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

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

***Abstract*—Performance improvements obtained by recent principled approaches for pulse rate (PR) estimation from face videos have typically been achieved by adding or modifying certain modules within a reconfigurable system. Yet, evaluations of such remote photoplethysmography (rPPG) are usually performed only at the system level. To better understand each module's contribution and facilitate future research in explainable learning and artificial intelligence for physiological monitoring, this paper conducts a comparative study of video-based, principled PR tracking algorithms, with a focus on challenging fitness scenarios. A review of the progress achieved over the last decade and a half in this field is utilized to construct the major processing modules of a reconfigurable remote pulse rate sensing system. Experiments are conducted on two challenging datasets—an internal collection of 25 videos of two Asian males exercising on stationary-bike, elliptical, and treadmill machines and 34 videos from a public ECG fitness database of 14 men and 3 women exercising on elliptical and stationary-bike machines. The signal-to-noise ratio (SNR), Pearson's correlation coefficient, error count ratio, error rate, and root mean squared error are used for performance evaluation. The top-performing configuration produces respective values of −0.8 dB, 0.86, 9%, 1.7%, and 3.3 beats per minute (bpm) for the internal dataset and 1.3 dB, 0.77, 28.6%, 6.0%, and 8.1 bpm for the ECG Fitness dataset, achieving significant improvements over alternative configurations. Our results suggest a synergistic effect between pulse color mapping and adaptive motion filtering, as well as the importance of a robust frequency tracking algorithm for PR estimation in low SNR settings.**

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

I. Introduction

PULSE rate (PR) is a vital, noninvasive, and time-efficient metric for monitoring training load and assessing an athlete's response [1], [2], [3], [4], [5]. Accurate PR estimation is essential for optimizing training effectiveness and safety, helping coaches and athletes achieve their training goals.

Conventional cardiac monitoring, such as chest-strap heart rate monitoring 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 an ECG 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 methods that involve electrodes, chest straps, or finger clips. Such a monitoring system extracts a one-dimensional (1-D) oscillating face color signal that has the same frequency as the heartbeat from a facial video. The ability to measure PR without direct contact is attractive and has potential applications in smart health, sports medicine, and cardiac rehabilitation.

In this paper, we ask and seek to answer the following research 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, stemming from every component of the rPPG sensing system, namely, the camera, the illumination conditions, and the subject [9]. In a fitness

Manuscript received 8 March 2024; revised 29 June 2024; accepted 4 August 2024. Date of publication 13 August 2024; date of current version 16 January 2025. This work was supported in part by the NSF under Grant 2030502, Grant 2030430, and Grant 2124291, and in part by the Maryland Innovation Initiative grant. *(Corresponding author: Min Wu.)*

Qiang Zhu and Mingliang Chen were with the Department of Electrical and Computer Engineering, University of Maryland, USA. They are now with Meta Inc., USA.

Chau-Wai Wong was with the Department of Electrical and Computer Engineering, University of Maryland, USA. He is now with the Department of Electrical and Computer Engineering, North Carolina State University, USA.

Zachary McBride Lazri is with the Department of Electrical and Computer Engineering and University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland, USA.

Min Wu is with the Department of Electrical and Computer Engineering and University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742 USA (e-mail: minwu@umd.edu).

Chang-Hong Fu was with the Department of Electrical and Computer Engineering, University of Maryland, China. He is now with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, China.

Digital Object Identifier 10.1109/TBME.2024.3442785

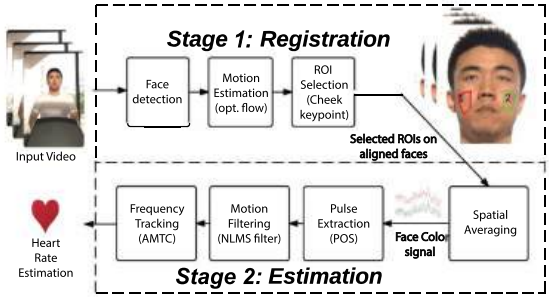

Fig. 1. The proposed modular system for studying the pulse rate monitoring for fitness exercise videos with candidate algorithms in parentheses.

setup, motion-induced changes in illumination intensity may dominate the light reflected from the skin of the face because pulse-induced color variations are usually much subtler. The measurement is also associated with nuisance sources such as the sensor and quantization noise. To handle these factors and perform reliable pulse signal extraction, dedicated algorithms need to be designed to address these challenges synergistically.

The last decade and a half has witnessed a rapid increase of works dedicated to pulse rate estimation for still/rest cases or with relatively little motion [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Previous works [25], [26], [27], [28], [29] on pulse rate estimation often overlook significant subject motion in fitness contexts [26], [27]. When subject motion is considered, these studies either fail to quantitatively assess performance [28] or show large deviations from reference values [29]. Meanwhile, the evaluation process provided in most works is reported 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 community's understanding of the design options of each system component and limit the progress of future research and development.

In this paper, we investigate techniques that can provide the best possible performance for fitness exercise videos. We particularly focus our analysis on principled methods—as opposed to primarily data-driven methods—since they are well suited to real-world situations that may involve many unique and unseen environmental conditions. Our analysis of these methods is performed by constructing a system serving as a platform to evaluate various modular configurations. The system shown in Fig. 1 contains the typical building blocks agreed upon within the literature. 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 and a half of research on rPPG and the skin reflection model adopted in this paper. In Section III, we describe a modular system for rPPG-based PR estimation specially designed for fitness exercises. In Section IV, we present the experimental conditions used to perform our experimental analysis. In Section V, we conduct a comparative study of PR estimation using different module combinations and provide a 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 here are in no way exhaustive. Nevertheless, the contributions of these works in addressing 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 signal extraction*.

A. ROI Selection

The purpose of ROI selection is to locate an exposed region on the human body that is easily trackable and contains pulsatile information in the visual form. For this reason, most methods typically use a person's face for ROI selection. Below, we summarize from the literature four main alternative approaches for ROI selection. All methods, aside from the manual selection approach, use automated face detection and differ in how they construct the ROI within the facial region.

Manual selection: A single ROI may be selected in the first frame of a video and used as the ROI of all subsequent frames [13], [32], [33]. While reviewed for completeness, we avoid manual selection in our comparative analysis given its sensitivity to small motions.

Frame-wise landmark localization: A face detection algorithm is applied frame-wise [10] to localize a person's face in each video frame. Facial landmarks are then detected in the facial region of each frame and used to construct polygons that represent the ROI.

Geometric transformation: Face detection and landmark localization algorithms are applied to construct polygons on a person's face in the first frame of a video. The ROI in each subsequent video frame is then constructed by applying a geometric projection to the ROI in the previous video frame. The frame-wise geometric projections are constructed using "good features for tracking" [15], [20]. However, in the presence of large motion displacement, more fine-grain local alignment is required to ensure the stability of the detected ROI region for accurate PR extraction.

Frame-wise skin detection: A face detection algorithm is applied frame-wise [10] to localize a person's face in each video frame. Since skin face pixels produce most, if not all,

pulsatile information in the facial region, an ROI is constructed only from skin pixels in the facial region of each frame. To distinguish between skin and non-skin face pixels, an approach proposed by Wang et al. [34] may be used to train a skin pixel detector using the first several video frames. While robust for different skin tones, it may generate false positive skin pixels when illumination conditions vary temporally.

