse Rate Measurement From Video

Hill et al. [36] proposed a hybrid-CAN-RNN framework by incorporating a bi-directional GRU on the top of CAN [65]. The original Hybrid-CAN has two branches: (1) the appearance branch, and (2) the motion branch. The appearance branch deploys 2D convolutions to extract the spatial skin features. The temporal module leverages 3D convolutions to capture the temporal relationships. An attention module is used to bridge the connection between motion and appearance branches, and the generated attention masks make the network focus on useful signals.

In addition, two layers of GRU (the first is bi-directional) are used for learning the longer-term rPPG waveform transitions, and the reason is that the convolutional layers can only capture local spatial and temporal features. Besides, a synthesized dataset is used to improve the model generalization ability, and other datasets such as AFRL [24] and UBFC [6] are used for training. The variety of the training data improves the model generalization ability, which helps to deal with more complicated environmental biases.

3.5.3 Automatic region-based heart rate measurement using remote photoplethysmography

Kossack et. al [49] utilize a popular classical method called "plane orthogonal to skin" (POS) [118] to measure the heart rate. The core idea of this V4V submission is to extract a signal from the region of the face that has the strongest signature of rPPG in it. In order to achieve this goal, the face region is divided into five subregions - forehead, right cheek, left cheek, nose, and the entire face is considered as the fifth subregion. For each of these ROIs, a score is computed and the ROI with the best score is used to determine the heart rate of the subject.

First, POS projection is applied to each subregion to extract the rPPG signal followed by FFT and bandpass filtering to obtain a power spectrum density for each subregion. Next, a scoring function is designed for the determination of the best region of interest. This function is dependent on two parameters, (1) Maximum magnitude ($Mmax$) and (2) Row wise sum of the correlation matrix of 5 rPPG signals ($Csum$). The sum of $Mmax$ and $Csum$ yields the final score which is used to determine the best region of interest. Since this method does not take a deep learning approach, the method can be used directly on the test set.

3.5.4 Results

Various approaches are evaluated for cMAE, cRMSE, cR as shown in Table 3.4. It is interesting to note that the self-supervised formulation used by [107] is on top of the leaderboard demonstrating good performance on continuous evaluation metrics. However, their approach also involves carefully handcrafted tricks. The method presented in [36] used additional datasets for training and obtained cMAE of 9.37. All participants chose to participate only in the HR sub-challenge. In summary, the results obtained by different methods indicate that continuous prediction is still a challenging task and there is scope for further improvements.

3.6 Conclusion

In this paper, we have presented the first Vision-for-Vitals challenge for benchmarking the performance of different rPPG methods on a newly curated large-scale dataset called *V4V dataset*. The dataset contains desirable attributes necessary for benchmarking various approaches including challenging elements such as spontaneous behavior and varied HR/RR signals. Further, in order to benchmark different methods effectively, we evaluate various approaches using granular frame-level error metrics rather than segment-level error metrics employed by previous methods. The results show that there is room for further improvement in methods and evaluation protocols used for non-contact video-based human physiological estimation.

Chapter 4

Transformer-based physiology estimation

The video-based physiological signal estimation has been limited primarily to predicting episodic scores in windowed intervals. While these intermittent values are useful, they provide an incomplete picture of patients' physiological status and may lead to late detection of critical conditions.

We propose a video Transformer for estimating instantaneous heart rate and respiration rate from face videos [97]. Physiological signals are typically confounded by alignment errors in space and time. To overcome this, we formulated the loss in the frequency domain.

We evaluated the method on the large-scale Vision-for-Vitals (V4V) benchmark. It outperformed both shallow and deep learning based methods for instantaneous respiration rate estimation. Our code is available here [1].

4.1 Introduction

Contact-based devices (e.g. pulse-oximeter) are prevalent among healthcare professionals for assessing and monitoring the vital signs of patients in hospital settings. These vital sign monitoring devices require physical contact and can cause discomfort to patients. As remote diagnosis is becoming increasingly common, partly due to the recent COVID-19 pandemic, there is a pressing demand for non-contact physiological estimation methods.

Over the years, conventional photoplethysmography (PPG), the contact-based optical estimation of microvascular blood volume changes, has evolved into contactless imaging PPG (iPPG). These methods utilize digital cameras and computer vision techniques for estimating heart-generated pulse waves and their respiratory modulation by means of peripheral blood perfusion measurements. Past research has demonstrated that these bio-signals can be extracted with high fidelity in a strictly controlled environment; i.e. with a prediction error, < 3 beats-per-min for Heart Rate extraction [10]. Intuitively, the camera captures subtle periodic color variations that result from the blood volume changes in the underlying skin tissues. A careful analysis of the subtle changes in the video reveals the physiological state.

[1] `https://github.com/revanurambareesh/instantaneous_transformer`

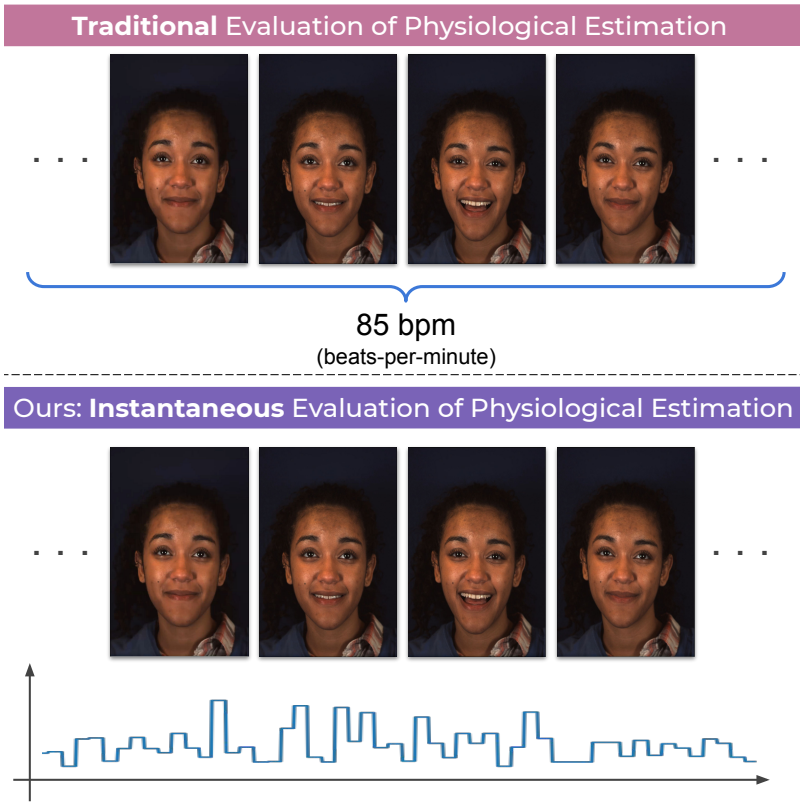

Figure 4.1: Traditional evaluation v/s instantaneous evaluation (Ours) for heart rate and respiration rate estimation

Previous video-based physiological extraction techniques have been limited to predicting episodic heart rate values over large, *non-overlapping* windows (e.g. 30 seconds), which influenced the choice of performance metrics to evaluate these methods. The window-based evaluation protocol does not provide complete insight into Heart Rate Variability (HRV), which plays an important role in understanding the physical and mental conditions of an individual [84]. This evaluation gap has been considered in a recent physiological challenge organized at the ICCV conference called "Vision-for-Vitals" (V4V). Instead of using non-overlapping windows to measure the error, the challenge organizers [96] proposed using an instantaneous (a.k.a. *continuous*) evaluation protocol that measures the performance of a method at a per-frame level. Hence, in our work, we aim to benchmark different methods including our proposed method using this continuous evaluation protocol (Fig. 4.1).

Recently, deep learning based solutions have been proposed for the task of human physiology estimation [10, 80]. One of the most popular methods in this direction is DeepPhys [10]. This method utilizes the optical principles of PPG to predict blood volume pulse (and respiratory wave for respiration rate estimation) from a facial video. Even though DeepPhys has made significant progress, it is still limited to episodic evaluation and is unable to fully exploit the temporal and periodic nature of the blood volume pulse. To remedy this, we draw motivation from recent work on video analysis [78] and utilize a Transformer based architecture for frame-level prediction.

There is a body of literature that shows that the Transformer is effective in modeling temporal sequences [8, 112].

The paper advances two main novelties:

- In our work we use a Transformer based architecture for instantaneous prediction and evaluation of human physiological signals from the facial video. We formulate the loss function in the frequency domain to minimize confounds introduced by the temporal misalignment of the video and the PPG signals.
- We evaluate the method on the challenging Vision-for-Vitals (V4V) dataset. We show that our approach reliably estimates Heart Rate (HR) and Respiration Rate (RR), outperforming shallow and deep-learning methods trained on the same data (V4V training set).

