UNIVERSITY OF CALIFORNIA

Los Angeles

Diverse Patient Heart Rate Monitoring Using Consumer Camera Systems

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Electrical and Computer Engineering

by

Pradyumna Venkatesh Chari

2021

ABSTRACT OF THE THESIS

Diverse Patient Heart Rate Monitoring Using Consumer Camera Systems

by

Pradyumna Venkatesh Chari
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2021
Professor Achuta Kadambi, Chair

Real world scenes and objects have diverse visual appearance. Such diversity stems from the fundamental physics in how light interacts with matter, across different weather conditions, object types, and even people. These appearance variations mesmerize human beings, but puzzle artificial vision systems, which cannot generalize to such diversity. Through this thesis, we look at one such case of biased performance over diversity- camera based remote heart rate (HR) estimation. HR is an essential clinical measure for the assessment of cardiorespiratory instability. The growing telemedicine market opens up the urgent requirement for scalable yet affordable remote HR estimation. However, existing computer vision methods that estimate HR from facial videos exhibit biased performance against dark skin tones. This is a major concern, since communities of color are disproportionately affected by both COVID-19 and cardiovascular disease. We identify and model the origin of this bias and present a novel physics-driven algorithm that boosts performance on darker skin tones in our reported data. We assess the performance of our method through the creation of the first telemedicine-focused remote vital signs dataset, the VITAL dataset. 432 videos (864 minutes) of 54 subjects with diverse skin tones are recorded under realistic scene conditions with corresponding vital sign data. Our method mitigates errors due environmental conditions and imparts unbiased performance gains across skin tones, setting the stage for making non-contact HR sensing technologies a viable reality for patients across skin tones.

The thesis of Pradyumna Venkatesh Chari is approved.

Jonathan Kao

Stefano Soatto

Achuta Kadambi, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to begin by acknowledging my advisor Prof. Achuta Kadambi. It has been through his guidance, motivation, kind help and hours or conversation regardless of the time, that this thesis has taken the exciting shape and direction it currently holds and for that, I am thankful to him.

I would next like to thank Prof. Laleh Jalilian. Her guidance and help in setting together as daunting a task as collecting physiological data on diverse participants was critical in us achieving the novelly diverse data that we currently have.

I would like to thank Prof. Jonathan Kao and Prof. Stefano Soatto, as part of my thesis committee, for their support of this thesis. I am grateful for their time and help in completing this work.

I would like to specifically thank Krish Kabra for all the additional hours of work that we have together put into getting this manuscript into shape and towards publication. Working with Krish has been a great learning experience for me, especially on a project of this scale with so many moving parts. He has played a very significant role in shaping this project.

I would also like to thank all my other co-authors and colleagues on this work (Doruk Karinca, Soumyarup Lahiri, Diplav Srivastava, Kimaya Kulkarni, Tianyuan Chen and Maxime Cannesson) for the help, assistance and guidance.

I would finally like to thank all my family, friends and colleagues, for their support throughout. The last year has been daunting for all of us in many different ways, and I am privileged to enjoy the love and care from all of you.

PREVIOUS PUBLICATIONS

This thesis revises the following publication:

P. Chari, K. Kabra, D. Karinca, S. Lahiri, D. Srivastava, K. Kulkarni, T. Chen, M. Cannesson, L. Jalilian, and A. Kadambi, "Diverse R-PPG: Camera-based heart rate estimation for diverse subject skin-tones and scenes," *arXiv preprint arXiv:2010.12769* (2020) [1].

CHAPTER 1

Introduction

Heart rate (HR) is an important clinical measure in the evaluation of cardiorespiratory and hemodynamic stability. Conventional HR assessment is performed in-person at a clinic or hospital using specialized monitoring equipment. However, the COVID-19 pandemic has accelerated the adoption of healthcare delivery to a remote model that uses telemedicine and mobile health (mHealth) technologies for patient evaluations [2, 3, 4] in order to protect patients and healthcare workers from infectious exposure in a pandemic setting. The assessment of HR in patients with suspected COVID-19 is particularly important as COVID-19 has been associated with pre-existing cardiovascular disease [5]. Given the clinical relevance of HR in triage decisions, diagnosis, prognosis, and as a criterion for transfer to higher-level medical care, there is a pressing need to develop HR sensing solutions that can facilitate the rapidly growing domain of telemedicine-based care and remote patient monitoring.

Presently, HR sensing solutions for telemedicine and remote patient monitoring have relied on the adoption of wearable sensors to make plethysmographic or electrocardiographic measurements [6, 7]. Although such wearable technologies have seen major advances in the past decade [8, 9], they still require major expenditure on production and distribution of hardware. This expense can create a barrier to adoption of mHealth technologies that disproportionately affects rural and socioeconomically burdened communities [10].

In contrast to wearable sensors, recent methods have proposed using camera-based hardware present on modern-day smartphones to estimate key vitals including HR. Contact-based methods, where the finger is typically placed overtop the camera module, have already seen widespread applications in major smartphones [11, 12]. Despite such methods showing good performance, their long-term practicality for telemedicine video-conferencing visits is potentially limited as the camera module is covered during measurement. This prevents continuous monitoring of patient HR, visual well-being, and collection of other vitals such as respiratory rate and spatial blood perfusion maps.

Contactless methods have also been proposed, in which computer vision algorithms and artificial intelligence (AI) tools are used to remotely extract a blood volume pulse (BVP) signal and corresponding HR estimate from facial videos [11, 13, 14]. Of these methods, remote photoplethysmography (r-PPG) is one of the most promising.

Early work conducted by Verkruysse *et al.* [15] showed that plethysmographic signals could be measured using ambient light and a consumer-grade digital camera. In order to accurately isolate and extract the correct BVP signal corresponding to the HR, several R-PPG algorithms have been proposed, including blind source separation (BSS) [13, 16, 17], model-based [18, 19, 20, 21], unsupervised data-driven [22, 23], and supervised deep learning [24, 25, 26, 27, 28, 29] methods. Unfortunately, the performance of existing R-PPG algorithms fluctuates with changes in illumination condition [30], subject motion [31, 21, 32], and skin tone [33]. We are specifically interested in the notion of skin tone dependent performance bias in r-PPG. Figure 1.1 shows the worldwide distribution of skin color among indigenous populations. This further establishes the fraction of worldwide populations that are inconvenienced and disadvantaged as a result of inequitable technology.

Moreover, assessment of these algorithms has typically been done on computer vision datasets that are not focused on telemedicine applications. Consequently, these datasets do not represent characteristics that are important for clinical translation such as a large population with diverse skin tone and gender representation and video data collection on end-user devices such as smartphones.

Through this thesis, we provide the first steps at telemedicine translation of contactless camera-based HR sensing technologies for smartphone deployment. We propose a novel r-

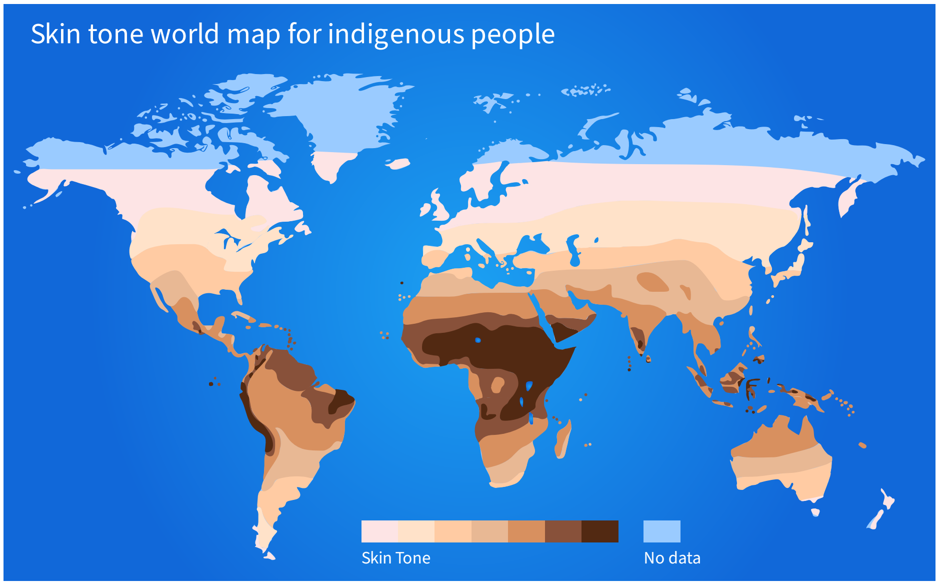


Figure 1.1: **A map of indigenous skin tone distribution across the world.** Any viable sensing technology must be able to work at comparable accuracy for the entire spectrum of skin tones.

PPG algorithm that specifically addresses mitigating bias for skin tone. In contrast to prior approaches, this work first establishes a theoretical framework to understand the unique physics that underlies the inconsistency in r-PPG measurement. We establish that the bias is due to imaging noise, and appropriately propose r-PPG denoising methods to alleviate performance losses. To assess the performance of the proposed method, we collect the first remote vital signs detection dataset focused on telemedicine applications that is demographically diverse.

CHAPTER 2

Theory

2.1 Light Transport for R-PPG

Plethysmographic estimation methods are enabled through the sensing of blood perfusion in the face. Specifically, the presence of varying volumes of blood under the skin manifest as minute changes in reflection properties of the overall skin system, as viewed by a camera. It is by identifying these changes that relevant physiological properties may be estimated.

In order to set up a novel light transport theory for r-PPG, we utilize existing biorealistic graphical rendering models [34] and extend them for r-PPG signal generation. Figure 2.1 shows the skin model assumed for our computations, similar to [35]. Specifically, a two layer skin model is assumed. The incident light undergoes attenuation while passing through the epidermis, while it undergoes scattering driven reflection at the dermis.

We start with describing the epidermal transmission. Following the Beer-Lambert Law,

$$\mathbf{T_{epi}}(\boldsymbol{\lambda}) = \mathbf{e}^{-\boldsymbol{\mu}_{\mathbf{a,epi}}(\boldsymbol{\lambda})}, \tag{2.1}$$

Where $\boldsymbol{\mu}_{\mathbf{a,epi}}(\boldsymbol{\lambda})$ is the absorption coefficient of the epidermis. Typically, this is modelled as a convex combination of skin tissue and melanin absorption,

$$\boldsymbol{\mu}_{\mathbf{a,epi}}(\boldsymbol{\lambda}) = \mathbf{f_{mel}}\boldsymbol{\mu}_{\mathbf{a,mel}}(\boldsymbol{\lambda}) + (\mathbf{1} - \mathbf{f_{mel}})\boldsymbol{\mu}_{\mathbf{a,ski}}(\boldsymbol{\lambda}). \tag{2.2}$$

$\boldsymbol{\mu}_{\mathbf{a,ski}}(\boldsymbol{\lambda})$, the skin tissue absorption coefficient, is a biological parameters which is known. $\boldsymbol{\mu}_{\mathbf{a,mel}}(\boldsymbol{\lambda})$ may be defined as,

$$\boldsymbol{\mu}_{\mathbf{a,mel}}(\boldsymbol{\lambda}) = \mathbf{f_{eum}}\boldsymbol{\mu}_{\mathbf{a,eum}}(\boldsymbol{\lambda}) + (\mathbf{1} - \mathbf{f_{eum}})\boldsymbol{\mu}_{\mathbf{a,phm}}(\boldsymbol{\lambda}), \tag{2.3}$$

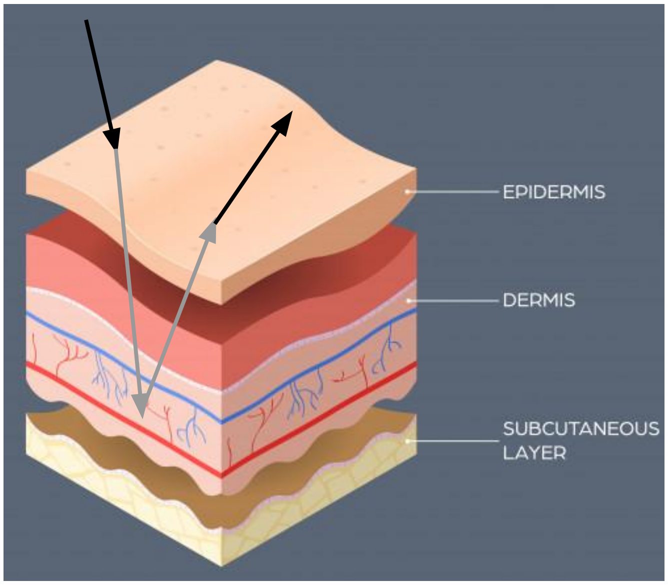


Figure 2.1: **A two-layer skin model used in prior biorealistic rendering works is used to develop the light transport theory for R-PPG.** The incident light ray attenuates through the epidermis. Following dermal reflection and another epidermal attenuation, the resultant ray properties are dependent on human physiological quantities.

Where $\boldsymbol{\mu}_{\mathbf{a,eum}}(\boldsymbol{\lambda})$ is the absorption coefficient of eumelanin and $\boldsymbol{\mu}_{\mathbf{a,phm}}(\boldsymbol{\lambda})$ is the absorption coefficient of pheomelanin, all biophysical known parameters. By combining Equations 2.1, 2.2 and 2.3, the epidermal transmission may be accurately modelled.

We move towards describing the dermal reflection. This model follows the Kubelka-Munk theory for scattering-dependent reflection. Specifically, the fraction of reflected light, as a function of wavelength, is given by,

$$\mathbf{R_d}(\boldsymbol{\lambda}) = \frac{(\mathbf{1} - \boldsymbol{\beta}(\boldsymbol{\lambda}))^{\mathbf{2}}(\mathbf{e}^{\mathbf{K}(\boldsymbol{\lambda})\mathbf{d_{der}}} - \mathbf{e}^{-\mathbf{K}(\boldsymbol{\lambda})\mathbf{d_{der}}})}{(\mathbf{1} + \boldsymbol{\beta}(\boldsymbol{\lambda}))^{\mathbf{2}}\mathbf{e}^{\mathbf{K}(\boldsymbol{\lambda})\mathbf{d_{der}}} - (\mathbf{1} - \boldsymbol{\beta}(\boldsymbol{\lambda}))^{\mathbf{2}}\mathbf{e}^{-\mathbf{K}(\boldsymbol{\lambda})\mathbf{d_{der}}}} \tag{2.4}$$

Here, $\boldsymbol{\beta}(\boldsymbol{\lambda})$ and $\mathbf{K}(\boldsymbol{\lambda})$ are deterministically related to $\boldsymbol{\mu}_{\mathbf{a,der}}(\boldsymbol{\lambda})$ (dermal absorption coefficient) and $\boldsymbol{\mu}_{\mathbf{s,der}}(\boldsymbol{\lambda})$ (reduced dermal scattering coefficient, known [36]). Similar to previously, the dermal absorption coefficient and the blood absorption coefficient are understood as convex combinations shown below:

$$\boldsymbol{\mu}_{\mathbf{a,der}}(\boldsymbol{\lambda}) = \mathbf{f_{bld}}\boldsymbol{\mu}_{\mathbf{a,bld}}(\boldsymbol{\lambda}) + (\mathbf{1} - \mathbf{f_{bld}})\boldsymbol{\mu}_{\mathbf{a,ski}}(\boldsymbol{\lambda}) \tag{2.5}$$

$$\boldsymbol{\mu}_{\mathbf{a,bld}}(\boldsymbol{\lambda}) = \mathbf{f_{oxy}}\boldsymbol{\mu}_{\mathbf{oxy}}(\boldsymbol{\lambda}) + (\mathbf{1} - \mathbf{f_{oxy}})\boldsymbol{\mu}_{\mathbf{dox}}(\boldsymbol{\lambda}) \tag{2.6}$$

Here, various factors include blood reflection, skin baseline reflection, oxygenated blood reflection and deoxygenated blood reflection respectively.

