IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **38** (2017) 1023–1044 https://doi.org/10.1088/1361-6579/aa6d02

# Robust heart rate from fitness videos

**Wenjin Wang**[1,3], **Albertus C den Brinker**[2], **Sander Stuijk**[1] **and Gerard de Haan**[1,2]

[1] Department of Electrical Engineering, Eindhoven University of Technology, 5600MB Eindhoven, Netherlands
[2] Philip Research Eindhoven, High Tech Campus 36, 5656AE Eindhoven, Netherlands

E-mail: w.wang@tue.nl

Received 9 January 2017, revised 6 April 2017
Accepted for publication 12 April 2017
Published 8 May 2017

## Abstract

Remote photoplethysmography (rPPG) enables contactless heart-rate monitoring using a regular video camera. *Objective:* This paper aims to improve the rPPG technology targeting continuous heart-rate measurement during fitness exercises. The fundamental limitation of the existing (multi-wavelength) rPPG methods is that they can suppress at most $n - 1$ independent distortions by linearly combining $n$ wavelength color channels. Their performance are highly restricted when more than $n - 1$ independent distortions appear in a measurement, as typically occurs in fitness applications with vigorous body motions. *Approach:* To mitigate this limitation, we propose an effective yet very simple method that algorithmically extends the number of possibly suppressed distortions without using more wavelengths. Our core idea is to increase the degrees-of-freedom of noise reduction by decomposing the $n$ wavelength camera-signals into multiple orthogonal frequency bands and extracting the pulse-signal per band-basis. This processing, namely Sub-band rPPG (SB), can suppress different distortion-frequencies using independent combinations of color channels. *Main results:* A challenging fitness benchmark dataset is created, including 25 videos recorded from 7 healthy adult subjects (ages from 25 to 40 yrs; six male and one female) running on a treadmill in an indoor environment. Various practical challenges are simulated in the recordings, such as different skin-tones, light sources, illumination intensities, and exercising modes. The basic form of SB is benchmarked against a state-of-the-art method (POS) on the fitness dataset. Using non-biased parameter settings, the average signal-to-noise-ratio (SNR) for POS varies in $[-4.18, -2.07]$ dB, for SB varies in $[-1.08, 4.77]$ dB. The ANOVA test shows that the improvement of SB over POS is statistically significant for almost all settings (p-value $<0.05$). *Significance:* The results suggest that the proposed SB method considerably increases the

[3] Author to whom any correspondence should be addressed.

1361-6579/17/061023+22$33.00

robustness of heart-rate measurement in challenging fitness applications, and outperforms the state-of-the-art method.

Keywords: biomedical monitoring, remote photoplethysmography, heart rate, video processing, fitness, healthcare

(Some figures may appear in colour only in the online journal)

## 1. Introduction

Remote photoplethysmography (rPPG) enables contactless monitoring of cardiac activity by measuring the pulse-induced subtle color variations of human skin using a color video camera (Takano and Ohta *et al* 2007, Verkruysse *et al* 2008). This measurement is based on the fact that the pulsatile blood propagating in the human cardiovascular system changes the blood volume in skin tissue. The oxygenated blood circulation leads to fluctuations in the amount of hemoglobin molecules and proteins thereby causing variations in the optical absorption and scattering across the light spectrum (Allen *et al* 2007). A single-/multi- wavelength camera can therefore be used to identify the phase of the blood circulation based on minute changes in skin reflections.

A thorough review on the development of rPPG can be found (McDuff *et al* 2015, Rouast *et al* 2016, Sikdar *et al* 2016, Sun and Thakor *et al* 2016). The fundamental use of rPPG leads to various applications for video health monitoring, enabling non-contact measurement of physiological parameters from a human body, such as heart-rate (Li *et al* 2014, Tarassenko *et al* 2014, Kumar *et al* 2015, Wang *et al* 2015a, Tulyakov *et al* 2016), heart-rate variability (Blackford *et al* 2016), respiration (Tarassenko *et al* 2014), $SpO_2$ (Guazzi *et al* 2015), pulse transit time (Shao *et al* 2014), blood pressure (Jeong *et al* 2016), atrial fibrillation (Couderc *et al* 2015), mental stress (McDuff *et al* 2014a), monitoring of neonates (Mestha *et al* 2014, Fernando *et al* 2015), living-skin detection for face anti-spoofing (Gibert *et al* 2013, Wang *et al* 2015b, Liu *et al* 2016), etc. In addition to the clinical and home-based applications, the rPPG technique would also be attractive in the gym. Particularly for vigorous exercise, rPPG allows a very convenient way to optimize the effectiveness of a workout. As compared to the wrist-based PPG function (Zhang *et al* 2015, Zhang 2015, Temko 2017) that is popular in current smart fitness bands or smart watches, the camera-based rPPG function is much less explored for the demanding fitness scenario.

In recent years, much progress has been reported in improving the robustness of rPPG in terms of skin-tone, body-motion and illumination challenges. The various proposed methods include: (i) Blind source separation based methods, which use different criteria (i.e. principal component analysis (PCA) (Lewandowska *et al* 2011) and independent component analysis (ICA) (Poh *et al* 2011, Tsouri *et al* 2012)) to unmix the RGB-signals obtained by a camera into uncorrelated or independent signal sources and select the most periodic one as the pulse; (ii) Color-space driven methods, which measure the pulse in different standard color-spaces (i.e. HUE color-space (Tsouri and Li 2015) and lab color-space (Yang 2016)), i.e. some optical disturbances (e.g. intensity variations) can be eliminated in the transformed color-spaces; (iii) Chrominance-based method (CHROM) (de Haan G and Jeanne 2013), which uses knowledge of the main distortion (e.g. specular variation) in a skin reflection model to deterministically extract the pulse; (iv) Blood-volume pulse signature method (PBV) (de Haan and Van Leest 2014), which uses the characteristic color absorption variations caused by the blood volume change as a signature to derive the pulse without assumptions regarding the optical distortions; (v) Spatial subspace rotation (2SR) (Wang *et al* 2016b), which measures the temporal hue-change of the subject-dependent skin subspace (e.g. body reflection) for pulse extraction, which is similar to the HUE-based approach (Tsouri and Li 2015) in essence; and (vi) Plane orthogonal to the Skin-tone method (POS) (Wang *et al* 2016a), which exploits the same skin reflection model as CHROM but uses a different color direction (i.e. a different distortion) for real-time projection tuning. All these methods use linear combinations of color channels to separate pulse and (motion-induced) distortions. They differ in the assumptions applied for deriving the combining weights. Their strength and weakness have been thoroughly benchmarked and discussed in (Wang *et al* 2016a).

However, there is a common fundamental limitation in existing rPPG methods: the one-dimensional pulse-signal extracted from three-dimensional RGB-signals can maximally be independent of two distortions by linear channel combination. Such a mathematical limitation highly restricts the rPPG performance when RGB-signals contain more than two independent distortions, which is the typical scenario in the challenging use-case of fitness exercises (see figure 1(a)). The reason is that in the non-homogeneous illumination conditions (e.g. different light sources with unequal spectrum, or reflections from nearby colored walls), the distribution of specular reflection on the skin surface (with 3D geometry) is non-uniform. During fitness exercises such as running, the vertical body motion has a frequency twice higher than that of the horizontal body motion, caused by the fact that the subject runs on two legs where each has to hit the ground in a full motion cycle. This implies that different motion frequencies may have different color variation directions in RGB space, which cannot be simultaneously eliminated by current rPPG methods (Lewandowska *et al* 2011, Poh *et al* 2011, de Haan G and Jeanne 2013, de Haan and Van Leest 2014, Wang *et al* 2016b) that exploit three degrees-of-freedom pulse extraction offered by the linear projection. One possible solution is to physically increase the dimensionality of measurement by using more color channels such as a five-band camera (RGBCO) (McDuff *et al* 2014b). Ignoring the application-related issues such as the availability of such devices and price points, it still imposes a clear-cut limit on the number of distortions, i.e. the number of distortions that can be eliminated is smaller than the number of channels.

In this paper, we propose a new strategy that algorithmically increases the dimensionality of pulse extraction given limited color-sensors. Our inspiration is based on the observation that different motion frequencies cause apparent distortions in different color variation directions in RGB space. This precludes treating them simultaneously, but treating them independently in different frequency bands may solve this problem. To this end, we decompose the RGB-signals into multiple orthogonal frequency bands for sub-band pulse extraction. Once the different motion frequencies are separated and suppressed in different frequency bands, we can synthesize a clean pulse-signal by combining the processing results from the individual sub-bands. This idea leads to a novel pulse extraction method called 'Sub-band rPPG' (SB). A benchmark is executed to evaluate the basic form of SB and compare its robustness to a state-of-the-art rPPG method. All the benchmark videos are recorded from the subjects running on a treadmill in an indoor environment, involving various challenges such as different skin-tones, illumination conditions, body-parts, and motion-types. Results from this benchmark show that SB has significant improvement in pulse-rate measurement in challenging fitness applications with their characteristic (vigorous and periodical) body movements (see figure 1(b)). The contributions made by this paper are threefold:

- it provides an in-depth analysis of the fundamental limitation in the existing rPPG methods, using a mathematical skin reflection model;
- it introduces a new strategy that uses the sub-band decomposition to extend the degrees-of-freedom for noise reduction, leading to a novel Sub-band rPPG method that particularly benefits the heart-rate measurement in fitness applications;

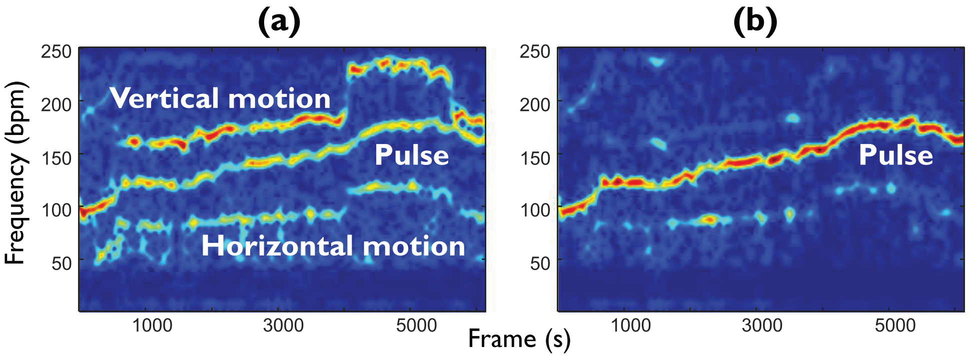

**Figure 1.** The spectrograms of the pulse-signals obtained by (a) the existing state-of-the-art method (POS (Wang *et al* 2016a)) and (b) the proposed SB method in this paper, from a subject running on a treadmill. The horizontal and vertical motion frequencies induced by running cannot be eliminated by POS, but are almost completely removed by SB.