4.2 Related Work

4.2.1 Video-based physiology extraction

Based on the principles of remote-photoplethysmography (rPPG), several methods have been advanced [13, 15, 85, 88, 111, 113, 118, 124] for the extraction of physiological signals from facial videos. In [113], the authors highlighted that the green channel of the RGB video can be used to compute the heart rate since the green channel has the strongest signature of photoplethysmography. Wang et. al [118] proposed a mathematical model for the reflection properties of skin and developed a novel rPPG method based on the model. Further, researchers have utilized face detection and tracking methods such as the Bounded Kalman Filter for extracting facial regions of interest [13, 88].

More recently, deep learning approaches [10, 65, 67] have been proposed for the task of physiological estimation. One of the main aspects of DeepPhys [10] and MTTS-CAN [65] architectures is the spatial attention mechanism which is used to determine the right regions of interest thereby enabling end-to-end trainability of the network. However, all of these methods are evaluated on episodic scores over windowed intervals. To tackle this limitation, the Vision-for-Vitals workshop [96] held at ICCV introduced multiple metrics to promote instantaneous prediction of HR and RR. In our work, we aim to evaluate all methods including ours over these metrics. We also aim to incorporate spatial attention masks and utilize Transformer for temporal learning.

4.2.2 Transformers

A transformer [112] consists of a transformer-encoder and a transformer-decoder which in turn are composed of several multi-headed self-attention layers. In [112], the authors demonstrated high accuracy of Transformer based architecture for multiple language translation tasks. With minor modifications to the proposed Transformer architecture, it has been successfully adapted to a wide range of research problems in computer vision and natural language processing. Par-

ticularly, the computer vision community has explored transformer-based architecture in two forms.

In one form, the Transformer based architecture includes a convolutional neural network (CNN) backbone that is used as a feature extractor [8, 61, 78]. In [8] the authors employed ImageNet pre-trained ResNet-based backbone as a spatial feature extractor for the task of object detection using Transformers in an end-to-end manner. In other related work, video Transformers [78] have been employed for the goal of temporal modeling of the videos. Here, the convolutional backbone extracts features, and the Transformer is used for temporal modeling. In our work, we use a DeepPhys [10] based convolutional backbone network and a Transformer for temporal modeling.

In another form, the architecture is developed purely using transformer layers. In [18] the authors trained a pure transformer for the task of image classification by dividing the image into multiple parts. In related work, [131] aimed at detecting fake 3D printed face masks using Transformer based architecture by drawing motivation from the principles of Photoplethysmography. This is achieved by feeding the Multi-scale Spatio-Temporal maps (MSTmaps) of the facial regions and background regions along with a positional embedding into a transformer network.

In this work, we focus on developing a method for the task of instantaneous evaluation of physiological signals by using a video transformer. We aim to train the network in an end-to-end manner by relying on the spatial attention module in DeepPhys.

4.3 Methods

The goal of remote PPG extraction is to effectively extract a bio-signal that contains HR (or RR) using a facial video. To this end, we propose a Transformer-based architecture, inspired by the principles of remote PPG. In this section, we first introduce the optical basis of our method by relying on the skin reflection model [10, 118]. Next, we propose the architecture for the HR/RR estimation and finally explain the loss formulation that we used for training the model.

4.3.1 Optical basis of video-based bio-signal extraction

The changes in the volume of blood flowing underlying the facial skin results in subtle color changes. In order to extract this bio-signal, we use the popular skin reflection model that is based on Shafer's dichromatic reflection [10, 118]. At a given time instance t in the video, the reflection of the light back to the camera can be considered as a function that varies in the RGB color space according to Eq. 4.1.

$$C_k(t) = i(t) \cdot (\mathbf{v}_s(t) + \mathbf{v}_d(t)) + \mathbf{v}_n(t) \tag{4.1}$$

Here, $C_k(t)$ is the color intensity of the RGB pixel k, $i(t)$ luminance intensity level which is regulated by specular reflectance $\mathbf{v}_s$ and diffuse reflectance $\mathbf{v}_d$. The term $\mathbf{v}_n$ is the camera quantization noise. The specular reflection is a mirror-like reflection that bounces the light right off

the facial skin while the diffuse component contains useful pulsatile signals. The components $\mathbf{v}_d(t)$ and $\mathbf{v}_s(t)$ can be expressed further in terms of stationary reflection strength, underlying physiological bio-signal $p(t)$ and motion induced changes $m(t)$ (e.g. facial movements, expressions).

$$\mathbf{v}_d(t) = \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t) \tag{4.2}$$

$$\mathbf{v}_s(t) = \mathbf{u}_s \cdot (s_0 + \Phi(m(t), p(t))) \tag{4.3}$$

Here $\mathbf{u}_d$ is the stationary skin reflection strength and $\mathbf{u}_p$ is pulsatile signal strength that varies according to the volume of hemoglobin. Notice how Eq. 4.2 does not depend on $m(t)$, while Eq. 4.3 depends on both $m(t)$ and $p(t)$. Further, $\mathbf{u}_s$ is the unit norm vector indicating the color of the light spectrum and s_0 is the stationary component of the specular reflection and Φ is a function of motion $m(t)$ and the physiological signals $p(t)$.

Next, $i(t)$ can be further decomposed into stationary and varying components according to,

$$i(t) = i_0 \cdot (1 + \Psi(m(t), p(t))) \tag{4.4}$$

Substituting Eq. 4.2, Eq. 4.3, and Eq. 4.4 into Eq. 4.1, we obtain Eq. 4.5.

$$\begin{aligned} C_k(t) \approx \mathbf{u}_c \cdot i_0 \cdot c_0 + \mathbf{u}_c \cdot i_0 \cdot c_0 \cdot \Psi(m(t), p(t)) + \\ \mathbf{u}_s \cdot i_0 \cdot \Phi(m(t), p(t)) + \mathbf{u}_p \cdot i_0 \cdot p(t) + \mathbf{v}_n(t) \end{aligned} \tag{4.5}$$

As a next step, we remove the dominant stationary signal by computing the first order derivative of $C_k(t)$ in line with [10]. Additionally, we reduce the image size to a size of 36×36 to suppress the quantization noise i.e. $\mathbf{v}_n \approx 0$.

$$\begin{aligned} C'_k(t) \approx \mathbf{u}_c \cdot i_0 \cdot c_0 \cdot (\frac{\partial \Psi}{\partial m} m'(t) + \frac{\partial \Psi}{\partial p} p'(t)) + \\ \mathbf{u}_s \cdot i_0 \cdot (\frac{\partial \Phi}{\partial m} m'(t) + \frac{\partial \Phi}{\partial p} p'(t)) + \mathbf{u}_p \cdot i_0 \cdot p'(t) \end{aligned} \tag{4.6}$$

As an additional normalization step, we also divide $C'_k(t)$ by $\mathbf{u}_c \cdot i_0 \cdot c_0$ to remove further stationary luminance intensity across all pixels. In practice, we approximate $C'_k(t)$ by computing the pixel-wise difference for consecutive pairs of frames and normalizing it by using a mean frame. We remove the constant factor of 2 from the resulting expression as it is just a scaling term. Therefore, we obtain $M_k(t)$, which we use for training our model (Eq. 4.7).

$$M_k(t) = (C_k(t+1) - C_k(t)) \oslash (C_k(t+1) + C_k(t)) \tag{4.7}$$

where $\oslash$ is the Hadamard division (element-wise) operation.

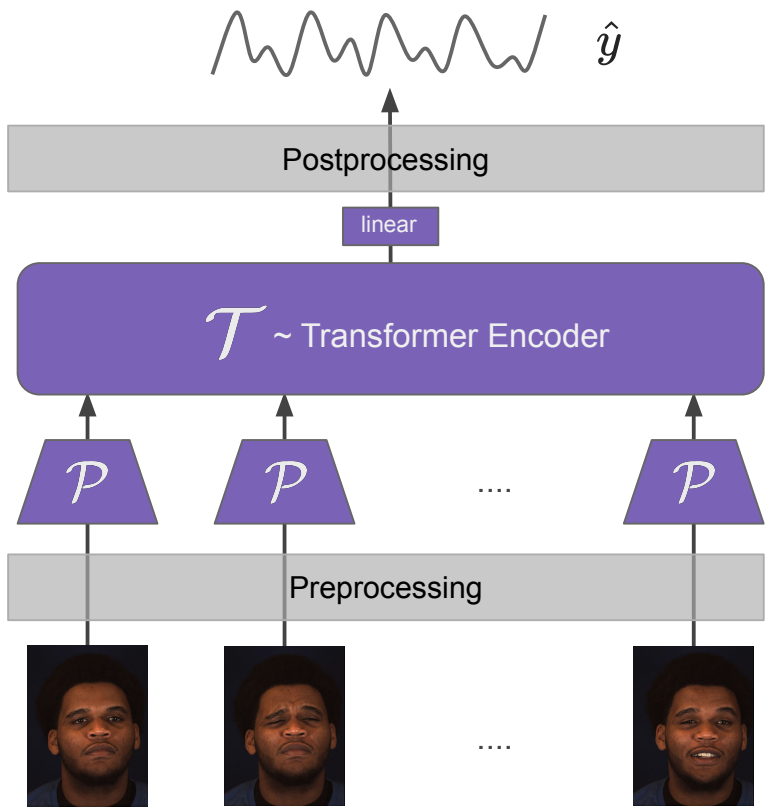

Figure 4.2: Video transformer for human physiological signal extraction

4.3.2 Video Transformer for physiological estimation

In this section, we propose a Video Transformer for the task of human physiological estimation. As shown in Fig. 4.2, the video transformer consists of a spatial backbone $\mathcal{P}$ and frame-level temporal aggregation module $\mathcal{T}$ to assist with learning temporal dependencies of the bio-signal waveform. For the spatial backbone, we use the popular DeepPhys-based architecture and for the temporal module, we use the Transformer-encoder architecture. Therefore, our video Transformer is an end-to-end trainable framework, unlike the approaches which use facial landmark detection for selecting pixels in the region of interest [67].