Given the expressions for epidermal transmission and dermal reflection, the expression for overall reflection is given by,

$$\mathbf{R}(\boldsymbol{\lambda}) = \mathbf{T^2}_{\mathbf{epi}}.\mathbf{R_d}(\boldsymbol{\lambda}). \tag{2.7}$$

Then, the overall intensity captured in channel $\mathbf{c}$ of the camera is given by,

$$\mathbf{I_c} = \int_{\boldsymbol{\lambda}} \mathbf{E}(\boldsymbol{\lambda})\mathbf{S_c}(\boldsymbol{\lambda})\mathbf{R}(\boldsymbol{\lambda})\mathbf{d}\boldsymbol{\lambda}, \tag{2.8}$$

Where $\mathbf{E}(\boldsymbol{\lambda})$ is the source spectral distribution and $\mathbf{S_c}(\boldsymbol{\lambda})$ is the camera spectral response for channel $\mathbf{c}$.

2.2 R-PPG Signal Strength

The R-PPG signal arises out of a variation in the blood volume fraction, $\mathbf{f_{bl}}$ under the skin. Our interest is in the signal strength across camera channels, $\boldsymbol{\Sigma}_c$, which can be defined as *the maximum variation in the captured intensity.* Mathematically,

$$\boldsymbol{\Sigma}_\mathbf{c} = \boldsymbol{\Delta}\mathbf{I_c} \approx \left|\frac{\boldsymbol{\partial}\mathbf{I_c}}{\boldsymbol{\partial}\mathbf{f_{bl}}}\right| \cdot \boldsymbol{\Delta}\mathbf{f_{bl}} \tag{2.9}$$

Since $\mathbf{R}(\boldsymbol{\lambda})$ is the only term dependent on $\mathbf{f_{bl}}$,

$$\boldsymbol{\Sigma}_\mathbf{c} \approx \left|\int_{\boldsymbol{\lambda}} \mathbf{E}(\boldsymbol{\lambda})\mathbf{S_c}(\boldsymbol{\lambda})\left.\frac{\boldsymbol{\partial}\mathbf{R}}{\boldsymbol{\partial}\mathbf{f_{bl}}}\right|_{\overline{\mathbf{f_{bl}}}}\mathbf{d}\boldsymbol{\lambda}\right| \cdot \boldsymbol{\Delta}\mathbf{f_{bl}}, \tag{2.10}$$

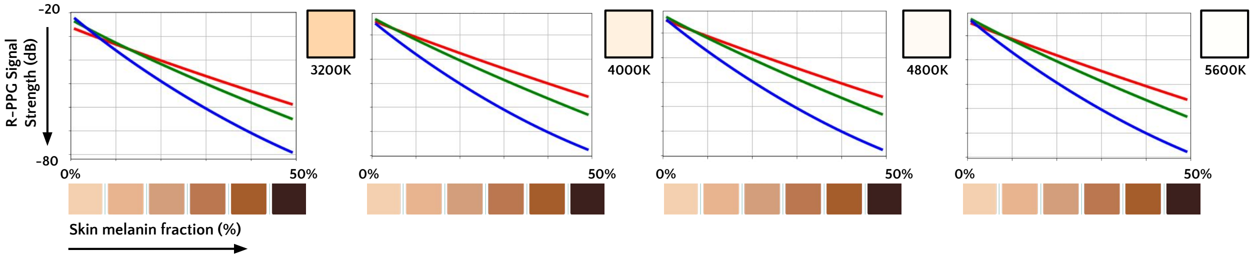


Figure 2.2: **The R-PPG signal strength is critically related to skin melanin fraction as well as scene lighting.** As opposed to previously accepted fact, the three channels may contain differing amounts of signal information, depending on regime of operation.

Where $\overline{\mathbf{f_{bl}}}$ is the average blood volume fraction, typically around 0.05. This approximation holds true since $\mathbf{f_{bl}}$ only varies by a small amount, typically around 0.05.

This plethysmographic signal rides on top of the average skin tone color, given by

$$\mathbf{\Gamma_c} = \int_{\boldsymbol{\lambda}} \mathbf{E}(\boldsymbol{\lambda})\mathbf{S_c}(\boldsymbol{\lambda})\mathbf{R}(\boldsymbol{\lambda})\Big|_{\overline{\mathbf{f_{bl}}}} \mathbf{d}\boldsymbol{\lambda}. \tag{2.11}$$

Since, $\mathbf{\Sigma}_c$ and $\mathbf{\Gamma}_c$ are both dependent on $\mathbf{f_{mel}}$, as a result of the dependence of $\mathbf{R}(\cdot)$ on the same, we refer to these as $\mathbf{\Sigma}(\mathbf{f_{mel}})$ and $\mathbf{\Gamma}(\mathbf{f_{mel}})$ subsequently.

Figure 2.2 shows the signal strength plots for the three camera color channels, across lighting conditions. We use average camera response functions $\mathbf{S_c}(\boldsymbol{\lambda})$ to identify responsiveness of each of the channels to incident light. We also generate signal strength across common light source characteristics. These plots provide incisive detail: the overall signal strength decays with increasing skin melanin fraction. Additionally, while previous works [15, 18, 19] have empirically determined that the green channel holds maximum R-PPG signal information, we show for the first time that this in-fact heavily depends on melanin fraction and scene lighting. While the green channel is dominant for light skin tones, for darker skin tones, the channel-wise signal strength depends significantly on lighting conditions and skin tone.

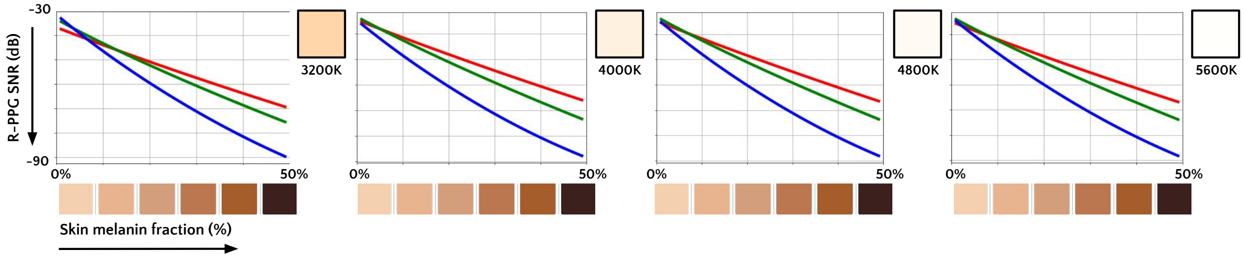


Figure 2.3: **The R-PPG SNR drastically worsens with increasing skin melanin fraction.** As expected, the R-PPG SNR reduces by orders of magnitude as the skin melanin fraction increases. Therefore, mitigating the skin tone bias present in R-PPG will require strategies that emphasize capturing more signal and reducing noise.

2.3 Effect of Imaging Noise on R-PPG

The goal of this subsection is to understand the relationship between imaging noise and R-PPG algorithm estimation. Imaging noise refers to the inherent noise that arises due to the image capture process in a commercial camera. This arises due to various effects related to photon arrival processes, thermal noise in electronics and the quantization noise associated with digitally capturing images [37]. For pixels below the saturation level, the noise can be modelled as follows:

$$\sigma_{pixel}^2 = \frac{\Phi t}{g^2} + \frac{\sigma_r^2}{g^2} + \sigma_q^2 \tag{2.12}$$

where Φ is the radiant power of light collect, t is the exposure time, g is the sensor gain (a constant for a given image), and σ_r and σ_q are camera noie parameters (also constant).

Using this noise model, we can the estimate the entire R-PPG signal to noise ratio (SNR) for a pixel of a particular intensity and color channel c as follows:

$$\mathbf{SNR_c} = \frac{\mathbf{\Sigma_c} t}{\sqrt{\frac{\mathbf{\Gamma_c} t}{g^2} + \frac{\sigma_r^2}{g^2} + \sigma_q^2}} \tag{2.13}$$

Here, we assume that the radiant power of light collected Φ is equal to the average skin tone color.

Figure 2.3 shows the R-PPG SNR plots for the three camera color channels, across lighting conditions. These observations are similar to those of the R-PPG signal strength, namely that the SNR decays with increasing skin melanin fraction. This leads us to the following inferences:

(i) **Imaging noise creates skin tone bias (and lighting bias):** The performance gap across skin tones, as well as across lighting differences, can be understood in terms of imaging noise. Darker skin regions have lower signal strength that manifest as lower pixel value changes in the video. This results in poorer SNRs. Note that this inference also holds true for shadowed regions, thereby extending this analysis towards understanding lighting bias.

(ii) **Imaging noise and specular reflections degrade the r-PPG signal:** Imaging noise, coupled with specular highlights due to lighting, are the major contributing factors to signal degradation. The corruption due to imaging noise depends on signal intensity. The corruption due to specular highlights depends on lighting conditions-regions with strong specular highlights have relatively lower PPG signal information. Combating the highlighted biases in existing r-PPG would therefore involve a principled approach towards reduction of the above highlighted imaging noise and specular highlight removal. Note that specular highlight removal, in addition to reducing lighting related biases, also indirectly affects skin tone bias: darker skin subjects are worse affected by these interferences, since the intensity difference between the signal and the highlight is much more.

We conclude that addressing this low-level light transport bias must occur in order to drastically mitigate the skin tone bias present in R-PPG. Biases higher up the chain of biases, such as algorithmic or dataset bias, must also be addressed, but may not necessarily overcome this fundamental physics-based problem. Therefore, image and signal processing strategies to increase signal capture and reduce noise may drastically improve performance for darker skin tone subjects as opposed to modifications to signal inference algorithms. With the inferences from this chapter in mind, we motivate our novel R-PPG algorithm outlined in the following Methods section.

CHAPTER 3

Methods

3.1 The typical R-PPG pipeline

There are four components to a typical r-PPG pipeline: (a) detection, which identifies facial regions of interest in the video frame, (b) combination, which condenses the information from regions of interest into a RGB time series signal, (c) signal inference, which uses the time series signal to estimate the pulse volume waveform, and (d) HR estimation, which estimates the HR from the pulse volume signal. This is visually described in Fig. 3.1.

The video is first passed through a neural network-based face detector [38], in order to identify the face region in the frame. Using feature point detectors [39], the eye and mouth regions are identified and explicitly removed from the videos (since these regions do not contribute to the pulsatile signal). This is the detection step. The next steps, namely combination, inference and HR step, are carried out for smaller video-windows of 10 seconds length with an overlap of 5 seconds.

For each video frame, the skin pixels are combined to get one RGB sample for that time instance (the methods for this combination vary across papers and is the crux of this work's novelty). Across all frames, after this combination, we obtain a time series RGB signal. This is the combination step.

These RGB signals are then put through an existing signal inference technique. In this paper, we use the CHROM algorithm [18] due to its versatility, as well as its easy access from openly available code [40]. The output obtained from this step results in a pulsatile waveform estimate for each window. This is the inference step.

The obtained pulsatile waveform is then processed to arrive at the final HR. This is the heart rate step. We first filter the waveform using a Butterworth bandpass filter with pass band frequencies of [0.7, 3.5] Hz. The power spectral density (PSD) is then computed. Temporal frequency artifacts were empirically observed in the original video as a result of aggressive compression, likely due to the unchanging green background. These erroneous peaks were appropriately removed. Next, the five highest peaks in the PSD are chosen. The peak with the highest combined fundamental and second harmonic power is chosen as the one corresponding to the HR. The final HR for the video is estimated as the average of the HR estimates for each 10 second window.

3.2 Analysis of existing methods

In order to understand the origin of the performance bias, for the first time, we theoretically analyze the r-PPG measurement process and the role of imaging noise using biophysical first principles (see Supplementary Materials for details). From this, we note three key observations: (i) Imaging noise creates skin tone bias (and lighting bias), (ii) imaging noise and specular reflections degrade the r-PPG signal, and (iii) denoising is to be done before signal inference.

This sets the stage for understanding how existing algorithms improve the noise performance in the combination step. The most straightforward approach is to simply average all face pixels in a frame to arrive at time samples of the RGB signal. We refer to this as facial aggregation [13, 18, 22, 19, 17, 21]. To improve upon this, previous approaches have sought to modify this averaging process. We describe the best performing result amongst these on the VITAL dataset. The face is gridded into smaller rectangular regions. Pixels within each region are averaged to arrive at individual time series for each region. Each of these gridded temporal signals is passed through the inference step, to obtain the corresponding blood volume signal estimate. Approaches use measures such as SNR at peak frequency of this blood volume signal to characterize the 'goodness' of each signal [18, 41, 42, 43, 44], with higher weights being assigned for better signals. As mentioned previously, in this paper we use the two harmonic SNR estimate, which was found to be more robust. That is, for a signal s (frequency domain S) with a HR p, the SNR at the HR frequency is given by:

$$SNR = \frac{\int_{p-w}^{p+w} |S(f)|^2 df + \int_{2(p-w)}^{2(p+w)} |S(f)|^2}{\int_{-\infty}^{\infty} |S(f)|^2 df - \int_{p-w}^{p+w} |S(f)|^2 df - \int_{2(p-w)}^{2(p+w)} |S(f)|^2} \tag{3.1}$$

where w is the peak window size for estimation (for this work's experiments, we use $w = 0.1Hz$). This resultant signal is passed to the HR step. We call this method SNR weighting [41, 42, 43, 44]. Finally, these weights are used to average the blood volume signals together.