B. Motion-Resilient Pulse Signal Extraction

Green channel methods [12], [15], [18], [32], [35] focus on using the green color channel for extracting pulse information as it produces the highest pulse-signal strength among the three color channels. This is because oxyhemoglobin and deoxyhemoglobin have greater absorption in the green wavelength compared with the red or blue wavelengths.

Blind source separation (BSS) methods [36], [37], [38] perform pulse extraction by demixing the pulse signal from the R, G, and B measurements. These methods assume that either the sources are uncorrelated [36] or independent [37], or use ensemble empirical mode decomposition to extract the intrinsic mode functions from multiple face ROIs to be passed to a BSS algorithm for demixing [38]. These methods perform well when the pulse signal, noise, and interfering components exhibit the aforementioned statistical behaviors but may break down when strong periodic motion artifacts enter the RGB signal sourced from the face.

Skin model-based methods [23], [25], [26], [27], [29], [34], [39], [40], [41], [42], [43] operate by providing a best-guessed color projection direction for extracting the pulse source. Multiple methods [23], [26], [29], [41], [42], [43] use or extend algorithms based on the dichromatic skin reflection model as prior knowledge for extracting the pulse signal by projecting temporally normalized RGB signals in some direction orthogonal to non-physiological information. Recognizing that the hue change on the skin is another useful feature for pulse extraction [39], the 2SR algorithm [40] exploits pulse-induced hue changes by tracking the principal direction of the hue channels. For a more detailed discussion about the strengths and weaknesses of the algorithms mentioned above, we referred the readers to [26].

Neural-network-based methods [21], [22], [44], [45], [46], [47], [48], [49], [50], [51] leverage the training data to perform PR estimation. These methods can either be constructed for end-to-end PR extraction [21], [22], [45], [49], or apply some form of prior knowledge to preprocess the data to be fed to the network [44], [46], [47], [48], [50], [51]. For example, Yu et al. [45] found that applying a 3D-CNN directly to the frames of a video can produce accurate results since this structure can jointly handle spatial and temporal information. Conversely, rather than inputting video frames directly into a neural network, Niu et al. [47] construct MSTmaps from the spatially averaged RGB and YCbCr signals taken from various regions of the face as the network input. However, for these trained models to generalize, the training and testing datasets need to be identically distributed. This makes it hard to perform PR extraction in different scenes and in fitness situations in which people have highly variable PR levels.

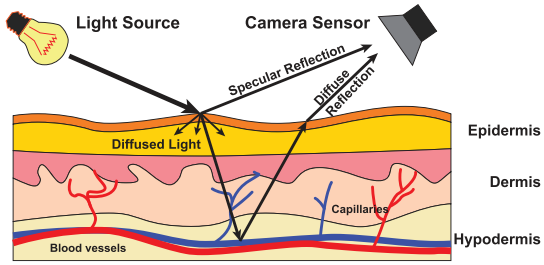

Fig. 2. Illustration of the composition of light reflected from human skin tissue and captured by an RGB camera sensor used for pulse signal modeling (adapted froms [26]).

C. Modeling the Skin Reflection and Motion

As illustrated in Fig. 2, when a light source illuminates a patch of skin, the reflected light can be characterized by two components–specular reflection and diffuse reflection. **Specular reflection** is a mirror-like reflection that contains no pulsatile information and is produced by light directly reflecting off of the outer layer of the skin [52], [53], [54]. **Diffuse reflection** is produced by light penetrating the skin's surface and reflecting off of the inner dermal skin layers carrying pulsatile information [52]. The aim of pulse signal extraction is to isolate the pulsatile information present in the diffuse reflection captured by a camera's sensor. To facilitate pulse signal extraction, the dichromatic skin reflection model (DRM)[1] provided in (1) is used to represent the components of light captured by a camera's sensor [26], [27], [54]:

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

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; $I(t)$ represents the intensity of the light that arrived at the corresponding skin surface; $\mathbf{v}_{\mathrm{s}}(t)$ and $\mathbf{v}_{\mathrm{d}}(t)$ denote the specular and diffuse reflection components, respectively; and $\mathbf{v}_{\mathrm{n}}^{\ell}(t)$ denotes camera sensing and compression noise. $\mathbf{v}_{\mathrm{s}}(t)$ and $\mathbf{v}_{\mathrm{d}}(t)$ may be decomposed as:

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

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

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, skin tissue, and pulse, respectively; s_0 and d_0 denote the strengths of the DC component of the specular and diffuse reflection, respectively; $s(t)$ and $p(t)$ denote the strengths of the AC component of the specular reflection and pulse signal, respectively. The temporal variations of $I(t)$ and $s(t)$ come from motion.

By letting $\mathbf{C}(t)$ denote the spatial average of all skin pixels in (1) and defining $I(t) \triangleq [1 + i(t)]I_0$ 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 change in illumination, we can modify (1) to obtain:

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

[1]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 is incorporated in this paper for consistency.

$$\approx I_0\,[\mathbf{u}_c\cdot c_0 + \mathbf{u}_c\cdot c_0\, i(t) + \mathbf{u}_s\cdot s(t) + \mathbf{u}_p\cdot p(t)], \quad (3b)$$

where second-order cross AC-terms are small, and thus dropped from the approximation, and spacial averaging effectively removes the noise term, $\mathbf{v}_n^{\ell}(t)$, when the number of pixels is large.

As pointed out in [27], a limiting assumption of model (3) is that a single light source produces a single specular variation direction, *i.e.*, $\mathbf{u}_s$, which is unrealistic in practice. To account for other sources of light in our model, we assume a total of J light sources present in the scene. Equation (3) therefore becomes:

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

where $i_j(t)$ and $s_j(t)$ denote the intensity variation and specular variation signals of the jth light source [27], respectively. The DC component $\sum_{j=1}^{J} \mathbf{u}_{c,j}\cdot I_{0,j}\, c_{0,j}$ can be estimated and subtracted from (4) by using the short-term smoothing approach introduced in [25], [26] or detrending methods introduced [55], [56]. Since both $i_j(t)$ and $s_j(t)$ come from 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. Denoting $\tilde{\mathbf{C}}(t)$ as the detrended signal after removing the DC component, we finally obtain:

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

where $\mathbf{u}_{m,k} \triangleq \sum_{j=1}^{J} \mathbf{u}_{c,j}\cdot a_{j,k}\, c_{0,j}\, I_{0,j} + \mathbf{u}_{s,j}\cdot b_{j,k}\, I_{0,j}$ is the color vector of the kth motion component, $\mathbf{u}'_p \triangleq \sum_{j=1}^{J} \mathbf{u}_{p,j}$, and $I_{0,j}$ 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 $\mathbf{u}'_p$ is simultaneously orthogonal to $\mathbf{u}_{m,1}, \ldots, \mathbf{u}_{m,K}$. 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 $\{\mathbf{u}_{m,k}\}_{k=1}^{K}$ is highly likely to have a nonnegligible component along the pulse color direction, making the pulse component $\mathbf{u}'_p\cdot p(t)$ not completely linearly separable from the motion.

To further alleviate the impact of motion artifacts, we use precise alignment of the face ROI in Section III-B(1), an adaptive motion filtering module in Section III-B(2), and a robust frequency-trace tracking algorithm in Section III-B(3) that leverages temporal correlations between consecutive human PR values. All these efforts jointly contribute to a robust and accurate extraction of PR signals.