- it contains a fitness video dataset that includes various practical challenges. This dataset has been used to evaluate the proposed method, and can be used in future benchmarking.

The remainder of this paper is structured as follows. In section 2, we analyze the considered problem in a mathematical context. In section 3, we describe the proposed method step-by-step. In section 4, we introduce the experimental setup. In sections 5 and 6, the proposed method is experimentally verified and discussed. Finally in section 7, we draw our conclusions.

## 2. Problem definition

Unless stated otherwise, we use the following mathematical conventions throughout the paper. Vectors and matrices are denoted as boldface characters, where the column vectors with unit-length are denoted as $\mathbf{u}$. The variable $t$ denotes the time and $\mathbf{1}$ denotes $(1, 1, 1)^\top$.

The goal of this paper is to improve the rPPG robustness in particularly challenging use-cases, e.g. pulse-rate monitoring in fitness exercises. As mentioned earlier, the fundamental mathematical limitation in all existing rPPG methods is that maximally two independent distortions can be eliminated in three color channels of an RGB camera. However, in practice, there are usually more than two independent distortions in RGB-signals, especially in a fitness application. The reason is that the unequal illumination spectra (emitted from different light sources or reflected from nearby colored walls) produce spatially-different specular reflections on the skin surface. The specular distortions, generated by different body motions, are varying in different movement directions and may have different color variations that cannot be simultaneously eliminated by current rPPG methods.

In order to design a robust solution, we first look into a skin reflection model (Wang *et al* 2016a) to define our problem. This model considers the pertinent optical and physiological properties of skin reflections in a mathematical context. It assumes a setup where a light source illuminates a piece of human skin-tissue containing pulsatile blood and an RGB camera records this image remotely and sequentially.

Averaging the skin-pixel values in individual video frames and concatenating the resulting values (e.g. spatial RGB mean) through the video, we can obtain the RGB-traces that describe the skin color changes over time. We denote the RGB-signals as $\mathbf{C}(t)$, which is a matrix with

RGB-channels sorted in rows and time-samples stacked in columns. Based on the dichromatic reflection model, we know that $\mathbf{C}(t)$ consists of the diffuse and specular reflections from the skin surface, where the diffuse reflection contains the target pulse-signal denoted as $p(t)$ and the specular reflection contains (non-pulsatile) specular variation signal denoted as $s(t)$. Both components are proportional to the light intensity level and thus modulated by the intensity variation signal denoted as $i(t)$. Note that $p(t)$, $s(t)$ and $i(t)$ are zero-mean AC-signals with the variation amplitudes much smaller than the overall DC component. According to Wang *et al* (2016a), the relation of $p(t)$, $s(t)$ and $i(t)$ in $\mathbf{C}(t)$ can be expressed as:

$$\mathbf{C}(t) = I_0\big(1 + i(t)\big)\big(\mathbf{u_c}c_0 + \mathbf{u_s}s(t) + \mathbf{u_p}p(t)\big), \tag{1}$$

where $I_0$ denotes the light intensity level; $c_0$ denotes the static reflection strength (i.e. the DC component); $\mathbf{u_c}$, $\mathbf{u_s}$ and $\mathbf{u_p}$ denote the unit color vectors associated with the skin reflection, lighting spectra and relative PPG-strengths in RGB channels (i.e. the blood volume pulse signature (de Haan and Van Leest 2014)). Equation (1) can be expanded as:

$$\mathbf{C}(t) \approx I_0\mathbf{u_c}c_0 + I_0\mathbf{u_c}c_0 i(t) + I_0\mathbf{u_s}s(t) + I_0\mathbf{u_p}p(t), \tag{2}$$

where the components involving the multiplication of two AC-signals (e.g. $p(t) \cdot i(t)$) are several orders of magnitude smaller than DC, and are therefore neglected in the approximation.

However, the model (2) has a clear limitation: it is restricted to a single light source and also the underlying assumption that motion only creates a single specular variation direction w.r.t. the light source (next to those in the intensity variation direction). It cannot be used to describe the complex situations with multiple light sources or multiple distortions, such as found in the fitness use-case. Therefore, we will consider the more general situation with multiple light sources and propose a method to handle this.

Since the effect of multiple lighting spectra on the same piece of skin-tissue is additive, (2) can be extended as:

$$\mathbf{C}(t) \approx \sum_{j=1}^{J} I_{0,j}\mathbf{u_{c,j}}c_{0,j} + \sum_{j=1}^{J} I_{0,j}\mathbf{u_{c,j}}c_{0,j}i_j(t) + \sum_{j=1}^{J} I_{0,j}\mathbf{u_{s,j}}s_j(t) + \left(\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\right)p(t), \tag{3}$$

where $j$ denotes the $j$th light source; $J$ denotes the total number of light sources in the setup; $i_j(t)$ and $s_j(t)$ denote the intensity variation signal and specular variation signal of the $j$-the light source; $p(t)$ still denotes a single pulse-signal, which has the average blood volume vector under multiple lighting spectra, i.e. mankind has only one cardiovascular system.

To eliminate the dependency of $\mathbf{C}(t)$ on the average skin reflection color (including the light source color and intrinsic skin-tone color), we temporally normalize the DC of $\mathbf{C}(t)$. The temporal mean of $\mathbf{C}(t)$ can be considered as the largest steady component in (3):

$$\bar{\mathbf{C}}(t) \approx \sum_{j=1}^{J} I_{0,j}\mathbf{u_{c,j}}c_{0,j}, \tag{4}$$

which is used to uniquely define a diagonal normalization matrix $\mathbf{N}$, such that:

$$\mathbf{N} \cdot \bar{\mathbf{C}}(t) = \mathbf{N} \cdot \sum_{j=1}^{J} I_{0,j}\mathbf{u_{c,j}}c_{0,j} = \mathbf{1}. \tag{5}$$

Then we use $\mathbf{N}$ to normalize $\mathbf{C}(t)$ and remove its mean (by subtracting $\mathbf{1}$) as:

$$\begin{aligned}\tilde{\mathbf{C}}(t) &= \mathbf{N}^{-1}\mathbf{C}(t) - \mathbf{1}\\ &\approx \mathbf{N}^{-1}\Big(\overbrace{\sum_{j=1}^{J} I_{0,j}\mathbf{u_{c,j}}c_{0,j}i_j(t)}^{\text{Intensity}} + \overbrace{\sum_{j=1}^{J} I_{0,j}\mathbf{u_{s,j}}s_j(t)}^{\text{Specular}} + \overbrace{\Big(\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\Big)p(t)}^{\text{Pulse}}\Big), \end{aligned} \tag{6}$$

where $\tilde{\mathbf{C}}(t)$ denotes the temporally normalized RGB-signals with zero-mean. Since skin-motion (or the relative position change between light source, camera and skin) is the source causing variations in $i_j(t)$ and $s_j(t)$, we can define both components in terms of the 'motion source'. Based on the fact that the intensity and specular variations due to the same motion source have the same frequency and phase, we write $i_j(t)$ and $s_j(t)$ as:

$$\begin{cases} i_j(t) = a_{j,1}m_1(t) + a_{j,1}m_2(t) + ...a_{j,K}m_K(t) = \sum\limits_{k=1}^{K} a_{j,k}m_k(t) \\ s_j(t) = b_{j,1}m_1(t) + b_{j,1}m_2(t) + ...b_{j,K}m_K(t) = \sum\limits_{k=1}^{K} b_{j,k}m_k(t), \end{cases} \tag{7}$$

where $m_k(t)$ denotes the $k$th motion source; $K$ denotes the total number of motion sources; $a_{j,k}$ and $b_{j,k}$ denote the intensity and specular variation strengths induced by the $k$th motion w.r.t. the $j$th light source. Substituting (7) and (6), (6) gives:

$$\begin{aligned}\tilde{\mathbf{C}}(t) &\approx \mathbf{N}^{-1}\Big(\sum_{j=1}^{J} I_0\mathbf{u_{c,j}}c_{0,j}\Big(\sum_{k=1}^{K} a_{j,k}m_k(t)\Big) + \sum_{j=1}^{J} I_{0,j}\mathbf{u_{s,j}}\Big(\sum_{k=1}^{K} b_{j,k}m_k(t)\Big) + \Big(\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\Big)p(t)\Big)\\ &= \mathbf{N}^{-1}\Big(\sum_{k=1}^{K}\Big(\sum_{j=1}^{J} a_{j,k}I_0\mathbf{u_{c,j}}c_{0,j}\Big)m_k(t) + \sum_{k=1}^{K}\Big(\sum_{j=1}^{J} b_{j,k}I_{0,j}\mathbf{u_{s,j}}\Big)m_k(t) + \Big(\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\Big)p(t)\Big)\\ &= \overbrace{\mathbf{N}^{-1}\sum_{k=1}^{K}\Big(\sum_{j=1}^{J} a_{j,k}I_0\mathbf{u_{c,j}}c_{0,j} + b_{j,k}I_{0,j}\mathbf{u_{s,j}}\Big)m_k(t)}^{\text{Motion}} + \overbrace{\Big(\mathbf{N}^{-1}\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\Big)p(t)}^{\text{Pulse}}.\end{aligned} \tag{8}$$