A few key issues arise with the SNR weighting method. Firstly, we empirically observe that the weight maps from previous methods (based on region-based SNR estimates) have the tendency to be sparse, especially for darker skin tones. Therefore, the expected improvements due to weighted averaging are lost to noise corruption for darker skin tone subjects since much lesser signal is being aggregated. This poorer denoising for darker skin subjects results in worse SNRs, thereby degrading performance. Datasets on which these previous methods were tested were not as diverse across skin tones: these performance caveats were therefore missed. Secondly, the previous method of SNR weighting may also fall prey to specular highlights. With these, the signal contains no information of the pulsatile signal, which gets buried in the light from the source. This is a considerable factor when looking at scene conditions, such as camera angle, lighting direction, lighting color and intensity, as well as skin tone. Previous weighting approaches do not explicitly take this into account and use the gridded weighting method to implicitly combat these highlights. However, since the nature of this gridding itself degrades for darker skin tones, we observe that specular effects must be directly addressed. Finally, the SNR weighting performs denoising after signal inference,

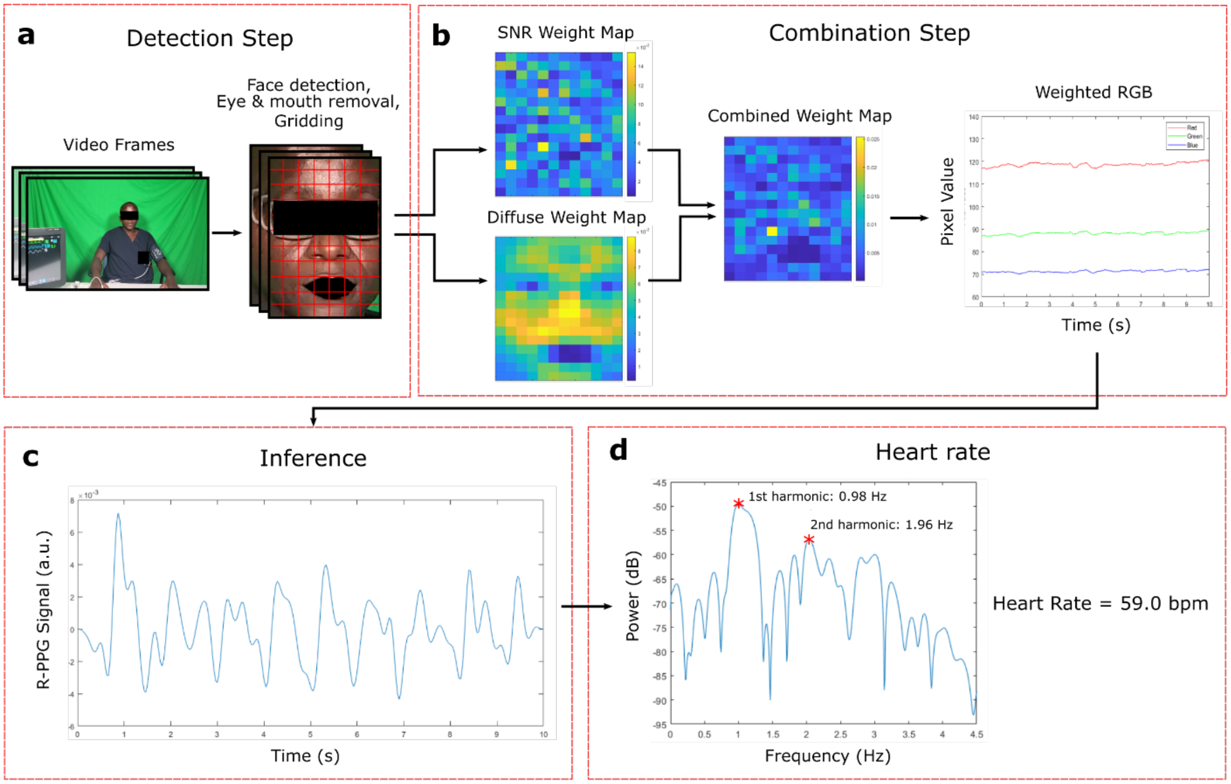


Figure 3.1: **The proposed heart rate estimation algorithm consists of four steps.** The proposed novelty in the combination step of the pipeline incorporates skin diffuse information weighting, in addition to SNR weighting in RGB space, to achieve robust r-PPG performance across skin tones. Written consent was obtained from the subject for using their image in the publication.

as opposed to before. Given that the inference method (CHROM [18]) is non-linear, such a weighting regime may not be the most optimal.

3.3 Novel modifications

Having identified the reasons for poor performance of existing methods, we propose novelties to be incorporated in the combination step, that look to achieve a performance gain in a manner that is fair across skin tones. We focus our novelties to this step since the origin of the performance bias is the image SNR. In order to move towards debiasing, it is critical that major modifications are applied during combination, so that the effect of noise during inference is minimized. This also allows for the proposed modifications to be applied independent of the inference algorithm, thereby making the modifications more generally applicable.

Specifically, we propose two major novelties: (i) weighting in RGB space, rather than blood volume signal space and (ii) skin diffuse component weighting.

- RGB-space weighting: Existing spatial averaging methods estimate weights for each grid region, based on the blood volume signal quality [41, 42, 43, 44]. Instead of using these estimated weights to average the blood volume signals, as done in previous methods, we propose using these weights to average in RGB space. As a result, we obtain one consolidated SNR weighted RGB signal, which is again passed through the inference step to obtain the final blood volume signal.

  The motivation for this can be understood in the context of noise. Averaging the RGB signal results in a less noisy signal passing through the inference step, enabling the inference method to provide better estimates, as compared to when noisier signals are passed through the method, to be averaged later. If the inference method is non-linear (such as CHROM [18]), a pre-weighting would lead to additional noise performance gain.

- Skin diffuse component weighting: An image can be split into two constituent components: the diffuse component, that arises out of transmission and reflection through the skin, and a specular component, that arises from mirror-like surface reflections. Since the diffuse component contains the signal of interest for us, we utilize gridded diffuse components as additional weights. For each frame, the diffuse component is estimated [45]. It is then gridded and averaged across the grid dimensions and time, in order to arrive at weights for each grid element.

The diffuse weights play two key roles in improving bias in performance as well as overall performance: first, they can remove specular affected regions from the average explicitly. Second, they combat the sparsity issue observed in traditional SNR weights, since the diffuse component is continuous and non-sparse. The SNR weights and the novel diffuse weights are multiplied together and renormalized to arrive at the final spatial weights for the gridded video.

The overall pipeline, therefore, involves using the novel weights together, to arrive at efficiently weighted RGB signals. These are averaged together and passed through the estimation step and HR step. This pipeline is visually highlighted as such in Fig. 3.1.

3.4 VITAL Dataset

To validate the performance of camera-based vital sign detectors, we construct the Vital-sign Imaging for Telemedicine AppLications (VITAL) dataset. The focus of this dataset is to represent diversity in factors that are relevant to telemedicine setups, including: (i) smartphone deployment, (ii) camera view angle, (iii) recording condition (lighting variation and talking), and (iv) patient demographic diversity. We address each of these aspects individually:

(i) **Smartphone deployment:** The ubiquity of smartphones globally has led to the development of patient portals, many of which can be accessed via smartphone applications that can be downloaded by patients [46, 47, 48]. Such applications have been used for hosting telemedicine appointments. A deployable remote HR estimation solution with a focus on telemedicine must be able to work efficiently on smartphone cameras by considering factors including video compression [26, 49, 50] and algorithmic complexity. Moreover, the solution must achieve success independent of camera type. Hence, the VITAL dataset uses different smartphone cameras for each view angle. The use of more than one smartphone imager inspires the development of algorithms that

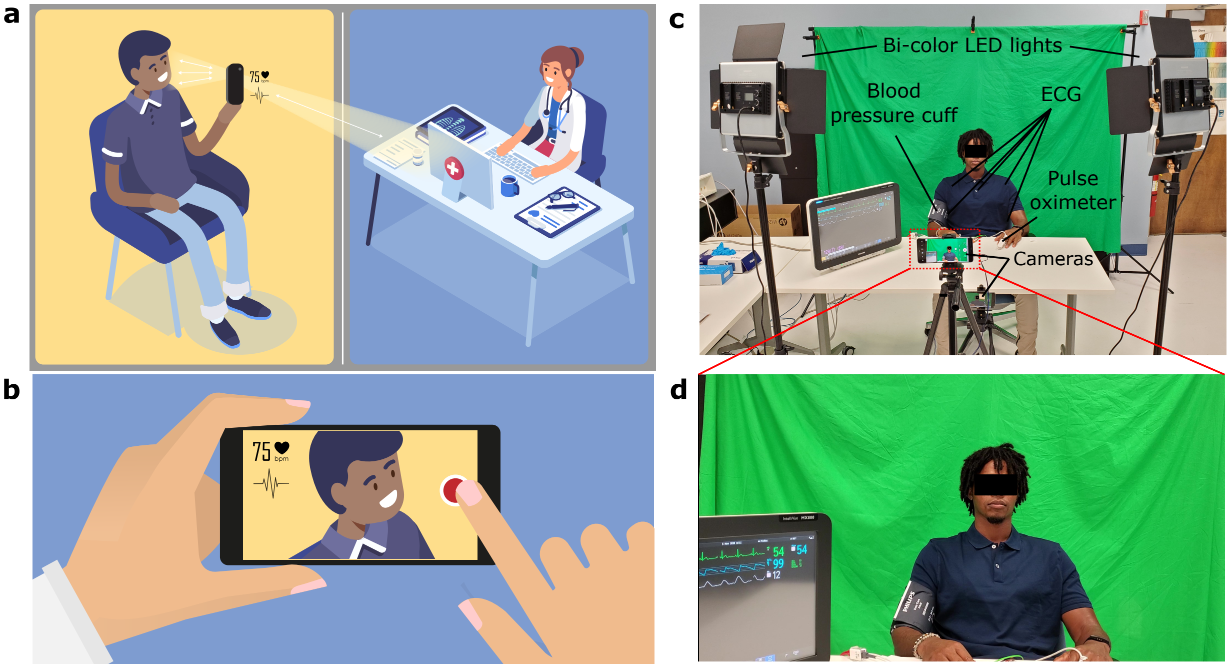


Figure 3.2: **Constructing a diverse remote vital sign monitoring dataset with a focus on telemedicine applications.** (a) Cartoon schematic depicting the telemedicine application for the proposed camera-based heart rate estimation. (b) Telemedicine video–conferencing applications can be integrated with a software toolkit to display patient BVP and HR. (c) Experimental setup employed during the construction of the VITAL dataset. Two bi-color LEDs are used for controlled illumination of the subject, and laboratory tube LEDs are used for ambient illumination. The Philips IntelliVue MX800 patient monitor is utilized for ground truth vital sign monitoring. Two smartphone cameras at differing viewing angles capture video of the subject. (d) Example frame from video captured by the smartphone camera. The subject wears a blood pressure cuff, 5-ECG leads, and a finger pulse oximeter, which is connected to the MX800 unit. Written consent was obtained from the subject for using their image in the publication.

can scale to a variety of device-agnostic telemedicine conditions.

(ii) **Camera view angle:** In a telemedicine setting, there can also be a variety of camera angles that the algorithm must work on. In order to facilitate this verification, the VITAL dataset consists of two camera view angles for all the videos of each subject (as seen in Figure 3.2).

(iii) **Recording condition:** Another essential factor involves testing algorithms across a range of recording conditions, to promote the development of algorithms that can operate in the "wild". The dataset consists of four recording conditions: (1) controlled lighting at 5600K ("cool" lighting) with the subject remaining stationary, (2) controlled lighting at 3200K ("warm" lighting) with the subject remaining stationary, (3) ambient room lighting- distributed white lighting- with the subject remaining stationary, and (4) ambient room lighting with the subject speaking. Additionally, a green screen backdrop is kept to potentially enable digital modification of background scenery.

(iv) **Patient demographic diversity:** The VITAL dataset consists of 54 subjects spread across skin tone, age, gender, race, and ethnic backgrounds. Subject characteristics (gender, age, height, weight, body mass index, race, and ethnicity) are summarized in Table 4.1 using mean (SD), median (IQR), or frequency (%), unless otherwise noted. For the purpose of this study, we split the subjects into three skin tone categories based on the Fitzpatrick (FP) skin type scale [51]: light, consisting of skin tones in the FP 1 and 2 scales, medium, consisting of skin tones in the FP 3 and 4 scales, and dark, consisting of skin tones in the FP 5 and 6 scales. This aggregation allows for more relevant trends, since any two consecutive FP scale categories are reasonably close.

The human study protocol was approved by the UCLA Institutional Review Board (IRB#20-001025-AM-00001), and participants provided written informed consent to take part in the study. Figure 3.2 shows the data collection setup. Each subject is made to sit on a height-adjustable chair, in the field of view of two cell-phone cameras (with different view angles): one camera (Samsung Galaxy S10) is perfectly front-on, while the other (Samsung Galaxy A51) is directly in front of the face, at a dip (lower) of 15 degrees. The front-on camera is placed approximately 130 cm from the subject, and the lower camera at a dip is approximately 90 cm from the subject. The height of the chair is chosen so that the subject is centered in the front-on frame. The controlled lights are set up on either side of the front-on camera, with a baseline of 100 centimeters between them.

As aforementioned, we record subjects using these cameras under four different scene conditions: (1) controlled lighting at 5600K ("cool" lighting) with the subject remaining stationary, (2) controlled lighting at 3200K ("warm" lighting) with the subject remaining stationary, (3) ambient room lighting (distributed white LED lighting) with the subject remaining stationary, and (4) ambient room lighting with the subject speaking. Controlled lighting is enabled by a pair of professional bi-color LED photography lights (Neewer Bi-Color 480 LED). The controlled lighting recording conditions were enabled with the room lights off, allowing for fine-tuned control over the illumination spectral properties. As incorporating controlled lighting only enables a front-facing illumination angle, two recording conditions in ambient room lighting were captured where the subject was lit more completely from several angles. The final recording condition involved variations in the subject, including talking, natural head movements, and facial expressions. Each scene recording session lasts for 2 minutes, for a total of 16 minutes of video footage across 8 videos.

During data collection, volunteers are fitted with standard anesthesiology cardiopulmonary monitors: pulse oximeter (Red DCI, Masimo), blood pressure cuff (Comfort Care, Philips), and 5-lead electrocardiogram (Philips IntelliVue). To collect vital sign data, we utilize the Philips IntelliVue MX800 patient monitor to perform real time monitoring of four vital signs- HR, respiratory rate, oxygen saturation, and non-invasive continuous blood pressure- of which three waveforms are collected (ECG, PPG and respiration). We use the open source tool VSCapture [52] to collect data onto a computer using the MX800's local area network communication protocol. The MX800's estimated numeric values for the vital signs are sampled every 1 second, while the waveforms are sampled at variable frequencies. The ECG signal is sampled between 400-600 Hz, the PPG signal between 100-150Hz and the respiration between 40-60Hz. Continuous non-invasive blood pressure estimates occur when the blood pressure cuff is activated, which is approximately once every 30 seconds.

3.5 Benchmark methods and techniques

To benchmark the performance of the proposed method, we compare the proposed method against previous remote HR estimation algorithms. All methods and techniques used are outlined in detail in the Materials and Methods section. We choose the CHROM [18] signal extraction method due to its versatility and open availability of code [40]. We compare with the two most common categories of algorithmic processing steps, which we refer to as facial aggregation [13, 18, 22, 19, 17, 21] and SNR weighting [41, 42, 43, 44]. We believe that these two processing steps regimes encapsulate the major processing philosophies used in existing r-PPG methods.

To ensure a fair comparison with the benchmark methods, we implement identical testing conditions across techniques. For each method, the input video is passed through the same face detection algorithm (convolutional neural network-based detector [38]), following which the eyes and mouth are cropped out using facial feature points [39]. Some methods also use skin segmentation algorithms [32, 53, 54], but we empirically found this to perform slightly worse on the VITAL dataset. We also use a consistent HR selection technique for each method.