III. A Modular System for Fitness rPPG

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

A. General rPPG System

The general rPPG system 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 system 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 the 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 system that achieve the best experimental results to be presented in Section V. The algorithms that optimize each module of the system 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*) refined removal of the remaining motion artifacts by using a normalized least mean square (NLMS) filter to "subtract" the motion information in the visual track [31] from the color-channel combined signal output by the POS algorithm [26], and (*iii*) the PR signal is extracted using a robust frequency tracker named the adaptive multi-trace carving (AMTC) algorithm [30], [57], [58].

B. Optimized System

1) Precise Face Registration Via Optical Flow [56]: We use the Viola–Jones face detector [59] 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 is widely adopted in the design of 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

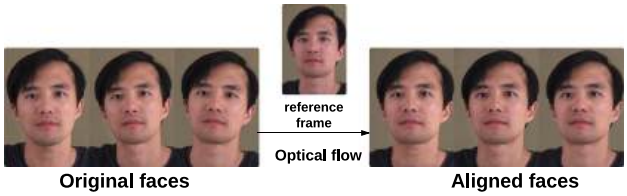

Fig. 3. Facial images from a video segment before and after optical-flow-based motion compensation, illustrating the use of the motion estimation module.

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 [56]. To alleviate potential bias, different strategies can be applied. For example, using a global flow regularization strategy [60] or a coarse-to-fine hierarchical searching strategy [60], [61] instead of doing one-shot Taylor-based local approximation. In this study, we use Liu's optical flow implementation [62] of Brox et al.'s method [60]. Modern deep-learning-based optical flow algorithms [63], [64] may also be used.

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. [65], 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 points of each segment. We apply a detrending algorithm [56] 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 the 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]. The goal of the NLMS problem is characterized as follows. Given an input and desired signal, determine a filter, $\hat{\mathbf{w}}(t)$, that minimizes the error between the filter output and the desired signal. In our application, we know that $c_{\text{pos}}(t)$ contains a mixture of a pulse and motion residue signal in (7). We also can obtain an isolated motion signal by tracking the movement of an individual exercising in a video. In particular, letting $m_x(t)$ and $m_y(t)$ denote the estimated face motion sequences in the horizontal and vertical directions obtained from the facial landmarks in each video frame, we construct a motion tap vector given by $\mathbf{m}(t) \triangleq [m_x(t-M+1), m_x(t-M+2), \ldots, m_x(t), m_y(t-M+1), m_y(t-M+2), \ldots, m_y(t)]^{\intercal}$. Assuming that the motion residue term in (7) can be represented as the output of a linear combination of the elements of $\mathbf{m}(t)$, then subtracting this output from $c_{\text{pos}}(t)$ will give us the pulse signal in (7). Thus, we can solve for this pulse signal by treating it as the error in the NLMS problem and optimizing for the filter weights that are used to construct the linear combination of the elements of $\mathbf{m}(t)$.

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 $\mathbf{m}(t) \triangleq [m_x(t-M+1), m_x(t-M+2), \ldots, m_x(t), m_y(t-M+1), m_y(t-M+2), \ldots, m_y(t)]^{\intercal}$ 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 iterative NLMS algorithm [31] as follows:

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

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

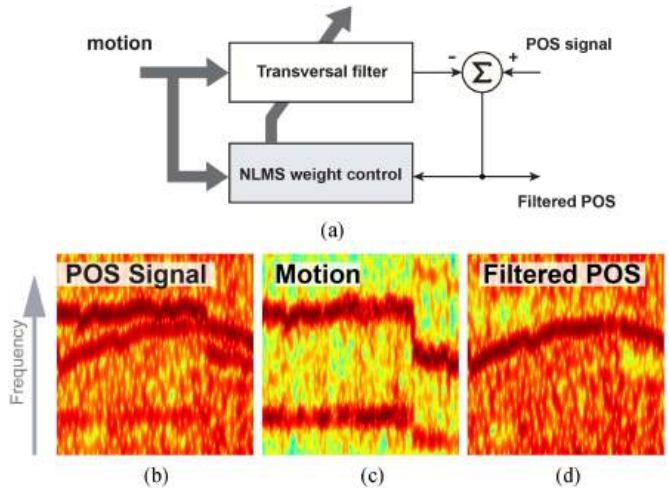

Fig. 4. (a) Adaptive motion compensation filter framework and spectrograms of (b) the POS signal $c_{\text{pos}}(t)$, (c) the combined normalized subject motion in horizontal and vertical directions, and (d) the filtered POS signal $\tilde{c}_{\text{pos}}(t)$. The NLMS filter removes the motion trace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

In (8a), $c_{\text{pos}}(t)$ represents the desired signal from the current time step in the NLMS formulation. $\hat{\mathbf{w}}^{\intercal}(t)\,\mathbf{m}(t)$ represents the filtered output signal in the current time step, which we use to model the motion residual in (7). Thus, the error signal, $\tilde{c}_{\text{pos}}(t)$, provides us with an estimate of the pulse signal from (7). An iterative update is performed in (8b) to obtain the weights needed to calculate $\tilde{c}_{\text{pos}}(t)$ in (8a) for each future time step. Specifically, the filter weights, $\hat{\mathbf{w}}(t)$, and motion tap vector, $\mathbf{m}(t)$, from the current time step are used to estimate the filter weights, $\hat{\mathbf{w}}(t+1)$, that minimize the error between the filtered output and desired signal. 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], [57], [58]. Its details are briefly described. Let $\mathbf{Z} \in \mathbb{R}_{+}^{M\times N}$ be the magnitude component 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 dominating *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

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

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 [57], [58].

IV. Experimental Conditions

We evaluate the reconfigurable system on two datasets to understand the different factors that affect principled PR estimation with fitness motions. The first dataset contains 25 self-collected videos consisting of subjects exercising on elliptical, treadmill, and stationary bike machines. The second dataset contains 34 videos of subjects exercising on elliptical and stationary bike machines from the ECG Fitness dataset [66]. The parameter settings, evaluation metrics, and dataset details are described in the following subsections.

A. Parameter Settings

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

1) The tap number for joint-channel NLMS is 8, and the NLMS learning rate/adaptation constant μ is 0.1.
2) Each video was empirically divided into segments of 1.5 seconds 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.
3) The spectrum analysis window length was set to 10 seconds 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 beats per minute (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 [57], [58] was a uniform random walk model with the width parameter k set to 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:

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

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):

$$E_{\text{RMSE}} = \left(\frac{1}{N} \sum_{n=1}^{N} \left[\hat{f}(n) - f(n) \right]^2 \right)^{\frac{1}{2}}, \quad (11)$$

2) Error rate:

$$E_{\text{rate}} = \frac{1}{N} \sum_{n=1}^{N} \left| \hat{f}(n) - f(n) \right| \Big/ f(n), \quad (12)$$

3) Error count ratio:

$$E_{\text{count}} = \frac{1}{N} \left| \{ n : |\hat{f}(n) - f(n)| / f(n) > \tau \} \right|, \quad (13)$$

4) Pearson's correlation coefficient:

$$\text{PCC} = \frac{\sum_{n=1}^{N} \left[\hat{f}(n) - \bar{\hat{f}} \right] \left[f(n) - \bar{f} \right]}{\left(\sum_{n=1}^{N} [\hat{f}(n) - \bar{\hat{f}}]^2 \sum_{n=1}^{N} [f(n) - \bar{f}]^2 \right)^{\frac{1}{2}}}, \quad (14)$$

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.

