A Comparative Study of rPPG-Based Pulse Rate Tracking Algorithms for Fitness Activities

Qiang Zhu [1], Chau-Wai Wong [1], Zachary Lazri [2], Mingliang Chen [1], Min Wu [1], and Chang-Hong Fu [1]

[1]Affiliation not available
[2]University or Maryland

October 30, 2023

Abstract

Recent studies have shown that subtle changes in human face color due to heartbeats can be captured by regular RGB digital video cameras. It is possible, though challenging, to track one's pulse rate when a video contains significant subject's body motions in a fitness setting. The robustness gain in the recently proposed systems is often achieved by adding or changing certain modules in the system's pipeline. Most existing works, however, only evaluate the performance of the pulse rate estimation at the system level of particular pipeline configurations, whereas the contribution from each module remains unclear. To gain a better understanding of the performance at the module level and facilitate future research in explainable learning and artificial intelligence (AI) in physiological monitoring, this paper conducts an in-depth comparative study at the module level for video-based pulse rate tracking algorithms; a special focus is placed on challenging fitness scenarios involving significant movement. The representative efforts over the past decade in the field are reviewed, upon which a reconfigurable rPPG framework/pipeline is constructed comprising of major processing modules. For performance attribution, different candidates for each module are evaluated while having the rest of modules fixed. The performance evaluation is based on a signal quality metric and four pulse-rate estimation metrics and uses the simultaneously recorded ECG-based heart rate measurement as a reference. Experimental results using a challenging fitness dataset reveals the synergy between pulse color mapping and adaptive motion filtering in obtaining accurate pulse rate estimates. The results also suggest the importance of robust frequency tracking for accurate PR estimation in low signal-to-noise ratio fitness scenarios.

SUPPLEMENTAL MATERIAL: DATA COLLECTION SETUPS

a) Environment: In order to test the robustness of the system in a fitness-in-the-wild setup, we conducted the experiments in two regular apartment fitting rooms. The illumination sources only involved the existing lighting equipment in each room including several over-the-top florescent lights and possible diffused sunlight passing into the room through glass walls or windows. No backdrop was placed during the recording. The presence of other subjects exercising or entering the scene is possible, as no regulation is placed to restrict people from entering the room.

b) Devices and Reference Signal: The first 5 stationary bike videos were captured by the rear camera of a Huawei P9 mobile phone. The other 20 videos involving the elliptical machine and treadmill motions were captured by the rear camera of an iPhone 6s mobile phone. The shutter speed of both sensors was set as constant to minimize the possibility of introducing artifacts by the built-in automatic features, e.g., automatic exposure. The aperture size was kept at a relatively low value to ensure the focus of the face at all times. We obtained the subject's reference heart rate by simultaneously measuring the subject's electrocardiogram (ECG) with a chest strap monitor (Model: Polar H7).

c) Placement of the Sensors: The mobile camera was placed on the holder of the stationary bike, affixed on a tripod, or held by the hands of a person other than the test subject. The camera is placed in front of the subject face at a distance of about 1 meter away at approximately the same height as the subject's face during the recording. The ECG chest strap was worn underneath the subject's cloth and in direct contact with the subject's skin to maximize the SNR of the reference ECG signal.

d) Participants: Two male Asian subjects are involved in the experiment. The skin tone of both subjects is classified as Type III according to the Fitzpatrick skin scale [1]. Among all the videos in the dataset, 5 treadmill videos and 5 elliptical machine videos belong to one subject. The remaining 15 belong to the other. Based on the most recent medical examination results, none of the subjects were diagnosed with any known CVDs or pulmonary diseases.

REFERENCES

[1] 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.

A Comparative Study of rPPG-Based Pulse Rate Tracking Algorithms for Fitness Activities

Qiang Zhu, Chau-Wai Wong, *Member, IEEE,* Zachary McBride Lazri, *Graduate Student Member, IEEE,* Mingliang Chen, Chang-Hong Fu, *Member, IEEE,* and Min Wu, *Fellow, IEEE*

***Abstract*—Recent studies have shown that subtle changes in human face color due to heartbeats can be captured by regular RGB digital video cameras. It is possible, though challenging, to track one's pulse rate when a video contains significant subject's body motions in a fitness setting. The robustness gain in the recently proposed systems is often achieved by adding or changing certain modules in the system's pipeline. Most existing works, however, only evaluate the performance of the pulse rate estimation at the system level of particular pipeline configurations, whereas the contribution from each module remains unclear. To gain a better understanding of the performance at the module level and facilitate future research in explainable learning and artificial intelligence (AI) in physiological monitoring, this paper conducts an in-depth comparative study at the module level for video-based pulse rate tracking algorithms; a special focus is placed on challenging fitness scenarios involving significant movement. The representative efforts over the past decade in the field are reviewed, upon which a reconfigurable rPPG framework/pipeline is constructed comprising of major processing modules. For performance attribution, different candidates for each module are evaluated while having the rest of modules fixed. The performance evaluation is based on a signal quality metric and four pulse-rate estimation metrics and uses the simultaneously recorded ECG-based heart rate measurement as a reference. Experimental results using a challenging fitness dataset reveals the synergy between pulse color mapping and adaptive motion filtering in obtaining accurate pulse rate estimates. The results also suggest the importance of robust frequency tracking for accurate PR estimation in low signal-to-noise ratio fitness scenarios.**

***Index Terms*—Heart/pulse rate, remote-photoplethysmography (rPPG), fitness exercise, pulse color mapping, motion compensation, frequency tracking, explainable AI**

I. Introduction

Pulse rate (PR) is an important noninvasive, time-efficient measure to monitor the training load and quantify the athletes' response to support the optimization of the effectiveness and safety of training [1]–[5]. PR monitoring during the training can help coaches and trainees to achieve individual and/or team training objectives.

The conventional cardiac monitoring, such as chest-strap heart rate monitor based on electrocardiography (ECG) [6] is not comfortable and may cause skin irritation during prolonged use; photoplethysmography (PPG) [7], [8] in the form of wristband or watch is prone to motion artifacts and has limited accuracy compared to chest strap. Contact-free monitoring of the PR using videos of human faces, known as remote-photoplethysmography (rPPG), is a user-friendly approach compared to conventional contact-based ones such as electrodes, chest straps, and finger clips. Such monitoring system extracts from a facial video a 1-D oscillating face color signal that has the same frequency as the heartbeat. The ability to measure PR without direct contact is attractive and gives it potentials in such applications as smart health and sports medicine, and cardiac rehabilitation.

In this paper, we ask and seek to answer the following questions: *(i) How can one's pulse rate be accurately tracked from facial videos captured in a typical fitness setup*? *(ii) How much impact does each major block of a pulse rate tracking pipeline have on the overall performance*? Addressing these questions requires us to understand and tackle multiple challenges in fitness rPPG sensing coming from every component of the rPPG sensing system, namely, the camera, the illumination conditions, and the subject [9]. In a fitness setup, motion-induced changes of illumination intensity may dominate the reflected light from the facial skin because pulse-induced color variations are usually much subtler. The measurement is also associated with such nuisance sources as the sensor and quantization noise. To extract the pulse signal that may have a much smaller magnitude than the dominating video components and may be affected by nuisance signals, dedicated algorithms need to be designed to tackle the challenges synergistically.

The last decade and a half has witnessed a rapidly increasing number of articles addressing the pulse rate estimation for still/rest cases or with relatively small motions [10]–[24]. Among the prior publications [25]–[29], the pulse rate estimation in the fitness scenario with significant subject motion is either not considered and reported [26], [27], the performance is not quantitatively examined [28], or the performance highly deviates from the reference [29]. Meanwhile, the evaluation process reported in most papers stays at the system level, whereas the contribution of the specific choice of each system module over other alternatives remains unclear. Such coarse evaluation may hinder the understandings of the design options of each system component, and limit the progress of research and development efforts.

In this paper, we investigate what techniques can provide the best possible performance for fitness exercise videos. We construct a framework as shown in Fig. 1 that contains typical

This work was conducted when Qiang Zhu and Mingliang Chen were with the University of Maryland, College Park.

Chau-Wai Wong was with the University of Maryland, College Park when the work was started, and is now with North Carolina State University (e-mail: chauwai.wong@ncsu.edu).

Chang-Hong Fu is with the Nanjing University of Science and Technology, China (e-mail: enchfu@njust.edu.cn).

Min Wu and Zachary Lazri are with the University of Maryland, College Park (e-mail: minwu@umd.edu, zlazri@umd.edu).

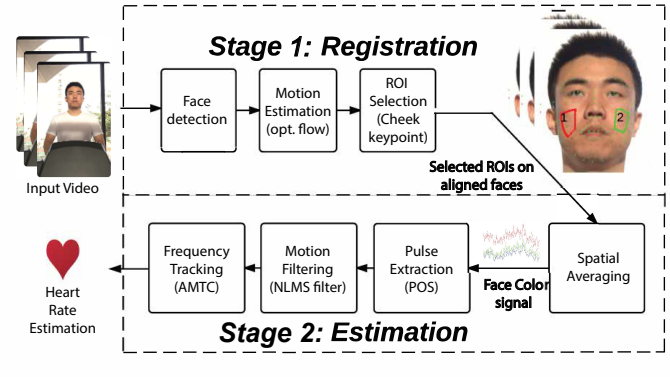

Fig. 1. The proposed modularized system for the heart rate monitoring for fitness exercise videos with the optimal module configurations in parentheses.

building blocks agreed within the literature as a platform to evaluate various configurations. Some key building blocks include **face registration**, **motion artifacts removal**, and **frequency tracking** [30]. A candidate algorithm for each module is listed in parentheses. For example, to accommodate fitness activities, motion artifacts within the intermediate face color signal can be removed by an adaptive filtering algorithm such as the normalized least mean squares (NLMS) [31]. An in-depth comparative study is conducted in the second half of the paper to examine the detailed contribution of each system module and determine the combination of modules that is likely to provide the best performance of the overall system.