Given (8), it is clear that if different motion signals $m_k(t)$, under different light sources, have different color vectors, they will not be fully eliminated by the three degrees-of-freedom noise suppression. Considering the two extreme scenarios with either the 'single light source' or 'single motion source', we have the following two observations:

- **Single light source:** $J = 1$ and (8) can be written as:

$$\tilde{\mathbf{C}}(t) = \mathbf{N}^{-1}\sum_{k=1}^{K}\Big(a_kI_0\mathbf{u_c}c_0 + b_kI_0\mathbf{u_s}\Big)m_k(t) + \mathbf{N}^{-1}I_0\mathbf{u_p}p(t), \tag{9}$$

  where different $m_k(t)$ may still have different color vectors, as the $k$th motion source may generate different intensity and specular variation amplitudes, typically when $a_k$ and $b_k$ are unequal.
- **Single motion source:** $K = 1$ and (8) can be written as:

$$\tilde{\mathbf{C}}(t) = \Big(\mathbf{N}^{-1}\sum_{j=1}^{J} a_jI_0\mathbf{u_{c,j}}c_{0,j} + b_jI_{0,j}\mathbf{u_{s,j}}\Big)m(t) + \Big(\mathbf{N}^{-1}\sum_{j=1}^{J} I_{0,j}\mathbf{u_{p,j}}\Big)p(t), \tag{10}$$

where the color vector of the single motion signal $m(t)$ is a single component averaged over multiple lighting spectra, which can be solved by existing rPPG methods.

Based on the above two extreme conditions, we recognize that the complexity of the model is essentially determined by the number of motion sources, instead of the number of light sources. Our strategy to solve (8) is, therefore, turning the 'multiple motion sources' problem into the 'single motion source' problem that already has a solution. This is equivalent to translating (8) and (10), separating different motion-sources into different units for independent processing.

Triggered by the observation that motion-sources in fitness exercises usually have different frequencies[4], we propose to apply the 'frequency' as a unit to separate $m_k(t)$. Once $m_k(t)$ are separated into different frequency bands, we can use existing rPPG methods in each sub-band to extract the pulse and eliminate distortions independently (see figure 2). Theoretically, when the number of sub-bands is not smaller than the number of motion sources, multiple motion distortions (with different color vectors) can be suppressed simultaneously. This strategy leads to a novel rPPG method that uses the sub-band decomposition to extend the dimensionality of pulse extraction. In the following section, we shall present the complete method based on this analysis step-by-step.

## 3. Method

This section presents the proposed Sub-band rPPG method and its implementation.

### 3.1. Spatial quantization and temporal normalization

The first step is common to that in the model-based rPPG methods (Wang *et al* 2016a, de Haan G and Jeanne 2013, de Haan and Van Leest 2014): given an input video sequence containing living skin-tissue, we spatially average the RGB values of skin-pixels in each frame as the *spatial RGB mean*, and then temporally concatenate these values obtained from consecutive frames into a matrix **C**, i.e. assuming a video interval has $N$ frames, the RGB traces are contained in a $3 \times N$ matrix. Each row represents a color-channel and three rows are sorted in R-G-B order. Given a video camera recording at 20 frames per second (fps), the time span of the data in **C** covers $N$/20 s. To eliminate the DC-color, each row of **C** is temporally normalized as:

$$\tilde{\mathbf{C}}_{\mathbf{i}} = \frac{\mathbf{C}_{\mathbf{i}}}{\mu(\mathbf{C}_{\mathbf{i}})} - 1, \tag{11}$$

where $\mathbf{C}_{\mathbf{i}}$ denotes the $i$th row (i.e. $i$th color-channel) of **C** and $\mu(\cdot)$ denotes the temporal averaging operator. The *spatial pixel averaging* reduces the camera quantization error, while the *temporal normalization* eliminates the dependency of **C** on the average skin reflection color (including the light source color and intrinsic skin-tone color).

### 3.2. Sub-band decomposition

The second step in current rPPG methods (Wang *et al* 2016a, Lewandowska *et al* 2011, Poh *et al* 2011, de Haan G and Jeanne 2013, de Haan and Van Leest 2014) is linearly combining

[4] In most fitness exercises (e.g. running, biking and stepping), body motions of a subject usually have different frequencies, as the subject exercises on two legs where each has to complete a full motion cycle.

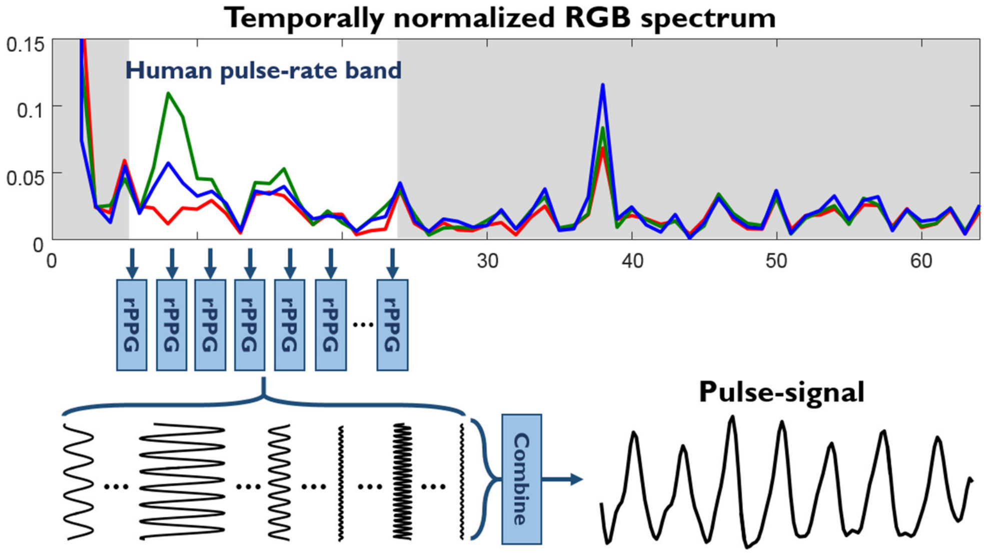

**Figure 2.** Illustration of the proposed Sub-band rPPG method. The temporally normalized RGB-signals are transformed into the frequency domain. Within a broad human heart-rate band (e.g. [40,240] beats per minute (bpm)), the RGB frequency spectrum are decomposed into multiple orthogonal sub-bands for local and parallel pulse extraction using an existing rPPG algorithm (e.g., POS). The extracted sub-band pulse-signals are combined into a global pulse-signal as the final output.

the RGB-signals into a pulse-signal. However, such a combination can maximally suppress two independent distortions, thus restricting its application to simple use-cases. Based on our earlier analysis, we propose to extend the degrees-of-freedom of noise suppression by decomposing the RGB-signals into different frequency bands for the sub-band pulse estimation.

Essentially, the frequency-based decomposition exploits an important property: *the pulsatile/motion component has the same frequency in different color channels*[5]. This property allows us to use the frequency band as a unit to group the different components for local processing. To this end, $\mathbf{C_i}$ is transformed into the frequency domain using the discrete fourier transform (DFT):

$$\mathbf{F_i} = \mathbf{DFT}(\tilde{\mathbf{C}}_\mathbf{i}), \tag{12}$$

where $\mathbf{F_i}$ denotes the frequency spectrum of the $i$th color channel (i.e. containing real and imaginary parts); $\mathbf{DFT}(\cdot)$ denotes the DFT operator.

Next, we decompose $\mathbf{F_i}$ into different sub-bands. Here the sub-band particularly refers to the *frequency bin* within a broad *human heart-rate band*. The rationale is: the frequency bins inside the human heart-rate band (e.g. 40–240 bpm) could all possibly contain the pulsatile content, and thus should be analyzed independently. In contrast, the frequency bins outside the human heart-rate band are clearly noise that can be safely ignored. Therefore, only the RGB components within the assumed broad heart-rate band need to be transformed back to the time-domain for analysis, using the inverse discrete Fourier transform (IDFT):

$$\tilde{\mathbf{C}}_\mathbf{i,k} = \mathbf{real}\big(\mathbf{IDFT}(\mathbf{F_{i,k}})\big),\ k \in \mathbf{b}, \tag{13}$$

[5] All color channels of a camera should sense the same heart-rate, i.e. it is impossible for the R-channel to have 70 bpm heart-rate and the G-channel to have 60 bpm heart-rate.

where $\mathbf{F_{i,k}}$ denotes the *k*th sub-band (frequency bin) of $\mathbf{F_i}$ selected between the human heart-rate range $\mathbf{b} = [b_1, b_2]$; $\tilde{\mathbf{C}}_{\mathbf{i,k}}$ denotes the *i*th channel signal in the *k*th sub-band; $\mathbf{IDFT}(\cdot)$ denotes the IDFT operator; $\mathbf{real}(\cdot)$ denotes the operator that takes the real part of a complex value. The reason why we need to take the real part of the transformed signal is that the conjugate symmetry of IDFT has been destroyed, since only a number of $\mathbf{F_{i,k}}$ within the assumed heart-rate band are selected for transformation.