We begin by describing the DeepPhys-based spatial backbone network depicted in Fig. 4.3. The network consists of two branches for modeling motion representation and spatial attention. The motivation for using the spatial attention branch is to learn regions of the face (via the attention masks) that could assist the motion branch for physiological estimation. The spatial attention branch is trained on a video frame $C_k(t)$ of dimension 36×36, and the motion branch is trained on the normalized frame differences $M_k(t)$ of dimension 36×36 (Eq. 4.7). The attention mask $\mathbf{q}$ is given by,

$$\mathbf{q} = \frac{(h_{\mathbf{z}} w_{\mathbf{z}}) \cdot \mathbf{z}_a}{2\|\mathbf{z}_a\|_1} \tag{4.8}$$

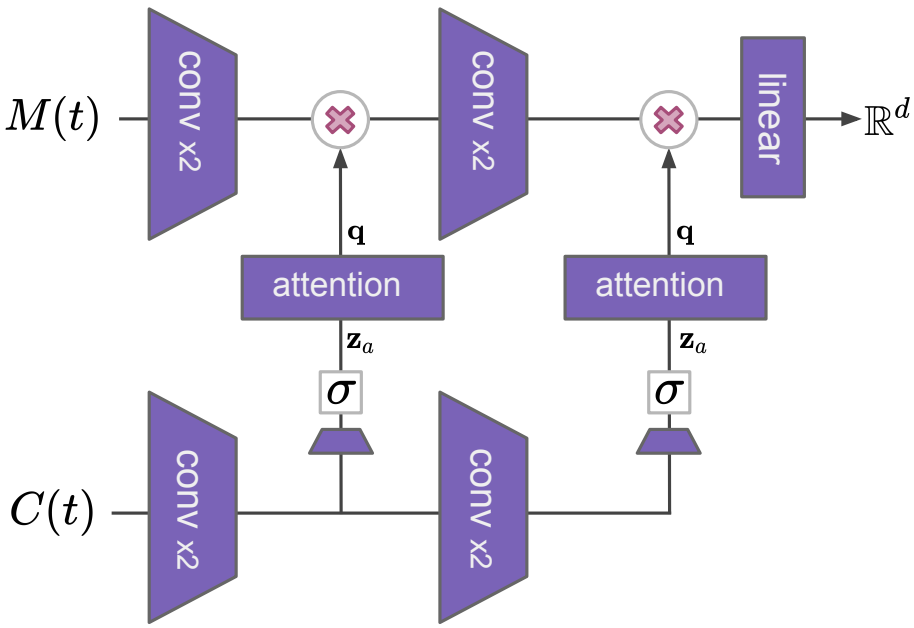

Figure 4.3: DeepPhys-based encoder $\mathcal{P}$ as a spatial backbone network for video Transformer. Here, σ denotes the sigmoid activation and q is the attention mask.

where $h_{\mathbf{z}}, w_{\mathbf{z}}$ are the dimensions of the input feature map, $\mathbf{z}_a$ are the sigmoid activations of the features from spatial attention branch (Fig. 4.3). These attention masks are multiplied element-wise across the features of the motion branch. Finally, the FC-layer of $\mathcal{P}$ encodes the spatial features into a vector of dimension d (Fig. 4.3)

We achieve frame-level temporal aggregation by utilizing a Transformer encoder $\mathcal{T}$ [112]. The encoder network is trained on the FC features of $\mathcal{P}$ using N frames of the video where N is the number of frames used for temporal aggregation. We pass all the features of $\mathcal{P}$ through a linear layer that reduces d dimensional vector into a $d_{\mathcal{T}} = 32$-dimensional embeddings. Along with these embeddings, we additionally include $[CLS]$ token for training the encoder network. Each encoder layer consists of a multi-headed self-attention block with 8 self-attention modules and an MLP layer. Further, at each input step of the Transformer, we also include positional encoding (PE) to inject the temporal information into the model. To this end, we utilize the position-encoding layer proposed by [112] and are controlled according to Eq. 4.9 and Eq. 4.10.

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\mathcal{T}}}) \tag{4.9}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\mathcal{T}}}) \tag{4.10}$$

where, pos is the position and i is the dimension in the position embedding. The output features of the encoder layers are then fed into a single MLP layer to obtain $\hat{y}$.

4.3.3 Loss formulation

We train our model end-to-end using a ground truth signal such as blood pressure (for HR) or respiratory wave (for RR). One straightforward way to train the model is by using a Mean-Squared Error (MSE) loss. However, MSE loss assumes that the ground truth is accurately synchronized with the bio-signal in the facial video. Unfortunately, it is challenging to perfectly synchronize the bio-signal with ground truth for two reasons. First, the devices used for ground-truth video capture and physiological signal recording are different. Therefore, one has to manually align the video frames with the ground-truth signal. Second, the ground truth is often collected at a peripheral site such as a finger. Therefore, there is an additional delay resulting from the Pulse-Transit Time (PTT). A related work [5] shows that the time delay for pulse transit between ear and finger is close to 150ms.

One of the other limitations of MSE loss is that it trains the model to learn both the amplitude and frequency of a wave. However, for the task of HR/RR estimation, we are interested only in the frequency of the underlying pulsatile signal and not the amplitude of the signal. Therefore, we make use of Maximum Cross-Correlation loss $l(y, \hat{y})$ [30] and perform the cross-correlation computation in the frequency domain instead of the time domain.

$$l(y, \hat{y}) = -c \cdot Max\left(\frac{F^{-1}\{\Omega(F(y) \cdot \overline{F(\hat{y})})\}}{\sigma_y \sigma_{\hat{y}}}\right) \tag{4.11}$$

where $\hat{y}$ is the ground-truth as computed by signal differences Δp and y is the predicted waveform. Further, Ω is a bandpass operator which retains only frequencies of interest, c is the ratio of power present inside the frequency range of heart rate to the total power, F is the Fourier-transform operator and $\overline{(\cdot)}$ is the conjugate operator.

4.4 Results

4.4.1 Implementation details

We reduced the input image size to 36×36 in line with [10] and computed the normalized input frame difference for the motion branch. For training the video Transformer, we fixed the number of input frames to N (where $N = 100$ for HR estimation and $N = 1000$ for RR estimation) and trained our network end-to-end. During inference, we used the predictions from $\mathcal{T}$ and computed cumulative sum to obtain the final waveform prediction. After that, we calculated the Fourier Transform of the waveform and applied a bandpass filter to limit the frequencies within the range of HR / RR and obtained $\hat{y}$. For the HR model, we used $d = 128$ and for the RR model, we used $d = 32$.

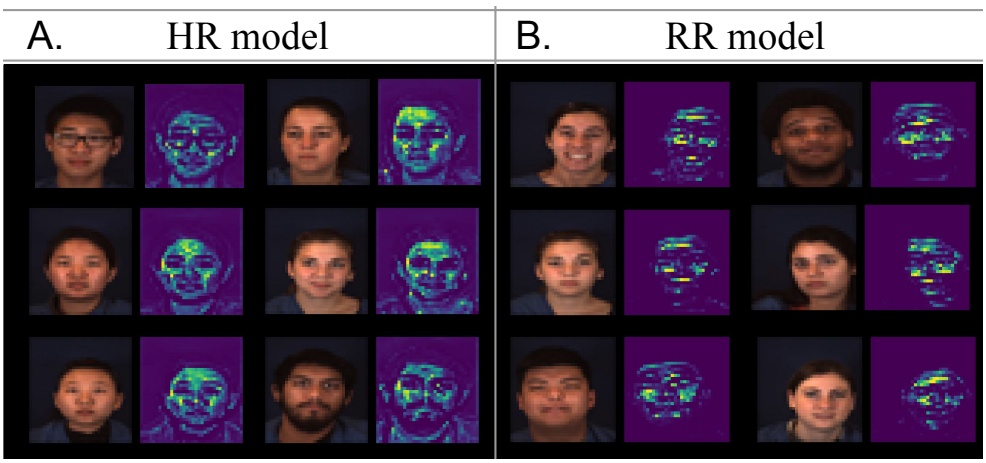

Figure 4.4: Spatial attention masks obtained for the HR model (left) and the RR model (right) on the V4V test set.