The rest of the paper is organized as follows. In Section II, we review the last decade's research efforts in rPPG and the face skin reflection model adopted in this paper. In Section III, we describe a modular framework of rPPG-based PR estimation method specially designed for fitness exercises. In Section IV, we present the experimental setups. In Section V, we conduct a comparative study of the PR estimation using different module combinations and provide discussion. In Section VI, we conclude the paper.

II. Related Work on Remote Pulse Rate Measurement

In this section, we review the recent progress made in the rPPG research for PR estimation. The works listed and discussed in this paper can in no way be exhaustive. Nevertheless, the contributions these works have brought to handling the various challenges associated with PR extraction from videos have enabled the design of the modular system proposed in this work. We extend our discussion on the prior art below from the perspectives of *region of interest (ROI) selection* and *motion-resilient pulse extraction*.

A. ROI Selection

ROI selection, aiming to locate the ROI consistently in each video frame in accordance with the subject's motion, is instrumental to obtain reliable rPPG signals. The selection of the face skin region as the ROI for pulse measurement is mainly due to the following two facts. First, compared with other parts of the human body, the face is less likely to be covered by other materials such as clothes. Second, owing to the development of the recent computer vision techniques, a subject's face can be accurately located and tracked from a video sequence, even when the background is complex and the video is noisy. We summarize three main approaches for ROI selection found in prior art as follows.

Manual Selection: When a subject is static in a video, it will be accurate to manually select a single ROI from the first frame of the video and extract face color signal using the same ROI in the subsequent video frames [13], [32], [33]. This may not be a viable solution even when the subject is instructed to remain still, because such involuntary motion as ballistocardiographic (BCG) and respiratory-induced motions are possible and may contaminate the desired rPPG signal.

Automatic Face Detection: Face detection process is necessary for automatically selecting the ROI when the video contains the subject's motion. This can be achieved by either frame-wise face detection [10] or face tracking via some "good features for tracking" [15], [20]. These methods have potential to accurately localize and track the face, but in the presence of large motion displacement a fine-grain local alignment method may need to be introduced to ensure the stability of the detected ROI region, which is critical for accurate PR extraction.

Skin Detection: The non-skin facial pixels (*e.g.*, lips and eyes) have little-to-no contributions to the pulse extraction and might bring additional motion artifacts when the subject is talking or blinking. It is thus reasonable to exclude those non-skin pixels in each frame. Wang et al. [34] proposed an online learning approach to train a skin pixel detector using the first several frames. This subject- and scene-specific learning approach is robust to the change of illumination source and the subject's skin tone. However, the system might generate false detection results when the illumination condition changes temporally.

B. Motion-Resilient Pulse Signal Extraction

Green channel generates the highest pulse-signal strength among the three color channels, as the oxyhemoglobin and deoxyhemoglobin have greater absorptivity in green light compared with red or blue. This fact motivated a series of works [12], [15], [18], [32], [35] to use the green channel for extracting the pulse information.

Blind source separation (BSS) methods improve the system robustness by incorporating additional information from other color channels. BSS is applied to demix the pulse signal from the R, G, and B measurements by assuming the sources are uncorrelated (PCA-based [36]) or independent (ICA-based [37]). A newer work [38] based on BSS uses ensemble empirical mode decomposition to extract the intrinsic mode functions from multiple ROIs defined on the face, which are then passed to a BSS algorithm for demixing. The most periodic component is selected as the pulse signal after the source separation is performed. Each BSS algorithm produces the optimal source separation results when the pulse signal, noise, and interfering components exhibit the aforementioned statistical behaviors. However, in a fitness scenario when strong periodic motion artifacts enter the RGB-signal sourced

from the face, such statistical assumptions might be violated, and the channel selection algorithm may mistakenly treat a motion component as the estimated pulse signal.

Skin model-based methods [25]–[27], [29], [34], [39] are proposed to avoid the difficulties in selecting the correct component in BSS methods by providing a best-guessed color projection direction for extracting the pulse source. With prior knowledge about the skin-tone color vector obtained from a large scale dataset, CHROM algorithm [29] maps the temporally normalized RGB signals to a color plane orthogonal to the specular component, and the pulse signal is obtained via an alpha-tuning operation. POS algorithm [26] adopts the same skin reflection model but instead maps the normalized RGB signals to the color plane orthogonal to the intensity variation direction aiming to eliminate the motion artifacts in that direction. The pulse color direction is then searched within a 90-degree sector bounded by two predefined color directions. The hue change on the skin is another useful feature for pulse extraction [40]. 2SR [41] exploits a pulse-induced hue change in a subject-dependent manner by tracking the principal direction of the hue channels. All these color mapping schemes except 2SR use linear combinations of RGB color channels to extract the pulse. The algorithmic differences concerning the assumptions of the relations of the source signals is reflected by the demixing weights applied to color channels. For a more detailed discussion about the strengths and weaknesses of the algorithms mentioned above, we referred the readers to [26]. Recognizing that the pulse signal is quasi-periodic due to subtle variability between consecutive heart beats, Pai et al. [39] use an amplitude-modulated-frequency-modulated (AM-FM) framework to model the pulse signal. By applying the CHROM algorithm and filtering strategies to the spatially averaged RGB skin pixels, they aim to isolate the fundamental frequency of the pulse signal.

Adapted skin model-based methods adapt one of the skin model-based methods to make it more robust. Tulyakov et al. [42] warp the face images to a grid and applied the CHROM algorithm to the spatially averaged RGB signals in each grid region. A matrix completion algorithm is then applied to extract the common pulse signal present in all grid regions. Demirezen et al. [43] apply nonlinear mode decomposition following the CHROM step to obtain more sinusoidal intermediate signal before applying Fourier analysis to extract the pulse rate. Song et al. [23] replace alpha tuning of CHROM with a semi-BSS algorithm to separate the pulse and chrominance information after observing that the alpha-tuning module breaks down when intensity and pulsatile information in the signal are of similar magnitude.

Neural-network-based methods [21], [22], [44]–[46] leverage the training data to perform PR estimation. Hsu et al. [44] treat the time-frequency representation of the extracted signal as an image and estimates the PR with a convolutional neural network (CNN). End-to-end rPPG learning systems [21], [22] which utilize the temporal and spatial attention modules for automatic channel weighting and signal selection, are appealing and outperform other non-learning-based methods in terms of the PR estimation accuracy. Yu et al. [45] test both ConvLSTM and 3D-CNN model structures and find that the 3D-CNN produces higher accuracy because of the structure's ability to better handle spatial and temporal information jointly. Rather than inputting video frames directly into the neural network, Niu et al. [47] input so-called MSTmaps constructed from spatially averaged RGB and YCbCr signals from various regions of the face into the network, which is designed to disentangle the pulse information from the noise sources contributing to these signals. However, for these trained models to generalize, the training and testing datasets need to be identically distributed. This makes it hard to analyze the PRs in the videos of people captured in different scenes and who may have highly variable PR levels, such as in resting versus exercise situations.

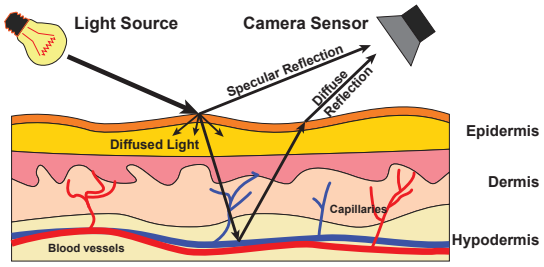

Fig. 2. Illustration of the composition of light reflected from human skin tissue and captured by an RGB camera sensor used for pulse signal modeling (modified based on [26]).

C. Modeling the Skin Reflection and Motion

Consider the situation when a piece of human skin containing pulsatile blood is illuminated by a light source as shown in Fig. 2. The reflected light from the skin surface can be characterized as the specular and diffuse reflections.[1] The **specular reflection** takes up approximately 4–7% of visible light reflected from the stratum corneum in the epidermis layer [49], [50]. The reflectance geometry among skin surface, light source, and the camera sensor determines the strength of the specular reflection [48]. The spectral distribution of the measured specular reflection component in a camera is a function of the spectral distribution of the light source and the spectral response of the camera. Thus for a single light source with fixed spectral distribution, the spectral distribution of the specular reflection will be constant regardless of the subject's motion. The **diffuse reflection** can be further decomposed into the epidermal reflection and dermal reflection [49]. The spectral distribution of the epidermal reflection is mostly determined by the concentration of the melanin in the epidermis layer. The dermal reflection, on the other hand, carries the blood pulse information. The variations of the blood volume, especially the amount of oxygenated and deoxygenated hemoglobin in the dermis layer, influence the color and intensity of the dermal reflection. The following two assumptions about the skin reflection are made to facilitate the modeling process: (*i*) Diffuse reflection from the skin surface

[1]Some literature [48] adopts the terms *interface* and *body* reflections rather than the *specular* and *diffuse* reflections. To avoid confusion and maintain the consistency of the terminology used in the rPPG community, we use the terms *specular* and *diffuse* reflections in this paper.

is isotropic with respect to rotation about the surface normal; and (*ii*) no interreflection is present among surface, as we approximately treat head as a convex shape.