C. Datasets for Evaluation

a) Internal Dataset: In order to test the robustness of the system in a fitness-in-the-wild setup, we conducted experiments in two typical apartment fitness rooms. The illumination sources involved only the existing lighting equipment in each room, including several overhead fluorescent lights and possibly diffused sunlight passing through a window. The environment was unconstrained so people were allowed to enter and exit the room during any video recording sessions. In total, 25 three-minute videos were recorded of two healthy Asian males between the ages of 25 and 35 exercising on a stationary bike, elliptical, and treadmill machine. The skin tone of both subjects is classified as Type III according to the Fitzpatrick skin scale [67]. Five elliptical and treadmill videos belong to each subject, and the final five stationary bike videos belong to one subject. All stationary bike videos were captured by a Huawei P9 mobile phone, whereas the other 20 videos were captured by the rear camera of an iPhone 6s mobile phone. All videos have a frame rate of 30 fps, a resolution of 1280×720, and an average bit rate is about 6 MB per second. Moreover, all videos were compressed with the H.264/AVC codec. The shutter speed of both sensors was set as constant to minimize the possibility of introducing artifacts. In each video, the camera was placed at eye level at a distance of approximately one meter from the subject's face. Each subject also wore a Polar H7 chest strap underneath his clothes to record electrocardiogram (ECG) signals to obtain a reference heart rate. A thorough analysis of each module in the principled modular system is conducted on this dataset in Sections V-B through V-E.

b) ECG Fitness Dataset: To further evaluate the modules of this principled modular system, we perform experiments on the ECG Fitness database created at the Czech Technical University [66]. This dataset contains one-minute videos of 17 subjects (14 males and 3 females) between the ages of 20 and 53 performing a variety of different activities that may produce various unconstrained motions. Though not reported, all subjects appear to have a Type I, II, or III skin type according to the Fitzpatrick skin scale [67]. All videos were captured at 30 fps with a resolution of 1920×1080 by one of two RGB Logitech C920 web cameras that were either directly attached to the fitness machine in the video or placed on a tripod as close as possible to the same position as the other camera. Each video was also stored in an uncompressed YUV planar pixel format. The environmental lighting conditions consist of three lighting setups—natural lighting coming from a nearby window, a 400 W halogen light, or a 30 W LED light. All PR references were obtained using a two-lead Viatom CheckME$^{\text{TM}}$ Pro ECG device. Since the focus of this paper is on analyzing exercise fitness videos, we analyze the videos in which the subjects are exercising on a stationary bike or elliptical machine. Each subject has one elliptical and stationary bike video leaving us with a total of 34 videos available for analysis. A detailed analysis of the principled modular system is conducted on this dataset in Section V-F. Since the subjects are allowed to perform unconstrained motions during their exercises, which have the potential to violate the underlying assumptions of OF-based motion compensation methods, we center our motion compensation analysis on OF methods, to emphasize the importance of satisfying these assumptions.

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 experimental benchmark 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. To analyze the contribution of a particular module, we vary the algorithms used in that module while fixing 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. For all results provided in tables in this section, values in parentheses represent sample standard deviations and the top-performing entry for each metric is highlighted in bold.

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 [65].

2) Face and skin detection (FSD): In each frame, the ROI is estimated by a color-based skin detection algorithm [68] operated in the face-detected rectangle region.
3) Geometric transform correction (GTC): We detect the face ROI in the first frame the same way as in FD. Then, we 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 2-D geometric transform. The geometric transform is estimated as in [15] by tracking a set of good-features-to-track [69].
4) Proposed optical flow framework as described in Section III-B(1), respectively, using Lucas and Kanade (OF-LK) [70], Horn and Schunk (OF-HS) [71], Farneback (OF-F) [61], and Brox et al. (OF-B) methods [60].

b) Compared Pulse Color Mapping Methods: Our second comparative analysis consists of comparing 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 for evaluating fitness videos. This is because two signals with the same SNR level might result in completely different PR estimates. For example, a pulse signal with high interference, originating from a subject's motion [see Fig. 4(b)], and low noise might confuse a frequency estimator/tracker more significantly than a signal with only white noise at the same SNR level. Thus, for evaluating the effect of the color mapping algorithm choice on the principled system, E_{rate} results are presented along with the SNR.

c) Compared Frequency Tracking/Estimation Methods: In order to isolate the contribution and demonstrate the effectiveness of the AMTC frequency estimation module presented in Section III-B(3), we compared it with three other commonly used frequency estimation methods:

1) Maximum energy (ME): The PR in each spectral frame is estimated as the frequency component with the highest spectral energy. This provides the maximum likelihood frequency estimate [72] when the noise component is independent of the source and is temporally independent.
2) Particle filter (PF) [73]: 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.
3) Yet Another Algorithm for Pitch Tracking (YAAPT) [74]: YAAPT estimates the frequency trace from a set of local spectral peaks in a spectrogram using a similar dynamic programming approach to the one detailed in Section III-B(3).

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.

B. Comparison Study for Motion Estimation Schemes

In Fig. 6, we provide examples of spectrograms generated from seven motion estimation schemes for four facial videos from the internal dataset.

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 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 an assumption can be easily violated by the pulse-induced color change on the face, and the resulting biased flow estimates distort the pulse information in return. The classic global optical flow estimation methods, such as OF-HS, also generate 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 significant performance gains in almost all measures.

C. Analysis of Pulse Color Mappings and Motion Filtering

We evaluate the pulse color mapping module by reconfiguring it with different algorithms described in Section II-B in situations in which the adaptive motion filter from Section III-B(2) is and is not applied. In doing so, we gain a better understanding of the possible synergistic strength of each pair of algorithms. We depicted the system's performance in terms of average SNR and E_{rate} using different pulse color mapping schemes in Fig. 7(a)–(b). Note that the blind source separation methods,

TABLE II
PERFORMANCE OF MOTION FILTERING WHEN OTHER MODULES ARE FIXED

	SNR (dB)	PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
No NLMS-1Ch	−3.5 (5.9)	0.72 (0.36)	48.1 (40.2)	10.2 (16.7)	20.9 (28.1)
NLMS-1Ch	**−0.8** (4.8)	**0.86** (0.21)	**9.0** (10)	**1.7** (2.2)	**3.3** (6.4)

TABLE III
PERFORMANCE OF TRACKING METHODS 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)	**9** (10)	**1.7** (2)	**3.3** (6)

Note: The average SNR of the associated spectrograms is −0.8 (4.8) dB.