4.4.2 Datasets and evaluation protocol

We use the Vision-for-Vitals (V4V) dataset that consists of 179 subjects and 1358 videos in total. The V4V dataset contains continuous blood pressure waveform recorded at 1KHz, frame-aligned HR, and frame-aligned RR. We use the V4V training dataset for training the model and report the performance of our model on both the V4V validation set and the V4V test set. We follow the evaluation protocol set forth in the V4V challenge [96] and report continuous MAE ($cMAE$) and continuous RMSE ($cRMSE$).

$$cMAE = \frac{\Sigma_i |\widehat{HR}_i - HR_i|}{N'} \tag{4.12}$$

$$cRMSE = \sqrt{\frac{(\Sigma_i |\widehat{HR}_i - HR_i|^2)}{N'}} \tag{4.13}$$

where, $\widehat{HR}_i$ and HR_i are the predicted HR and ground-truth HR for the frame i respectively and N' is total number of frames in test-set. We use the same evaluation protocol for benchmarking RR results. Further, in order to enable evaluation of all methods on continuous evaluation protocol, we use a short moving window over the predicted blood volume pulse for HR (and predicted respiratory wave for RR) and employed FFT to predict continuous HR.

4.4.3 Heart rate estimation results

Table 4.1 shows the comparison of our method against traditional non-deep learning methods (implemented in [74]) - Green [113], POS [118] and recent deep-learning methods - DeepPhys [10] and TS-CAN [65]. It is important to note, that for a fair comparison, we excluded studies

Table 4.1: Comparison of our method against previous works for HR estimation on the V4V validation set and V4V test set. Note that lower $cMAE$ and lower $cRMSE$ are better (↓).

Name	$cMAE$ Val. set (↓)	$cRMSE$ Val. set (↓)	$cMAE$ Test set (↓)	$cRMSE$ Test set (↓)
Green [113]	16.5	21.4	15.5	21.9
POS [118]	17.3	21.2	15.3	21.8
ICA [85]	13.9	20.0	15.1	20.6
DeepPhys [10]	13.6	18.1	14.7	19.7
TS-CAN [65]	11.7	17.8	13.9	19.2
Ours	**10.3**	**16.1**	**13.0**	**18.8**

Table 4.2: Comparison of our method against previous works for RR estimation on the V4V validation set and V4V test set. Note that lower $cMAE$ and lower $cRMSE$ are better (↓).

Name	$cMAE$ Val. set (↓)	$cRMSE$ Val. set (↓)	$cMAE$ Test set (↓)	$cRMSE$ Test set (↓)
Green [113]	5.9	6.8	7.0	7.5
POS [118]	6.1	6.9	6.5	6.9
ICA [85]	6.4	7.2	5.8	6.2
DeepPhys [10]	5.0	6.1	5.5	**5.9**
Ours	**4.8**	**5.6**	**5.4**	6.0

that either utilize external training data [36] or access the test set for domain adaptation [107] on this benchmark. For computing the HR, we used a bandpass filter with a range of $[0.7, 2.5]$.

4.4.4 Spatial attention mask

The spatial attention mask offers visibility into where the model is extracting the HR and RR in a given frame. As shown in Fig. 4.4, the base encoder is able to focus on the regions corresponding to the facial skin for extraction of physiological signals. Notice how the model is excluding facial accessories such as eye-glasses (top-left subject in Fig. 4.4A) and facial hair (bottom-right subject in Fig. 4.4A).

4.4.5 Respiration rate estimation results

We train our proposed model on the continuous respiration waveform of the V4V training dataset and we report the results on the V4V validation set and the V4V test set in Table 4.2. We compare our results with traditional and deep-learning based approaches. For computing the RR, we extracted the biosignal [111] and used a bandpass filter with a range of $[0.13, 0.34]$. Results indicate that our method performs better than the other approaches on the V4V dataset.

4.5 Conclusion

In this paper, we take a step towards instantaneous prediction of physiological signals by utilizing a Transformer based architecture for extracting the heart rate and respiration rate from facial videos. We train the video Transformer model in an end-to-end manner using a cross-correlation loss in the frequency domain. The results of our approach over continuous evaluation metric using the Vision-for-Vitals (V4V) dataset shows that the model is able to outperform both shallow and deep learning methods on the task of heart rate and respiration rate estimation. As part of future work Video Transformers can be used to tackle the domain shift problem (laboratory to real-world) [50, 53] and can be used to extract other physiological signals such as Oxygen Saturation (SpO_2) [105, 111].

Chapter 5

Extraction of novel vital signs using smartphone

Inspired by Photoplethysmography and related technologies, several works have utilized facial videos to extract the blood volume pulse and subtle motion information for estimating vital signs such as heart rate and respiration rate. Although heart rate and respiration rate have been well studied in the context of vital sign estimation, blood pressure estimation is still an open research question and is an active area of research [77]. Further, blood pressure is an extremely important vital sign and is a primary indicator of cardiological problems such as Hypertension [35]. However, smartphone-based blood pressure estimation is more challenging than heart rate estimation and respiration rate estimation due to several reasons. *First*, unlike heart rate and respiration rate estimation, blood pressure is not directly revealed in the power spectral analysis of facial RGB signal. *Second*, collecting research datasets for RGB-video based blood pressure estimation is more challenging since it typically requires multi-site videos recorded with a high-frequency camera for better precision.

RGB-video based blood pressure estimation has been broadly studied under two umbrellas - Pulse Transit Time (PTT) and PPG signal morphology (See Fig. 5.1). Pulse Transit Time is generally defined as the difference in the amount of time it takes for the pulse signal to arrive at two different arterial sites [5]. This has long been known to directly relate to blood pressure. More recently, the study of human blood pressure estimation has started to focus on PPG signal morphology. For an example of a morphological feature, the time difference between systolic and diastolic peaks in the PPG sub-wave is correlated with blood pressure [63]. These studies are able to extract the blood pressure value since it shows that these morphological aspects of the PPG signal are altered by the blood pressure. Some of these methods are shown to require periodic calibration owing to the natural changes induced to the PPG signal as a result of aging and other factors. Therefore, we develop a data collection that uses a dual-camera setup of a mobile phone and records multi-site videos to effectively utilize PTT and multi-site PPG information.

In this chapter, we introduce two datasets collected using smartphones, (1) Large-scale multi-physiology dataset and (2) Dual-camera based blood pressure data collection.

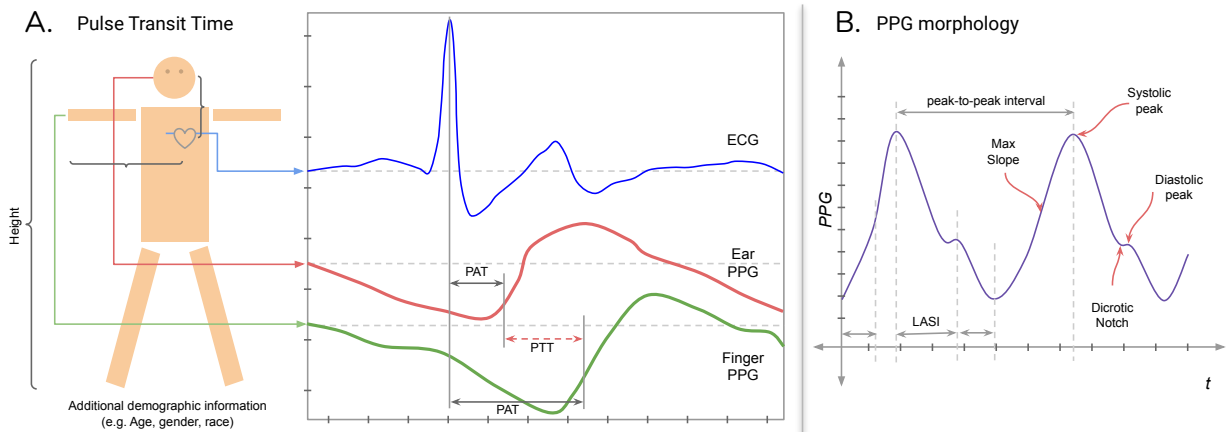

Figure 5.1: Video-based blood pressure estimation has been estimated based on two methodologies. **(a)** Pulse Transit Time **(b)** Morphological information of Photoplethysmography.