Based on the analysis and assumptions made above, one can arrive at the reflection formulation[2] based on the skin characteristics and the dichromatic reflection model (DRM) [26], [27], [48]:

$$\mathbf{C}^{\ell}(t) = I(t)\,[\mathbf{v}_{\mathrm{s}}(t) + \mathbf{v}_{\mathrm{d}}(t)] + \mathbf{v}_{\mathrm{n}}^{\ell}(t), \tag{1}$$

where $\mathbf{C}^{\ell}(t) \in \mathbb{R}^3$ denotes the vector of the intensity values of the R, G, and B channels of the ℓth skin-pixel at time t; $I(t)$ denotes the intensity of the light source arrived at the corresponding skin surface; $\mathbf{v}_{\mathrm{s}}(t)$ and $\mathbf{v}_{\mathrm{d}}(t)$ denote the specular and diffuse reflection, respectively; $\mathbf{v}_{\mathrm{n}}^{\ell}(t)$ denotes camera's sensor noise and the video or image compression noise. Specifically, $\mathbf{v}_{\mathrm{s}}(t)$ and $\mathbf{v}_{\mathrm{d}}(t)$ can be decomposed as:

$$\mathbf{v}_{\mathrm{s}}(t) = \mathbf{u}_{\mathrm{s}} \cdot [s_0 + s(t)]\,, \tag{2a}$$

$$\mathbf{v}_{\mathrm{d}}(t) = \mathbf{u}_{\mathrm{d}} \cdot d_0 + \mathbf{u}_{\mathrm{p}} \cdot p(t), \tag{2b}$$

where $\mathbf{u}_{\mathrm{s}}$, $\mathbf{u}_{\mathrm{d}}$, and $\mathbf{u}_{\mathrm{p}} \in \mathbb{R}^3$ denote the unit color vectors of the light spectrum, the skin tissue, and the pulse, respectively; s_0 and d_0 denote the strengths of the DC component of the specular and diffuse refection, respectively; $s(t)$ and $p(t)$ denote the strengths of the AC component of the specular reflection and pulse signal, respectively. Note that the temporal variations of both $I(t)$ and $s(t)$ come from the subject's motion; the variation of $I(t)$ is affected by the distance of the light source to the skin surface, whereas $s(t)$ is influenced by the variation of the surface normal direction.

We use $\mathbf{C}(t) = \sum_{\ell=1}^{L} \mathbf{C}^{\ell}(t)/L$ to denote the spatially averaged RGB vector, where L is the number of skin pixels involved. We substitute (1) into $\mathbf{C}(t)$ and let $I(t) \triangleq [1 + i(t)]\, I_0$ and $\mathbf{u}_{\mathrm{c}}\, c_0 \triangleq \mathbf{u}_{\mathrm{s}}\, s_0 + \mathbf{u}_{\mathrm{d}}\, d_0$, where $i(t)$ indicates the illuminance change. We further assume zero phase difference of the pulse signal at any point of the face, and $\{\mathbf{v}_{\mathrm{n}}^{\ell}(t)\}_{\ell}$ for any fixed t is a zero-mean white Gaussian process. We therefore obtain:

$$\mathbf{C}(t) \approx I_0\,[1 + i(t)]\,[\mathbf{u}_{\mathrm{c}} \cdot c_0 + \mathbf{u}_{\mathrm{s}} \cdot s(t) + \mathbf{u}_{\mathrm{p}} \cdot p(t)] \tag{3a}$$

$$\approx I_0\,[\mathbf{u}_{\mathrm{c}} \cdot c_0 + \mathbf{u}_{\mathrm{c}} \cdot c_0\, i(t) + \mathbf{u}_{\mathrm{s}} \cdot s(t) + \mathbf{u}_{\mathrm{p}} \cdot p(t)], \tag{3b}$$

where the absence of the noise term $\mathbf{v}_{\mathrm{n}}^{\ell}(t)$ from (3a) is due to spatial averaging when a large number of skin pixels are used, and the approximation of (3b) is because the second-order cross AC-terms are much smaller than the remaining DC terms and first-order AC terms.

As pointed out in [27], the limitations of model (3) include the assumption of the single light source and the assumption that the subject's motion only creates a single specular variation direction, *i.e.*, $\mathbf{u}_{\mathrm{s}}$, in the RGB space. This is unfortunately unrealistic because the skin surface might receive reflected light from other objects with nonuniform light spectrum absorbance in the scene, and the spectrum of such a reflected light differs from that of the light source. To capture this complication in our model, we assume a total of J light sources present in the scene, including the reflected light from other objects in the scene. Equation (3) therefore becomes:

$$\begin{aligned}\mathbf{C}(t) \approx & \underbrace{\sum_{j=1}^{J} \mathbf{u}_{\mathrm{c},j} \cdot I_{0,j} \cdot c_{0,j}}_{\textbf{DC}} + \underbrace{\sum_{j=1}^{J} \mathbf{u}_{\mathrm{c},j} \cdot I_{0,j} \cdot c_{0,j} \cdot i_j(t)}_{\textbf{Intensity}} \\ & + \underbrace{\sum_{j=1}^{J} \mathbf{u}_{\mathrm{s},j} \cdot I_{0,j} \cdot s_j(t)}_{\textbf{Specular}} + \underbrace{\left(\sum_{j=1}^{J} \mathbf{u}_{\mathrm{p},j} \cdot I_{0,j}\right) \cdot p(t)}_{\textbf{Pulse}},\end{aligned} \tag{4}$$

where j denotes the jth light source; $i_j(t)$ and $s_j(t)$ denote the intensity variation signal and specular variation signal of the jth light source [27], respectively. The DC component $\sum_{j=1}^{J} \mathbf{u}_{\mathrm{c},j} \cdot I_{0,j}\, c_{0,j}$ can be estimated and subtracted from (4) by using the short-term smoothing approach introduced in [25], [26] or detrending methods introduced [51], [52]. Since both $i_j(t)$ and $s_j(t)$ come from the subject's motion, they can be approximated as different linear combinations of the motion components, *i.e.*, $i_j(t) = \sum_{k=1}^{K} a_{j,k}\, m_k(t)$ and $s_j(t) = \sum_{k=1}^{K} b_{j,k}\, m_k(t)$, where $m_k(t)$ denotes the kth motion component. If we denote $\tilde{\mathbf{C}}(t)$ as the detrended signal after removing the DC component, we finally arrive at

$$\tilde{\mathbf{C}}(t) = \underbrace{\sum_{k=1}^{K} \mathbf{u}_{\mathrm{m},k} \cdot m_k(t)}_{\textbf{Motion}} + \underbrace{\mathbf{u}'_{\mathrm{p}} \cdot p(t)}_{\textbf{Pulse}}, \tag{5}$$

where $\mathbf{u}_{\mathrm{m},k} \triangleq \sum_{j=1}^{J} (\mathbf{u}_{\mathrm{c},j} \cdot a_{j,k}\, c_{0,j}\, I_{0,j} + \mathbf{u}_{\mathrm{s},j} \cdot b_{j,k}\, I_{0,j})$ is the color vector of the kth motion component, and $\mathbf{u}'_{\mathrm{p}} \triangleq \sum_{j=1}^{J} \mathbf{u}_{\mathrm{p},j} \cdot I_{0,j}$ is the color vector of the pulse component. Equation (5) reveals that it is possible to completely separate the pulse term from the motion term via linear projection only if $\mathbf{u}'_{\mathrm{p}}$ is simultaneously orthogonal to $\mathbf{u}_{\mathrm{m},1}, \ldots, \mathbf{u}_{\mathrm{m},K}$. This is almost never the case when a subject is performing physical exercises in an uncontrolled environment. In this scenario, the motion subspace spanned by $\{\mathbf{u}_{\mathrm{m},k}\}_{k=1}^{K}$ is highly likely to have a nonnegligible component along the pulse color direction, making the pulse component $\mathbf{u}'_{\mathbf{p}} \cdot p(t)$ not completely linearly separable from the motion.

To alleviate the drawback of not being able to completely remove motion components through such linear projection based algorithms as POS, in Section III-B, we use an adaptive motion filtering module to further remove motion artifacts. Additional efforts to combat fitness motion include (*i*) reducing the source of motion in the RGB intensity signal $\mathbf{C}^{\ell}(t)$ through a precise alignment of the face ROI, and (*ii*) using a robust frequency-trace tracking algorithm that leverages temporal correlation between consecutive human PR values. All these efforts jointly contribute to a robust and accurate extraction of PR signals.

III. A Modular Framework for Fitness rPPG

In this section, we first present the general modularized fitness rPPG framework for PR extraction, followed by a detailed discussion of the module setup that leads to the highest accuracy of the overall system.

[2]For the completeness of this paper, we briefly review the modeling process that has been presented in detail in [26], [27]. The terminology used in the two papers are incorporated in this paper for consistency.

A. General rPPG Framework

The general rPPG framework for PR extraction as shown in Fig. 1 consists of seven modules, five of which are considered to be customizable with different candidate algorithms. The pipeline starts with face detection since only the skin pixels on the face are useful for extracting the pulse signal. The next two modules include motion estimation and ROI selection. The ROI is used to define exact regions on the face from which we will aim to extract the pulse signal. Since there may be displacement in a region from frame to frame due to the movement of the subject, a motion estimation module is used to align the face in each frame before defining the ROI to ensure that the face is stabilized throughout the video. A spatial averaging module is applied to the pixels inside the stabilized ROI of each frame to obtain temporal R, G, and B signals $\mathbf{C}(t)$ with boosted signal-to-noise ratio (SNR) levels. The pulse extraction module uses a channel combination algorithm, as described in Section II-B, to obtain a 1-D channel combined signal $c_{\text{pos}}(t)$ with most lighting and motion artifacts removed. This signal can be further processed to obtain a cleaner pulse signal $\tilde{c}_{\text{pos}}(t)$ through additional motion filtering. In the final module of the system, the estimated PR signal can be obtained by applying a frequency-tracking algorithm.