3.6 Statistical analysis tools

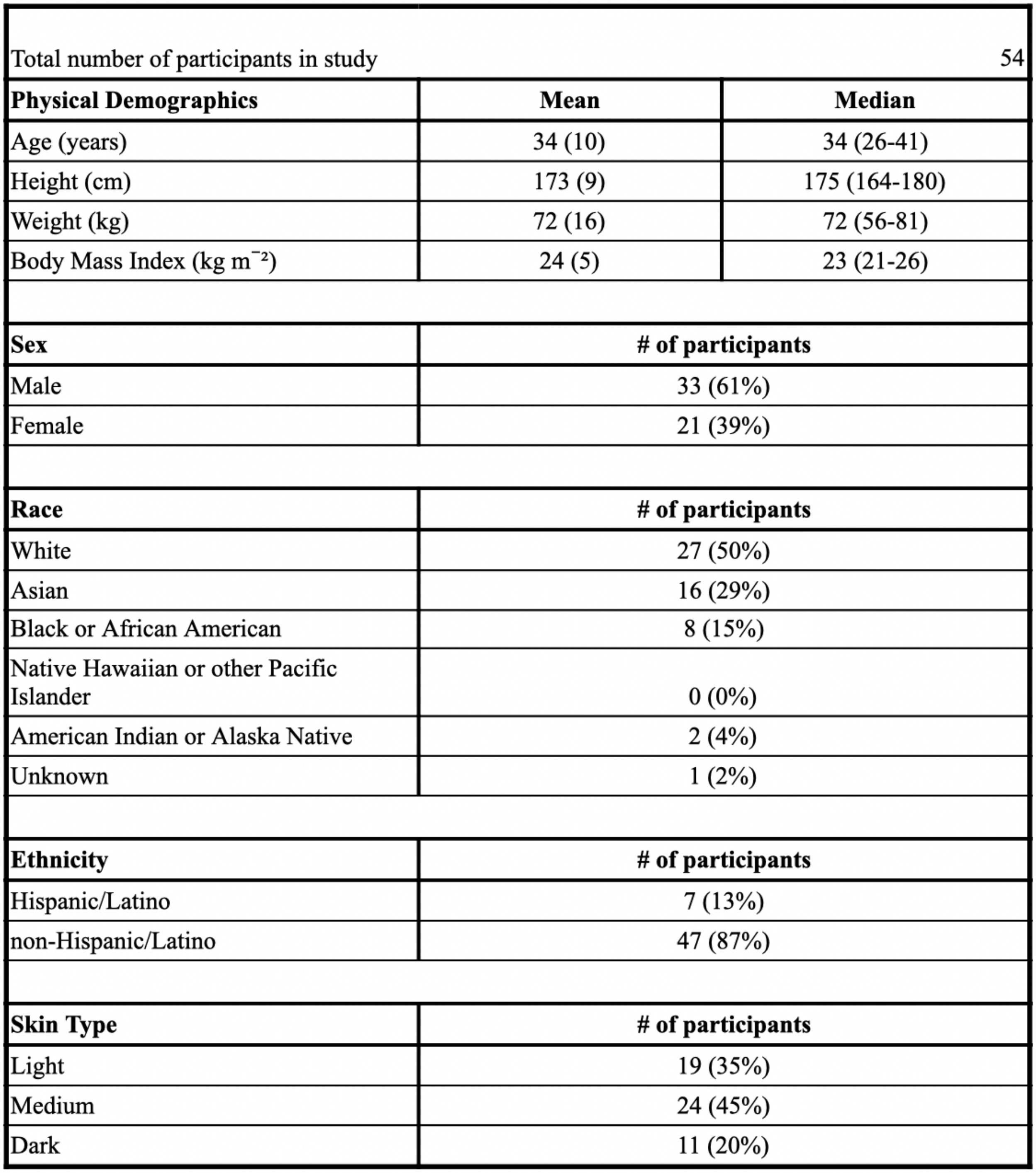
To quantitatively assess the performance of the proposed method, the following statistical metrics are used: (i) Mean Absolute Error (MAE), (ii) Standard deviation of the error (SE) and the correlation coefficient (r) between the estimated r-PPG average HR and the ground truth PPG average HR for the entire video. We also employ Bland-Altman (B&A) plots [55] to compare differences in the benchmark and proposed method's HR estimates and MX800 PPG HR measurements. These plots are labelled with the corresponding mean difference (m) that shows the systematic bias, and the limits of agreement (LoA) within which 95% of the differences are expected to lie, estimated as LoA = m ± 1.96 σ, assuming a normal distribution.

CHAPTER 4

Results

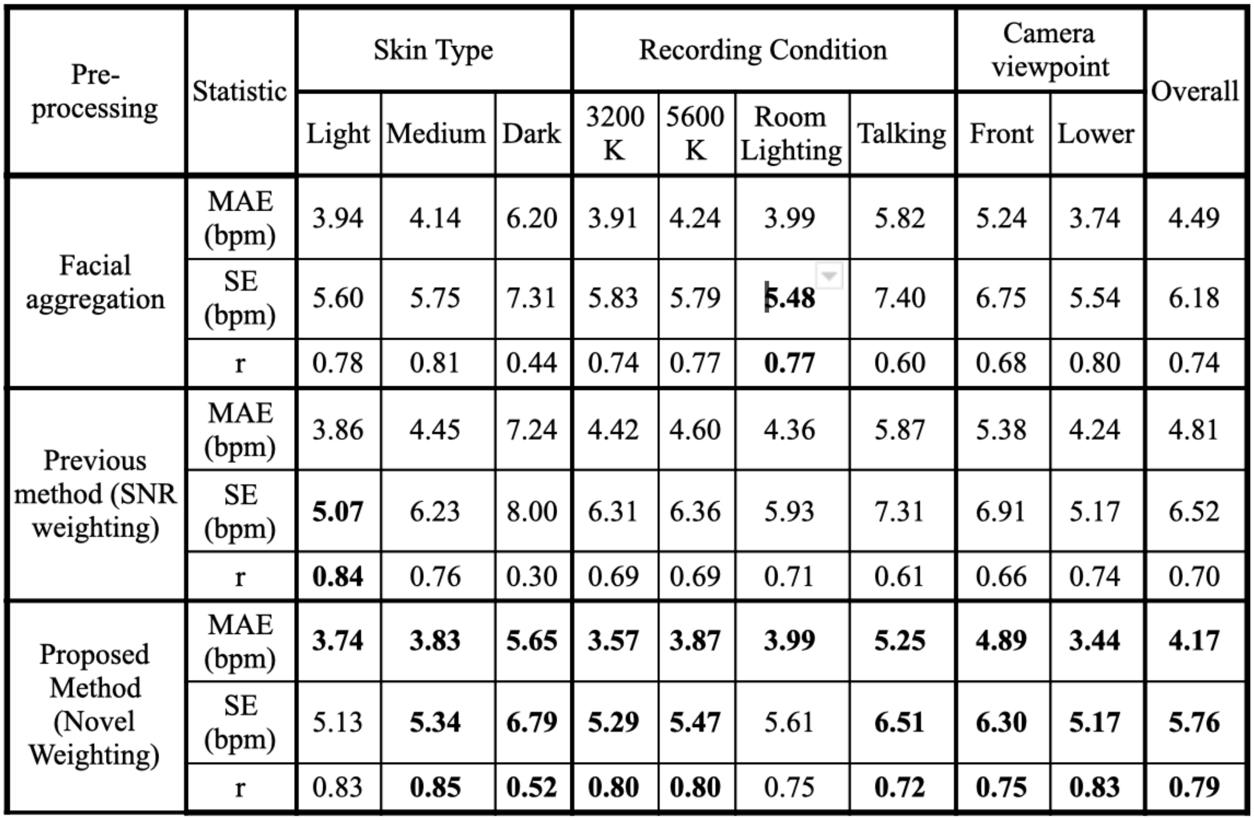
Table 4.1 describes the distribution of subjects across various demographic metrics. Overall, remote HR estimation performance was compared across 54 subjects, across 4 scene conditions and 2 camera angles, resulting in a total of 432 videos with an average length of 2 minutes. HR estimation is carried out for windows of duration 10 seconds, with an overlap of 5 seconds. The overall HR for the subject is then estimated by averaging these window-estimated HR. Table 4.2 contains a performance summary across all statistical metrics employed- namely the Mean Absolute Error (MAE), Standard deviation of the error (SE) and the correlation coefficient (r) (details in the Methods section). In addition, Table 4.3 contains information about improvement in the Mean Absolute Error (MAE) metric for the SNR weighting and proposed methods, over the facial aggregation method.

The experiments highlight that the proposed method: (i) shows an overall performance increase on the skin tone diverse VITAL dataset, (ii) shows debiased performance gain across skin tones, which is shown to not be the case with existing methods, (iii) is robust to recording conditions such as lighting and talking, and (iv) is robust to camera placement with respect to the subject. Secondary observations include the nature of bias in existing methods, the accuracy under best performing conditions, and the nature of performance differentials across scene conditions and camera angles.



Total number of participants in study		54
Physical Demographics	**Mean**	**Median**
Age (years)	34 (10)	34 (26-41)
Height (cm)	173 (9)	175 (164-180)
Weight (kg)	72 (16)	72 (56-81)
Body Mass Index (kg m^{-2})	24 (5)	23 (21-26)
Sex	**# of participants**	
Male	33 (61%)	
Female	21 (39%)	
Race	**# of participants**	
White	27 (50%)	
Asian	16 (29%)	
Black or African American	8 (15%)	
Native Hawaiian or other Pacific Islander	0 (0%)	
American Indian or Alaska Native	2 (4%)	
Unknown	1 (2%)	
Ethnicity	**# of participants**	
Hispanic/Latino	7 (13%)	
non-Hispanic/Latino	47 (87%)	
Skin Type	**# of participants**	
Light	19 (35%)	
Medium	24 (45%)	
Dark	11 (20%)	

Table 4.1: **Demographic characteristics of volunteers in the VITAL dataset.**



Pre-processing	Statistic	Skin Type			Recording Condition				Camera viewpoint		Overall
		Light	Medium	Dark	3200 K	5600 K	Room Lighting	Talking	Front	Lower	
Facial aggregation	MAE (bpm)	3.94	4.14	6.20	3.91	4.24	3.99	5.82	5.24	3.74	4.49
	SE (bpm)	5.60	5.75	7.31	5.83	5.79	**5.48**	7.40	6.75	5.54	6.18
	r	0.78	0.81	0.44	0.74	0.77	**0.77**	0.60	0.68	0.80	0.74
Previous method (SNR weighting)	MAE (bpm)	3.86	4.45	7.24	4.42	4.60	4.36	5.87	5.38	4.24	4.81
	SE (bpm)	**5.07**	6.23	8.00	6.31	6.36	5.93	7.31	6.91	5.17	6.52
	r	**0.84**	0.76	0.30	0.69	0.69	0.71	0.61	0.66	0.74	0.70
Proposed Method (Novel Weighting)	MAE (bpm)	**3.74**	**3.83**	**5.65**	**3.57**	**3.87**	**3.99**	**5.25**	**4.89**	**3.44**	**4.17**
	SE (bpm)	5.13	**5.34**	**6.79**	**5.29**	**5.47**	5.61	**6.51**	**6.30**	**5.17**	**5.76**
	r	0.83	**0.85**	**0.52**	**0.80**	**0.80**	0.75	**0.72**	**0.75**	**0.83**	**0.79**

Table 4.2: **Performance of proposed method as compared to benchmark methods.** The table shows the performance comparison of the proposed method and the chosen benchmark methods. The metrics shown are Mean Absolute Error (MAE), Standard Deviation of Error (SE) and correlation coefficient (r). Both MAE and SE are given in beats per minute. The best results across methods have been bolded for each skin type, recording condition, and camera viewpoint.

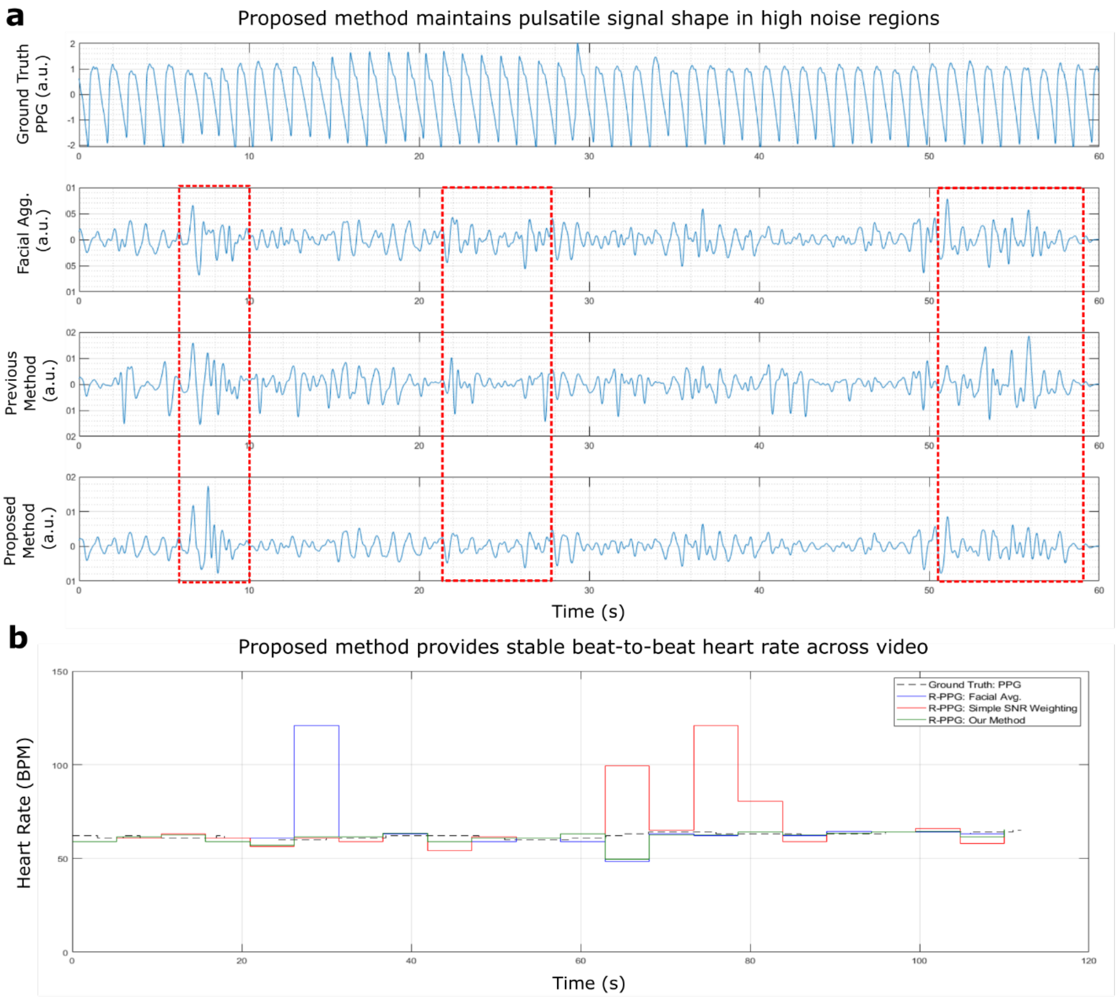
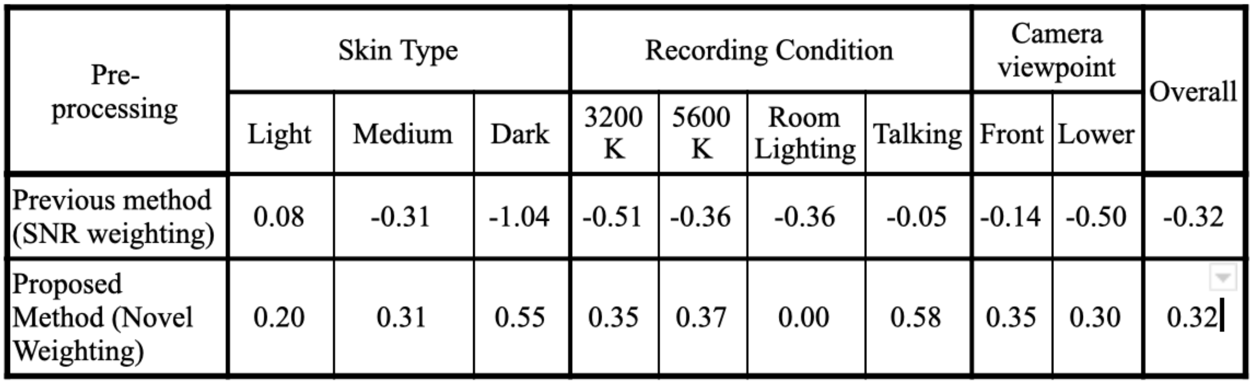


Figure 4.1: **The proposed method qualitatively recovers the pulsatile signal in a more stable manner compared to prior methods.** (A) Example pulsatile waveforms, including the ground truth PPG, facial aggregation r-PPG, previous method's (SNR weighting) r-PPG, and the proposed method's (novel weighting) r-PPG waveform (labelled from top to bottom). The dashed red windows show noisy regions where the r-PPG signal deteriorates. The proposed method maintains pulsatile signal shape, with pulsatile peaks seen more clearly and distinctly. (B) Beat-to-beat heart rate numerics over time are captured by the proposed method in a more stable manner, consistently staying within 5 bpm of the ground truth PPG.



