Algorithmic Principles of Remote-PPG

Wenjin Wang, Bert den Brinker, Sander Stuijk, and Gerard de Haan

***Abstract*—This paper introduces a mathematical model that incorporates the pertinent optical and physiological properties of skin reflections with the objective to increase our understanding of the algorithmic principles behind remote photoplethysmography (rPPG). The model is used to explain the different choices that were made in existing rPPG methods for pulse extraction. The understanding that comes from the model can be used to design robust or application-specific rPPG solutions. We illustrate this by designing an alternative rPPG method where a projection plane orthogonal to the skin-tone is used for pulse extraction. A large benchmark on the various discussed rPPG methods shows that their relative merits can indeed be understood from the proposed model.**

***Index Terms*—Biomedical monitoring, photoplethysmography, remote sensing, colors.**

I. Introduction

REMOTE photoplethysmography (rPPG) enables contactless monitoring of human cardiac activities by detecting the pulse-induced subtle color variations on human skin surface using a multi-wavelength RGB camera [1]. In recent years, several core rPPG methods have been proposed for extracting the pulse-signal from a video. These include: (i) *Blind Source Separation* (e.g., PCA-based [2] and ICA-based [3]), which use different criteria to separate temporal RGB traces into uncorrelated or independent signal sources to retrieve the pulse; (ii) *CHROM* [4], which linearly combines the chrominance-signals by assuming a standardized skin-color to white-balance the images; (iii) *PBV* [5], which uses the signature of blood volume changes in different wavelengths to explicitly distinguish the pulse-induced color changes from motion noise in RGB measurements; and (iv) *2SR* [6], which measures the temporal rotation of the spatial subspace of skin-pixels for pulse extraction. The essential difference between these rPPG methods is in the way of combining RGB-signals into a pulse-signal. A better understanding of the core rPPG methods could benefit many systems/applications for video health monitoring, such as the monitoring of heart-rate [7]–[11], respiration [8], SpO_2 [8], [12], blood pressure [13], neonates [14], [15], and the detection of atrial fibrillation [16] and mental stress [17].

In this paper, we investigate the algorithmic principles of rPPG in a mathematical context with optical and physiological reasoning. Our exploration based on the skin reflection model shows that different characteristic properties of rPPG can be used to design algorithmic solutions for pulse extraction. As such, this study not only gives an integral view on and insight into the core rPPG methods [2]–[6], but also leads to a new alternative that demonstrates tractable algorithm development based on understanding. The new method defines a plane orthogonal to the skin-tone in the temporally normalized RGB space for pulse extraction, and is therefore referred to as the "Plane-Orthogonal-to-Skin" (POS). The main contribution of this paper is its in-depth analysis to the working principles of rPPG (in a mathematical context), which benefits the development of novel rPPG methods in the future, as demonstrated by the POS algorithm introduced in this paper.

The remainder of this paper is structured as follows. In Section II, we present a skin reflection model. In Section III and IV, we analyze the existing rPPG methods in the model and describe a new method to demonstrate our understanding. In Section V and VI, we use a benchmark to verify our analysis. Finally in Section VII, we draw the conclusions.

II. Skin reflection model

Unless stated otherwise, we use the following mathematical conventions throughout the paper. Vectors and matrices are denoted as boldface characters, where the column vectors are denoted as $\mathbf{v}$, except for the ones with unit-length which are denoted as $\mathbf{u}$. The variable t denotes the time; $\top$ denotes the transposition; $\mathbb{E}\{\cdot\}$ denotes the expectation operator; and the vector $\mathbf{1}$ denotes $(1,1,1)^\top$.

To thoroughly understand the principles for pulse extraction in rPPG methods, we start from the basics by defining an rPPG model that considers the pertinent optical and physiological properties of skin reflections. This model allows us to analyze the problems in detail and point out how these problems are addressed in various rPPG methods.

Consider a light source illuminating a piece of human skin-tissue containing pulsatile blood and a remote color camera recording this image, as illustrated in Fig. 1. We further assume that the light source has a constant spectral composition but varying intensity, and the intensity observed at the camera depends on the distance from the light source to the skin tissue and to the camera sensor. The skin measured by the camera has a certain color[1] that varies over time, due to the motion-induced intensity/specular variations and pulse-induced subtle color changes. These temporal changes are proportional to the luminance intensity level.