In the next subsection, we provide detailed descriptions of the algorithms used in this framework that achieve the best experimental results presented in Section V. The algorithms that optimize each module of the framework are shown in parentheses in Fig. 1. Specifically, (*i*) an optical flow-based motion estimation and compensation algorithm is used to minimize face registration error, (*ii*) the POS algorithm [26] is used to remove the remaining motion artifacts from the channel combined signal by "subtracting" the motion information available from the visual track using a normalized least mean square (NLMS) filter [31], and (*iii*) the PR signal is extracted using a robust frequency tracker named the adaptive multi-trace carving (AMTC) algorithm [30], [53], [54].

B. Optimized Framework

1) ***Precise Face Registration via Optical Flow [52]:*** We use the Viola–Jones face detector [55] to obtain rough estimates of the location and scale of the face, effectively generating a pre-aligned video for the facial region. Optical flow is applied next to fine-tune the facial alignment.

In our problem, two facial images likely have a global color difference due to the heartbeat, making it imprecise to use the illumination consistency assumption that widely adopted in designing standard optical flow algorithms. Instead, to ensure that an optical flow algorithm can precisely align two facial images with a subtle color difference, one has to assume more generally that the intensity I of a point in two frames is related by an affine model, namely,

$$I(x + \Delta x_t, y + \Delta y_t, t + 1) = (1 - \epsilon_t)\, I(x, y, t) + b_t, \quad (6)$$

where $(\Delta x_t, \Delta y_t)$ is the motion vector tracking the point (x, y) from frame index t to $t + 1$, and ϵ_t and b_t control the scaling and bias of the intensities between two frames, respectively. When $\epsilon_t = b_t = 0$ for all t, the model degenerates to fulfill the illumination consistency assumption. Applying a standard optical flow algorithm will result in a mismatch between the modeling assumption and the characteristics of the rPPG facial images. The bias of the estimated motion vectors is reported to be at the same order of magnitude compared to the intrinsic error of the optical flow system [52]. To alleviate potential bias, different strategies can be applied. For example, using a global flow regularization strategy [56] or a coarse-to-fine hierarchical searching strategy [56], [57] instead of doing one-shot Taylor-based local approximation. In this study, we use Liu's optical flow implementation [58] of Brox et al.'s method [56]. Modern deep learning based optical flow algorithms [59], [60] may also be used.

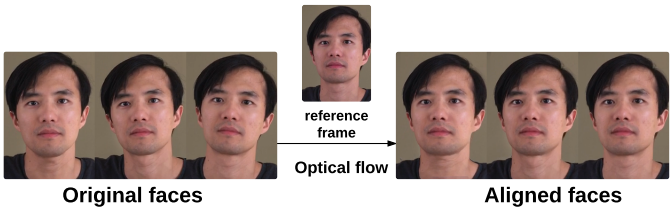

Fig. 3. Face images from a video segment before and after optical-flow-based motion compensation, illustrating the use of the motion estimation module.

To avoid potential occlusion issues when applying optical flow-based motion compensation, we divide each video into small temporal segments with one frame overlapping for successive segments and use the frame in the middle of the segment as the reference. Fig. 3 shows a few facial images from the same segment before and after the application of optical flow. The faces are precisely aligned. Using facial landmarks identified by the method proposed by Yu et al. [61], we construct a polygon on each cheek to represent an ROI and perform spatial averaging for each of the R, G, and B channels to obtain three 1-D time-series signals for each segment. We then temporally concatenate these signals, removing the discontinuities between consecutive segments by taking the difference between the first and last point of each segment. We apply a detrending algorithm [52] to remove the DC and slowly varying components for each color channel. Finally, we temporally normalize each of the resulting 1-D time series to obtain the standardized vector-valued RGB time-series signal, $\tilde{\mathbf{C}}(t)$, to be further processed in the next module.

2) ***Motion Artifacts Removal via Adaptive Filtering:*** This module begins by linearly mapping $\tilde{\mathbf{C}}(t)$ to a specific color direction in the RGB space to generate a 1-D pulse signal. The pulse color mapping schemes have been extensively investigated in [26] and [27]. We note that the design of the pulse color mapping algorithms discussed in this paper is not within the contributions of this work, although different pulse color mapping approaches [26], [27], [29], [37] are implemented and evaluated in the Section V.

Without loss of generality, we assume $\tilde{\mathbf{C}}(t)$ will be mapped to the POS direction [26], which is one of the most robust color feature representations, containing highest relative pulse strength. We denote the projected 1-D channel combined

signal as $c_{\text{pos}}(t)$. According to (5), we have

$$c_{\text{pos}}(t) = \mathbf{p}^{\intercal}\tilde{\mathbf{C}}(t) = \underbrace{\mathbf{p}^{\intercal}\mathbf{u}'_{\text{p}} \cdot p(t)}_{\textbf{Pulse}} + \underbrace{\sum_{k=1}^{K} \mathbf{p}^{\intercal}\mathbf{u}_{\mathbf{m},\mathbf{k}} \cdot m_k(t)}_{\textbf{Motion Residue}}, \quad (7)$$

where $\mathbf{p} \in \mathbb{R}^3$ denotes the projection vector of the POS algorithm. The motion residue term in (7) is negligible when the illumination source is single, as the POS direction is orthogonal to the color direction of the motion-induced intensity change, and the specular change is suppressed via alpha tuning [29]. However, if the video is captured in an uncontrolled environment, the motion residue is often nonnegligible, and may even have a higher strength than the pulse term.

To adaptively track and decouple the possibly time-varying signal correlation between the motion residue and pulse signal in (7), we apply the normalized least mean square (NLMS) filter [31]. We denote the estimated face motion sequence in horizontal and vertical directions as $m_x(t)$ and $m_y(t)$. The structure of the filtering framework is shown in Fig. 4(a). We treat $c_{\text{pos}}(t)$ as the filter's observed response at time instant t. We treat the motion tap vector $\mathbf{m}(t) \triangleq [m_x(t-M+1), m_x(t-M+2), ..., m_x(t), m_y(t-M+1), m_y(t-M+2), ..., m_y(t)]^{\intercal}$ as the input and $\tilde{c}_{\text{pos}}(t)$ as the output of the system and also the error signal. The estimated tap-weight vector of the transversal filter is denoted as $\hat{\mathbf{w}}(t)$, and the weight control mechanism follows the NLMS algorithm [31] as follows:

$$\tilde{c}_{\text{pos}}(t) = c_{\text{pos}}(t) - \hat{\mathbf{w}}^{\intercal}(t)\,\mathbf{m}(t), \quad (8a)$$

$$\hat{\mathbf{w}}(t+1) = \hat{\mathbf{w}}(t) + \frac{\mu}{\|\mathbf{m}(t)\|^2}\,\mathbf{m}(t) \cdot \tilde{c}_{\text{pos}}(t), \quad (8b)$$

where μ denote the adaptation constant.Fig. 4(b)–(d) give an example of the adaptive filtering result using this approach. Note that the NLMS filter has successfully removed almost all the motion residue components from the channel combined signal $c_{\text{pos}}(t)$ while protecting the pulse information $p(t)$.

*3) **PR Signal Estimation via Frequency Tracking**:* Noting that two temporally consecutive heart/pulse rate measurements may not deviate too much from each other, we propose to exploit this PR continuity property to improve the estimation quality of PR signals by searching for the dominating frequency trace appearing in the signal's spectrogram image using the adaptive multi-trace carving (AMTC) algorithm [30], [53], [54]. Its details are briefly described. Letting $\mathbf{Z} \in \mathbb{R}_{+}^{M \times N}$ be the magnitude of a signal's spectrogram image, with N discrete bins along the time axis and M bins along the frequency axis, we aim to find the dominant *frequency trace*, $\mathbf{f} \triangleq \{(f(n), n)\}_{n=1}^{N}$, inside the image. Defining the energy of a trace to be $E(\mathbf{f}) \triangleq \sum_{n=1}^{N} \mathbf{Z}(f(n), n)$ and modeling the transition probability of the pulse rate, $P_m = \mathbb{P}[f(1) = m]$ and $P_{m'm} = \mathbb{P}[f(n) = m | f(n-1) = m']$, by a discrete-time Markov chain, the tracking problem is formulated as follows

$$\mathbf{f}^* = \underset{\mathbf{f}}{\text{argmax}} \quad E(\mathbf{f}) + \lambda P(\mathbf{f}), \quad (9)$$

where $P(\mathbf{f}) \triangleq \log P(f(1)) + \sum_{n=2}^{N} \log P(f(n)|f(n-1))$ controls the trace smoothness. This regularized tracking problem (9) can be solved by using dynamic programming to recursively track the path that leads to the highest point in *accumulated regularized maximum energy map* at the most recent time instant n [53], [54].