Pre-processing	Skin Type			Recording Condition				Camera viewpoint		Overall
	Light	Medium	Dark	3200 K	5600 K	Room Lighting	Talking	Front	Lower	
Previous method (SNR weighting)	0.08	-0.31	-1.04	-0.51	-0.36	-0.36	-0.05	-0.14	-0.50	-0.32
Proposed Method (Novel Weighting)	0.20	0.31	0.55	0.35	0.37	0.00	0.58	0.35	0.30	0.32

Table 4.3: **Performance improvement with respect to facial aggregation benchmark of the previous (SNR weighting) method and the proposed method.** The metric shown is the Mean Absolute Error (MAE) improvement, in beats per minute.

4.1 Overall performance

Fig. 4.1 shows the qualitative performance of the proposed method in comparison to the ground truth PPG and benchmark methods. The estimated pulse volume signal for the proposed method is found to visually contain peaks at the same frequency as the ground truth PPG signal. In some instances, the dicrotic notch is also present, although less prominent. Particularly noisy regions of the video are highlighted by the dashed red lines in Fig. 4.1A. In these time windows, the proposed method is found to visually recover peaks more distinctly with less high frequency artifacts in comparison to the benchmark r-PPG methods. Additionally, Fig. 4.1B shows the beat-to-beat time evolution of the HR estimate, across the 10 second windows. Both the estimates from the ground truth signal and the output of the proposed method follow similar trends, consistently staying within 5 beats per minute (bpm) of each other. However, because of the high frequency noise artifacts in existing methods, the estimated HR suffers from large errors in localized regions, worsening the overall HR estimate across the 2-minute video. Such qualitative improvements also translate quantitatively, where the proposed method shows a sub-6 beats per minute MAE for all skin tones, with an overall average MAE of 4.17 beats per minute.

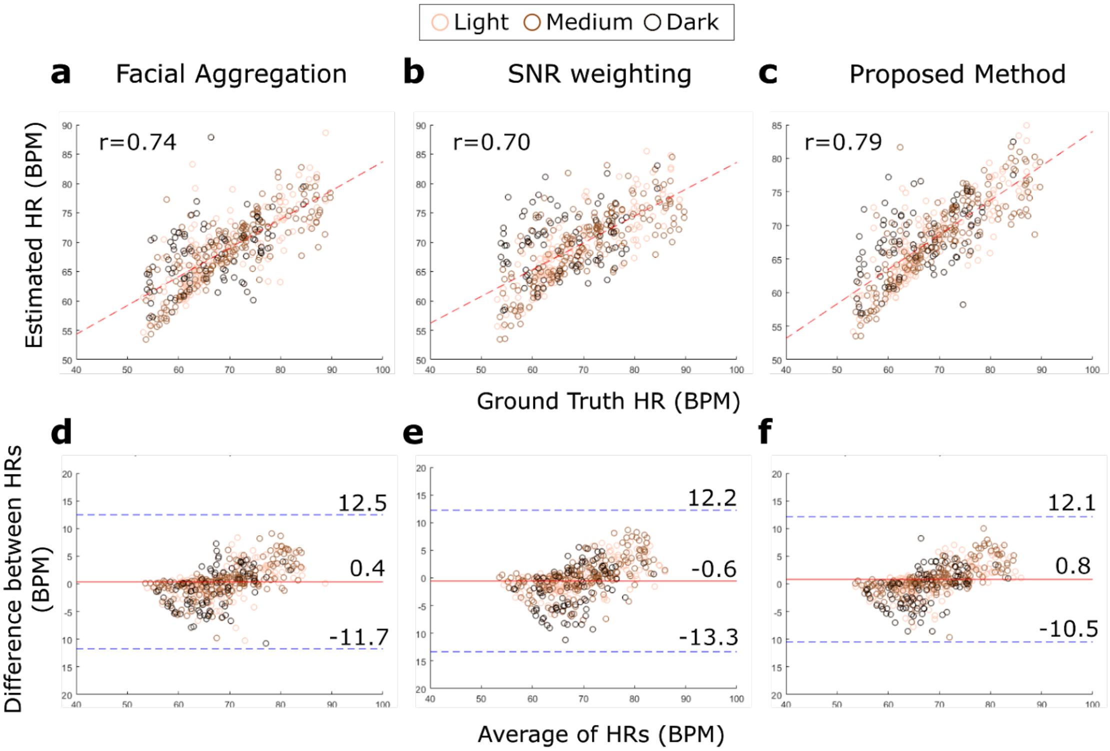


Figure 4.2: **Scatter and Bland Altman plots for benchmark and proposed heart rate recovery methods.** The label shows a marker for each skin type. (A-C) Scatter plots for different methods. The proposed method shows strong correlation with respect to ground truth heart rates from the Philips IntelliVue MX800, denoted by the Pearson Correlation Coefficient r. (D-F) Bland-Altman plots for different methods. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA = m $\pm$ 1.96 σ) by the upper and lower dotted blue lines.

Fig. 4.2 shows the corresponding scatter and B&A plots for the proposed method, facial aggregation method and SNR weighting methods across all collected videos. The proposed method shows a higher correlation (r = 0.79) in comparison to the benchmark facial aggregation (r = 0.74) and SNR weighting (r = 0.70) methods. The B&A plots show a less than 1 bpm bias across all methods. The proposed method shows the best limits of agreement with almost all videos falling within 10 bpm of the ground truth HR.

4.2 Skin tone performance

For all three methods, performance degrades from light to dark skin. The facial aggregation approach obtains a MAE of 3.94, 4.14 and 6.20 bpm for light, medium and dark skin tone subjects, resulting in an overall average performance of 4.49 bpm. When comparing the facial aggregation results to the SNR weighting approach, a MAE improvement of +0.08 bpm is obtained for light skin tones, and a successive MAE degradation of -0.31 bpm and -1.04 bpm is obtained for medium and dark skin tones respectively. Hence, on a skin tone diverse dataset such as VITAL, this leads to a comparative decrease in overall performance of -0.32 bpm. In contrast, the proposed method shows significant improvement across all skin tones when compared to the facial aggregation method, with a MAE improvement of +0.20 bpm, +0.31 bpm and +0.55 bpm obtained for light, medium and dark skin tones respectively. Consequently, the overall performance of the proposed method on the VITAL dataset improves by +0.32 bpm.

Fig. B.1A-C highlights the high correlation between the proposed method's r-PPG HR estimates and ground truth PPG HR for light (r = 0.83) and medium skin tones (r = 0.85), and moderate correlation for dark skin tones (r = 0.52). The B&A plots in Fig. B.1D-F show a less than 2 bpm bias across all skin tones, and that all the proposed method's r-PPG HR estimates are mostly within 10 bpm of the ground truth. These correlation metrics are an improvement to the benchmark methods of facial aggregation and SNR weighting. Fig. B.2

and Fig. B.3 show the corresponding scatter and B&A plots for the facial aggregation and SNR weighting methods respectively.

4.3 Recording condition performance

Each of the three methods performs similarly across the three lighting conditions. The facial aggregation method shows an average MAE of 4.05 bpm across the lighting conditions, while the SNR weighting method shows an average performance of 4.46 bpm. In contrast to this, the proposed method shows an average performance of 3.81 bpm across the three lighting conditions, representing an improvement of +0.24 bpm MAE.

The performance on the 'talking' activity is worse as compared to that on other scene conditions for all three methods. Similar to other trends, the SNR weighting method shows a performance reduction of -0.05 bpm over the facial aggregation benchmark. However, the proposed method shows a large improvement of +0.57 bpm when compared to the facial aggregation benchmark.

Fig. B.4A-D highlights the high correlation between the proposed method's r-PPG HR estimates and ground truth PPG HR across the various recording conditions. The dark skin tone markers across all recording conditions make up the majority of outlying data. The B&A plots in Fig. B.4E-G show a bias of less than 1 bpm across the three lighting conditions, and Fig. B.4H shows a bias of less than 2 bpm during subject talking. These figures also show that the proposed method's r-PPG heart estimates are mostly within 10 bpm of the ground truth across all recording conditions. These correlation metrics are an improvement to the benchmark methods of facial aggregation and SNR weighting. Fig. B.5 and Fig. B.6 show the corresponding scatter and B&A plots for the facial aggregation and SNR weighting methods respectively.

4.4 Camera viewpoint performance

For all three methods, the bottom camera viewpoint performs the best. The facial aggregation method shows a MAE of 5.24 bpm for the front setting, and 3.74 bpm for the bottom setting. The SNR weighting method on the other hand shows MAE of 5.38 bpm and 4.24 bpm, while the proposed method shows a MAE of 4.89 bpm and 3.44 bpm. Fig. B.7, B.8, and B.9 show the corresponding scatter and B&A plots for the proposed method, facial aggregation method and SNR weighting method respectively. The correlation between estimated and ground truth HR seen by the proposed method for the front and bottom viewpoints (0.75 and 0.83) is a clear improvement over the same for the facial aggregation (0.68 and 0.80) and the SNR weighting (0.66 and 0.74) methods.

4.5 Best performance

The best performing camera viewpoint and recording condition on the VITAL dataset is using the bottom camera angle with lighting at 5600K, where the label "best performing" is chosen with respect to both overall performance and skin tone bias. Fig. 4.3 highlights that the proposed method achieves a MAE performance of below 3 bpm across all skin tone categories. Specifically, a MAE of 1.97, 2.86 and 3.01 bpm, and correlation of 0.93, 0.91, and 0.87, is achieved for the light, medium and dark skin tones respectively. This is a significant improvement over the two existing methods with regards to both overall performance and skin tone bias. The facial averaging method shows an MAE of 2.40, 3.47 and 4.09 bpm, and correlation of 0.89, 0.84, and 0.75, while the SNR weighting method shows an MAE of 1.48, 3.30 and 5.66 bpm, and correlation of 0.98, 0.85, and 0.58, for the same respective skin tone categories.

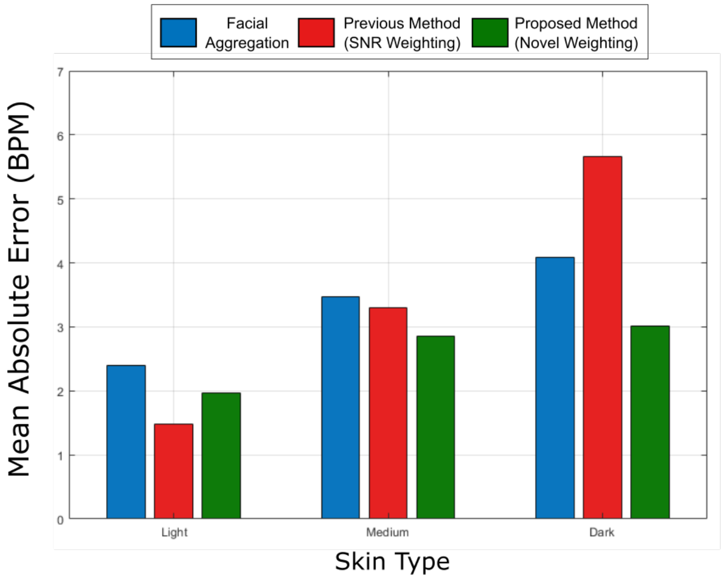


Figure 4.3: **Bar plot highlighting algorithmic comparison for the best-performing scene configuration.** It is seen that the proposed method shows increasing performance gains over both the facial aggregation and the SNR weighting methods. Specifically, for the best-performing scene configuration using the bottom camera angle viewpoint with 5600K lighting the proposed method is the only method able to have a close to 3 or below 3 bpm MAE performance in the best case, thereby establishing its capability towards medically relevant HR measurements.

CHAPTER 5

Conclusions

In this work, we propose a novel r-PPG algorithm to estimate subject HR in a contactless manner using only a smartphone camera. Several r-PPG algorithms have been proposed to extract the BVP signal from videos. However, these algorithms exhibit a performance gap, and therefore a bias, for certain types of skin tones [33], subject motions [31, 21, 32], and illumination conditions [30]. Addressing these biases is essential for successful deployment of r-PPG technology in telemedicine applications, yet it remains a challenge. For example, dark skin, which contains higher amounts of melanin, fundamentally reduces the signal to noise ratio of all existing r-PPG algorithms. The important work of Nowara et al. [33] highlights this reduction, thereby conclusively determining that current r-PPG algorithms have markedly worse performance on darker skin tones. The work also highlights the issue of biased skin tone and gender representation in computer vision datasets, which is especially true for the comparatively small datasets used in r-PPG analyses. This dataset bias further prevents underlying algorithmic biases, such as skin tone bias, from being addressed. Should contactless HR sensing using video be implemented in a clinical setting, the development of r-PPG computer vision algorithms and datasets that improve the accuracy and reduce the bias of HR measurements for patients of all skin tones (especially the darker skin tones) is critically necessary for high-quality telemedicine care.

A key contribution of this work is the creation of the VITAL dataset, which is a first effort towards collecting a demographically diverse video vital sign database for telemedicine applications. While societal demographics are skewed largely towards light skin tone persons, it is essential to have diversely represented computer vision healthcare datasets to understand performance limitations that may otherwise be masked within biased data [56]. Although the VITAL dataset is not entirely unbiased itself, it achieves a much higher degree of skin tone diversity as compared to existing datasets. Moreover, the VITAL dataset records facial videos using smartphone cameras, which introduces significant video compression and imaging noise artifacts. Typically, r-PPG methods are developed and tested using uncompressed videos. However, deployment of r-PPG technology for telemedicine will ultimately require a robustness to video compression noise artifacts. Therefore, the VITAL dataset enables a more realistic evaluation of remote video-based vital sign monitoring methods for telemedicine translation, which contrasts from previous works. Finally, the VITAL dataset captures four ground truth vital signs: HR, respiratory rate, oxygen saturation, and non-invasive continuous blood pressure, of which three waveforms are collected: ECG, PPG and respiration. Although this work only utilizes the HR obtained from the PPG waveform for testing, we anticipate future work capturing all four vital signs simultaneously from the facial videos. Overall, we envision the VITAL dataset to be an essential resource for upcoming related research and, in addition, to set the tone for future data collection endeavors for similar interdisciplinary clinical cum technological applications.