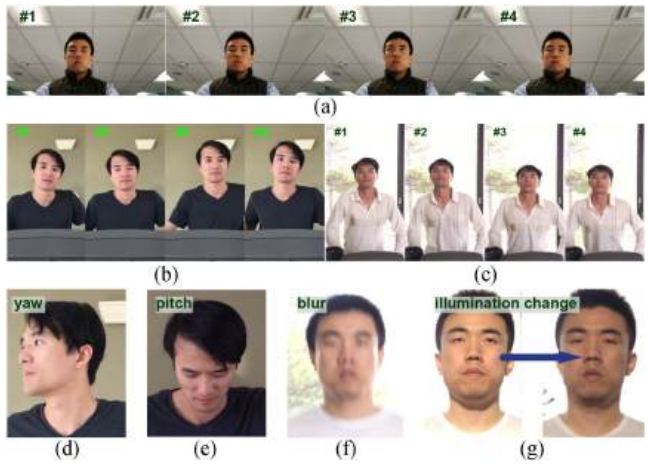

Fig. 5. Sample frames in fitness video dataset with three types of fitness motion: (a) stationary bike, (b) elliptical machine, and (c) treadmill. The challenges in the dataset include head rotation in (d) yaw and (e) pitch, (f) motion blurred frames, and (g) significant illumination change on the face.

i.e., ICA and PCA, typically produce less accurate PR estimates compared with the model-based methods such as POS and SB. This mainly occurs when the presence of a dominant motion frequency in the normal 50–240 bpm PR range causes the source selection method to choose a motion component instead of the PR component from the three demixed source components. 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 of about 2 dB with almost every color mapping scheme can be achieved. This is mainly due to the successful removal of excess motion residue from the signal produced by the color mapping operation. Of the three model-based methods—CHROM, POS, and SB—SB performed the best when the NLMS filter was turned off, whereas POS performed slightly better than SB when the NLMS filter was turned on. The improvement in the quality of the processed signal has naturally led to the improvement in the pulse estimation accuracy. Specifically, applying the NLMS filter improved the performance of the system in the E_{rate} metric by about 8% for almost all the pulse color mapping schemes.

To obtain an understanding of the overall effectiveness of the NLMS motion filtering module, Table II provides results for the optimal modular system when the NLMS filter is and is not applied. A clear improvement in all listed metrics can be observed by using the NLMS filter. This highlights the importance of accounting for the motion residue term in (7). Failing to do so can cause the PR extraction system to mistake the PR trace for the motion trace in videos that contain severe quasi-periodic motions.

D. Comparison Study for Frequency Estimation Methods

To study the contribution of different frequency tracking algorithms for robust PR estimation, we compare the performances of four frequency estimation algorithms. Experimental results for these algorithms are provided in Table III. AMTC significantly outperforms the other three methods in the PCC, E_{count}, E_{rate}, and E_{RMSE} 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 results in the final extracted PR signal remaining relatively noisy around the PR frequency, indicated by the average SNR of −0.8 dB for the videos processed with the optimized system. This is evidenced in the top right spectrogram 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 enforces temporal continuity through regularization in the cost function, it is less susceptible to noise influence, generating a smoother frequency trace.

E. Impact of the Fitness Motion Type

To study the effect of the subject's exercise motion on the pulse signal and the PR estimation accuracy, we show the averaged SNR and E_{rate} using bar plots in Fig. 7(c) and (d), respectively. We note that the highest pulse signal quality and the PR estimation accuracy are achieved in the stationary bike scenario, whereas 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.

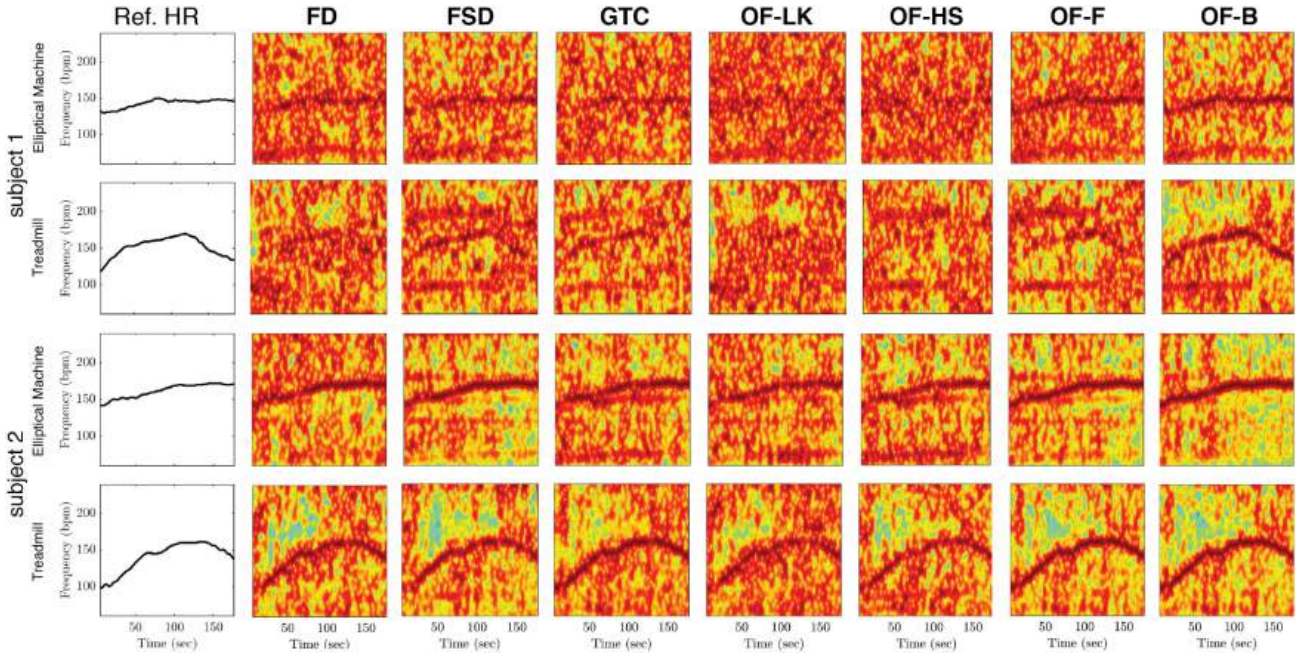

Fig. 6. Comparison of seven motion estimation schemes for four test videos. (Column 1) The reference heart rate measured by the ECG-based chest strap. (Columns 2–8) Spectrograms of the extracted pulse signals using the proposed system with the motion estimation schemes FD, FSD, GTC, OF-LK, OF-HS, OF-F,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

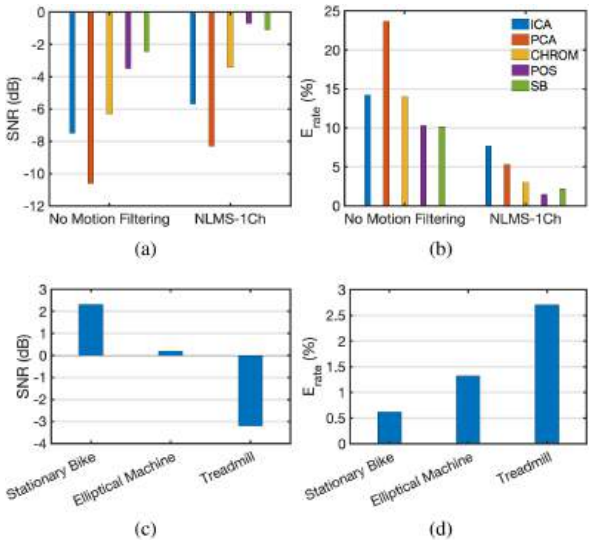

Fig. 7. System performance using different pulse color mappings in terms of (a) SNR and (b) E_{rate} when motion filtering is and is not applied. Optimal system performance in terms of (c) SNR and (d) E_{rate} under different forms of exercise. Motion filtering improves the system performance regardless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

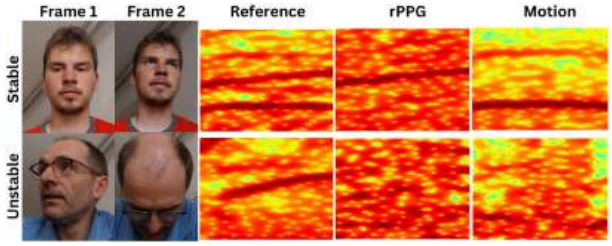

Fig. 8. Illustration of different motion types in the ECG Fitness dataset and the spectrograms produced for the reference ECG signals, the estimated pulse signals obtained from the modular system, and the motion signals obtained from face tracking.