5.1 Large scale India-Sierra Leone Smartphone dataset

There have been several datasets collected for the task of video-based physiology estimation. These datasets have accelerated scientific progress in this research area and have improved the accuracy of video-based physiology estimation methods. However, these datasets have certain limitations. First, some of these datasets are small and are exclusively collected for a specific physiological signal. For e.g., PFF [39], and MMSE-HR [135] datasets include a limited number of subjects (13 and 40 respectively) and have been collected exclusively for heart rate estimation. Therefore the benchmarking has often been limited to a small pool of subjects. On the other hand, VIPL-HR [79] is a larger dataset containing 107 subjects, but it is also exclusively collected for heart rate estimation. Further, the V4V dataset described in this dissertation is a relatively large-scale dataset and has 179 subjects along with HR and RR ground truth signals for each subject. However, the V4V dataset does not include other vital signs such as oxygen saturation. To maximize the impact of video-based physiological estimation, it becomes imperative to conduct research on smartphone videos for multiple vital signs beyond Heart Rate.

5.1.1 Dataset description

We collected a new dataset from two locations that are likely to significantly benefit from video-based technology (See Applications section in Chapter 1.1.) This dataset includes smartphone videos recorded from over 1000 subjects from India and Sierra Leone respectively. For each subject, we recorded a portrait video of the subject along with physiological information such as heart rate, respiration rate, systolic, and diastolic blood pressure, thermal video, and skin temperature. While heart rate and respiration rate can be directly extracted from the video, other vital signs such as blood pressure are still under research and the literature suggests that recording videos of an additional site (such as a finger) could possibly improve the performance of blood pressure estimation. Finally, the estimation of skin temperature is a difficult problem and has

Figure 5.2: Smartphone-based multi-physiology data collection setup. **(a)** Collecting heart rate, respiration rate, and oxygen saturation using Masimo PPG device. **(b)** Skin temperature using Kinsa IR thermometer **(c)** Blood pressure measurement using Omron blood pressure device **(d)** Thermal video using FLIR camera **(e)** Artificial Skin Heating with IR heat lamp. The subject's face has been blurred to respect privacy.

not been well studied in the literature. In summary, our data collection protocol is designed based on current research and also hopes to enable future advancements in human physiological estimation.

The dataset collection, correction, and processing are still ongoing at the time of writing. In the data collected so far, there is a total of 2250 subjects in the dataset. We study the subject distribution by analyzing the histogram of subjects according to age, gender, height, and weight. In our data collection, subjects self-identified into two gender categories - male and female. Most of our subjects have ages between 20 and 40. Based on the histogram, the height and weight of subjects approximately follow a normal distribution. There were more than 1400 males and around 600 female subjects who participated in the data collection process. Certain subjects chose not to disclose their gender. Our data collection protocol was approved by Carnegie Mellon University's Institutional Review Board (IRB).

5.1.2 Data collection protocol

As depicted in Fig. 5.2, we recorded multiple physiological signals for each participant while recording facial video using a smartphone. We follow the below steps for each participant. The videos recorded have a resolution of 1280x720.

- Subject completes a questionnaire and fills in the information fields such as age and gender.
- Height and weight of the subject are measured.
- An image of the subject is captured in a standing pose (head-to-toe).
- Heart rate, respiration rate, and oxygen saturation are measured while recording an RGB video with Google Pixel 6.
- Blood pressure is recorded while the subject raises an arm. We request the subject to raise

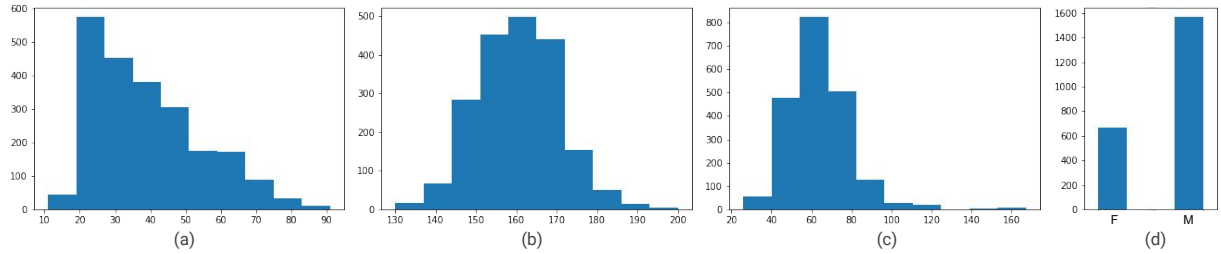

Figure 5.3: Histograms of the collected dataset. **(a)** Age distribution (in years) **(b)** Height distribution (in cm) **(c)** Weight distribution (in Kg) and **(d)** Gender distribution (F stands for female and M stands for male).

the hand as shown in Fig. 5.2(c) so we have at least two sites for extracting the PPG signal. This will allow researchers to extract Pulse Transit Time. We continue to record the facial video in a background process on the phone.

- Next, we record a thermal video using a FLIR camera and also measure skin temperature. We also artificially heat the skin using a Philips IR bulb as depicted in Fig. 5.2(e). We hope that this data would allow researchers to explore the possibility of extracting skin temperature using IR video or an RGB video based on a smartphone in the future.

5.2 Dual-camera based blood pressure dataset

We collect a new dataset for the task of video-based blood pressure estimation. The goal of this dataset collection is to record videos simultaneously from facial and finger locations. This is possible with a typical smartphone that comes with front and rear cameras which allows multi-site video recording. The front camera can be used to record facial video while the rear camera can be used to record the finger video by placing the finger on the flashlight and rear camera. Therefore, our data collection setup is developed to record multi-site videos simultaneously. However, in our setup, we use two phones with fixed positions to limit any noise caused due to motion.

5.2.1 Setup

We used iPhone 12 and iPhone 7 smartphones in this data collection for recording facial and finger videos. As shown in Fig. 5.4, one phone is placed right in front of the face to record facial video while the other phone is placed on the table for the subject to place the finger comfortably on the rear camera with flash turned on. The position of the light source is predetermined and is held constant throughout the data collection to limit any confounds due to the lighting variations. The subject is comfortably seated in an upright sitting position in front of a blue screen to prevent sources of noise from the background.

For recording the ground truth physiological signal, we use the Biopac NIBP100D system

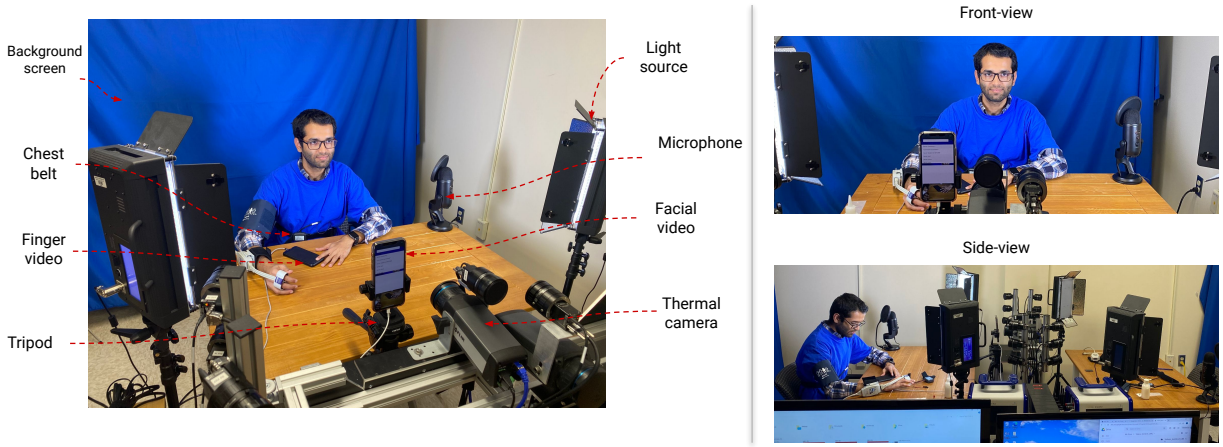

Figure 5.4: Data collection setup for the dual-camera based blood pressure estimation

which continuously measures systolic blood pressure, diastolic blood pressure, and heart rate. Although respiration rate has not been used to estimate blood pressure, future research could benefit from exploiting respiratory sinus arrhythmia [2] for improving heart rate estimation which is known to be correlated with blood pressure. As seen in Fig. 5.4, we collect respiration rate data using the Biopac Respiration Belt.

5.2.2 Software

We developed an in-house iPhone application software and a central development server to synchronize the recordings of ground truth devices and the iPhones. The central server identifies each of the devices through the network and connects to them. Once the physiological ground truth recording devices are ready, the central server simultaneously triggers all the devices to start recording the videos and physiological signals. We also play a chime on an iPhone for audio validation and save timestamps to video file names to validate that the videos are recorded with minimal synchronization error. In case of any minor mismatch, we drop the frames from the video that contains excessive frames to match the trigger timestamp.

5.2.3 Data Collection Protocol

In our data collection setup, we aim to record videos of the person under normal resting conditions and also when the subject has elevated blood pressure. Drawing inspiration from studies on Valsalva Maneuver and blood pressure [76], we ask the subject to take a deep breath and hold the breath as long as possible. We see that this activity disturbs the blood pressure activity and often modifies the blood pressure values.

We repeat this activity again after waiting for a minute for the subject to relax. In total, for each subject, we collect two baseline video segments with ground truth physiology signals and two video segments with altered physiology signal values.