With respect to algorithmic development, this work addresses the aforementioned biases in skin-tone, illumination conditions, and subject motions using physics-rooted knowledge and camera noise analysis. From our theory, we derive 3 key conclusions: (i) imaging noise creates skin tone bias (and lighting bias), (ii) imaging noise and specular reflections degrade the r-PPG signal, and (iii) denoising is to be done before signal inference. Therefore, we primarily focus our attention to signal processing strategies as opposed to signal extraction modifications. The first attempted work to reduce r-PPG skin tone bias was done by Kumar et al. (DistancePPG) [43], in which a weighted average of BVP signals from various facial regions-of-interest (ROI). However, to the best of the authors' knowledge, no work yet has continued development of r-PPG algorithms that tackle the important issue of performance bias on darker skin tones. The proposed r-PPG algorithm draws from existing r-PPG denoising methods that use a similar weighted ROI philosophy as in DistancePPG [41, 42, 44]. Specifically, it modifies the strategy by weighting in RGB space rather than blood volume signal space, and by introducing a skin diffuse component weighting. This enables the proposed algorithm to mitigate performance losses for subjects with darker skin tones, subjects in varying illumination conditions, and subjects who may be moving their face such as when they are talking.

The proposed method achieves the best overall average MAE performance across the VITAL dataset of 4.17 bpm, as opposed to 4.49 bpm by the facial aggregation method [13, 18, 22, 19, 17, 21] and 4.81 bpm for the SNR weighting method [41, 42, 43, 44]. This achievement can be attributed to the performance gains seen across all skin tones in comparison to the facial aggregation method. The SNR weighting method shows performance gain only for the light skin tone subjects (+0.08 bpm) and a performance drop for the medium and dark skin tones (-0.31 and -1.04 bpm respectively), thereby actually increasing the skin tone performance bias. Consequently, the method's overall performance suffers on a more diversely represented dataset such as VITAL. This illustrates the importance for the need of a truly diverse dataset when developing r-PPG technology.

Nevertheless, as with previous methods, the performance of the proposed method still exhibits a skin-tone bias. However, we highlight that the proposed method achieves the largest MAE improvements over the facial aggregation method of +0.55 bpm for the traditionally worse performing dark skin tone in comparison with the light (+0.20 bpm) and medium (+0.31 bpm) skin tones. This outcome attests to the fairness of the method. The proposed method is the only method able to perform with an overall MAE less than 6 bpm across all skin tones. For the best performing setting (bottom camera viewpoint with 5600K lighting), the proposed method obtains a less than 3 bpm MAE across all skin tones. This establishes the viability and performance accuracy of the proposed method for medically relevant HR estimation. These inferences are further enforced by the largest increase in the correlation coefficient and largest decrease in the SE for dark skin tones by the proposed method, as opposed to the SNR weighting method which sees performance reduction for medium and dark skin tones. Hence, in addition to the overall improvement in performance across all skin tones, the proposed method successfully steps towards reducing the performance bias that exists between skin tones. If the VITAL dataset were to have even more equal representation in terms of skin tone, the overall average performance measures are further expected to improve.

Large improvements in performance of the proposed method are also observed for the talking activity over the facial aggregation benchmark, as compared to the SNR weighting method which shows an overall performance drop. This technology may one day allow for real-time continuous contact-less HR monitoring during a telemedicine visit, which would provide greater information to outpatient clinicians. This advance may also be relevant for in-hospital continuous contactless monitoring in ICU settings or hospital floor care.

Improvements in performance are also observed across camera viewpoints. The proposed method shows considerable improvements for the front and bottom angles. A typical telemedicine visit, through a cell phone platform, may involve the patient holding the camera at varying angles with respect to the face. The shown robustness and performance improvement of the proposed method therefore makes it increasingly amenable to such tasks. Interestingly, for all methods tested (existing and novel), the bottom angle shows improved performance as compared to the front angle. This could be because interfering factors such as hair, spectacles and so on occupy a smaller portion of the usable frame in the bottom angle, as well as differing face scales in the two angles.

In relation to the clinical significance of this work, remote vital sign monitoring has risen in prominence over recent years, with an acceleration in clinical development due to the COVID-19 pandemic. In response to the pandemic, health systems across the country implemented a large-scale restriction of non-urgent in-person appointments [57], transitioned many outpatient services to telemedicine visits [4], and developed remote monitoring care

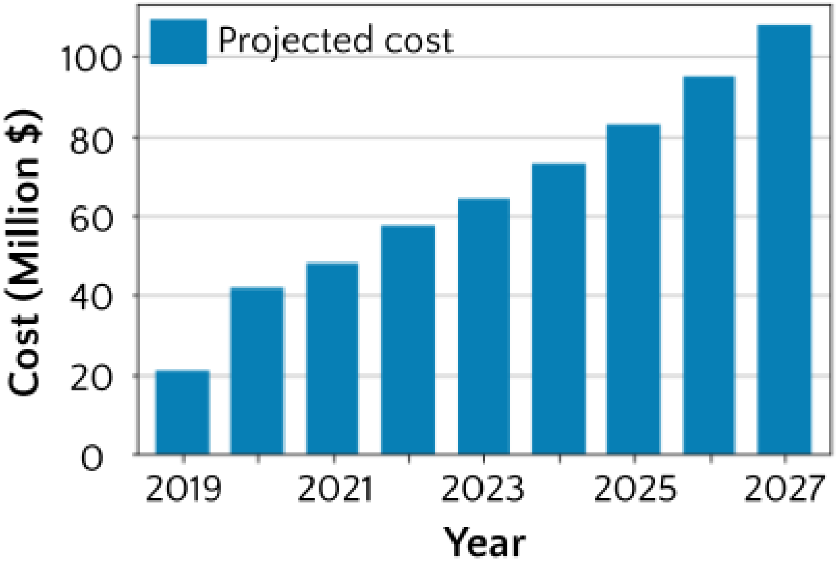


Figure 5.1: **Projected cost of deploying finger pulse oximeters for telemedicine application.** HR sensing solutions for telemedicine and remote patient monitoring have relied on the adoption of wearable sensors. Currently, the most viable and inexpensive existing wearable solution to assess patient HR and oxygen saturation are finger pulse oximeters. For the scales at which telemedicine is projected to grow, such a solution would involve a deployment cost in excess of $700 million in the US alone. In contrast, a smartphone camera-based method offers a purely algorithmic solution that can be integrated into existing healthcare system telemedicine video-conferencing applications.

pathways [2] in order to facilitate social distancing yet maintain continuity of care. To remotely monitor COVID-19 patients, many health systems shipped home vital sign equipment to patients in order to obtain quantitative physiological data that could facilitate high quality remote management via telemedicine. At a population level, however, supplying and shipping vital sign monitoring devices to patients is expensive and not scalable, making such a solution nonviable. Fig. 5.1 shows the projected cost of deploying finger pulse oximeters for telemedicine application, the most viable and inexpensive existing solution to assess patient HR and oxygen saturation. For the scales at which telemedicine is projected to grow, even this solution would involve a deployment cost in excess of $700 million in the US alone (see Appendix A for calculation details). Given the high penetration of mobile phone technology globally, there is great interest in transforming smartphones into low-cost portable HR, respiratory rate, and pulse oximeter monitors, thereby increasing accessibility to vital monitoring equipment and alleviating healthcare inequity. Using in-built camera modules and computer vision algorithms to obtain quantitative vital sign data remotely offers a purely algorithmic solution with potentially zero marginal cost.

Outside of a pandemic situation, knowledge of vital signs is also important information for clinicians who are managing medical conditions that require such data for health management, and remotely obtaining vital signs may allow care teams to perform remote surveillance and home monitoring of patients with greater confidence. Notably, several minority and lower socioeconomic status patient populations may benefit from more remote care, especially as it has been established that the COVID-19 pandemic has disproportionately affected such communities, both nationally and in states the most affected by the pandemic [58, 59]. In New York City and Michigan, African American and Latino residents have the highest age-adjusted rates of hospitalized and non-hospitalized COVID-19, and age-adjusted death rates for African Americans are more than twice those for white and Asian residents [60, 61]. African American communities have also been found to have higher prevalence of cardiovascular and related complications, when compared with traditionally light skin toned people [62]. These patient populations may therefore stand to benefit the most from skin tone robust contactless vital sign (specifically heart rate) sensing technologies that facilitate high-quality remote care pathways. Finally, we believe contactless vital sign sensing technology would be useful at the start of in-person clinic or hospital encounters or for continuous patient monitoring in a hospital floor or ICU setting. Cameras, as opposed to hospital staff, may one day obtain key vital signs without contact, thereby reducing exposure of patients to staff, enabling improved infection control, and freeing up hospital staff to attend to other important patient care needs.

With regards to limitations and future work, while our method has been tested on an adult population, additional work is needed to enable clinical adoption. Further research investigating HR estimation using our proposed method is still needed in pediatric and geriatric populations and patient populations with known cardiopulmonary disease. Future work must also focus on improving computer vision methods to detect extremes of HR and discern heart arrhythmias. Additionally, the proposed method does not obviate skin tone bias but rather is the first work that can be demonstrated to mitigate skin tone bias in the VITAL dataset. Therefore, research must be undertaken to further reduce bias and assure fairness by building upon our work, as well as to continue improving overall performance on subjects and videos in real life scenarios.

From an algorithmic perspective, we believe that one of the most important factors towards large scale deployment of such methods for clinical use is the inherent fairness of the algorithm. As healthcare increasingly accelerates towards a digitally connected and virtual future, early consideration must be given to developing equitable health technology that does not exacerbate healthcare disparities or create new disparities. Ultimately, we hope this work motivates the community towards exciting and essential research avenues looking into inherent system biases associated with r-PPG. By reducing biases, we move a step closer towards deploying high-quality, medically inclusive non-contact vital sensing techniques that can aid clinicians in delivering remote patient care, during times of peace and pandemic alike.

At a broader level, such research into both algorithmic and hardware bias holds a lot of relevance moving forward. Understanding existing and potential biases in widely used technologies is key towards equitable deployment and performance. Additionally, fixing biases in performance may not only lead to better performance for the biased class, but potentially move towards across the board improved performance. We therefore hope to use this work as a starting point as we delve into better understanding the origins of bias in vision and imaging based technologies and elegant fixes to address these biases.

APPENDIX A

Deployment Cost Projections for Telemedicine

In order to calculate the estimated average deployment cost for the cheapest existing method (finger pulse oximeters), we use the following methodology:

1. We identify the estimated user base numbers for telemedicine in the US using the numbers from [63] and extend these up to 2027 using the compound annual growth rate (CAGR) of 15.8% as suggested in [64].

2. We make the conservative assumption that all members of a given family would be active users of telemedicine services. Therefore, an estimate of the number of families using telemedicine services is given by:

   $$No.\ of\ Families = \frac{Number\ of\ Telemedicine\ Users}{Avg.\ Family\ Size\ in\ the\ US} \tag{A.1}$$

   We use the average family size of 3.15 from the U.S. Census Bureau's Current Population Survey [65].

3. Assuming that one pulse oximeter costs \$20 (as observed from a survey of available units in the market), and assuming conservatively that one pulse oximeter has to be deployed per family, the cost of deployment is given by:

   $$Cost\ of\ Deployment = No.\ of\ Families\ \times\ Cost\ per\ Pulse\ Oximeter\ Unit \tag{A.2}$$

APPENDIX B

Additional Bland-Altman Plots

For the sake of completeness, this appendix highlights various Bland-Altman plots that analyze the performance of the proposed as well as baseline methods across various factors and scene parameters. These are referenced at various places in the thesis.

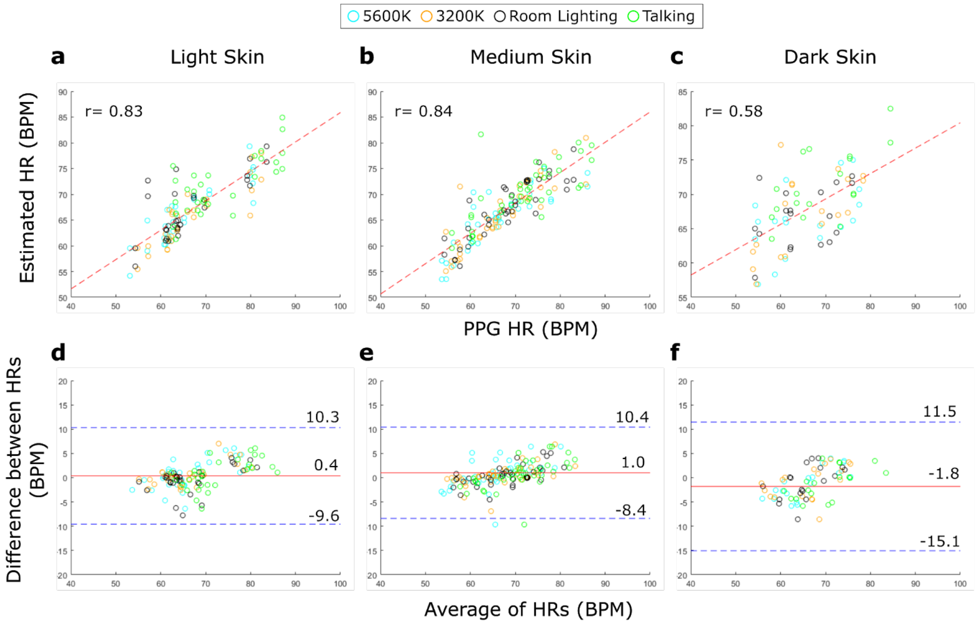


Figure B.1: **Scatter and Bland Altman plots for proposed method, varied across skin tone categories.** The label shows a marker for each video recording condition. (A-C) Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (D-F) Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \text{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

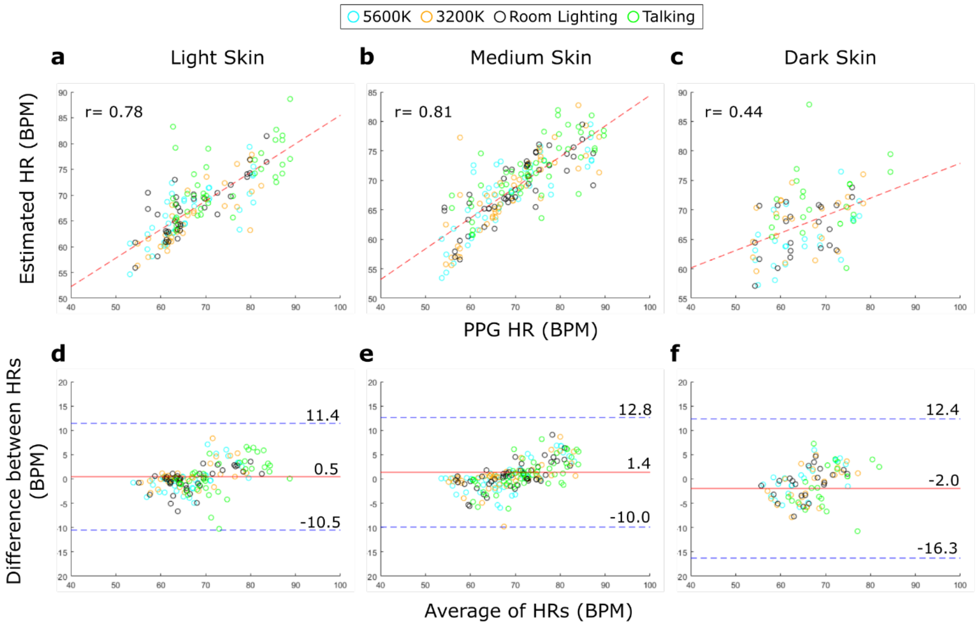


Figure B.2: **Scatter and Bland Altman plots for facial aggregation method, varied across skin tone categories.** The label shows a marker for each video recording condition. (A-C) Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (D-F) Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA = m $\pm$ 1.96 σ) by the upper and lower dotted blue lines.