W. Wang and S. Stuijk are with the Electronic Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, e-mail: (W.Wang@tue.nl, S.Stuijk@tue.nl).

B. den Brinker and G. de Haan are with the Philips Innovation Group, Philips Research, Eindhoven, The Netherlands, e-mail: (Bert.den.Brinker@philips.com, G.de.Haan@philips.com).

[1]The skin color measured by the camera is a combination of the light source (e.g., intensity and spectrum), intrinsic skin color, and sensitivities of color channels of the camera.

Based on the dichromatic model, the reflection of each skin-pixel in a recorded image sequence can be defined as a time-varying function in RGB channels:

$$\mathbf{C_k}(t) = I(t) \cdot \big(\mathbf{v_s}(t) + \mathbf{v_d}(t)\big) + \mathbf{v_n}(t), \quad (1)$$

where $\mathbf{C_k}(t)$ denotes the RGB channels (i.e., ordered in column) of the k-th skin-pixel; $I(t)$ denotes the luminance intensity level, which absorbs the intensity changes due to the light source as well as to the distance changes between the light source, skin tissue and camera; $I(t)$ is modulated by two components in the dichromatic model: specular reflection $\mathbf{v_s}(t)$ and diffuse reflection $\mathbf{v_d}(t)$. The time dependency is due to the body motion and pulsatile blood; the last component $\mathbf{v_n}(t)$ denotes the quantization noise of the camera sensor.

The specular reflection is a mirror-like light reflection from the skin surface, which does not contain any pulsatile information. As such, its spectral composition is equivalent to that of the light source. It is time-dependent in the sense that body-motion will influence the geometric structure between the light source, skin surface and camera. We write $\mathbf{v_s}(t)$ as:

$$\mathbf{v_s}(t) = \mathbf{u_s} \cdot \big(s_0 + s(t)\big), \quad (2)$$

where $\mathbf{u_s}$ denotes the unit color vector of the light spectrum; s_0 and $s(t)$ denote the stationary and varying parts of specular reflections, more specifically, $s(t)$ is induced by motion.

The diffuse reflection is associated with the absorption and scattering of the light in skin-tissues. The hemoglobin and melanin contents in skin-tissues lead to a specific chromaticity for $\mathbf{v_d}$. Meanwhile, $\mathbf{v_d}$ is varied by blood volume changes and is thus time-dependent. We write $\mathbf{v_d}(t)$ as:

$$\mathbf{v_d}(t) = \mathbf{u_d} \cdot d_0 + \mathbf{u_p} \cdot p(t), \quad (3)$$

where $\mathbf{u_d}$ denotes the unit color vector of the skin-tissue; d_0 denotes the stationary reflection strength; $\mathbf{u_p}$ denotes the relative pulsatile strengths in RGB channels; $p(t)$ denotes the pulse-signal. Substituting (2) and (3) into (1), we arrive at:

$$\mathbf{C_k}(t) = I(t) \cdot \Big(\mathbf{u_s} \cdot \big(s_0 + s(t)\big) + \mathbf{u_d} \cdot d_0 + \mathbf{u_p} \cdot p(t)\Big) + \mathbf{v_n}(t). \quad (4)$$

The stationary parts in specular and diffuse reflections can be combined into a single component representing the stationary skin reflection:

$$\mathbf{u_c} \cdot c_0 = \mathbf{u_s} \cdot s_0 + \mathbf{u_d} \cdot d_0, \quad (5)$$

where $\mathbf{u_c}$ denotes the unit color vector of the skin reflection and c_0 denotes the reflection strength. Thus (4) is rewritten as:

$$\mathbf{C_k}(t) = I_0 \cdot \big(1 + i(t)\big) \cdot \big(\mathbf{u_c} \cdot c_0 + \mathbf{u_s} \cdot s(t) + \mathbf{u_p} \cdot p(t)\big) + \mathbf{v_n}(t), \quad (6)$$

where $I(t)$ is also expressed as the combination of a stationary part I_0 and a time-varying part $I_0 \cdot i(t)$, i.e., the (motion-induced) intensity variation strength observed by the camera is proportional to the intensity level; $i(t)$, $s(t)$ and $p(t)$ are zero-mean signals. Note that the specular reflection can be the largest component by far, overshadowing all other components. We assume there are means (e.g., a skin classifier) to reject areas where the specular reflection is dominant. Therefore, we only consider the pixels k where $\mathbf{u_d}$ is to a non-negligible degree determined by the diffuse reflection. In terms of the model (6), the task of an rPPG method is clear: extracting $p(t)$ from $\mathbf{C_k}(t)$.