Chapter 6

Conclusion

In summary, we studied existing video-based human physiology estimation methods and have made the following contributions to the research area.

- **Vision-4-Vitals (V4V) dataset**.

  We introduced a new dataset for video-based estimation of HR and RR. Unlike existing datasets, this dataset contains challenging physiological states resulting from emotion elicitation.
- **Vision-4-Vitals (V4V) challenge and workshop**.

  In total, 48 teams responded to our call and participated in the challenge. We invited the authors of the top 4 scoring methods to present their methods at a workshop in conjunction with the International Conference on Computer Vision. Over 30 participants attended the virtual event.
- **Evalution metrics**.

  We introduced a more challenging evaluation protocol for benchmarking various methods at a higher granularity. (more fine-grained)
- **Transformed-based physiology estimation**.

  We introduced a new method for physiological estimation using transformer-based architecture. This approach was able to outperform existing methods on continuous evaluation protocol for HR and RR estimation.
- **Large-scale smartphone multi-physiology dataset**.

  We are collecting a large-scale dataset using smartphones at two sites - India and Sierra Leone. This dataset contains more than 2000 subjects and has measurements of multiple physiological data - heart rate, respiration rate, blood pressure, oxygen saturation, and skin temperature.
- **Data collection for smartphone-based blood-pressure estimation**. We are collecting another dataset exclusively for blood pressure estimation in a more controlled environment. We use two high-frequency smartphone cameras to record the facial and finger videos. A central server and in-house software are used to trigger and synchronize recordings of all devices including the ground truth device.

6.1 Future work

In our work, we have focused on improving video-based human physiology estimation by introducing new datasets, metrics, and a deep learning method. One of our datasets was collected exclusively from traditionally underrepresented groups (India and Sierra Leone). We are hopeful that this dataset will improve the accessibility of new methods to subjects of all skin types including the underrepresented skin types. Further, we are hopeful that the introduced evaluation metrics encourage researchers in the community to focus on fine-grained vital sign extraction.

More research is required before deploying video-based methods in critical care settings. First, these methods have not been tested on videos recorded in critical care. Second, most of these methods are yet to be reviewed and approved by an independent organization such as the FDA on a wide range of patients. Finally, more recent deep learning methods tend to perform worse on samples in a new environment condition that was not present in training samples. Furthermore, predictions of current deep learning based approaches are hard to interpret.

There is also a lot of room for optimism. With growing interest in this research area, more vital signs such as blood pressure are being studied. Further, research on Generative Adversarial Networks and Diffusion models could enable the generation of synthetic video-based physiological datasets.

Appendix

Abbreviation	Description
PTT	Pulse Transit Time
ECG	Electrocardiogram
PPG	Photophlethysmography
PPGm	Photophlethysmography waveform morphology
ANN	Artificial Neural Network
SVM	Support Vector Machine
LiR	Linear Regression
LiM	Linear Model
LasR	Lasso Regression
RF	Random Forest
DBN	Deep Belief Network
RBM	Restricted Boltzmann Machine
LSTM	Long Short Term Memory
CART	Classification and Regression Tree
DT	Decision Tree
MAE	Mean Absolution Error
ME	Mean Error
r	Correlation Coefficient
MAD	Mean Absolute Difference
STD	Standard Deviation
MAPD	Mean Absolute Percentage Difference
RMSE	Root Mean Squared Error
Acc	Accuracy
PSD	Power Spectral Density
rPPG	Remote PPG (*a.k.a.* video based PPG signal)
n/a	Not applicable
n/r	Not reported

Table 6.1: Notations

Bibliography

[1] Stephen R Alty, Natalia Angarita-Jaimes, Sandrine C Millasseau, and Philip J Chowienczyk. Predicting arterial stiffness from the digital volume pulse waveform. *IEEE Transactions on Biomedical Engineering*, 54(12):2268–2275, 2007. 2.2.2

[2] Alfonso Angelone and Norman A Coulter JR. Respiratory sinus arrhythmia: a frequency dependent phenomenon. *Journal of Applied Physiology*, 19(3):479–482, 1964. 5.2.1

[3] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Convolutional neural networks for mammography mass lesion classification. In *2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 797–800. IEEE, 2015. 1

[4] Rohan Banerjee, Avik Ghose, Anirban Dutta Choudhury, Aniruddha Sinha, and Arpan Pal. Noise cleaning and gaussian modeling of smart phone photoplethysmogram to improve blood pressure estimation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 967–971. IEEE, 2015. 2.1

[5] Robert C Block, Mohammad Yavarimanesh, Keerthana Natarajan, Andrew Carek, Azin Mousavi, Anand Chandrasekhar, Chang-Sei Kim, Junxi Zhu, Giovanni Schifitto, Lalit K Mestha, et al. Conventional pulse transit times as markers of blood pressure changes in humans. *Scientific Reports*, 10(1), 2020. 2.2.1, 4.3.3, 5

[6] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124, 2019. 3.5.2

[7] Zeineb Bouzid, Salah S Al-Zaiti, Raymond Bond, and Ervin Sejdic. Remote and wearable ecg devices with diagnostic abilities in adults: A state-of-the-science scoping review. *Heart Rhythm*, 2022. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020. 4.1, 4.2.2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 2020. 1, 3.5.1

[10] 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. 1, 2.1.5, 3.1, 3.2, 4.1, 4.1, 4.2.1, 4.2.2, 4.3, 4.3.1, 4.3.1, 4.4.1, 4.4.3, 4.1, 4.2

[11] Santenna Chenchula, Padmavathi Karunakaran, Sushil Sharma, and Madhavrao Chavan. Current evidence on efficacy of covid-19 booster dose vaccination against the omicron variant: A systematic review. *Journal of Medical Virology*, 94(7):2969–2976, 2022. 1

[12] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1.1

[13] Ananyananda Dasari, Sakthi Kumar Arul Prakash, Lszl A. Jeni, and Conrad Tucker. Evaluation of biases in remote photoplethysmography methods. *NPJ Digital Medicene*, 2021. 3.2, 4.2.1

[14] Shreyasi Datta, Rohan Banerjee, Anirban Dutta Choudhury, Aniruddha Sinha, and Arpan Pal. Blood pressure estimation from photoplethysmogram using latent parameters. In *2016 IEEE International conference on communications (ICC)*, pages 1–7. IEEE, 2016. 2.1

[15] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10), 2013. 3.1, 4.2.1

[16] DePristo, M. and Poplin, R. M, depristo and r, poplin. `https://research.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html`, 2017. Accessed: 2018-12-06. 1

[17] Jishnu Dey, Aman Gaurav, and Vijay N Tiwari. Instabp: cuff-less blood pressure monitoring on smartphone using single ppg sensor. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5002–5005. IEEE, 2018. 2.1

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 4.2.2

[19] Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012. 2.2.2

[20] Mohamed Elgendi, Richard Fletcher, Yongbo Liang, Newton Howard, Nigel H Lovell, Derek Abbott, Kenneth Lim, and Rabab Ward. The use of photoplethysmography for assessing hypertension. *NPJ digital medicine*, 2(1):1–11, 2019. 2.2.2

[21] Heesang Eom, Dongseok Lee, Seungwoo Han, Yuli Sun Hariyani, Yonggyu Lim, Illsoo Sohn, Kwangsuk Park, and Cheolsoo Park. End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism. *Sensors*, 20(8):2338, 2020. 2.2.2

[22] Fikret Ercal, Anurag Chawla, William V Stoecker, Hsi-Chieh Lee, and Randy H Moss. Neural network diagnosis of malignant melanoma from color images. *IEEE Transactions on biomedical engineering*, 41(9):837–845, 1994. 1

[23] Jamal Esmaelpoor, Mohammad Hassan Moradi, and Abdolrahim Kadkhodamohammadi. A multistage deep neural network model for blood pressure estimation using photoplethysmogram signals. *Computers in Biology and Medicine*, 120:103719, 2020. 2.2.2

[24] Justin R Estepp, Ethan B Blackford, and Christopher M Meier. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014. 3.5.2

[25] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 1

[26] Mohamad Forouzanfar, Hilmi R Dajani, Voicu Z Groza, Miodrag Bolic, Sreeraman Rajan, and Izmail Batkin. Oscillometric blood pressure estimation: past, present, and future. *IEEE reviews in biomedical engineering*, 8:44–63, 2015. 2.2

[27] Jürgen Fortin, Dorothea E Rogge, Christian Fellner, Doris Flotzinger, Julian Grond, Katja Lerche, and Bernd Saugel. A novel art of continuous noninvasive blood pressure measurement. *Nature communications*, 12(1):1–14, 2021. 2.2

[28] Shi Chao Gao, Peter Wittek, Li Zhao, and Wen Jun Jiang. Data-driven estimation of blood pressure using photoplethysmographic signals. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 766–769. IEEE, 2016. 2.1