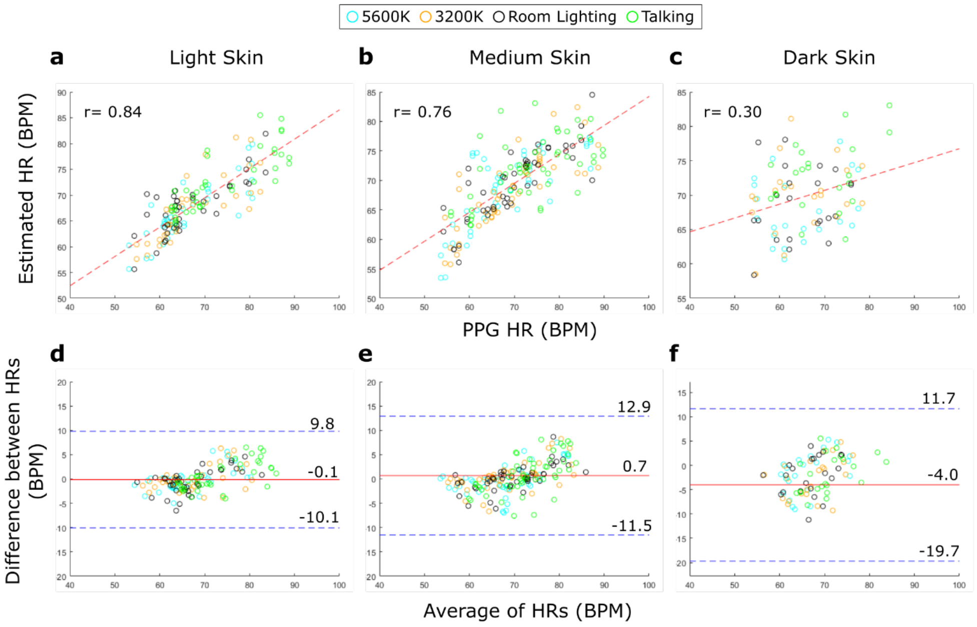


Figure B.3: **Scatter and Bland Altman plots for SNR weighting method, varied across skin tone categories.** The label shows a marker for each video recording condition. (A-C) Scatter plots for different skin types highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (D-F) Bland-Altman plots for different skin types. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA = m $\pm$ 1.96 σ) by the upper and lower dotted blue lines.

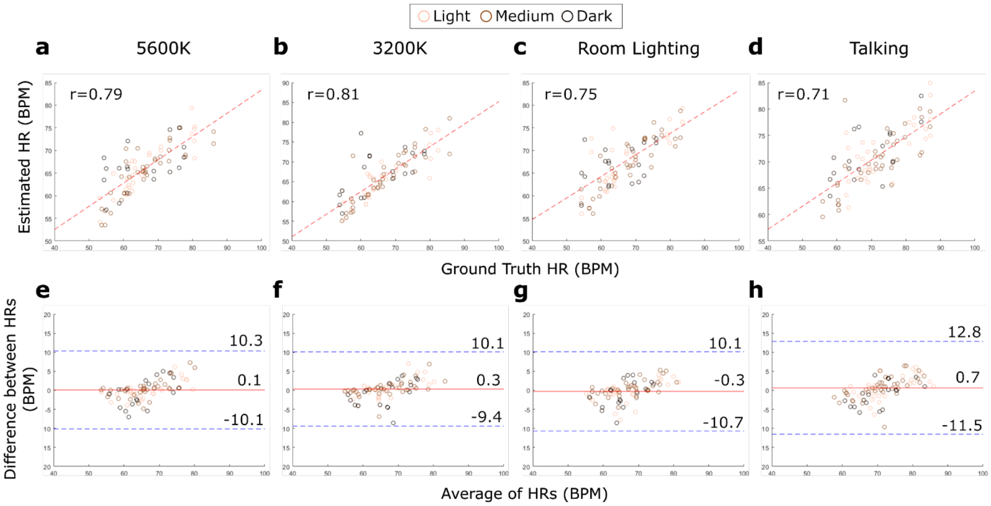


Figure B.4: **Scatter and Bland Altman plots for proposed method, varied across scene condition categories.** The label shows a marker for each skin type. (A-D) Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (E-H). Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \mathrm{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

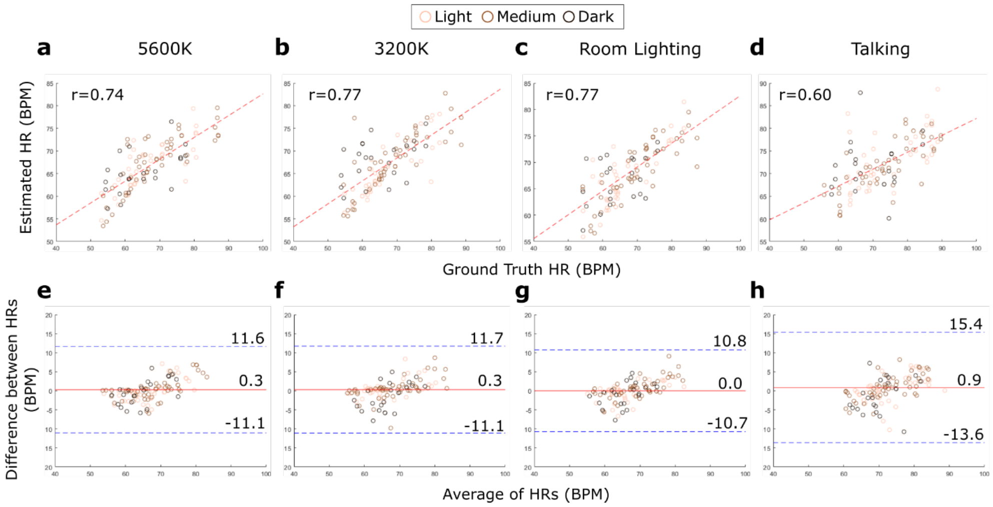


Figure B.5: **Scatter and Bland Altman plots for facial aggregation method, varied across scene condition categories.** The label shows a marker for each skin type. (A-D) Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (E-H) Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA = m ± 1.96 σ) by the upper and lower dotted blue lines.

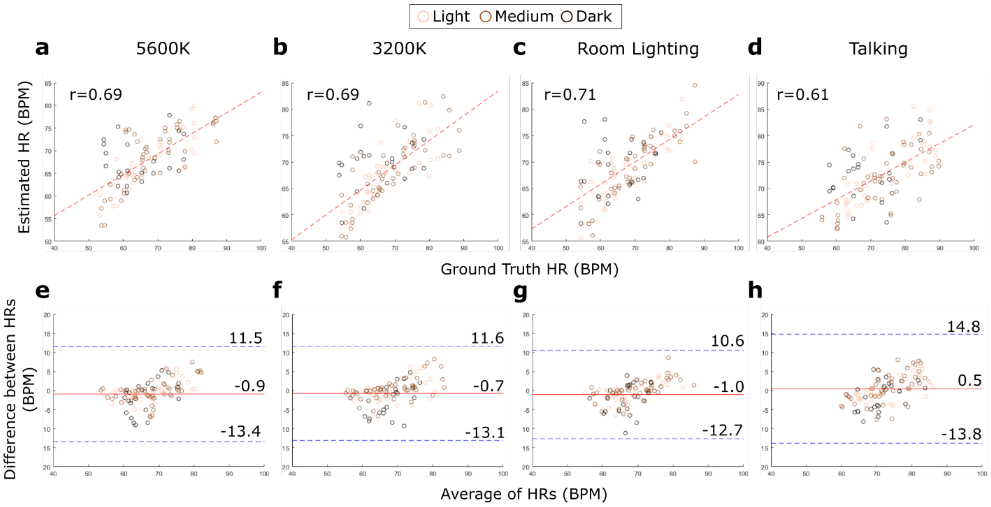


Figure B.6: **Scatter and Bland Altman plots for SNR weighting method, varied across scene condition categories.** The label shows a marker for each skin type. (A-D) Scatter plots for different recording conditions highlighting the correlation between estimated and ground truth heart rate, denoted by the Pearson Correlation Coefficient r. (E-H) Bland-Altman plots for different recording conditions. The bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \text{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

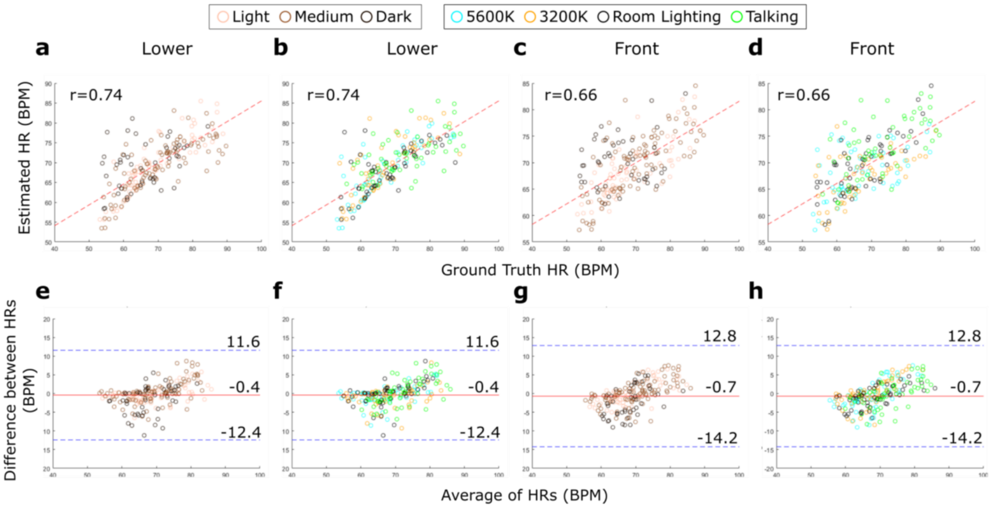


Figure B.7: **Scatter and Bland Altman plots for proposed method's dependence on camera angle, varied across skin tone categories and recording conditions.** (A-B) Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (C-D) Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. (E-F) Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (G-H) Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \text{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

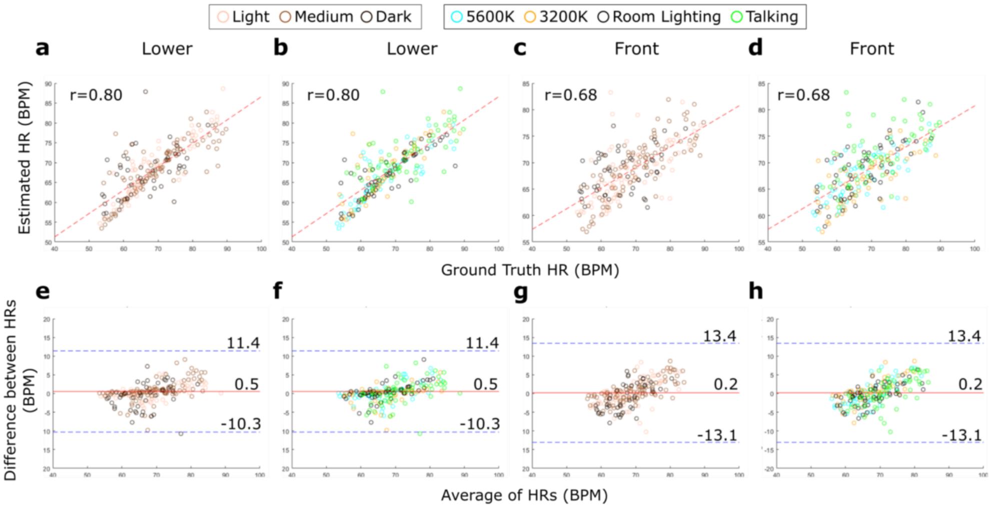


Figure B.8: **Scatter and Bland Altman plots for the facial aggregation method's dependence on camera angle, varied across skin tone categories and recording conditions.** (A-B) Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (C-D) Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. (E-F) Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (G-H) Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \text{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

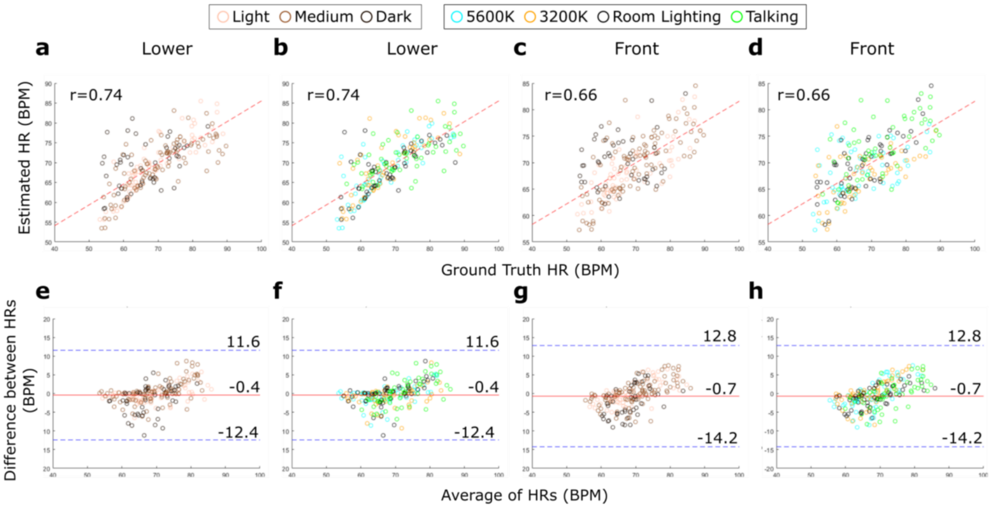


Figure B.9: **Scatter and Bland Altman plots for the SNR weighting method's dependence on camera angle, varied across skin tone categories and recording conditions.** (A-B) Scatter plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (C-D) Scatter plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. (E-F) Bland Altman plots for the lower camera angle, varying across skin tone categories and recording conditions, respectively. (G-H) Bland Altman plots for the front camera angle, varying across skin tone categories and recording conditions, respectively. For all Bland Altman plots, the bias (m) is shown by the middle solid red line, and the limits of agreement (LoA $= \text{m} \pm 1.96\ \sigma$) by the upper and lower dotted blue lines.

REFERENCES

[1] P. Chari, K. Kabra, D. Karinca, S. Lahiri, D. Srivastava, K. Kulkarni, T. Chen, M. Cannesson, L. Jalilian, and A. Kadambi, “Diverse R-PPG: Camera-based heart rate estimation for diverse subject skin-tones and scenes,” *arXiv preprint arXiv:2010.12769*, 2020.