F. Evaluation on ECG Fitness Dataset

In this section, we extend our analysis of the OF methods described in this paper to emphasize the importance of satisfying their underlying assumptions to obtain accurate results. OF methods rely on two major assumptions:

1) Brightness constancy: The observed brightness of an object is constant over time.
2) Motion smoothness: Local image motion can be approximated by local derivatives.

These assumptions are reasonably satisfied when a person's face remains in a frontal position with respect to a camera during exercise. However, large unconstrained motions such as head rotations and tilting may violate these assumptions. That is, as a person's face deviates from its frontal position with respect to the camera, the angle between the light source, skin, and camera also changes. This violates the first assumption. The second assumption may also be violated since such unconstrained motions may cause a large frame-by-frame displacement of nearby points, and may even cause certain points on a person's face to become occluded.

We focus our analysis on the ECG Fitness dataset [66] since subjects are allowed to perform unconstrained motions in these videos, which could potentially lead to violations of the OF assumptions. Fig. 8 provides examples of stable and unstable head movements present in this dataset and the effects they have on PR estimation. The first two columns provide sample frames from

TABLE IV
OF MOTION COMPENSATION PERFORMANCE SCHEMES FOR EXTERNAL DATABASE

		SNR (dB)	PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
All Videos	OF-LK	0.9 (2.3)	0.39 (0.72)	50.9 (37.9)	11.5 (15.4)	14.7 (18.8)
	OF-HS	0.9 (2.3)	0.41 (0.72)	49.5 (38.1)	11.5 (15.4)	14.5 (18.8)
	OF-F	0.9 (1.8)	0.53 (0.59)	47.1 (41.7)	12.0 (16.2)	15.2 (19.7)
	OF-B	**1.3** (1.9)	**0.77** (0.43)	**28.6** (32.1)	**6.0** (12.2)	**8.1** (16.2)
Stable Videos	F-LK	0.9 (2.4)	0.45 (0.70)	46.1 (36.5)	9.5 (13.9)	11.6 (16.8)
	OF-HS	0.9 (2.4)	0.47 (0.70)	44.6 (36.6)	9.3 (14.0)	11.6 (16.8)
	OF-F	1.1 (1.8)	0.54 (0.60)	42.2 (40.6)	11.1(16.4)	13.8 (20.5)
	OF-B	**1.4** (1.9)	**0.83** (0.36)	**20.1** (22.8)	**2.2** (1.8)	**3.1** (2.6)

two videos. Columns three, four, and five respectively provide the spectrograms of the reference ECG signal, estimated pulse signal, and motion signal obtained from face tracking. We can see clear PR traces in the rPPG spectrogram of the first subject, whereas the second subject's occlusion-based movements lead the rPPG spectrogram to produce noise.

Notably, we have observed that some of the ECG reference signals provided in this dataset are susceptible to motion artifacts. This issue can be observed from the presence of multiple strong traces appearing in the top reference ECG spectrogram in Fig. 8. In particular, multiple strong traces can be seen (corresponding to the motion and PR harmonics) in the first reference spectrogram. Thus, identifying the correct trace associated with a person's PR becomes critical for analyzing a PR estimation system on this dataset. To determine the correct trace, we track the horizontal and vertical face motions in each video to obtain m_x and m_y as in Section III-B(2). By analyzing the spectrogram of the sum of these two signals, we obtain an estimate of the motion present in the video, which enables us to distinguish between PR and motion (when present) spectrogram traces of the reference ECG signals. Once this is done, we manually label all PR traces in the spectrograms of each reference ECG signal. Referring back to the spectrograms in Fig. 8, we can clearly observe the motion traces present in the first subject's reference spectrogram. In particular, the bottom trace in this reference spectrogram corresponds with the strong trace in the corresponding motion spectrogram.

Of the 34 elliptical and stationary bike machines available for analysis, the ECG reference and motion spectrograms associated with the elliptical videos for subjects 1 and 4 show strong traces in the same locations. This suggests that the motion interference significantly degrades these ECG reference signals, so we omit these videos from analysis leaving us with 32 videos to analyze. In Table IV, we provide results that demonstrate the impact that violating the OF assumptions has on the performance of the PR system. The top set of results was obtained by analyzing the optimal modular system configuration while varying the optical flow implementation for all 32 elliptical and stationary bike videos. The bottom set of results is provided for the subset of videos for which there are no significant violations of the OF assumptions. That is, we removed the elliptical videos for subject 5 and the bike videos for subjects 9 and 14, leaving us with 29 remaining videos. The OF-B method still tends to produce the most stable results for all analyzed OF methods. Noticeably, the results provided for all videos are moderately worse than those provided in Table I. This is because none of the videos in our internal dataset severely violate the OF assumptions. As expected, the results for the subset of stable videos show a distinct improvement over the results for all 32 videos, with particularly noticeable improvements in the metrics associated with the top-performing OF-B method. These performances are more in line with what we would expect for videos that do not violate OF based on the results observed in Table I. The PCC, E_{RMSE} and E_{rate} values for OF-B, in particular, are similar to those produced in Table I, meaning that the heart rate estimates are close to the ground truth estimates and capture the overall PR trends.

TABLE V
PERFORMANCE OF MOTION FILTERING WHEN OTHER MODULES ARE FIXED FOR EXTERNAL DATABASE

		SNR (dB)	PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
All Videos	No NLMS-1Ch	1.2 (1.7)	0.64 (0.54)	44.3 (44.1)	15.8 (20.5)	19.0 (25.1)
	NLMS-1Ch	**1.3** (1.9)	**0.77** (0.43)	**27.6** (32.1)	**6.0** (12.3)	**8.1** (16.2)
Stable Videos	No NLMS-1Ch	1.3 (1.9)	0.76 (0.40)	36.6 (41.5)	12.9 (19.9)	15.4 (23.7)
	NLMS-1Ch	**1.4** (1.9)	**0.83** (0.36)	**20.1** (22.8)	**2.2** (1.8)	**3.1** (2.6)

TABLE VI
PERFORMANCE OF TRACKING METHODS WHEN OTHER MODULES ARE FIXED FOR EXTERNAL DATABASE

		PCC	E_{count} (%)	E_{rate} (%)	E_{RMSE} (bpm)
All Videos	ME	0.31 (0.49)	40.2 (34.2)	14.6 (13.1)	25.0 (20.0)
	PF	0.42 (0.63)	44.8 (44.8)	14.5 (20.6)	19.2 (25.9)
	YAPPT	0.66 (0.42)	34.5 (34.1)	8.6 (11.3)	13.3 (16.8)
	AMTC	**0.77** (0.43)	**28.6** (32.1)	**6.0** (12.2)	**8.1** (16.2)
Stable Videos	ME	0.35 (0.50)	34.7 (31.0)	12.8 (12.2)	22.4 (18.9)
	PF	0.49 (0.59)	39.1 (43.1)	12.5 (19.4)	16.2 (22.9)
	YAPPT	0.73 (0.37)	27.7 (27.9)	5.6 (6.7)	9.4 (11.7)
	AMTC	**0.83** (0.36)	**20.1** (22.8)	**2.2** (1.8)	**3.1** (2.6)

We compare the performance of the modular system with and without motion filtering, along with different methods used for tracking the pulse rate (PR). These comparisons are provided in Tables V and VI, respectively. The results in Table V show that the overall system performs better when motion filtering is applied, while all other optimal modules are held constant across every performance metric. Additionally, as presented in Table VI, the AMTC method demonstrates the most stable performance for PR tracking, outperforming all other methods across all performance metrics. These findings are consistent with the results and discussion provided for the internal dataset presented in previous sections of this paper.