[29] Aman Gaurav, Maram Maheedhar, Vijay N Tiwari, and Rangavittal Narayanan. Cuff-less ppg based continuous blood pressure monitoringa smartphone based approach. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 607–610. IEEE, 2016. 2.1

[30] John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4.3.3

[31] John Gideon and Simon Stent. Remote photoplethysmography from un- labelled video. In *Proceedings of the IEEE international conference on computer vision*, 2021. 3.5.1

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1.1

[33] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3.5.1

[35] Kjeld Hermansen. Diet, blood pressure and hypertension. *British Journal of Nutrition*, 83

(S1):S113–S119, 2000. 5

[36] Brian Hill, Xin Liu, and Daniel McDuff. Beat-to-beat cardiac pulse rate measurement from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 3.4, 3.5.2, 3.5.4, 4.4.3

[37] Dean Ho. Artificial intelligence in cancer therapy. *Science*, 367(6481):982–983, 2020. 1

[38] Manish Hosanee, Gabriel Chan, Kaylie Welykholowa, Rachel Cooper, Panayiotis A Kyriacou, Dingchang Zheng, John Allen, Derek Abbott, Carlo Menon, Nigel H Lovell, et al. Cuffless single-site photoplethysmography for blood pressure monitoring. *Journal of clinical medicine*, 9(3):723, 2020. 2.2.2

[39] Gee-Sern Hsu, ArulMurugan Ambikapathi, and Ming-Shiang Chen. Deep learning with time-frequency representation for pulse estimation from facial videos. In *2017 IEEE international joint conference on biometrics (IJCB)*, pages 383–389. IEEE, 2017. 1.2, 5.1

[40] Yan-Cheng Hsu, Yung-Hui Li, Ching-Chun Chang, and Latifa Nabila Harfiya. Generalized deep neural network model for cuffless blood pressure estimation with photoplethysmogram signal only. *Sensors*, 20(19):5668, 2020. 2.1, 2.2.1, 2.2.2

[41] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters*, 471:61–71, 2020. 1

[42] DJ Hughes, Charles F Babbs, LA Geddes, and JD Bourland. Measurements of young's modulus of elasticity of the canine aorta with ultrasound. *Ultrasonic imaging*, 1(4):356–367, 1979. 2.2.1

[43] Mohammed H Iqbal, Abdullatif Aydin, Oliver Brunckhorst, Prokar Dasgupta, and Kamran Ahmed. A review of wearable technology in medicine. *Journal of the Royal Society of Medicine*, 109(10):372–380, 2016. 1

[44] Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard De Haan. Video-based respiration monitoring with automatic region of interest detection. *Physiological measurement*, 2015. 3.1

[45] In Cheol Jeong and Joseph Finkelstein. Introducing contactless blood pressure assessment using a high speed video camera. *Journal of medical systems*, 40(4):1–10, 2016. 2.1, 2.2.1

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, et al. Alphafold 2. *In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book*, 2020. 1

[47] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1

[48] Mohammad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and Mahdi Shabany. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. *IEEE Transactions on Biomedical Engineering*, 64(4):859–869, 2016. 2.1, 2.2.2

[49] Benjamin Kossack. Automatic region-based heart rate measurement using remote photoplethysmography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 3.4, 3.5.3

[50] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4.5

[51] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017. 1

[52] Eugene Lee, Evan Chen, and Chen-Yi Lee. Meta-rppg: Remote heart rate estimation using a transductive meta-learner. In *European Conference on Computer Vision*, 2020. 3.1, 3.2

[53] Eugene Lee, Annie Ho, Yi-Ting Wang, Cheng-Han Huang, and Chen-Yi Lee. Cross-domain adaptation for biometric identification using photoplethysmogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. doi: 10.1109/ICASSP40776.2020.9053604. 3.2, 4.5

[54] Joonnyong Lee, Seungman Yang, Saram Lee, and Hee Chan Kim. Analysis of pulse arrival time as an indicator of blood pressure in a large surgical biosignal database: Recommendations for developing ubiquitous blood pressure monitoring methods. *Journal of clinical medicine*, 8(11):1773, 2019. 2.1, 2.2.1

[55] Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS biology*, 15(1):e2001402, 2017. 1

[56] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018. 3.2

[57] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st challenge on remote physiological signal sensing (repss). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3.1

[58] Hao Lin, Wenyao Xu, Nan Guan, Dong Ji, Yangjie Wei, and Wang Yi. Noninvasive and continuous blood pressure monitoring using wearable body sensor networks. *IEEE intelligent systems*, 30(6):38–48, 2015. 2.2.2

[59] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2.1.5

[60] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision, 2019. 3.2

[61] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, June 2021. 4.2.2

[62] Wan-Hua Lin, Hui Wang, Oluwarotimi Williams Samuel, Gengxing Liu, Zhen Huang, and Guanglin Li. New photoplethysmogram indicators for improving cuffless and continuous blood pressure estimation accuracy. *Physiological measurement*, 39(2):025005, 2018. 2.1, 2.2.2

[63] Wan-Hua Lin, Xiangxin Li, Yuanheng Li, Guanglin Li, and Fei Chen. Investigating the physiological mechanisms of the photoplethysmogram features for blood pressure estimation. *Physiological Measurement*, 41(4):044003, 2020. 2.1, 2.2.2, 5

[64] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 2.1.5

[65] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*, 2020. 3.1, 3.1, 3.2, 3.5.2, 4.2.1, 4.4.3, 4.1

[66] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. *MetaPhys: Few-Shot Adaptation for Non-Contact Physiological Measurement*. Association for Computing Machinery, 2021. URL `https://doi.org/10.1145/3450439.3451870`. 3.2

[67] Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3.1, 4.2.1, 4.3.2

[68] Hong Luo, Deye Yang, Andrew Barszczyk, Naresh Vempala, Jing Wei, Si Jia Wu, Paul Pu Zheng, Genyue Fu, Kang Lee, and Zhong-Ping Feng. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circulation: Cardiovascular Imaging*, 12(8):e008857, 2019. 2.1, 2.2.2

[69] Heather Ting Ma. A blood pressure monitoring method for stroke management. *BioMed research international*, 2014, 2014. 2.2.1

[70] Sumit Majumder, Tapas Mondal, and M Jamal Deen. Wearable sensors for remote health monitoring. *Sensors*, 17(1):130, 2017. 1

[71] Michael S Manak, Jonathan S Varsanik, Brad J Hogan, Matt J Whitfield, Wendell R Su, Nikhil Joshi, Nicolai Steinke, Andrew Min, Delaney Berger, Robert J Saphirstein, et al. Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via machine learning. *Nature biomedical engineering*, 2(10):761–772, 2018. 1

[72] Bernard Marr. First fda approval for clinical cloud-based deep learning in healthcare. *Forbes. Forbes Publishing Company, New York City*, 2017. 1

[73] Michela Mase, Walter Mattei, Roberta Cucino, Luca Faes, and Giandomenico Nollo. Feasibility of cuff-free measurement of systolic and diastolic arterial blood pressure. *Journal*

of electrocardiology, 44(2):201–207, 2011. 2.2.1

[74] Daniel McDuff and Ethan Blackford. iphys: An open non-contact imaging-based physiological measurement toolbox. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019. 3.5, 4.4.3

[75] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[76] Ramakrishna Mukkamala, Jin-Oh Hahn, Omer T Inan, Lalit K Mestha, Chang-Sei Kim, Hakan Töreyin, and Survi Kyal. Toward ubiquitous blood pressure monitoring via pulse transit time: theory and practice. *IEEE Transactions on Biomedical Engineering*, 62(8): 1879–1901, 2015. 5.2.3

[77] Ramakrishna Mukkamala, George S Stergiou, and Alberto P Avolio. Cuffless blood pressure measurement. *Annual Review of Biomedical Engineering*, 24:203–230, 2022. 2.2, 5

[78] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 4.1, 4.2.2

[79] 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*. Springer, 2018. 1.2, 3.2, 5.1

[80] 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*, 2020. 3.1, 3.1, 3.2, 4.1

[81] Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The benefit of distraction: Denoising remote vitals measurements using inverse attention, 2020. 3.2

[82] Yassine Ouzar, Djamaleddine Djeldjli, and Frdric Bousefsaf. Lcoms labs approach to the vision for vitals (v4v) challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 3.4

[83] Rong-Chao Peng, Wen-Rong Yan, Ning-Ling Zhang, Wan-Hua Lin, Xiao-Lin Zhou, and Yuan-Ting Zhang. Cuffless and continuous blood pressure estimation from the heart sound signals. *Sensors*, 15(9):23653–23666, 2015. 2.1

[84] Tania Pereira, Nate Tran, Kais Gadhoumi, Michele M. Pelter, Duc H. Do, Randall J. Lee, Rene Colorado, Karl Meisel, and Xiao Hu. Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digital Medicene*, 2020. 3.4, 4.1

[85] 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), 2010. 2.1.2, 3.1, 3.2, 4.2.1, 4.1, 4.2