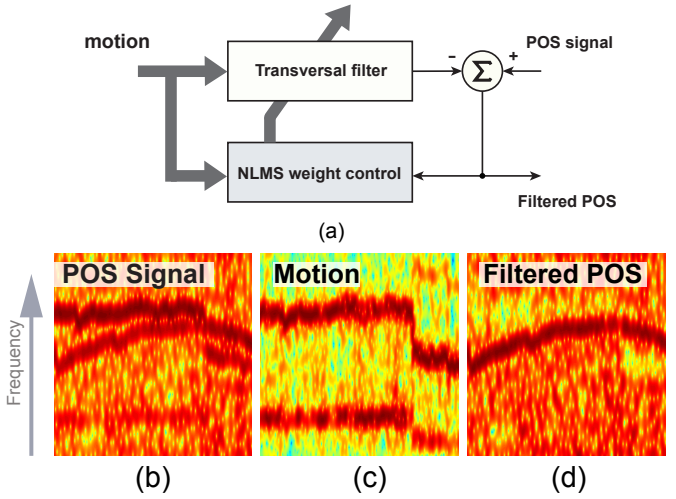

Fig. 4. (a) Adaptive motion compensation filter framework and spectrograms of (b) the POS signal $c_{\text{pos}}(t)$, (c) the combined normalized subject motion in horizontal and vertical directions, and (d) the filtered POS signal $\tilde{c}_{\text{pos}}(t)$. The NLMS filter removes the motion trace in the spectrogram of the POS signal, allowing for easier pulse tracking.

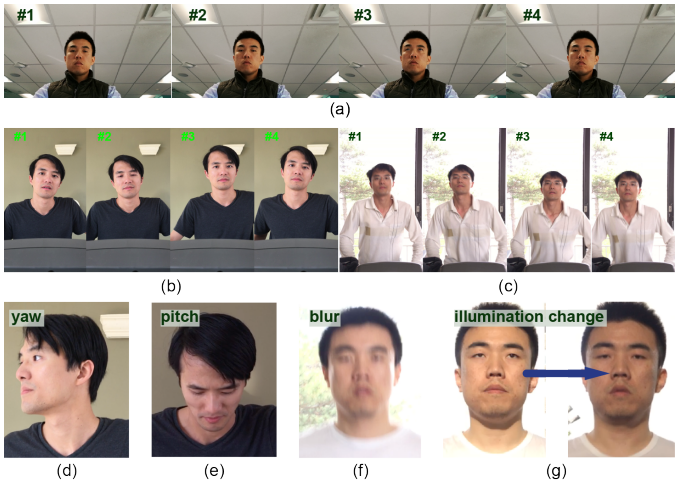

Fig. 5. Sample frames in fitness video dataset with three types of fitness motion: (a) stationary bike, (b) elliptical machine, (c) treadmill. The challenges in the dataset include head rotation in (d) yaw and (e) pitch, (f) motion blurred frames, and (g) significant illumination change on the face.

IV. Experimental Conditions

We evaluate the reconfigurable pipeline on a self-collected fitness exercise dataset to understand the factors in the PR estimation with fitness motions. The dataset has 25 videos in which 10 contain human motions on an elliptical machine, 10 contain motions on a treadmill, 5 contain motions on a stationary bike. The parameter settings and compared methods are described in the ensuing subsections.

A. Parameter Settings

The following parameters are used in our investigation unless otherwise stated:

1) The length of each video is about 3 minutes.
2) The frame rate is 30 frame per second. The resolution is 1280×720. The average bit rate is about 6 MB per second. The video codec is H.264/AVC.
3) The tap number for joint-channel NLMS is 8, and the NLMS learning rate/adaptation constant μ is 0.1.
4) Each video was empirically divided into segments of 1.5 secs with one frame overlap to ensure two frames being aligned by the optical flow method do not have significant occlusion due to long separation in time.
5) The spectrum analysis window length was set to 10 secs with 98% overlap to balance the trade-off between the resolution in the frequency and time domains. A Hamming window was applied in each analysis window, and the number of frequency bins in the normal PR range (50 to 240 bpm) was set as 1024 via padding zeros at the end of the analysis signal sequence. The transitional probability model used in the frequency tracking algorithm [53], [54] was a uniform random walk model with the width parameter $k = 1$ bpm.

B. Metrics of Performance Evaluation

a) Pulse Signal Quality: As in other papers, we use SNR as the pulse signal quality metric [26], [27], [29], [34]. The SNR in each spectral frame is defined as the ratio between the spectral energy around the first two harmonics of the reference PR and the remaining energy of the power spectrum. We express the SNR measure using the logarithmic decibel scale:

$$\text{SNR} = 10\log_{10}\left(\frac{\sum_{f\in\mathcal{F}} S_n(f)P(f)}{\sum_{f\in\mathcal{F}}\left[1-S_t(f)\right]P(f)}\right), \qquad (10)$$

where $S_n(f)$ is a defined binary window to select the frequency bins belong to the two-harmonics region; $P(f)$ is the power spectrum of the pulse signal; set $\mathcal{F} \triangleq \{f \mid 50 \text{ bpm} \le f \le 240 \text{ bpm}\}$

b) PR Estimation Accuracy: Three well-adopted metrics for pulse rate estimation accuracy are used in this study:

1) Root mean squared error (RMSE):

$$E_{\text{RMSE}} = \left(\frac{1}{N}\sum_{n=1}^{N}\left[\hat{f}(n) - f(n)\right]^2\right)^{\frac{1}{2}},$$

2) Error rate:

$$E_{\text{rate}} = \frac{1}{N}\sum_{n=1}^{N}\left|\hat{f}(n) - f(n)\right| \Big/ f(n),$$

3) Error count ratio:

$$E_{\text{count}} = \frac{1}{N}\left|\{n : |\hat{f}(n) - f(n)|/f(n) > \tau\}\right|,$$

4) Pearson's correlation coefficient:

$$\text{PCC} = \frac{\sum_{n=1}^{N}\left[\hat{f}(n) - \bar{\hat{f}}\right]\left[f(n) - \bar{f}\right]}{\left(\sum_{n=1}^{N}[\hat{f}(n) - \bar{\hat{f}}]^2 \sum_{n=1}^{N}[f(n) - \bar{f}]^2\right)^{\frac{1}{2}}},$$

where $|\{\cdot\}|$ denotes the cardinality of a countable set; N denotes the total number of the PR estimates; $\hat{f}(n)$, $f(n)$, $\bar{\hat{f}}$, and $\bar{f}$ denote the PR estimate at time instant n, ground-truth PR at time instant n, average PR estimate, and average reference PR, respectively. τ was empirically chosen to be 3%, determined from the spread of the frequency components.

V. Results and Discussions

As our proposed system consists of multiple modules with each focusing on a specific task, a holistic end-to-end system-level test would be insufficient to evaluate the contribution of each system component. In this section, we discuss the benchmark experimental results based on fine-level comparisons in terms of the motion estimation schemes, the pulse color mapping algorithms, the motion adaptive filtering operations, and the frequency estimation methods. As it is infeasible to exploit all possible combinations of alternative modules, we show a subset of comparisons at each module level. For example, when different motion estimation schemes are evaluated, we fix all other modules according to the top-performing algorithms introduced in Section III, namely, OF-B for motion estimation, POS algorithm for pulse color mapping, NLMS filtering for motion filtering, and AMTC for pulse frequency tracking.

A. Modules for Comparison

a) Compared Registration Methods: In order to test the efficacy of the optical flow-based motion estimation method, we compared it with other possible alternatives listed below for a thorough evaluation.

1) Face detection and landmark localization (FD): in each frame, the facial rectangle region is first estimated, and the two cheek regions are localized according to the facial landmarks estimated by [61].
2) Face and skin detection (FSD): in each frame, the ROI is estimated by a color-based skin detection algorithm [62] operated in the face detected rectangle region.
3) Geometric transform correction (GTC): we first detect the face ROI in the first frame the same way as in FD. We then estimate the ROI in the next frame by projecting each point in the ROI of the previous frame to the next frame using the estimated 2D geometric transform. The geometric transform is estimated in the same way as in [15] by tracking a set of good-features-to-track [63].
4) Proposed optical flow framework as described in Section III-B1, respectively, using Lucas and Kanade (OF-LK) [64], Horn and Schunk (OF-HS) [65], Farneback (OF-F) [57], and Brox et al. (OF-B) methods [56].

b) Compared Pulse Color Mapping Methods: As another comparison study, we evaluated the state-of-the-art pulse color mapping algorithms including the blind source separation (BSS) based approaches (ICA [10] and PCA [36]) and skin model-based approaches (CHROM [29], POS [26], and SB [27]). Each method maps the RGB face color signal to a specific direction aiming to provide the highest relative pulse strength based on its model/source-observation assumptions.

A detailed discussion of these approaches based on the human skin reflection model can be found in [26] and [27].

However, the evaluations and the conclusions in both papers are only based on the SNR metric, which may be insufficient in the fitness scenario that this paper focuses on. This is because two signals with the same SNR level might result in completely different PR estimation accuracy. For example, a pulse signal with high interference originated from the subject's motion [see Fig. 4(b)] and low noise might confuse a frequency estimator/tracker much more significantly than a signal with only white noise at the same SNR level. In this paper, we reevaluate these color mapping approaches using our proposed estimation framework and present the result in terms of both the signal quality metric and three PR estimation accuracy metrics.

c) Compared Frequency Tracking/Estimation Methods: In order to single out the contribution and demonstrate the effectiveness of our proposed frequency estimation method used in this paper, we compared it with three other trending frequency estimation methods listed below.

1) Maximum energy (ME): the pulse rate in each spectral frame is estimated as the frequency component with the highest spectral energy. This highest peak selection scheme yields the maximum likelihood frequency estimation result [66] when the noise component is independent with the source and is temporally independent.
2) Particle filter (PF) [67]: PF first approximates the posterior distribution of the frequency state via the sequential Monte Carlo method. The pulse rate is then estimated by the maximum a posteriori estimator.
3) Yet Another Algorithm for Pitch Tracking (YAAPT) [68]: YAAPT estimates the frequency component from a set of local spectral peaks in the spectrogram using a similar dynamic programming approach that has been detailed in Section III-B3.