G. Discussion: Incorporating DNNs in a Modular System

Much recent 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 produced highly accurate results on benchmark datasets. In

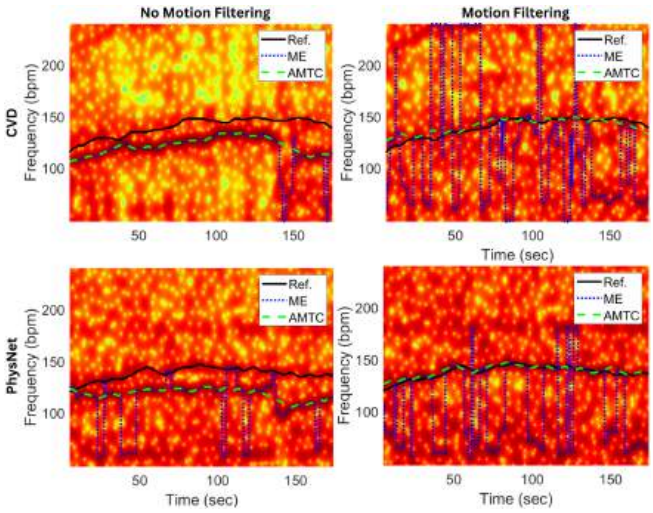

Fig. 9. Example of the spectrograms of the rPPG signals generated from CVD (first row) and PhysNet (second row) without (left column) and with (right column) the NLMS motion filtering. The reference PR trace and the estimated PR traces generated by the AMTC and ME methods are plotted on top of each spectrogram. The combination of motion filtering and AMTC-based tracking produces the closest estimated pulse signal to the reference.

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 region selection, spatial averaging, and pulse color mapping modules of our system. For PhysNet, this means that we feed in motion-aligned 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 [75], 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 used 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, respectively. These values verify the networks' high performance on the PURE dataset.

To verify our optimized system's utility, we compared PR estimation performance with and without the NLMS filter and with AMTC and ME for pulse extraction on our fitness exercise dataset. The visual results in Fig. 9 show that motion artifacts can dominate the spectrograms of neural network rPPG signals without motion filtering, degrading AMTC and ME tracking quality due to strong traces from subjects' motions. When NLMS filtering is applied, it effectively eliminates motion artifacts, improving the processing of these rPPG signals. AMTC, being robust in frequency tracking, can track the PR signal frequency once the motion trace is removed, while ME, less robust to noise, produces unstable PR estimates even after filtering.

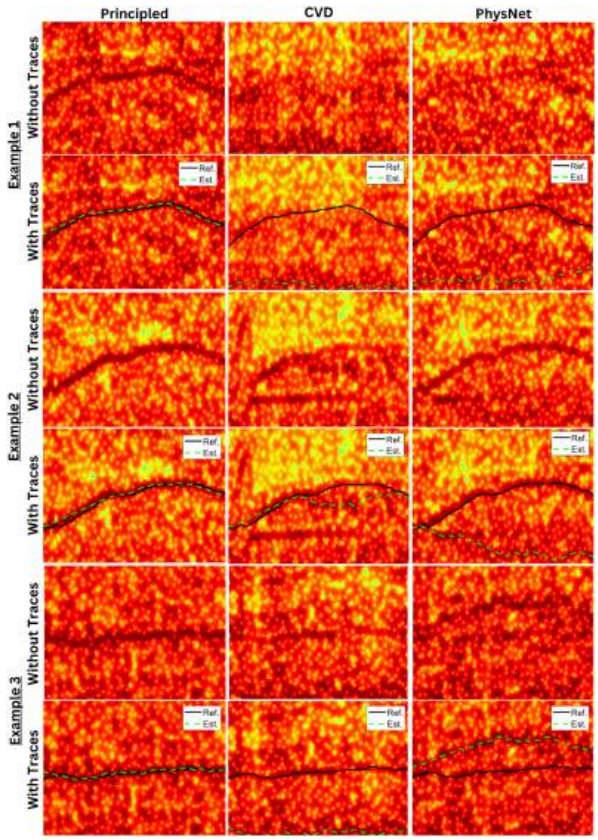

Fig. 10. Three example neural network failure cases (right two columns) versus principled system (left column). Spectrograms with and without overlaid traces are provided for reference. It is revealed that weak PR traces produced by the neural networks prevent AMTC from precisely tracking the PR signals.

Neural networks often struggle to generalize when the characteristics of the training and testing data differ significantly. We illustrate three examples of failure cases from the neural network methods that were trained on the PURE dataset in the right two columns of spectrograms in Fig. 10, and the corresponding success cases from the principled system in the left column. The raw spectrograms without overlaying traces reveal that the traces that appear in the spectrograms produced by the optimized neural network systems for the CVD and PhysNet models are weaker relative to the noise of the signal or non-existent around the ground-truth PR estimates. This makes it difficult, if not impossible, for any pulse extraction methods to extract the PR accurately. This issue is not present when using the optimized principled system, as evident from the precisely plotted PR estimates seen in the left column of Fig. 10. 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 should 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 [76] to handle motion.

VI. Conclusion

In this paper, we have carried out a quantitative review of the last decade and a half's representative efforts in the rPPG field, and have built a robust principled 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