Different sub-band RGB-signals $\tilde{\mathbf{C}}_{\mathbf{k}}$ are orthogonal to each other, as they are sampled from independent DFT frequency-bins. Thus the signal is split into $K$ independent signal components, each of which can be used as input to a pulse extraction algorithm. This allows us to address the color distortions associated with each distortion-frequency individually, which is impossible in a single channel system.

### 3.3. Local pulse extraction

Given the sub-band RGB-signals $\tilde{\mathbf{C}}_{\mathbf{k}}$, we can now use the rPPG method to extract the pulse-signal from it. This step can be generally expressed as:

$$\mathbf{P_k} = \mathbf{rPPG}(\tilde{\mathbf{C}}_{\mathbf{k}}), \tag{14}$$

where $\mathbf{P_k}$ denotes the pulse-signal extracted from the *k*th sub-band; $\mathbf{rPPG}(\cdot)$ denotes the rPPG function that converts the input 3D RGB-signals into the 1D pulse-signal. As a matter of fact, not all rPPG methods (Lewandowska *et al* 2011, Poh *et al* 2011, de Haan G and Jeanne 2013, de Haan and Van Leest 2014, Wang *et al* 2016b) can be used for this task. The covariance-related methods (e.g. PCA-based (Lewandowska *et al* 2011), ICA-based (Poh *et al* 2011), PBV (de Haan and Van Leest 2014)) should be avoided. The reason is that the covariance matrix could be singular/near-singular in sub-band RGB-signals, especially when different color-signals in $\tilde{\mathbf{C}}_{\mathbf{k}}$ have no phase-shift, i.e. $\tilde{\mathbf{C}}_{\mathbf{k}}$ may contain purely noise or pulse. Since the sub-band color signals are not spatially redundant, 2SR (Wang *et al* 2016b) cannot be applied. Thus only G-R (Hülsbusch 2008), HUE (Tsouri and Li 2015), Lab (Yang 2016), CHROM (de Haan G and Jeanne 2013) and POS (Wang *et al* 2016a) are considered as candidates here.

Since POS reports the overall best performance in general use-cases in a large benchmark of Wang *et al* (2016a), we choose POS to demonstrate the sub-band pulse extraction, even though its performance does not significantly differ from CHROM in the fitness use-case (Wang *et al* 2016a). Accordingly, $\mathbf{P_k}$ in (14) can be derived, for example, by:

$$\mathbf{P_k} = \mathbf{X_k} + \frac{\sigma(\mathbf{X_k})}{\sigma(\mathbf{Y_k})} \cdot \mathbf{Y_k} \quad \text{with} \quad \begin{cases} \mathbf{X_k} = \mathbf{G_k} - \mathbf{B_k} \\ \mathbf{Y_k} = \mathbf{G_k} + \mathbf{B_k} - 2\mathbf{R_k} \end{cases}, \tag{15}$$

where $\mathbf{R_k}$, $\mathbf{G_k}$ and $\mathbf{B_k}$ denote the RGB-signals of $\tilde{\mathbf{C}}_{\mathbf{k}}$, respectively; $\sigma(\cdot)$ denotes the standard deviation operator.

### 3.4. Global pulse combination

In order to output a single pulse-signal, we need to combine individual sub-band pulse-signals. This can be done either by averaging (i.e. take the mean of different $\mathbf{P_k}$) or by weighting. Since the sub-bands influenced by motion frequencies have typically large energies, their estimated $\mathbf{P_k}$ remain to have large variations in their amplitudes. To further suppress the $\mathbf{P_k}$ from motion-dominated sub-bands, we introduce a weighted summation to combine $\mathbf{P_k}$ into a final pulse-signal

**Algorithm 1.** Sub-band rPPG (SB)

```
Input: The raw RGB-signals $\mathbf{RGB}$ with dimension $3 \times N$
Initialize: $l = 128$ (for example), $\mathbf{B} = [6, 24]$ (adapted to $l$), $\hat{\mathbf{P}} = \mathbf{zero}(1, N)$
for $n = 1, 2, ..., N - 1 + 1$ do
    $\mathbf{C} = \mathbf{RGB}(:, n : n + l - 1)$;
    $\tilde{\mathbf{C}} = \mathbf{diag}(\mathbf{mean}(\mathbf{C}, 2))^{-1} * \mathbf{C} - 1$;
    $\mathbf{F} = \mathbf{fft}(\tilde{\mathbf{C}}, [\,], 2)$;
    $\mathbf{S} = [0, 1, -1; -2, 1, 1] * \mathbf{F}$; $\mathbf{Z} = \mathbf{S}(1, :) + \mathbf{abs}(\mathbf{S}(1, :))./\mathbf{abs}(\mathbf{S}(2, :)). * \mathbf{S}(2, :)$;
    $\bar{\mathbf{Z}} = \mathbf{Z}. * (\mathbf{abs}(\mathbf{Z})./\mathbf{abs}(\mathbf{sum}(\mathbf{F}, 1)))$;
    $\bar{\mathbf{Z}}(:, 1 : \mathbf{B}(1) - 1) = 0$; $\bar{\mathbf{Z}}(:, \mathbf{B}(2) + 1 : \mathrm{end}) = 0$;
    $\bar{\mathbf{P}} = \mathbf{real}(\mathbf{ifft}(\bar{\mathbf{Z}}, [\,], 2))$;
    $\hat{\mathbf{P}}(1, n : n + l - 1) = \hat{\mathbf{P}}(1, n : n + l - 1) + (\bar{\mathbf{P}} - \mathbf{mean}(\bar{\mathbf{P}}))/\mathbf{std}(\bar{\mathbf{P}})$;
end for
Output: The pulse-signal $\hat{\mathbf{P}}$ with dimension $1 \times N$
```

$$\bar{\mathbf{P}} = \sum_{k=b_1}^{b_2} w_k \cdot \mathbf{P_k}, \tag{16}$$

where $\bar{\mathbf{P}}$ denotes the combined pulse-signal; $w_k$ denotes the weighting factor for the $k$th sub-band within $[b_1, b_2]$. Since large motion distortions usually present large *intensity variations* (Wang *et al* 2016a), we use the ratio between the pulsatile amplitude and intensity variation amplitude to define the combining weight:

$$w_k = \frac{\sigma(\mathbf{P_k})}{\sigma(\mathbf{R_k} + \mathbf{G_k} + \mathbf{B_k})}, \tag{17}$$

where $\mathbf{R_k} + \mathbf{G_k} + \mathbf{B_k}$ is the projection of the color-signals onto the direction of $\mathbf{1}$, which is the intensity variation direction in the temporally normalized RGB space (Wang *et al* 2016a, de Haan G and Jeanne 2013). According to (17), the sub-bands suffering from large intensity variations w.r.t. the pulsatile variations will receive a lower weight in (16). Note that the presented solution is one way of creating the weighted combination. Although there could be alternatives, we do not aim to find the optimal combination in this work.

Consequently, we arrive at a long-term pulse-signal $\hat{\mathbf{P}}$ by overlap-adding $\bar{\mathbf{P}}$ (after removing its mean and normalizing its standard deviation) estimated in each short video interval using a sliding window (with one time-sample shift similar to (Wang *et al* 2016a, de Haan G and Jeanne 2013, de Haan and Van Leest 2014, Wang *et al* 2016b)), and output $\hat{\mathbf{P}}$ as the final pulse-signal.

### 3.5. Algorithm

*3.5.1. Overview.* The overview of the complete method is illustrated in figure 2. The novelty of our method is an algorithmic extension of the degrees-of-freedom for pulse extraction using sub-band analysis. Thus it is named Sub-band rPPG (SB). In order to show the fundamental behavior of SB and facilitate its replication, we keep it as clean and simple as possible, although we acknowledge that there are various known techniques that could improve it further, i.e. using dedicated post-processing (e.g. adaptive band-pass filtering (Wang *et al* 2015a) or singular spectrum analysis for motion de-noising (Wang *et al* 2016c)) to improve the time-consistency of the outcome or using other filter-banks (e.g. wavelet-based) for the sub-band decomposition. The bare algorithm of SB is shown in algorithm 1, which can be implemented in a few lines of Matlab code.

*3.5.2. Parameter selection.* The proposed SB method has only two parameters: the processing window length $l$ and the human heart-rate band $\mathbf{b}$. Since $\mathbf{b}$ (a broad band representing [40,240] bpm) is adapted to $l$, we only need to define $l$. Given a video camera recording at 20 fps, the time scale of $l$ is in seconds. We expect that a longer window is better for pulse and motion separation. The reason is that a longer time-signal has higher frequency resolution. It allows more dense sub-band segmentation, thus increasing the chance for the pulsatile component and motion component to be separated. However, it may not be suitable for instantaneous pulse-rate measurement, as a longer window is less sensitive to beat-to-beat variations. Different parameter settings leading to the final method will all be benchmarked and discussed in our experiments.

*3.5.3. Limitation.* For fair assessment, we also indicate the limitation in our SB method. We expect it to have quality drops when using the short processing window, especially in the case that the pulsatile component cannot be clearly separated from the motion components. Besides, the effect of motion suppression could be sub-optimal when the number of distortions is larger than the number of used sub-bands.

## 4. Experimental setup

This section presents the experimental setup for the benchmarking. First, we introduce the recording setup in fitness. Next, we present the evaluation metrics. Finally, two rPPG methods are adopted for comparison, i.e. one is the state-of-the-art POS and the other is the proposed SB.

### 4.1. Benchmark dataset

We create a benchmark dataset containing 25 videos (with 169 998 frames) recorded in the fitness scenario, and categorize them into individual challenges to study different use-cases. The videos are recorded using a regular RGB camera[6] in an *uncompressed* bitmap format and constant frame-rate. The ground-truth is the contact-based ECG-signal sampled by the NeXus device[7] and synchronized with the video frames. All videos are recorded from the subjects exercising on a treadmill. This study has been approved by the Internal Committee Biomedical Experiments of Philips Research, and informed consent has been obtained from each subject.