B. Comparison Study for Motion Estimation Schemes

In Fig. 6, we show comparison examples for four facial videos, each containing spectrograms resulted from seven motion estimation schemes. We listed the averaged SNR estimates of the processed pulse signals and the PR estimation accuracy in terms of PCC, E_{count}, E_{rate}, and E_{RMSE} in TABLE I. As observed from Fig. 6, the pulse signal obtained using the OF-B motion estimation scheme has the highest signal quality when compared with the other schemes especially for the videos of the subject 1 (first two rows). This observation is consistent with the quantitative results listed in TABLE I. Specifically, when compared with the second best results, OF-B improves the SNR by about 0.4 dB, E_{rate} by about 3.4%, and E_{RMSE} by about 4 bpm. These results suggest the importance of a precise face alignment for the video-based heart-rate monitoring method for fitness scenarios.

Nonetheless, not all optical flow-based motion estimation schemes generate as good results as OF-B. OF-LK estimates the pixel displacement between two images by assuming a local parameterized flow structure with the linearized gray value constancy assumption. However, such assumption can be easily violated by the pulse-induced color change on the face, and the resulting biased flow estimates in return cancels the pulse information. The classic global optical flow estimation methods, such as OF-HS, also generates highly biased flow estimates due to the large head motion in the fitness scenarios. By incorporating the coarse-to-fine flow searching strategy to tackle the large motion problem, both OF-F and OF-B have significantly performance gains in almost all measures.

TABLE I
PERFORMANCE OF MOTION COMPENSATION SCHEMES WHEN OTHER MODULES ARE FIXED.

	SNR (dB)	PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
FD	−5.0 (4.0)	0.73 (0.38)	23 (25)	6.4 (8.9)	9.0 (16.8)
FSD	−1.6 (4.3)	0.86 (0.21)	14 (28)	5.3 (12.3)	7.3 (15.8)
GTC	−3.1 (2.9)	0.78 (0.33)	28 (34)	7.5 (3.0)	12.5 (15.8)
OF-LK	−7.6 (3.2)	0.67 (0.42)	36 (40)	11.9 (14.9)	12.6 (20.6)
OF-HS	−6.6 (3.6)	0.78 (0.34)	40 (47)	7.6 (13.0)	18.6 (20.9)
OF-F	−1.2 (5.0)	0.82 (0.28)	15 (26)	5.1 (12.5)	8.9 (12.4)
OF-B	**−0.8** (4.8)	**0.86** (0.21)	**9** (10)	**1.7** (2.2)	**3.3** (6.4)

Note: Values in parentheses are sample standard deviations; the top performing entry for each metric is highlighted in bold.

C. Comparison Study for Pulse Color Mapping Algorithms

We evaluate the pulse color mapping modules by exhausting all the alternatives and by turning on and off the adaptive motion filtering method introduced in Section III-B2. By this means, we could gain a better understanding of the possible synergistic strength of each pair of algorithms. We depicted the system performance in terms of averaged SNR and E_{rate} using different pulse color mapping schemes in Fig. 7(a)–(b). Note that the blind source separation methods, i.e., ICA and PCA, in general output less accurate PR estimates compared with the model-based methods such as POS and SB. This is mainly due to the occasional failure of the pulse source selection out of the three demixed source components when face color measurement contains stronger motion components with the dominating frequency in the normal human PR range, for example, 50–240 bpm. Unfortunately, the violation of the assumption that pulse is the dominating component in the measurement is commonly seen in fitness scenarios.

By turning on the NLMS motion filtering module, an SNR improvement by about 2 dB with almost every color mapping scheme can be achieved. This is mainly due to the successful further motion-component removal after the color mapping operation. Out of the three model-based methods, namely, CHROM, POS, and SB, SB generated the best performance when the NLMS was turned off, whereas the POS performed slightly better than SB when NLMS was turned on. The improvement in the quality of the processed signal has naturally led to the improvement in the pulse estimation accuracy. Specifically, NLMS successfully improved about 8% in E_{rate} for almost all the pulse color mapping schemes.

D. Comparison Study for Frequency Estimation Methods

To study the contribution of the proposed frequency tracking algorithm for robust PR estimation, we evaluate the

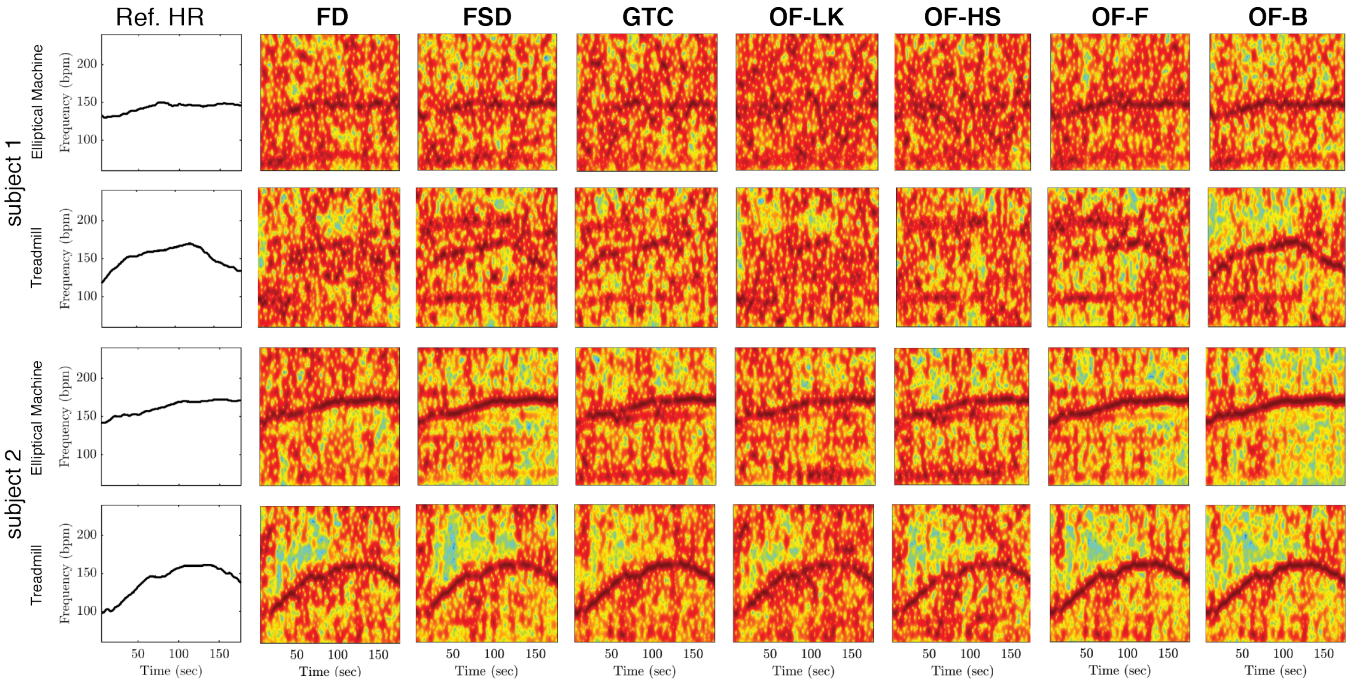

Fig. 6. Comparison of seven motion estimation schemes for four test videos. (Column 1) The reference HR measured by the ECG based chest strap. (Columns 2–8) Spectrograms of the extracted pulse signal using the proposed system with the motion estimation schemes FD, FSD, GTC, OF-LK, OF-HS, OF-F, and OF-B, respectively. Using OF-B produces spectrograms with the cleanest PR traces.

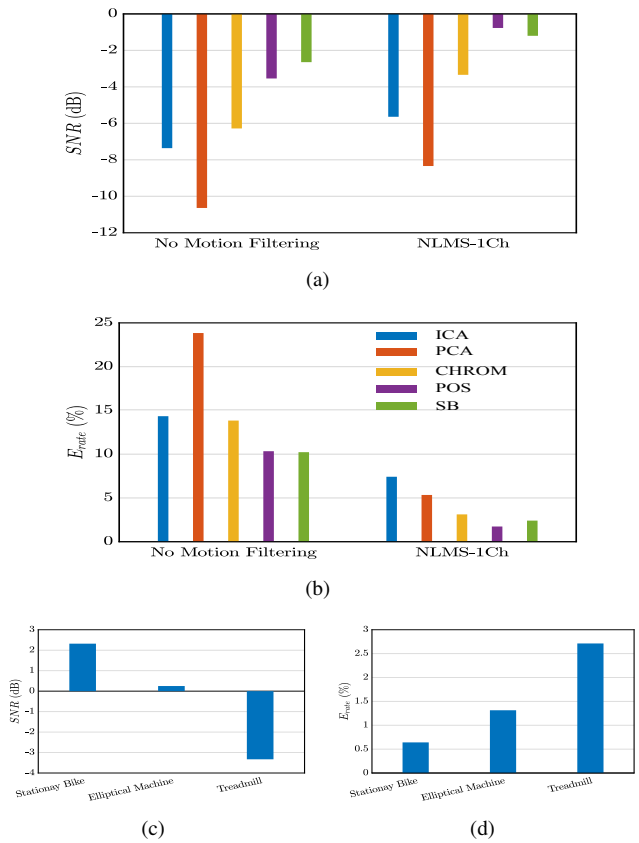

Fig. 7. System performance using different pulse color mappings in terms of (a) SNR and (b) E_{rate} when motion filtering is and is not applied. Optimal system performance in terms of (c) SNR and (d) E_{rate} under different forms of exercise. Motion filtering improves the system performance regardless of the selected pulse color mapping, while exercises involving less nonrigid motion lead to the highest system performance.

frequency estimation accuracy of AMTC with three other algorithmic alternatives. The pulse signals for analyzing the frequency estimation/tracking algorithms are generated using the optimal configuration in TABLE I. The performance results of four different frequency estimation/tracking methods are listed in TABLE II. The proposed AMTC tracking method significantly outperforms the other three methods in the PCC, E_{count}, E_{rate}, and E_{RMSE} performance metrics with respective performance gains of 0.26, 24.1%, 9.3%, and 15.7 bpm over the second best performing algorithm in each of these metrics. The superior performance of AMTC highlights the challenge of frequency tracking under extremely noisy conditions. Even though motion estimation, pulse color mapping, and adaptive motion filtering are designed to mitigate motion artifacts, they cannot completely remove such artifacts. This leads the final extracted PR signal to remain relatively noisy around the PR frequency, indicated by the average SNR of -0.8 dB for the videos processed with the optimized pipeline. This is clearly evidenced in the top right spectrogram image in Fig. 6, in which the PR trace signal is visible, but surrounded by noise. The influence of outliers in PR extraction methods that rely on local peak finding may thus result in biased estimates under such conditions. Since AMTC directly accounts for temporal continuity via regularizing in the cost function, it is less susceptible to noise influence, generating a smoother frequency trace. Hence, optimizing this module is critical for accurate PR estimation under motion intensive conditions.