The authors would like thank Prof. James M. Hagberg of the University of Maryland 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 Med.*, vol. 5, no. 5, pp. 303–311, May 1988.
[2] M. P. Tulppo et al., "Vagal modulation of heart rate during exercise: Effects of age and physical fitness," *Amer. J. Physiol.-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?," *Front. Physio.*, vol. 5, Feb. 2014, Art. no. 73.
[4] C. Schneider et al., "Heart rate monitoring in team sports–A conceptual framework for contextualizing heart rate measures for training and recovery prescription," *Front. Physio.*, vol. 9, 2018, Art. no. 639.
[5] H. A. Daanen et al., "A systematic review on heart-rate recovery to monitor changes in training status in athletes," *Int. J. Sports Physiol. Perform.*, vol. 7, no. 3, pp. 251–260, Sep. 2012.
[6] W. Einthoven et al., "Galvanometrische registratie van het menschelijk electrocardiogram," *Herinneringsbundel Professor SS Rosenstein*, pp. 101–107, 1902. [Online]. Available: https://www.christies.com/en/lot/lot-5067294
[7] A. B. Hertzman, "The blood supply of various skin areas as estimated by the photoelectric plethysmograph," *Amer. J. Physiol.-Legacy Content*, vol. 124, no. 2, pp. 328–340, Oct. 1938.
[8] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, p. R1, Feb. 2007.
[9] W. Wang, "Robust and automatic remote photoplethysmography," Ph.D. dissertation, Eindhoven University of Technology, 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 et al., "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, 2012, Art. no. 65.
[12] C. G. Scully et al., "Physiological parameter monitoring from optical recordings with a mobile phone," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 303–306, Feb. 2012.
[13] F. Zhao et al., "Remote measurements of heart and respiration rates for telemedicine," *PLoS One*, vol. 8, no. 10, Oct. 2013, Art. no. e71384.
[14] L. A. Aarts et al., "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit–A pilot study," *Early Hum. Dev.*, vol. 89, no. 12, pp. 943–948, Dec. 2013.
[15] X. Li et al., "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 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 *Proc. IEEE 23rd Int. Symp. Robot Hum. Interact. Commun.*, Edinburgh, U.K., Aug. 2014, pp. 1056–1062.
[17] S.-C. Huang et al., "A new image blood pressure sensor based on PPG, RRT, BPTT, and harmonic balancing," *IEEE Sensors J.*, vol. 14, no. 10, pp. 3685–3692, Oct. 2014.
[18] L. Tarassenko et al., "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, Mar. 2014, Art. no. 807.
[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, Oct. 2014.
[20] L. Feng et al., "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 *Eur. Conf. Comput. Vis.*, 2018, pp. 349–365.
[22] X. Niu et al., "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proc. IEEE 2019 14th Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–8.
[23] R. Song et al., "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Comput. Bio. Med.*, vol. 116, 2020, Art. no. 103535.
[24] A. Gudi et al., "Efficient real-time camera based estimation of heart rate and its variability," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1570–1579.
[25] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiol. Meas.*, vol. 35, no. 9, Aug. 2014, Art. no. 1913.
[26] W. Wang et al., "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.
[27] W. Wang et al., "Robust heart rate from fitness videos," *Physiol. Meas.*, vol. 38, no. 6, May 2017, Art. no. 1023.
[28] Y. Sun et al., "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *J. Biomed. Opt.*, vol. 16, no. 7, Jul. 2011, Art. no. 077010.
[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, MD, USA, 2020.
[31] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2002.
[32] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Exp.*, vol. 16, no. 26, pp. 21434–45, Dec. 2008.
[33] L. Kong et al., "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light," *Opt. Exp.*, vol. 21, no. 15, pp. 17464–17471, 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 *Proc. IEEE 2020 11th Int. Conf. Comput., Comm. Netw. Tech.*, 2020, pp. 1–5.
[36] M. Lewandowska et al., "Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity," in *Federated Conf. Comput. Sci. Info. Syst.*, 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," *Opt. Exp.*, vol. 18, no. 10, pp. 10762–10774, May 2010.
[38] R. Song et al., "Remote photoplethysmography with an EEMD-MCCA method robust against spatially uneven illuminations," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13484–13494, Jun. 2021.

[39] 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. Biomed. Opt.*, vol. 20, no. 4, Apr. 2015, Art. no. 048002.
[40] 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, Sep. 2016.
[41] A. Pai, A. Veeraraghavan, and A. Sabharwal, "HRVCam: Robust camera-based measurement of heart rate variability," *J. Biomed. Opt.*, vol. 26, no. 2, 2021, Art. no. 022707.
[42] S. Tulyakov et al., "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 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 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 *Proc. IEEE 2017 Int. Joint Conf. Biometrics*, 2017, pp. 383–389.
[45] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proc. Brit. Mach. Vis. Conf*., 2019, pp. 1–12.
[46] V. R. Shenoy et al., "Unrolled iPPG: Video heart rate estimation via unrolling proximal gradient descent," in *Proc. IEEE 2023 Int. Conf. Image Process.*, 2023, pp. 2715–2719.
[47] X. Niu et al., "Video-based remote physiological measurement via cross-verified feature disentangling," in *Eur. Conf. Comput. Vision*, 2020, pp. 295–310.
[48] Y.-Y. Tsou et al., "Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos," in *Annu. ACM Symp. Appl. Comput*., 2020, pp. 2066–2073.
[49] M. Hu et al., "Robust heart rate estimation with spatial–temporal attention network from facial videos," *IEEE Trans. Cogn. Devlop. Syst.*, vol. 14, no. 2, pp. 639–647, Jun. 2022.
[50] Y. Napolean et al., "Heart rate estimation in intense exercise videos," in *Proc. IEEE 2022 Int. Conf. Image Process*., 2022, pp. 3933–3937.
[51] X. Niu et al., "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video," in *Computer Vis.–ACCV 2018: 14th Asian Conf. Computer Vis.*, Perth, Australia, Dec. 2–6, 2018, 2019, pp. 562–576.
[52] R. R. Anderson and J. A. Parrish, "The optics of human skin," *J. Invest. Dermatol.*, vol. 77, no. 1, pp. 13–19, Jul. 1981.
[53] H. Takiwaki et al., "Measurement of skin color: Practical application and theoretical considerations," *J. Med. Investigation*, vol. 44, pp. 121–126, Feb. 1998.
[54] S. A. Shafer, "Using color to separate reflection components," *Color Res. App.*, vol. 10, no. 4, pp. 210–218, Dec. 1985.
[55] 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, Feb. 2002.
[56] Q. Zhu et al., "Fitness heart rate measurement using face videos," in *Proc. IEEE 2017 Int. Conf. Image Process.*, Beijing, China, Sep. 2017, pp. 2000–2004.
[57] Q. Zhu et al., "Adaptive multi-trace carving based on dynamic programming," in *Proc. IEEE 52nd Asilomar Conf. Signal Syst. Comput*., Pacific Grove, CA, USA, Oct. 2018, pp. 1716–1720.
[58] Q. Zhu et al., "Adaptive multi-trace carving for robust frequency tracking in forensic applications," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1174–1189, 2021.
[59] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[60] T. Brox et al., "High accuracy optical flow estimation based on a theory for warping," in *Eur. Conf. Comput. Vis.*, May 2004, pp. 25–36.
[61] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scand. Conf. Image Anal.*, Jun. 2003, pp. 363–370.
[62] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.
[63] S. Jiang et al., "Learning to estimate hidden motions with global motion aggregation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9772–9781.
[64] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.
[65] X. Yu et al., "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Dec. 2013, pp. 1944–1951.
[66] R. Špetlík, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Brit. Mach. Vis. Conf.*, Newcastle, U.K., 2018, pp. 3–6.
[67] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, Jun. 1988.
[68] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, Jan. 2002.
[69] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE/CVF 1994 Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 593–600.
[70] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, Vancouver, Canada, Aug. 1981, pp. 674–679.
[71] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, Aug. 1981.
[72] 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.
[73] Y. Shi and E. Chang, "Spectrogram-based formant trackin via particle filters," in *Proc. IEEE 2003 Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, Apr. 2003, pp. 168–171.
[74] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proc. IEEE 2002 Int. Conf. Acoust., Speech, Signal Process.*, May 2002, pp. I-361–I-364.
[75] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. IEEE 23rd Int. Symp. Robot Hum. Interact. Commun.*, 2014, pp. 1056–1062.
[76] A. K. Maity et al., "RobustPPG: Camera-based robust heart rate estimation using motion cancellation," *Biomed. Opt. Exp.*, vol. 13, no. 10, pp. 5447–5467, Oct. 2022.