To thoroughly investigate the practical functionality of SB, we simulate various challenges in recordings by changing the experimental setup (i.e. monitoring conditions). Unless mentioned otherwise, each recording uses the following default settings: the camera is placed around 2 meters in front of the subject running on a treadmill, which, with the used optics, results in approximately $20\,000$ skin-pixels. The default subject has a skin-type III according to the Fitzpatrick scale and the face region is recorded for pulse extraction. The subject is illuminated by the office ceiling light with an illumination direction oblique to the skin-normal, which is the typical illumination condition in a fitness setting. During the recording, the subject varies the running speed between low-intensity (3 km h$^{-1}$) and high-intensity (12 km h$^{-1}$) within 5–8 min, depending on his endurance. The background is a skin-contrasting cloth to

[6] Global shutter RGB CCD camera USB UI-2230SE-C from IDS, with $640 \times 480$ pixels, 8 bit depth, and 20 frames per second (fps).
[7] The wireless physiological monitoring and feedback device. The type of the device is NeXus-10 MKII.

facilitate the skin-detection/segmentation, which we regard as an independent research challenge outside the scope of this paper.

Based on this default experiment, we vary selected parameters to study their effects (the bold number in brackets denotes the number of frames recorded for each category):

- **Skin-tone (40 951)** A total of 6 subjects (ages 25–45 yrs, 5 male and 1 female) with various skin-tones are recorded and categorized into three different skin-types based on the Fitzpatrick scale: 2 Western European subjects (skin-type I–II), 2 Eastern Asian subjects (skin-type III), and 2 Southern Asian subjects (skin-type IV–V).
- **Light source (52 085)** The type and position of the light source influence the rPPG performance. This is because different lighting spectra may result in different specular reflections on the skin surface and also different relative PPG-strength in RGB channels (de Haan and Van Leest 2014), while the position of the light source w.r.t. the skin determines the motion artifact (Moco *et al* 2016). To investigate this challenge, we use 3 light sources, i.e. oblique fluorescent light (from ceiling), frontal fluorescent light (fluo), and frontal halogen light, to create 7 different illumination conditions, which are respectively the single light (ceiling), single light (fluo), single light (halogen), double lights (ceiling + fluo), double lights (fluo + halogen), double lights (ceiling + halogen), and triple lights (ceiling + fluo + halogen).
- **Luminance level (61 089)** The luminance intensity, determining the amount of skin reflections that can be received by the camera, also affects the rPPG performance. High intensities may cause clipping on the skin surface, while low intensities may lead to low pulsatile amplitude in RGB channels while the camera sensor noise is not reduced. To vary the intensity, we adjust the camera aperture to increase/decrease the amount of light entering the camera shutter. A total of 8 intensity-levels, from level-1 (low intensity) to level-8 (high intensity), are defined to study this challenge.
- **Miscellaneous (15 873)** To improve our understanding to the studied topic of rPPG applications in fitness, we define four challenges in this category: (i) different running paces, where the subject runs at a constant speed with different paces; (ii) different running slopes, where the subject runs at a constant speed with different slopes, i.e. the gradient of the treadmill is adjusted from 0° (flat) to 15° (maximum); (iii) running-hand, where the subject's hand is recorded instead of the face during running, i.e. the raised hand introduces more erratic motion components than the face; and (iv) fake-face, where the running subject wears a skin-mask (with skin-similar color) for false positive/negative assessment.

Figure 3 exemplifies the snapshots of some recordings in our benchmark dataset. Since a skin-contrasting background is used in the setup, we apply a simple thresholding method in YCrCb space (Bousefsaf *et al* 2013) to detect and segment the skin-region across the video and save the temporal RGB traces of spatially averaged skin-pixels for processing. In this way, we ensure that the experimental results are minimally affected by non-rPPG techniques, and thus the essence of the proposed method is highlighted and the replication of the experiment is facilitated.

### *4.2. Evaluation metrics*

We use the following two metrics to evaluate the rPPG performance:

- **SNR** In line with (de Haan G and Jeanne 2013), the rPPG-signal is quantitatively measured by the signal-to-noise-ratio (SNR). Given an rPPG frequency spectrum, the SNR is calculated by the ratio between the energy around the fundamental pulse frequency components and remaining components, where the fundamental pulse frequency is determined from the reference ECG spectrum using its frequency peak within [40,240] bpm

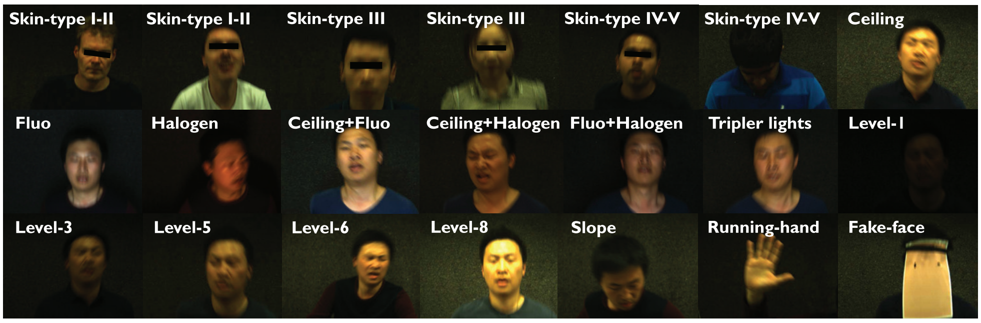

**Figure 3.** Snapshots of some recordings in the benchmark dataset, which show different challenges simulated in our recordings.

(see figure 4). Since the pulse-frequency of an exercising subject is time-varying, we use a sliding window to measure the SNR of short-term pulse spectrum within a time-interval and concatenate the subsequent SNR values into a long SNR-trace. Finally, we take the average/mean of the SNR-trace as the output metric value. More specifically, the length of the sliding window is 256 frames (corresponding to 6.4 s in 20 fps camera), and its sliding step is 1 frame per measurement.

- **ANOVA** the analysis of variance (ANOVA) is applied to compare the statistical performance (SNR) of benchmarked methods in the entire dataset. It reflects whether the difference between the compared methods is significant. In ANOVA, the p-value is used as the indicator for statistical significance and a widely used threshold 0.05 is specified as the criterion, i.e. if p-value $<$0.05, the difference is considered to be significant. Through the ANOVA test, we can clearly see whether the proposed method introduces significant improvement to the heart-rate measurement in fitness applications, as compared to the state-of-the-art method.

### *4.3. Compared methods*

The SB method proposed in this paper is an independent algorithmic component in an rPPG monitoring-system. Thus we compare it with the direct algorithmic alternative in the same system where SB is replaced, i.e. the input RGB-signals and used parameters remained the same. Since we use POS for the sub-band pulse extraction, the most straightforward comparison is between SB and POS. This essentially compares the 'sub-band POS' (i.e. the SB method introduced in this paper) with the 'full-band POS' (i.e. the POS method introduced in Wang *et al* (2016a)), which shall show the independent advantage of the sub-band strategy. Both methods have been implemented in Matlab and run on a laptop with an Intel Core i7 processor (2.70 GHz) and 8 GB RAM. The implementation of SB strictly follows algorithm 1.

According to Wang *et al* (2016a), the difference between model-based rPPG methods (CHROM, PBV and POS) is non-significant in fitness. Therefore their comparison is not repeated in this paper, although CHROM could be integrated into SB as well. Since a broad human heart-rate band (e.g. [40,240] bpm) is used in SB, we use the same heart-rate band to band-pass filter the POS-signal for fair comparison. This is to show that the actual benefit of the sub-band processing is from the dimensionality extension and not from the band-limitation. To draw solid conclusions on the comparison, we show the results obtained by

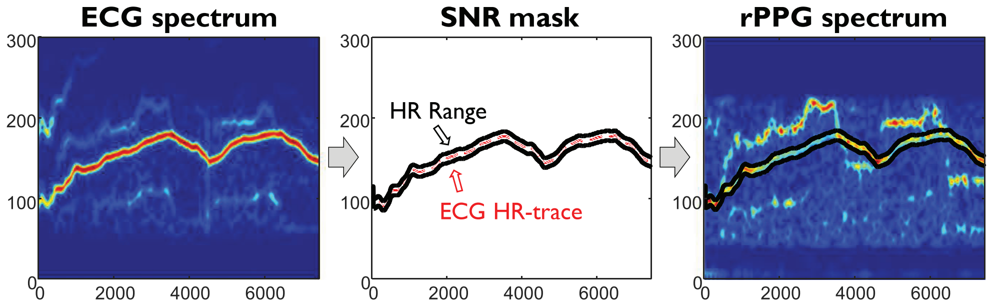

**Figure 4.** Procedure of calculating the SNR for an rPPG extraction. First, the ECG HR-trace (dashed red line) is derived from the ECG spectrum, using the maximum frequency peak within [40,240] bpm. Next, a reference HR range (solid black lines) is determined based upon the ECG HR-trace (±6 beats). In the end, the HR range is used to create a binary mask (i.e. the values inside the range are 1, otherwise 0) to measure the SNR of the rPPG spectrum.

both methods using different parameters without biasing[8]. Considering a 20 fps recording camera, we define four groups of parameters: (i) $l = 32$ (1.6 s), $\mathbf{b} = [3, 6]$, (ii) $l = 64$ (3.2 s), $\mathbf{b} = [4, 12]$, (iii) $l = 128$ (6.4 s), $\mathbf{b} = [6, 24]$, and (iv) $l = 256$ (12.8 s), $\mathbf{b} = [10, 50]$. For each method, each video has been processed four times using these settings.