E. Impact of the Fitness Motion Type

To study the effect of the subject's exercise motion to the pulse signal and the PR estimation accuracy, we show the

TABLE II
PERFORMANCE OF PULSE COLOR MAPPING SCHEMES WHEN OTHER MODULES ARE FIXED.

	PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
ME	0.17 (0.38)	39 (28)	14 (12)	34 (17)
PF	0.37 (0.33)	34 (25)	13 (9)	23 (16)
YAAPT	0.60 (0.21)	33 (34)	11 (3)	19 (16)
AMTC	**0.86** (0.21)	**8.9** (10)	**1.7** (2)	**3.3** (6)

Note: Values in parentheses are sample standard deviations. The top performing entry for each metric is highlighted in bold. The average SNR of the associated spectrograms is −0.8 (4.8) dB.

averaged SNR and E_{rate} using bar plots in Fig. 7(c) and (d) respectively. Notice that the highest pulse signal quality and the PR estimation accuracy are achieved in the stationary bike scenario while the PR estimation in the treadmill scenario is overall the least accurate. As seen in the sample video frames shown in Fig. 5(a)–(c), there is only minor face rigid motion when a subject is exercising on a stationary bike, especially in a sitting position. On the other hand, the subject motion is much more significant in the elliptical machine and the treadmill scenarios. The experimental results are therefore consistent with the intuition that the more significant the subject exercising motion is, the more difficult it becomes to extract precise PRs from the facial videos.

F. Discussion

Much of the latest effort has been devoted to the development of neural-network-based approaches, typically designed to be as close to end-to-end as possible to avoid the tuning of many hyperparameters in intermediate modules. Such methods have been able to show highly accurate results on benchmark datasets. In this subsection, we illustrate the benefit that a modularized system can have on the performance of two such networks—PhysNet [45] and CVD [47]. Specifically, we incorporate them in place of the motion estimation, cheek regions selection, spatial averaging, and pulse color mapping modules of our system. For PhysNet, this means that we feed in motion align face clips into the network before outputting rPPG signals, while for CVD this means extracting MSTmaps from the aligned face clips before feeding them into the network for rPPG extraction. Since our fitness exercise dataset only provides ground truth heart rate data instead of pulse data, we train these models on the PURE dataset [69], which contains six videos, each under different types of face motions (still, talking, slow rotation, fast rotation, slow translation, fast translation), for ten subjects. We trained the models using the publicly available source code provided by the authors on eight of the subjects' data and use the remaining two subjects' data for testing. The results from using the optimized system on the leave-two-out participants from the PURE dataset show respective E_{rate} and E_{RMSE} values of 0.07 and 4.81 for CVD and 0.04 and 2.59 for PhysNet. These values verify that the networks' high performance on the PURE dataset.

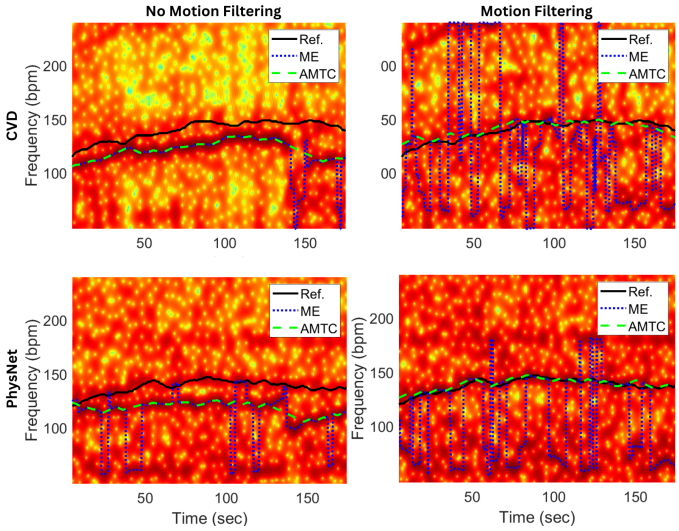

Fig. 8. Example of the spectrograms of the rPPG signals generated from CVD (first row) and PhysNet (second row) without (left column) and with (right column) the NLMS motion filtering. The reference PR trace and the estimated PR traces generated by the AMTC and ME methods are plotted on top of each spectrogram. The combination of motion filtering and AMTC-based-tracking produces the closest estimated pulse signal to the reference.

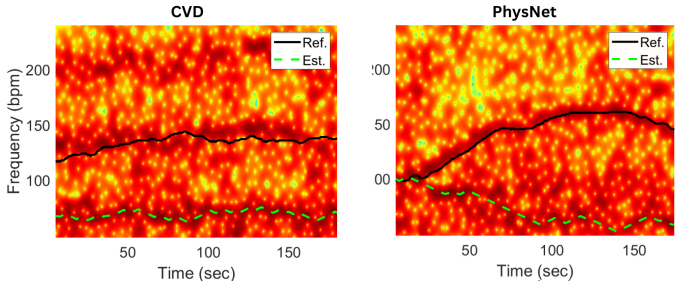

Fig. 9. Failure cases for each of the two neural network based systems. A weak PR trace produced by the neural network prevents AMTC from tracking the PR signal.

To verify the utility of our optimized system, we compare the PR estimation performance when the NLMS filter is and is not applied, and also compare the performance when AMTC and ME are used for pulse extraction, on our fitness exercise dataset. Visual results from these experiments are displayed in Fig. 8. We can see that motion artifacts can still dominate the spectrograms of the rPPG signals produced by both neural network methods from the left column of plots in this figure in which rPPG spectrograms without motion filtering are displayed. This degrades the quality of the AMTC and ME tracking methods since the subjects' motions lead to strong traces in these spectrograms. In the right column of plots, NLMS filtering is applied to the rPPG signals produced by these methods prior to spectrogram generation. It can be observed that the NLMS filter effectively eliminates the motion artifacts in both spectrograms, indicating that this module has utility in processing neural network rPPG signals. Since AMTC is robust in frequency tracking, it is able to track the frequency of the PR signal in both spectrograms once the motion trace is removed. However, because the ME method is less robust to noise, the PR estimates produced by this signal

are still extremely unstable for both spectrograms in the right column of plots.

Neural networks often struggle to generalize when the characteristics of the training and testing data differ significantly. We illustrate an example of a failure case in Fig. 9 in which the dominating traces that appear in the spectrograms of rPPG signals produced the optimized neural network framework for the CVD and PhysNet models are significantly different from the ground truth PR estimates, making it impossible for the pulse extraction methods to extract the heart rate accurately. While domain adaptation and transfer learning techniques may help address the data mismatch between training and deployment, it is challenging to automatically identify the mismatch, gather necessary additional data, and perform additional training or adaptation. That said, a more thorough analysis must be conducted to verify the generalization capabilities (especially for end-to-end neural network systems); gain broader insights into the roles that a system with principled, explainable approaches such as ours can have on these neural network methods; and use these insights to guide the future design of neural networks for PR extraction under challenging fitness scenarios. Such efforts can lead to the design and optimization of explainable neural-network-based modules in a systematic pipeline, for example, to understand the roles of adaptive filtering versus the recurrent neural network adopted in Maity et al.'s design [70] to handle motion.

Moreover, it has been shown that traditional oximeters and video-based-oximetry methods are less accurate for people with darker skin tones [71]. Mantri and Jokerst [71] propose a compensation method to de-bias the results of people with different skin tones. Insights from such methods can be taken to design a neural network that explicitly accounts for such a bias term in the color mapping module, thereby improving its precision and the overall robustness the proposed system. We plan to devote future research efforts to these tasks.

VI. Conclusion

In this paper, we have carried out a quantitative review of the last decade's representative efforts in the rPPG field, and have built a robust PR monitoring system for fitness exercise videos. We focused on building a high-precision motion compensation scheme with the help of the localized facial optical flow, and used motion information as a cue to adaptively remove ambiguous frequency components for improving the PR estimates. We have compared different methods at each module level by examining four representative performance measures. The results demonstrate the synergistic strength of the POS pulse color mapping and NLMS motion compensation schemes. The results also suggest the importance of robust frequency tracking for accurate PR estimation in low SNR fitness scenarios.

Acknowledgment: We thank Prof. James M. Hagberg for an enlightening discussion on chest strap based heart rate monitoring in sports medicine and Jiahao Su for his contributions to the initial phase of this project.