[86] CCY Poon and YT Zhang. Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. In *2005 IEEE engineering in medicine and biology 27th annual conference*, pages 5877–5880. IEEE, 2006. 2.1

[87] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst,

Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018. 1

[88] Sakthi Kumar Arul Prakash and Conrad S Tucker. Bounded kalman filter method for motion-robust, non-contact heart rate estimation. *Biomedical optics express*, 9(2), 2018. 3.2, 4.2.1

[89] Jiří Přibil, Anna Přibilová, and Ivan Frollo. Comparative measurement of the ppg signal on different human body positions by sensors working in reflection and transmission modes. *Engineering Proceedings*, 2(1):69, 2020. 2

[90] Jorge Proença, Jens Muehlsteff, Xavier Aubert, and Paulo Carvalho. Is pulse transit time a good indicator of blood pressure changes during short physical exercise in a young population? In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 598–601. IEEE, 2010. 2.2.1

[91] Yuliya S Putyatina. Measurement of arterial blood pressure by processing pulse wave data. In *Proceedings. 3rd Annual Siberian Russian Workshop on Electron Devices and Materials*, volume 2, pages 77–78. IEEE, 2002. 2.2.2

[92] Daniel Quang, Yifei Chen, and Xiaohui Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 2015. 1

[93] Iman Rahimi, Fang Chen, and Amir H Gandomi. A review on covid-19 forecasting models. *Neural Computing and Applications*, pages 1–11, 2021. 1

[94] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[95] Ambareesh Revanur and Laszlo A. Jeni. Camera-based non-contact physiology sensing. In *Affective Computing and Intelligent Interaction demos*, 2022. 1

[96] Ambareesh Revanur, Zhihua Li, Umur A Ciftci, Lijun Yin, and László A Jeni. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 1, 4.1, 4.2.1, 4.4.2

[97] Ambareesh Revanur, Ananyananda Dasari, Conrad Tucker, and László A Jeni. Instantaneous physiological estimation using video transformers. *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence, Studies in Computational Intelligence*, 2022. 4

[98] Mercedes Rigla, Gema García-Sáez, Belén Pons, and Maria Elena Hernando. Artificial intelligence methodologies and their application to diabetes. *Journal of diabetes science and technology*, 12(2):303–310, 2018. 1

[99] Mohammad Sabokrou, Masoud Pourreza, Xiaobai Li, Mahmood Fathy, and Guoying Zhao. Deep-hr: Fast heart rate estimation from face video under realistic conditions. *Expert Systems with Applications*, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.115596. URL `https://www.sciencedirect.com/`

science/article/pii/S0957417421009969. 3.2

[100] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical physics*, 43(12):6654–6666, 2016. 1

[101] Patrick Schoettker, Jean Degott, Gregory Hofmann, Martin Proença, Guillaume Bonnier, Alia Lemkaddem, Mathieu Lemay, Raoul Schorer, Urvan Christen, Jean-François Knebel, et al. Blood pressure measurements with the optibp smartphone app validated against reference auscultatory measurements. *Scientific Reports*, 10(1):1–12, 2020. 2.1

[102] Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of non-invasive blood pressure prediction from ppg and rppg signals using deep learning. *Sensors*, 21(18):6022, 2021. 2.2.2

[103] Almina Šećkanović, Marijana Šehovac, Lemana Spahić, Irma Ramić, Nuraiym Mamatnazarova, Lejla Gurbeta Poković, Almir Badnjević, and Mirsad Kacila. Review of artificial intelligence application in cardiology. In *2020 9th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–5. IEEE, 2020. 1

[104] Dangdang Shao, Yuting Yang, Chenbin Liu, Francis Tsow, Hui Yu, and Nongjian Tao. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. *IEEE Transactions on Biomedical Engineering*, 61(11):2760–2767, 2014. 2.2.1

[105] Dangdang Shao, Chenbin Liu, and Francis Tsow. Noncontact physiological measurement using a camera: A technical review and future directions. *ACS Sensors*, 6(2):321–334, 2021. doi: 10.1021/acssensors.0c02042. URL https://doi.org/10.1021/acssensors.0c02042. PMID: 33434004. 4.5

[106] Gašper Slapničar, Nejc Mlakar, and Mitja Luštrek. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19(15):3420, 2019. 2.2.2

[107] Simon Stent and John Gideon. Estimating heart rate from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021. 3.4, 3.5.1, 3.5.4, 4.4.3

[108] James V Stone. Independent component analysis: an introduction. *Trends in cognitive sciences*, 6(2):59–64, 2002. 2.1.2

[109] Hayato Tada, Noboru Fujino, Akihiro Nomura, Chiaki Nakanishi, Kenshi Hayashi, Masayuki Takamura, and Masa-aki Kawashiri. Personalized medicine for cardiovascular diseases. *Journal of Human Genetics*, 66(1):67–74, 2021. 1

[110] Md Sayed Tanveer and Md Kamrul Hasan. Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ann-lstm network. *Biomedical Signal Processing and Control*, 51:382–392, 2019. 2.2.1

[111] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, João Jorge, DA Clifton, and Chris Pugh. Non-contact video-based vital sign monitoring using ambient light and autoregressive models. *Physiological measurement*, 35(5), 2014. 3.1, 3.2, 4.2.1, 4.4.5, 4.5

[112] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4.1, 4.2.2, 4.3.2

[113] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26), 2008. 1, 2, 3.1, 3.2, 3.4, 3.5, 4.2.1, 4.4.3, 4.1, 4.2

[114] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008. 2.1.1

[115] Aishwarya Visvanathan, Aniruddha Sinha, and Arpan Pal. Estimation of blood pressure levels from reflective photoplethysmograph using smart phones. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–5. IEEE, 2013. 2.1

[116] Johannes Wagner, Jonghwa Kim, and Elisabeth André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE international conference on multimedia and expo*, pages 940–943. IEEE, 2005. 1.1

[117] Edward Jay Wang, Junyi Zhu, Mohit Jain, Tien-Jui Lee, Elliot Saba, Lama Nachman, and Shwetak N Patel. Seismo: Blood pressure monitoring using built-in smartphone accelerometer and camera. In *Proceedings of the 2018 CHI conference on human factors in computing Systems*, pages 1–9, 2018. 2.1, 2.2.1

[118] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7), 2016. (document), 1, 2.1.3, 3.1, 3.5.3, 4.2.1, 4.3, 4.3.1, 4.4.3, 4.1, 4.2

[119] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 1

[120] Naoki Watanabe, Yasuko K Bando, Taiji Kawachi, Hiroshi Yamakita, Kouki Futatsuyama, Yoshikazu Honda, Hisae Yasui, Kazuyuki Nishimura, Takahiro Kamihara, Takahiro Okumura, et al. Development and validation of a novel cuff-less blood pressure monitoring device. *Basic to Translational Science*, 2(6):631–642, 2017. 2.1

[121] CDC WONDER Online Database website. About multiple cause of death, 19992020. In *Centers for Disease Control and Prevention, National Center for Health Statistics*, 2022. 1

[122] Joel A Wolf, Jacqueline F Moreau, Oleg Akilov, Timothy Patton, Joseph C English, Jonhan Ho, and Laura K Ferris. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA dermatology*, 149(4):422–426, 2013. 1

[123] Jedrek Wosik, Marat Fudim, Blake Cameron, Ziad F Gellad, Alex Cho, Donna Phinney, Simon Curtis, Matthew Roman, Eric G Poon, Jeffrey Ferranti, et al. Telehealth transformation: Covid-19 and the rise of virtual care. *Journal of the American Medical Informatics*

Association, 27(6):957–962, 2020. 1.1

[124] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, 31(4), 2012. 2, 4.2.1

[125] Yan Wu, Ruolei Gu, Qiwei Yang, and Yue-jia Luo. How do amusement, anger and fear influence heart rate and heart rate variability? *Frontiers in Neuroscience*, 13, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.01131. URL `https://www.frontiersin.org/article/10.3389/fnins.2019.01131`. 3.3.2

[126] Seungman Yang, Jangjay Sohn, Saram Lee, Joonnyong Lee, and Hee Chan Kim. Estimation and validation of arterial blood pressure using photoplethysmogram morphology features in conjunction with pulse arrival time in large open databases. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1018–1030, 2020. 2.1, 2.2.1

[127] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1

[128] Kun-Hsing Yu and Michael Snyder. Omics profiling in precision oncology. *Molecular & Cellular Proteomics*, 15(8):2525–2536, 2016. 1

[129] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018. 1

[130] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019. 2.1.4, 3.5.1

[131] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 2021. 4.2.2

[132] Aaron M Zakrzewski, Athena Y Huang, Rebecca Zubajlo, and Brian W Anthony. Real-time blood pressure estimation from force-measured ultrasound. *IEEE Transactions on Biomedical Engineering*, 65(11):2405–2416, 2018. 2.2

[133] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, 2017. 3.5.1

[134] 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*, 2016. 3.3

[135] 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. 5.1

[136] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. De-

formable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1