## 5. Results

This section presents the experimental results of POS and SB. Table 1 lists the SNR values obtained by both methods on all benchmark videos[9] using different window lengths. Figures 5–7 exemplify the spectrograms of the rPPG-signals (and also the motion-signals in figure 6) obtained by POS and SB[10]. Figure 8 compares the average SNR between POS and SB per category per window length. Figure 9 shows the statistical comparison between POS and SB over the entire dataset as a function of window length setting.

From table 1, we can clearly see that the average SNR (last row) of POS is improved by SB for all window lengths. By increasing the window length $l$ (from 32 to 256), the average SNR for SB increases from $-1.08$ dB to 4.77 dB, whereas for POS drops from $-2.07$ dB to $-4.18$ dB. This is further confirmed by figures 5 and 6, where SB demonstrates much cleaner spectrograms. Figure 8 shows that, on average, SB outperforms POS in almost all categories with all settings, except for the ‘skin-tone’ category with $l = 32$. Figure 9 shows the increased statistical improvements of SB over POS with the increased window lengths.

## 6. Discussion

In this section, we will perform a detailed comparison and discussion on the benchmarked methods, first considering each category and then the overall dataset.

[8] The performance of SB is tested under different parameter settings, without selecting the optimal setting for this benchmark.

[9] The ‘fake-face’ challenge under the ‘miscellaneous’ category is not considered in our SNR comparison. It does not contain any living skin-pixels and thus no pulse-signal can be extracted.

[10] Due to the limited space of the paper, we only show the spectrograms obtained by POS and SB using $l = 128$, i.e. a window length offering a compromise between the robustness and latency.

**Table 1.** SNR (dB) of each method per video per setting.

| Category | Challenge | POS(32) | SB(32) | POS(64) | SB(64) | POS(128) | SB(128) | POS(256) | SB(256) |
|---|---|---|---|---|---|---|---|---|---|
| Skin-tone | Type I–II (subject 1) | **7.51** | 5.88 | **5.82** | 5.81 | 5.32 | **7.49** | 5.24 | **8.71** |
| | Type I–II (subject 2) | **−5.53** | −5.59 | **−5.32** | −6.52 | −5.40 | **−3.75** | −5.73 | **−2.42** |
| | Type III (subject 3) | 1.09 | **2.69** | 0.86 | **6.33** | 0.53 | **9.42** | 0.23 | **11.20** |
| | Type III (subject 4) | 0.41 | **0.49** | −0.14 | **2.87** | −0.74 | **5.27** | −1.18 | **8.39** |
| | Type IV–V (subject 5) | **1.20** | 0.23 | −0.02 | **2.73** | −0.85 | **6.65** | −1.35 | **9.27** |
| | Type IV–V (subject 6) | **−6.14** | −7.59 | **−7.40** | −8.93 | **−8.01** | −10.01 | **−8.79** | −10.32 |
| Light source | Ceiling | −1.50 | **0.94** | −2.58 | **2.12** | −3.47 | **5.86** | −3.97 | **7.46** |
| | Fluo | 2.77 | **4.44** | 1.75 | **5.70** | 1.13 | **8.48** | 0.79 | **10.68** |
| | Halogen | −0.45 | **0.62** | −1.63 | **2.91** | −2.59 | **5.52** | −3.31 | **7.26** |
| | Ceiling + Fluo | −1.16 | **0.31** | −2.87 | **2.27** | −3.87 | **4.47** | −4.26 | **6.45** |
| | Ceiling + Halogen | −4.13 | **−2.80** | −5.07 | **−0.45** | −5.79 | **2.89** | −6.42 | **4.94** |
| | Fluo + Halogen | −1.03 | **0.00** | −2.10 | **2.08** | −2.88 | **4.38** | −3.71 | **6.86** |
| | Ceiling + Fluo + Halogen | **−1.42** | −1.43 | −2.47 | **−0.71** | −3.47 | **1.02** | −4.42 | **1.96** |
| Luminance level | Luminance level-1 | −8.15 | **−6.76** | −8.69 | **−4.82** | −8.98 | **−2.87** | −9.26 | **−0.78** |
| | Luminance level-2 | −4.67 | **−3.07** | −4.55 | **−0.42** | −4.98 | **3.23** | −5.32 | **6.04** |
| | Luminance level-3 | −1.26 | **0.82** | −2.53 | **3.90** | −3.09 | **7.61** | −3.51 | **9.87** |
| | Luminance level-4 | −2.43 | **−0.40** | −3.13 | **1.60** | −3.54 | **5.18** | −4.28 | **7.50** |
| | Luminance level-5 | −2.79 | **−1.15** | −3.25 | **0.78** | −3.62 | **3.78** | −3.77 | **7.11** |
| | Luminance level-6 | −3.52 | **−0.40** | −4.50 | **1.71** | −4.95 | **4.64** | −5.56 | **5.90** |
| | Luminance level-7 | −4.07 | **−2.77** | −5.61 | **−2.39** | −6.55 | **−0.69** | −7.15 | **1.22** |
| | Luminance level-8 | −5.99 | **−5.33** | −7.01 | **−5.21** | −7.76 | **−4.77** | −8.34 | **−4.18** |
| Miscellaneous | Pace | −4.22 | **−2.43** | −4.95 | **−1.54** | −5.25 | **2.91** | −5.78 | **4.54** |
| | Slope | −2.55 | **−1.33** | −3.71 | **−0.39** | −4.54 | **2.05** | −4.84 | **4.10** |
| | Running-hand | −1.76 | **−1.17** | −3.38 | **−0.73** | −4.74 | **1.21** | −5.72 | **2.81** |
| | Fake-face | — | — | — | — | — | — | — | — |
| Overall | Average | −2.07 | **−1.08** | −3.02 | **0.36** | −3.67 | **2.92** | −4.18 | **4.77** |

*Note:* Bold entry denotes the best method compared per challenge per setting. The number in brackets denotes the window length used for this method, i.e. POS(32) means the POS method with $l = 32$.

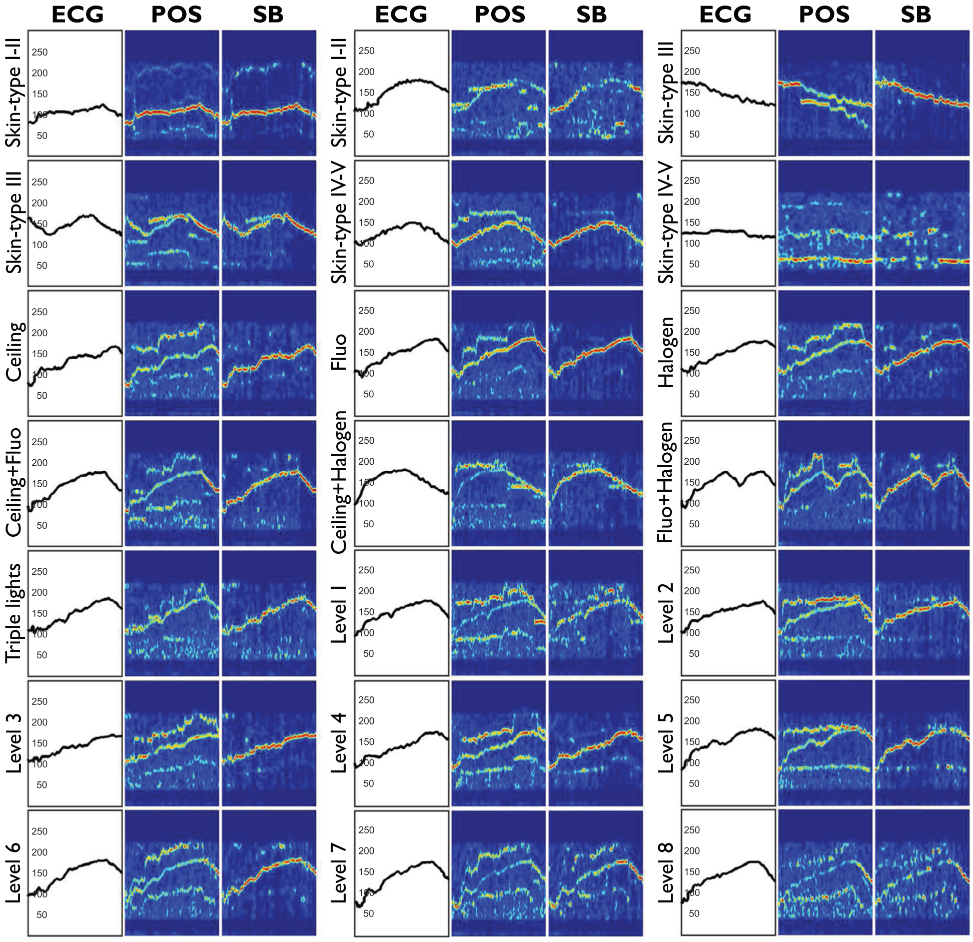

**Figure 5.** The ECG reference signal and spectrograms of the pulse-signals of POS and SB obtained on videos in the categories of 'skin-tone', 'light source' and 'luminance level'. The *x*-axis and *y*-axis denote the frame number and frequency, respectively.