References

[1] J. Karvonen and T. Vuorimaa, "Heart rate and exercise intensity during sports activities," *Sports Medicine*, vol. 5, no. 5, pp. 303–311, May 1988.

[2] M. P. Tulppo, T. H. Makikallio, T. Seppänen, R. T. Laukkanen, and H. V. Huikuri, "Vagal modulation of heart rate during exercise: Effects of age and physical fitness," *American J. Physio.-Heart Circulatory Physio.*, vol. 274, no. 2, pp. H424–H429, Feb. 1998.

[3] M. Buchheit, "Monitoring training status with HR measures: Do all roads lead to rome?" *Frontiers Physio.*, vol. 5, p. 73, Feb. 2014.

[4] C. Schneider, F. Hanakam, T. Wiewelhove, A. Döweling, M. Kellmann, T. Meyer, M. Pfeiffer, and A. Ferrauti, "Heart rate monitoring in team sports—a conceptual framework for contextualizing heart rate measures for training and recovery prescription," *Frontiers Physio.*, vol. 9, 2018.

[5] H. A. Daanen, R. P. Lamberts, V. L. Kallen, A. Jin, and N. L. Van Meeteren, "A systematic review on heart-rate recovery to monitor changes in training status in athletes," *Int. J. Sports Physio. Perf.*, vol. 7, no. 3, pp. 251–260, Sep. 2012.

[6] W. Einthoven, A. Jaffe, P. Venge, and B. Lindahl, "Galvanometrische registratie van het menschelijk electrocardiogram," *Herinneringsbundel Professor SS Rosenstein*, pp. 101–107, 1902.

[7] A. B. Hertzman, "The blood supply of various skin areas as estimated by the photoelectric plethysmograph," *Ame. J. Physio.-Legacy Content*, vol. 124, no. 2, pp. 328–340, Oct. 1938.

[8] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physio. Meas.*, vol. 28, no. 3, p. R1, Feb. 2007.

[9] W. Wang, "Robust and automatic remote photoplethysmography," Ph.D. dissertation, Eindhoven, The Netherlands, Oct. 2017.

[10] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[11] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graphics*, vol. 31, no. 4, 2012.

[12] C. G. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon, "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 303–306, Jul. 2011.

[13] F. Zhao, M. Li, Y. Qian, and J. Z. Tsien, "Remote measurements of heart and respiration rates for telemedicine," *PloS one*, vol. 8, no. 10, p. e71384, Oct. 2013.

[14] L. A. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. B. Oetomo, and W. Verkruysse, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—a pilot study," *Early Human Dev.*, vol. 89, no. 12, pp. 943–948, Dec. 2013.

[15] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *IEEE/CVF Conf. Comput. Vision Pattern Recog. (CVPR)*, Columbus, OH, Jun. 2014, pp. 4264–4271.

[16] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *IEEE Int. Symp. Robot Human Interact. Commun.*, Edinburgh, UK, Aug. 2014, pp. 1056–1062.

[17] S.-C. Huang, P.-H. Hung, C.-H. Hong, and H.-M. Wang, "A new image blood pressure sensor based on ppg, rrt, bptt, and harmonic balancing," *IEEE Sensors J.*, vol. 14, no. 10, pp. 3685–3692, Jun. 2014.

[18] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological measurement*, vol. 35, no. 5, p. 807, Mar. 2014.

[19] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2593–2601, May 2014.

[20] L. Feng, L. M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 879–891, May 2015.

[21] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *European Conf. on Comput. Vision (ECCV)*, 2018, pp. 349–365.

[22] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *IEEE Int. Conf. Automatic Face & Gesture Recog. (FG)*, 2019, pp. 1–8.

[23] 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," *Comput. Bio. Med.*, vol. 116, p. 103535, 2020.

[24] A. Gudi, M. Bittner, R. Lochmans, and J. van Gemert, "Efficient real-time camera based estimation of heart rate and its variability," in *IEEE Int. Conf. Comput. Vision Workshops (ICCVW)*, 2019.

[25] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physio. Meas.*, vol. 35, no. 9, p. 1913, Aug. 2014.

[26] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, 2016.

[27] ——, "Robust heart rate from fitness videos," *Physio. Meas.*, vol. 38, no. 6, p. 1023, May 2017.

[28] Y. Sun, S. Hu, V. Azorin-Peris, S. Greenwald, J. Chambers, and Y. Zhu, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *J. Biom. Optics*, vol. 16, no. 7, pp. 077 010:1–9, Jul. 2011.

[29] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[30] Q. Zhu, "Robust and analytical cardiovascular sensing," Ph.D. dissertation, University of Maryland, College Park, 2020.

[31] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ: Prentice Hall, 2002.

[32] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21 434–45, Dec. 2008.

[33] L. Kong, Y. Zhao, L. Dong, Y. Jian, X. Jin, B. Li, Y. Feng, M. Liu, X. Liu, and H. Wu, "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light," *Optics express*, vol. 21, no. 15, pp. 17 464–17 471, Jul. 2013.

[34] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rPPG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

[35] K. B. Jaiswal and T. Meenpal, "Continuous pulse rate monitoring from facial video using rPPG," in *Int. Conf. Comput., Comm. Netw. Tech. (ICCCNT)*, 2020, pp. 1–5.

[36] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," in *Federated Conf. Comput. Sci. Info. Syst. (FedCSIS)*, Sep. 2011, pp. 405–410.

[37] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics Express*, vol. 18, no. 10, pp. 10 762–74, May 2010.

[38] R. Song, J. Li, M. Wang, J. Cheng, C. Li, and X. Chen, "Remote photoplethysmography with an eemd-mcca method robust against spatially uneven illuminations," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13 484–13 494, 2021.

[39] A. Pai, A. Veeraraghavan, and A. Sabharwal, "HRVCam: Robust camera-based measurement of heart rate variability," *J. Biom. Optics*, vol. 26, no. 2, p. 022707, 2021.

[40] G. R. Tsouri and Z. Li, "On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras," *J. Biom. Optics*, vol. 20, no. 4, p. 048002, Apr. 2015.

[41] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 9, pp. 1974–1984, Dec. 2015.

[42] 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 *IEEE/CVF Conf. Comput. Vision Pattern Recog. (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 2396–2404.

[43] H. Demirezen and C. Eroglu Erdem, "Heart rate estimation from facial videos using nonlinear mode decomposition and improved consistency check," *Signal, Image and Video Process.*, vol. 15, no. 7, pp. 1415–1423, 2021.

[44] G.-S. Hsu, A. Ambikapathi, and M.-S. Chen, "Deep learning with time-frequency representation for pulse estimation from facial videos," in *IEEE Int. Joint Conf. on Biomet. (IJCB)*, 2017, pp. 383–389.

[45] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *British Mach. Vision Conf.*, 2019.

[46] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, and S.-H. Chang, "Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos," in *Annual ACM Symp. Applied Comput.*, 2020, pp. 2066–2073.

[47] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," in *European Conf. on Comput. Vision (ECCV)*. Springer, 2020, pp. 295–310.

[48] S. A. Shafer, "Using color to separate reflection components," *Color Res. & App.*, vol. 10, no. 4, pp. 210–218, Dec. 1985.

[49] R. R. Anderson and J. A. Parrish, "The optics of human skin," *J. Investigative Dermatology*, vol. 77, no. 1, pp. 13–19, Jul. 1981.

[50] H. Takiwaki *et al.*, "Measurement of skin color: Practical application and theoretical considerations," *J. Med. Invest.*, vol. 44, pp. 121–126, Feb. 1998.

[51] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 2, pp. 172–175, Aug. 2002.

[52] Q. Zhu, C.-W. Wong, C.-H. Fu, and M. Wu, "Fitness heart rate measurement using face videos," in *IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 2000–2004.

[53] Q. Zhu, M. Chen, C.-W. Wong, and M. Wu, "Adaptive multi-trace carving based on dynamic programming," in *Asilomar Conf. Signal Syst. Comput.*, Pacific Grove, CA, Oct. 2018, pp. 1716–1720.

[54] ——, "Adaptive multi-trace carving for robust frequency tracking in forensic applications," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1174–1189, 2020.

[55] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[56] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conf. on Computer Vision*, May 2004, pp. 25–36.

[57] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, Jun. 2003, pp. 363–370.

[58] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[59] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2021, pp. 9772–9781.

[60] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conf. on Comput. Vision (ECCV)*, 2020, pp. 402–419.

[61] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE Int. Conf. Comput. Vision (ICCV)*, Sydney, Australia, Dec. 2013, pp. 1944–1951.

[62] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vision*, vol. 46, no. 1, pp. 81–96, Jan. 2002.

[63] J. Shi *et al.*, "Good features to track," in *IEEE/CVF Conf. Comput. Vision Pattern Recog. (CVPR)*, 1994, pp. 593–600.

[64] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[65] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, Aug. 1981.

[66] D. Rife and R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, Sep. 1974.

[67] Y. Shi and E. Chang, "Spectrogram-based formant tracking via particle filters," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Apr. 2003.

[68] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, May 2002, pp. I–361.

[69] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *IEEE Int. Symp. Robot Human Interact. Commun.*, 2014, pp. 1056–1062.

[70] A. K. Maity, J. Wang, A. Sabharwal, and S. K. Nayar, "RobustPPG: camera-based robust heart rate estimation using motion cancellation," *Biomed. Opt. Express*, vol. 13, no. 10, pp. 5447–5467, Oct 2022.

[71] Y. Mantri and J. V. Jokerst, "Impact of skin tone on photoacoustic oximetry and tools to minimize bias," *Biom. Optics Express*, vol. 13, no. 2, pp. 875–887, 2022.