[2] T. Annis, S. Pleasants, G. Hultman, E. Lindemann, J. A. Thompson, S. Billecke, S. Badlani, and G. B. Melton, “Rapid implementation of a COVID-19 remote patient monitoring program,” *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1326–1330, Aug. 2020, publisher: Oxford Academic. [Online]. Available: https://academic.oup.com/jamia/article/27/8/1326/5835871

[3] D. Ford, J. B. Harvey, J. McElligott, K. King, K. N. Simpson, S. Valenta, E. H. Warr, T. Walsh, E. Debenham, C. Teasdale, S. Meystre, J. S. Obeid, C. Metts, and L. A. Lenert, “Leveraging health system telehealth and informatics infrastructure to create a continuum of services for COVID-19 screening, testing, and treatment,” *Journal of the American Medical Informatics Association*, vol. 27, no. 12, Dec. 2020. [Online]. Available: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocaa157/5865271

[4] S. L. Connolly, K. L. Stolzmann, L. Heyworth, K. R. Weaver, M. S. Bauer, and C. J. Miller, “Rapid Increase in Telemental Health Within the Department of Veterans Affairs During the COVID-19 Pandemic,” *Telemedicine and e-Health*, Sep. 2020, publisher: Mary Ann Liebert, Inc., publishers. [Online]. Available: https://www.liebertpub.com/doi/10.1089/TMJ.2020.0233

[5] M. Nishiga, D. W. Wang, Y. Han, D. B. Lewis, and J. C. Wu, “COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives,” *Nature Reviews Cardiology*, vol. 17, no. 9, pp. 543–558, Sep. 2020, number: 9 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41569-020-0413-9

[6] C. Dinh-Le, R. Chuang, S. Chokshi, and D. Mann, “Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions,” *JMIR mHealth and uHealth*, vol. 7, no. 9, p. e12861, Sep. 2019, company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://mhealth.jmir.org/2019/9/e12861

[7] H. Lukas, C. Xu, Y. Yu, and W. Gao, “Emerging Telemedicine Tools for Remote COVID-19 Diagnosis, Monitoring, and Management,” *ACS Nano*, vol. 14, no. 12, pp. 16 180–16 193, Dec. 2020, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/acsnano.0c08494

[8] S. Kumar, W. Nilsen, M. Pavel, and M. Srivastava, "Mobile Health: Revolutionizing Healthcare Through Transdisciplinary Research," *Computer*, vol. 46, no. 1, pp. 28–35, Jan. 2013, conference Name: Computer.

[9] S. R. Steinhubl, E. D. Muse, and E. J. Topol, "The emerging field of mobile health," *Science Translational Medicine*, vol. 7, no. 283, pp. 283rv3–283rv3, Apr. 2015, publisher: American Association for the Advancement of Science Section: Review. [Online]. Available: https://stm.sciencemag.org/content/7/283/283rv3

[10] J. Sawyer, "Wearable Internet of Medical Things Sensor Devices, Artificial Intelligence-driven Smart Healthcare Services, and Personalized Clinical Care in COVID-19 Telemedicine," *American Journal of Medical Research*, vol. 7, no. 2, pp. 71–77, 2020, publisher: Addleton Academic Publishers. [Online]. Available: https://www.ceeol.com/search/article-detail?id=906655

[11] T. Proesmans, C. Mortelmans, R. V. Haelst, F. Verbrugge, P. Vandervoort, and B. Vaes, "Mobile Phone–Based Use of the Photoplethysmography Technique to Detect Atrial Fibrillation in Primary Care: Diagnostic Accuracy Study of the FibriCheck App," *JMIR mHealth and uHealth*, vol. 7, no. 3, p. e12284, Mar. 2019, company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://mhealth.jmir.org/2019/3/e12284

[12] K. H. C. Li, F. A. White, T. Tipoe, T. Liu, M. C. Wong, A. Jesuthasan, A. Baranchuk, G. Tse, and B. P. Yan, "The Current State of Mobile Phone Apps for Monitoring Heart Rate, Heart Rate Variability, and Atrial Fibrillation: Narrative Review," *JMIR mHealth and uHealth*, vol. 7, no. 2, p. e11606, Feb. 2019, company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://mhealth.jmir.org/2019/2/e11606

[13] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Opt. Express*, vol. 18, no. 10, pp. 10 762–10 774, May 2010. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-18-10-10762

[14] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting Pulse from Head Motions in Video," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3430–3437, iSSN: 1063-6919.

[15] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, Dec. 2008, publisher: Optical Society of America. [Online]. Available: https://www.osapublishing.org/oe/abstract.cfm?uri=oe-16-26-21434

[16] G. R. Tsouri, S. Kyal, S. A. Dianat, and L. K. Mestha, "Constrained independent component analysis approach to nonobtrusive pulse rate measurements," *Journal of Biomedical Optics*, vol. 17, no. 7, p. 077011, Jul. 2012, publisher: International Society for Optics and Photonics. [Online]. Available: https://www.spiedigitallibrary.org/journals/journal-of-biomedical-optics/volume-17/issue-7/077011/Constrained-independent-component-analysis-approach-to-nonobtrusive-pulse-rate-measurements/10.1117/1.JBO.17.7.077011.short

[17] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity," in *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2011, pp. 405–410.

[18] G. d. Haan and V. Jeanne, "Robust Pulse Rate From Chrominance-Based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013, conference Name: IEEE Transactions on Biomedical Engineering.

[19] W. Wang, A. C. d. Brinker, S. Stuijk, and G. d. Haan, "Algorithmic Principles of Remote PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017, conference Name: IEEE Transactions on Biomedical Engineering.

[20] R. Song, S. Zhang, J. Cheng, C. Li, and X. Chen, "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Computers in Biology and Medicine*, vol. 116, p. 103535, Jan. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482519303944

[21] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, 2014.

[22] W. Wang, S. Stuijk, and G. d. Haan, "A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 9, pp. 1974–1984, Sep. 2016, conference Name: IEEE Transactions on Biomedical Engineering.

[23] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2396–2404, iSSN: 1063-6919.

[24] W. Chen and D. McDuff, "DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks," in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 356–373.

[25] X. Niu, S. Shan, H. Han, and X. Chen, "RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2020, conference Name: IEEE Transactions on Image Processing.

[26] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 151–160, iSSN: 2380-7504.

[27] Z. Yu, X. Li, and G. Zhao, "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks," *arXiv e-prints*, vol. 1905, p. arXiv:1905.02419, May 2019. [Online]. Available: http://adsabs.harvard.edu/abs/2019arXiv190502419Y

[28] E. Nowara, D. McDuff, and A. Veeraraghavan, "The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention," *arXiv:2010.07770 [cs, eess]*, Oct. 2020, arXiv: 2010.07770. [Online]. Available: http://arxiv.org/abs/2010.07770

[29] R. Spetlík, V. Franc, J. Cech, and J. Matas, "Visual Heart Rate Estimation with Convolutional Neural Network," in *BMVC*, 2018.

[30] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote Heart Rate Measurement from Face Videos under Realistic Situations," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 4264–4271, iSSN: 1063-6919.

[31] A. V. Moço, S. Stuijk, and G. de Haan, "Motion robust ppg-imaging through color channel mapping," *Biomedical Optics Express*, vol. 7, no. 5, pp. 1737–1754, 2016.

[32] W. Wang, S. Stuijk, and G. d. Haan, "Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, Feb. 2015, conference Name: IEEE Transactions on Biomedical Engineering.

[33] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "A Meta-Analysis of the Impact of Skin Tone and Gender on Non-Contact Photoplethysmography Measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 284–285. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w19/Nowara_A_Meta-Analysis_of_the_Impact_of_Skin_Tone_and_Gender_CVPRW_2020_paper.html

[34] T. Igarashi, K. Nishino, and S. K. Nayar, *The appearance of human skin: A survey.* Now Publishers Inc, 2007.

[35] S. Alotaibi and W. A. P. Smith, "A Biophysical 3D Morphable Model of Face Appearance," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2017, pp. 824–832, iSSN: 2473-9944.

[36] R. R. Anderson and J. A. Parrish, "The optics of human skin," *Journal of Investigative Dermatology*, vol. 77, no. 1, pp. 13–19, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022202X15461251

[37] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 553–560, iSSN: 1063-6919.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, conference Name: IEEE Signal Processing Letters.

[39] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1867–1874, iSSN: 1063-6919.

[40] D. McDuff and E. Blackford, "iPhys: An Open Non-Contact Imaging-Based Physiological Measurement Toolbox," *arXiv:1901.04366 [cs]*, Jan. 2019, arXiv: 1901.04366. [Online]. Available: http://arxiv.org/abs/1901.04366

[41] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, and K.-W. Cheung, "Block-based adaptive ROI for remote photoplethysmography," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6503–6529, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s11042-017-4563-7

[42] P. Li, Y. Benezeth, K. Nakamura, R. Gomez, and F. Yang, "Model-based Region of Interest Segmentation for Remote Photoplethysmography," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, vol. 4, Prague, Czech Republic, Oct. 2020, pp. 383–388. [Online]. Available: https://www.scitepress.org/PublicationsDetail.aspx?ID=2AmX9UVJhmY=&t=1

[43] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, May 2015, publisher: Optical Society of America. [Online]. Available: https://www.osapublishing.org/boe/abstract.cfm?uri=boe-6-5-1565

[44] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, Jun. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865517303860

[45] Q. Yang, S. Wang, and N. Ahuja, "Real-Time Specular Highlight Removal Using Bilateral Filtering," in *Computer Vision – ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer, 2010, pp. 87–100.

[46] A. S. M. Mosa, I. Yoo, and L. Sheets, "A Systematic Review of Healthcare Applications for Smartphones," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 67, Jul. 2012. [Online]. Available: https://doi.org/10.1186/1472-6947-12-67

[47] C. L. Ventola, "Mobile Devices and Apps for Health Care Professionals: Uses and Benefits," *Pharmacy and Therapeutics*, vol. 39, no. 5, pp. 356–364, May 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4029126/

[48] M. N. K. Boulos, S. Wheeler, C. Tavares, and R. Jones, "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX," *BioMedical Engineering OnLine*, vol. 10, no. 1, p. 24, Apr. 2011. [Online]. Available: https://doi.org/10.1186/1475-925X-10-24

[49] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "Systematic analysis of video-based pulse measurement from compressed videos," *Biomedical Optics Express*, vol. 12, no. 1, pp. 494–508, Jan. 2021, publisher: Optical Society of America. [Online]. Available: https://www.osapublishing.org/boe/abstract.cfm?uri=boe-12-1-494

[50] E. Nowara and D. McDuff, "Combating the impact of video compression on non-contact vital sign measurement using supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, p. 1706–1712.

[51] T. B. Fitzpatrick, "The Validity and Practicality of Sun-Reactive Skin Types I Through VI," *Archives of Dermatology*, vol. 124, no. 6, pp. 869–871, Jun. 1988, publisher: American Medical Association. [Online]. Available: https://jamanetwork.com/journals/jamadermatology/fullarticle/549509

[52] J. G. Karippacheril and T. Y. Ho, "Data acquisition from S/5 GE Datex anesthesia monitor using VSCapture: An open source.NET/Mono tool," *Journal of Anaesthesiology, Clinical Pharmacology*, vol. 29, no. 3, pp. 423–424, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3788264/

[53] C. Tang, J. Lu, and J. Liu, "Non-contact heart rate monitoring by combining convolutional neural network skin detection and remote photoplethysmography via a low-cost camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1309–1315.

[54] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit," *npj Digital*

Medicine, vol. 2, no. 1, pp. 1–18, Dec. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-019-0199-5

[55] D. G. Altman and J. M. Bland, "Measurement in Medicine: The Analysis of Method Comparison Studies," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 32, no. 3, pp. 307–317, 1983, publisher: [Royal Statistical Society, Wiley]. [Online]. Available: https://www.jstor.org/stable/2987937

[56] E. M. Cahan, T. Hernandez-Boussard, S. Thadaney-Israni, and D. L. Rubin, "Putting the data before the algorithm in big data addressing personalized healthcare," *npj Digital Medicine*, vol. 2, no. 1, pp. 1–6, Aug. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-019-0157-2

[57] F. Jm, J. J, Y. M, G. L, H. L, and Z. Dm, "Virtual Care Expansion in the Veterans Health Administration During the COVID-19 Pandemic: Clinical Services and Patient Characteristics Associated with Utilization." *Journal of the American Medical Informatics Association : JAMIA*, Oct. 2020. [Online]. Available: https://europepmc.org/article/med/33125032

[58] V. Abedi, O. Olulana, V. Avula, D. Chaudhary, A. Khan, S. Shahjouei, J. Li, and R. Zand, "Racial, Economic, and Health Inequality and COVID-19 Infection in the United States," *Journal of Racial and Ethnic Health Disparities*, Sep. 2020.

[59] K. M. J. Azar, Z. Shen, R. J. Romanelli, S. H. Lockhart, K. Smits, S. Robinson, S. Brown, and A. R. Pressman, "Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California," *Health Affairs*, vol. 39, no. 7, pp. 1253–1262, May 2020, publisher: Health Affairs. [Online]. Available: https://www.healthaffairs.org/doi/10.1377/hlthaff.2020.00598

[60] D. R. Holtgrave, M. A. Barranco, J. M. Tesoriero, D. S. Blog, and E. S. Rosenberg, "Assessing racial and ethnic disparities using a COVID-19 outcomes continuum for New York State," *Annals of Epidemiology*, vol. 48, pp. 9–14, Aug. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047279720302076

[61] T. Gu, J. A. Mack, M. Salvatore, S. Prabhu Sankar, T. S. Valley, K. Singh, B. K. Nallamothu, S. Kheterpal, L. Lisabeth, L. G. Fritsche, and B. Mukherjee, "Characteristics Associated With Racial/Ethnic Disparities in COVID-19 Outcomes in an Academic Health Care System," *JAMA network open*, vol. 3, no. 10, p. e2025197, 2020.

[62] G. A. Mensah, "Cardiovascular Diseases in African Americans: Fostering Community Partnerships to Stem the Tide," *American Journal of Kidney Diseases*, vol. 72, no. 5, Supplement 1, pp. S37–S42, Nov. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272638618308321

[63] R. Kats, “US telemedicine users will surpass 40 million this year,” *eMarketer*, Nov. 2020. [Online]. Available: https://www.emarketer.com/content/us-telemedicine-users-will-surpass-40-million-this-year

[64] “U.S. Telemedicine Market Share, Size, Trends, Industry Analysis Report By Component (Hardware, Software & Services); By Application (Teleradiology, Telepsychiatry, Telestroke, Tele-ICU, Teledermatology, Teleconsultation); Mode of Delivery (Mobile Health Apps, Virtual, Telehealth Portals & Kiosks), By End User (Providers, Payers, Patients); Segment Forecast, 2020 - 2027,” Polaris Market Research, New York, Tech. Rep. PM1672, Aug. 2020. [Online]. Available: https://www.polarismarketresearch.com/industry-analysis/us-telemedicine-market

[65] “Current population survey (CPS),” The United States Census Bureau, Nov 2020. [Online]. Available: https://www.census.gov/programs-surveys/cps.html