### *6.1. Discussion per category*

- **Skin-tone category** Since SB is not designed for addressing the low pulsatility problem of dark skin, we expect that its improvement in various skin-types is mainly from the motion suppression. All recordings are performed on a treadmill, so the skin-tone challenge is accompanied with the motion challenge in this category. For example, SB obtains much higher SNR in subject 5 (skin-type IV-V) than subject 2 (skin-type I-II). This is not because the dark skin is easier for SB to extract the pulse than the bright skin, but the running motion of subject 2 is much more vigorous than that of subject 5. We notice that both methods obtain high SNR ($>$5 dB) on subject 1, who is jogging during the largest part of the recording (i.e. the speed is around 7 km h$^{-1}$) and therefore this subject shows less significant body motions. In contrast, both methods obtain rather low SNR ($<-6$ dB) on subject 6. This is not only due to the dark skin, but also the limited number of skin-pixels, i.e. subject 6 bows his face during the largest part of the running (see figure 3).

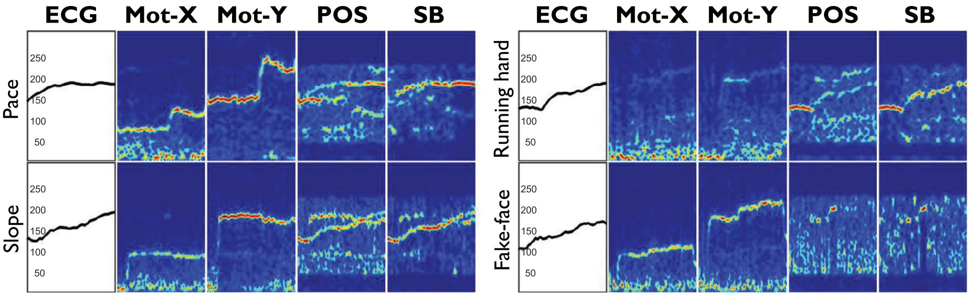

**Figure 6.** The ECG reference signal and the spectrograms of the horizontal motion (motion-x, denoted as 'Mot-X'), vertical motion (motion-y, denoted as 'Mot-Y'), and pulse-signals of POS and SB obtained from the 'miscellaneous' category. The signals of Mot-X and Mot-Y are measured using the center of gravity of segmented skin-regions across the video.

- **Light source category** As compared to POS, SB shows clearly improved robustness in various light source conditions. Both methods perform better under a single light source than under multiple light sources. The reason is: when the subject is moving under the multiple light sources, the motion-induced specular variations have different color vectors due to different lighting spectra. In such cases, SB enabling the multi-dimensional noise suppression is expected to attain most gain relative to the existing methods.
  Also as we expected, both methods achieve their best performance under the frontal fluorescent light source. The 'Fluo' challenge in figure 5 shows that the high frequency component (i.e. the specular distortion) corresponding to the vertical body motion is much less significant than the ones in the ceiling light conditions. This observation has been studied and verified in (Moco *et al* 2016): the position of the light source has a large impact on the motion distortion, and thus the rPPG performance. The frontal, diffuse and homogeneous illumination is suggested for an rPPG setup to improve the rPPG-signal quality (Moco *et al* 2016), which also holds for our fitness setup. Additionally, both methods (especially POS) have better performance under the frontal fluorescent lamp than the frontal halogen lamp. This could due to the different spectra of two light sources, i.e. the blue channel is much weaker in the halogen illumination.
- **Luminance level category** SB demonstrates a clear advantage in motion suppression over the tested light intensity range. Figure 5 shows that both methods obtain better results at mid-level intensities (e.g. level 3–6) than low-level intensities (e.g. level 1–2) and high-level intensities (e.g. level 7–8). The reasons are the following. (i) The skin-pixels at low-level intensities have much lower pulsatile amplitudes in RGB channels, i.e. the skin pulsatility is proportional/multiplicative to the intensity. However, the camera sensor noise is not reduced at lower intensity levels. These together lead to a low signal-to-noise ratio of the measured rPPG-signal. (ii) The skin-pixels at high-level intensities may contain clipping (see the second snapshot starting from the left side in the last row of figure 3), which obviously constitutes a major deviation from the signal model (10). In contrast, the mid-level intensities, providing sufficient amount of intensity energies and also avoiding the clipping, is recommended for the setup.
- **Miscellaneous category** SB also demonstrates improved motion robustness over POS in this category, except the 'fake-face' challenge that does not contain a living skin-pixel. When the subject is running at constant speed (e.g. 9 km $h^{-1}$) but varying the pace or the treadmill slope, SB is consistently better than POS for all the window lengths. Even when

**Table 2.** ANOVA test between POS and SB over the entire dataset per window length setting.

| Comparison | POS & SB (32) | POS & SB (64) | POS & SB (128) | POS & SB (256) |
|---|---|---|---|---|
| *p*-value | 0.2937 | **0.0018** | $\mathbf{6.15 \times 10^{-7}}$ | $\mathbf{3.90 \times 10^{-9}}$ |

*Note*: Smaller *p*-values suggest larger differences between POS and SB. If p-value <0.05 (denoted by the **bold entry**), the difference between POS and SB is considered to be significant.

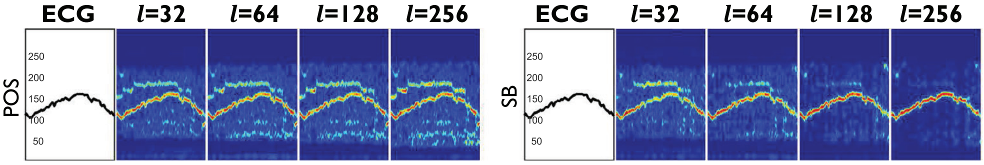

**Figure 7.** The spectrograms of the pulse-signals of POS and SB obtained on the video of skin-type IV-V (subject 5) using different window lengths. By increasing the window length, the SNR of SB is significantly increased from 0.23 dB to 9.27 dB, whereas the SNR of POS decreases from 1.20 dB to −1.35 dB.

the hand is recorded during running, SB can still provide a reasonable pulse-signal as compared to POS.

Figure 6 shows more insights of the fitness application: (i) when the subject is running at the constant speed but doubling the pace frequency, the motion frequencies are increased accordingly but the pulse frequency remains relatively constant. This can be explained by the law of conservation of energy: the constant input energy/speed of the treadmill does not change the total energy consumption (and therefore the pulse-rate) of the subject; (ii) when the subject is running at the same constant speed but the slope of the treadmill is increased (from 0° to 15°), the pulse frequency is increased whereas the motion frequencies remain stable. This is because the elevated treadmill slope introduces the gravitational potential energy to the subject. Thus the subject needs to consume more energy to stay on the treadmill, i.e. the pulse-rate will be raised; (iii) when the subject's hand is recorded, the hand movement introduces erratic and irregular motion components to RGB-signals, which makes the pulse extraction more challenging. We also notice that the vertical motion frequency of the hand is no longer twice higher than its horizontal motion frequency (see figure 6). The reason is that the raised hand has more space to move in the air and its moving trace is more arbitrary as compared to that of the head. Besides, a hand has less skin-pixels than a face, thus larger quantization noise that makes the pulse extraction even more difficult; and (iv) when the subject is wearing a (skin-color similar) mask during running (see figure 3), both methods cannot measure a pulse-signal as there is no real skin-pixels. This experiment is to confirm that when the living skin-tissue is absent in a video, no false pulse-signal is created.

### 6.2. Overall discussion

Figure 9 shows the statistical comparison between POS and SB per window setting: the median SNR of SB is higher than that of POS for all window lengths, although the amount of improvement of SB is much more obvious for the longer windows. To verify this, we perform the ANOVA test between POS and SB over the entire dataset per window setting. Table 2 shows that the improvement of SB over POS is statistically significant with longer

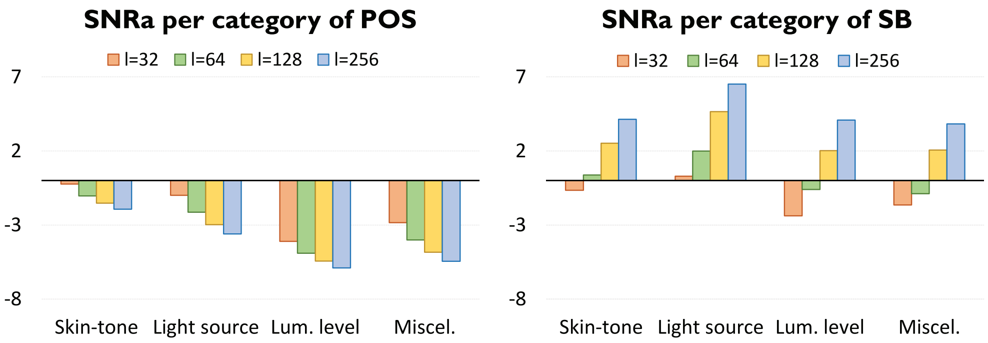

**Figure 8.** The average SNR (SNRa) of POS and SB per challenge category.

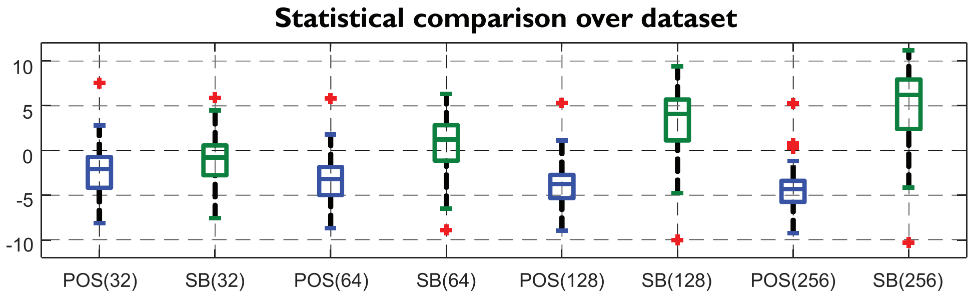

**Figure 9.** Statistical comparison of the SNR values obtained by POS and SB as a function of window length setting. The median values are indicated by horizontal bars inside boxes, the quartile range by boxes (POS by blue, SB by green), the full range by whiskers, disregarding the outliers (red crosses).

window lengths (e.g. l = 64 128 256), as the *p*-values for these settings are smaller than 0.05. Moreover, we show a qualitative spectrogram comparison in figure 7. The pulse spectrum of POS does not vary much when changing $l$, i.e. it is a bit cleaner at $l = 32$. In contrast, the pulse spectrum of SB becomes much cleaner when increasing $l$, and such improvement is monotonic, i.e. it obtains the cleanest spectrum at $l = 256$.

The proceeding observation is in line with our expectation: (i) SB performs better with the longer window, as the longer window provides higher frequency resolution that allows dense sub-band segmentation. The chance that the pulsatile component can be separated from the motion component is increased, i.e. SB has only 4 sub-bands at $l = 32$, but 41 sub-bands at $l = 256$; (ii) POS performs slightly better with the shorter window, as it can quickly adapt the alpha-tuning (Wang *et al* 2016a) to suppress the instant motion distortions. Note that even in the worst case of SB (e.g. $l = 32$), it is still better than POS in the overall comparison, although the improvement is not as large as that obtained with the longer windows. Meanwhile, we have to recognize that increasing the window length means increasing the processing latency. The purpose of benchmarking with different parameters is to provide readers a full-view on SB under different settings. Based on our benchmark results, ones can choose their preferred settings in a specific application scenario. Considering both the robustness and latency in a fitness setup, we recommend $l = 128$ for SB, in case of a 20 fps camera.

Notwithstanding the improvements and overall good results, there are still certain limitations to the usage of rPPG in fitness scenarios. Since our SB method uses the sub-band decomposition to separate and suppress motion frequencies, it cannot deal with the case that motion has exactly the same frequency as the pulse. This is an inherent restriction of the method originated from design, irrespective of the used sliding window length. Next to that, some other physical restrictions may preclude the heart-rate extraction. Based on our experiments, we find that the extremely low light intensity conditions or extremely dark skins (or the skin with coverage like make-up) will be very tough[11], as no (or only a limited amount of) light can penetrate deep into the skin and is reflected and received by the camera. Although a near infrared (NIR) camera could be used in these cases to bypass the problem, the PPG-absorption (and thus the skin pulsatility) in NIR wavelengths is much lower than that in visible wavelengths, especially the pulsatility in the G-channel is much higher than that in NIR channels. Thus the robustness of applying the NIR-camera in a fitness setup is still questionable.

## 7. Conclusion

In this paper, we improve the robustness of rPPG in fitness applications that measures continuous heart-rate. We analyze the fundamental limitation of the existing rPPG methods and propose a novel method to overcome it. Our strategy is using the sub-band decomposition to extend the degrees-of-freedom of noise reduction. This process, namely Sub-band rPPG (SB), enables the independent suppression of multiple motion-frequencies. The basic form of SB is benchmarked against a state-of-the-art method (POS) on a challenging fitness video dataset using non-biased parameter settings. The results clearly show that SB outperforms POS in all-round statistical comparisons, and in particular shows significant improvements at longer sliding window lengths.

## Acknowledgments

The authors would like to thank Dr Ihor Kirenko at Philips Research for his support, and also the volunteers from Eindhoven University of Technology and Philips Research for their efforts in creating the benchmark dataset.

## References

Allen J 2007 Photoplethysmography and its application in clinical physiological measurement *Physiol. Meas.* **28** R1

Bousefsaf F *et al* 2013 Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate *Biomed. Signal Process. Control* **8** 568–74

Blackford E B *et al* 2016 Measuring pulse rate variability using long-range, non-contact imaging photoplethysmography *Proc. IEEE Conf. Engineering in Medicine and Biology Society* (Orlando, FL, USA) pp 3930–6

Couderc J-P *et al* 2015 Detection of atrial fibrillation using contactless facial video monitoring *Heart Rhythm* **12** 195–201

de Haan G and Jeanne V 2013 Robust pulse rate from chrominance-based rPPG *IEEE Trans. Biomed. Eng.* **60** 2878–86

de Haan G and Van Leest A 2014 Improved motion robustness of remote-PPG by using the blood volume pulse signature *Physiol. Meas.* **35** 1913–22

[11] Other challenges, such as the high light intensity (causing image clipping), non-white illumination spectra, and body motions, can be addressed by either tuning the camera settings (e.g. aperture or gain) or designing better algorithms (e.g. motion tracking).

Fernando S *et al* 2015 Feasibility of contactless pulse rate monitoring of neonates using google glass *Proc. EAI Conf. Wireless Mobile Communication Healthcare (London, UK)* pp 198–201

Gibert G *et al* 2013 Face detection method based on photoplethysmography *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance* (Krakow, Poland) pp 449–53

Guazzi A R *et al* 2015 Non-contact measurement of oxygen saturation with an RGB camera *Biomed. Opt. Express* **6** 3320–38

Hülsbusch M 2008 An image-based functional method for opto-electronic detection of skin perfusion *PhD Dissertation* Dept. Elect. Eng., RWTH Aachen Univ., Aachen, Germany (in German)

Jeong I C and Finkelstein J 2016 Introducing contactless blood pressure assessment using a high speed video camera *J. Med. Syst.* **40** 1–10

Kumar M *et al* 2015 DistancePPG: robust non-contact vital signs monitoring using a camera *Biomed. Opt. Express* **6** 1565–88

Li X *et al* 2014 Remote heart rate measurement from face videos under realistic situations *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (Columbus, OH, USA)* pp 4264–71

Liu S *et al* 2016 3D Mask face anti-spoofing with remote photoplethysmography *European Conf. Computer Vision (Amsterdam, Netherlands)* pp 85–100

Lewandowska M *et al* 2011 Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity in *Proc. Federated Conf. Computer Science and Information Systems (Szczecin, Poland)* pp 405–10

McDuff D J *et al* 2015 A survey of remote optical photoplethysmographic imaging methods *Proc. IEEE Conf. Engineering in Medicine and Biology Society (Milan, Italy)* pp 6398–404

McDuff D *et al* 2014a Remote measurement of cognitive stress via heart rate variability *Proc. IEEE Conf. Engineering in Medicine and Biology Society (Chicago, IL, USA)* pp 2957–60

McDuff D *et al* 2014b Improvements in remote cardiopulmonary measurement using a five band digital camera *IEEE Trans. Biomed. Eng.* **61** 2593–601

Mestha L K *et al* 2014 Towards continuous monitoring of pulse rate in neonatal intensive care unit with a webcam *Proc. IEEE Conf. Engineering in Medicine and Biology Society (Chicago, IL, USA)* pp 3817–20

Moco A V *et al* 2016 Ballistocardiographic artifacts in PPG imaging *IEEE Trans. Biomed. Eng.* **63** 1804–11

Poh M-Z *et al* 2011 Advancements in noncontact, multiparameter physiological measurements using a webcam *IEEE Trans. Biomed. Eng.* **58** 7–11

Rouast P V *et al* 2016 Remote heart rate measurement using low-cost RGB face video: a technical literature review *Front. Comput. Sci.* (https://doi.org/10.1007/s11704-016-6243-6)

Shao D *et al* 2014 Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time *IEEE Trans. Biomed. Eng.* **61** 2760–7

Sikdar A *et al* 2016 Computer-vision-guided human pulse rate estimation: a review *IEEE Rev. Biomed. Eng.* **9** 91–105

Sun Y and Thakor N 2016 Photoplethysmography revisited: from contact to noncontact, from point to imaging *IEEE Trans. Biomed. Eng.* **63** 463–77

Takano C and Ohta Y 2007 Heart rate measurement based on a time-lapse image *Med. Eng. Phys.* **29** 853–7

Tarassenko L *et al* 2014 Non-contact video-based vital sign monitoring using ambient light and auto-regressive models *Physiol. Meas.* **35** 807

Temko A 2017 Accurate wearable heart rate monitoring during physical exercises using PPG *IEEE Trans. Biomed. Eng.* 1

Tsouri G R *et al* 2012 Constrained independent component analysis approach to nonobtrusive pulse rate measurements *J. Biomed. Opt.* **17** 077011

Tsouri G R and Li Z 2015 On the benefits of alternative color spaces for noncontact heart rate measurements using standard red-green-blue cameras *J. Biomed. Opt.* **20** 048002

Tulyakov S *et al* 2016 Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (Las Vegas, NV, USA)* pp 2396–404

Verkruysse W *et al* 2008 Remote plethysmographic imaging using ambient light *Opt. Express* **16** 21434–45

Wang W *et al* 2015a Exploiting spatial redundancy of image sensor for motion robust rPPG *IEEE Trans. Biomed. Eng.* **62** 415–25

Wang W *et al* 2015b Unsupervised subject detection via remote PPG *IEEE Trans. Biomed. Eng.* **62** 2629–37

Wang W *et al* 2016a Algorithmic principles of remote-PPG *IEEE Trans. Biomed. Eng.* (https://doi.org/10.1109/TBME.2016.2609282)

Wang W *et al* 2016b A novel algorithm for remote photoplethysmography: spatial subspace rotation *IEEE Trans. Biomed. Eng.* **63** 1974–84

Wang W *et al* 2016c Quality metric for camera-based pulse rate monitoring in fitness exercise *Proc. IEEE Int. Conf. Image Processing (Phoenix, AZ, USA)* pp 2430–4

Yang Y *et al* 2016 Motion robust remote photoplethysmography in CIELab color space *J. Biomed. Opt.* **21** 117001

Zhang Z *et al* 2015 A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise *IEEE Trans. Biomed. Eng.* **62** 522–31

Zhang Z 2015 Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction *IEEE Trans. Biomed. Eng.* **62** 1902–10