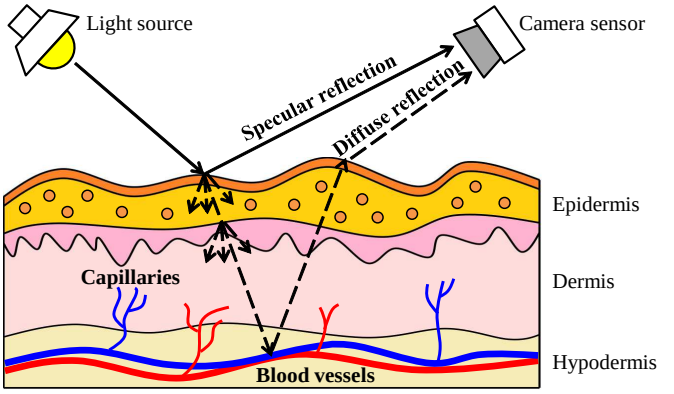

Fig. 1. The skin reflection model that contains specular and diffuse reflections, where only the diffuse reflection contains pulsatile information.

III. Existing rPPG methods

In this section, we review the existing rPPG methods using the model defined in Section II and analyze their strengths and weaknesses as a function of pulse extraction.

Existing rPPG methods [2]–[5] (except 2SR [6]) use the spatially averaged RGB values of skin-pixels to generate temporal RGB-signals for pulse extraction. The spatial pixel averaging step can reduce the camera quantization error. Based on (6), we assume that a sufficient amount of pixels (i.e., sensor arrays) are focused on comparable skin-tissues, and average $\mathbf{C_k}$ over the observed skin-pixels as:

$$\mathbf{C}(t) \approx I_0 \cdot \big(1 + i(t)\big) \cdot \big(\mathbf{u_c} \cdot c_0 + \mathbf{u_s} \cdot s(t) + \mathbf{u_p} \cdot p(t)\big), \quad (7)$$

which provides a $\mathbf{C}(t)$ where the quantization noise $\mathbf{v_n}(t)$ is negligible when the number of skin-pixels is sufficiently large. However, we note that when this step is performed on a small skin patch/area with a limited number of pixels, the camera quantization noise remains large and is thus non-negligible. We also note that this step assumes that various color vectors are not dependent on the skin-pixel positions in an image. The obtained $\mathbf{C}(t)$ is essentially the spatial RGB mean at time t. (7) can be further expanded and simplified to:

$$\begin{aligned}\mathbf{C}(t) =& \mathbf{u_c} \cdot I_0 \cdot c_0 + \mathbf{u_s} \cdot I_0 \cdot s(t) + \mathbf{u_p} \cdot I_0 \cdot p(t) + \\ & \mathbf{u_c} \cdot I_0 \cdot c_0 \cdot i(t) + \mathbf{u_s} \cdot I_0 \cdot s(t) \cdot i(t) + \\ & \mathbf{u_p} \cdot I_0 \cdot p(t) \cdot i(t) \\ \approx & \mathbf{u_c} \cdot I_0 \cdot c_0 + \mathbf{u_c} \cdot I_0 \cdot c_0 \cdot i(t) + \mathbf{u_s} \cdot I_0 \cdot s(t) + \\ & \mathbf{u_p} \cdot I_0 \cdot p(t)\end{aligned} \quad (8)$$

where the approximation holds because all AC-modulation terms are much smaller (i.e., orders of magnitude) than the DC term and thus the product modulation terms (e.g., $p(t) \cdot i(t)$) can be neglected. The approximation (8) shows that the observation $\mathbf{C}(t)$ is a linear mixture of three source-signals $i(t)$, $s(t)$ and $p(t)$. This implies that by using the linear projection, we are able to separate these source-signals. Thus the task of extracting the pulse-signal from the observed RGB-signals can be translated into defining a projection-system to decompose $\mathbf{C}(t)$.

A. BSS-based methods (PCA/ICA)

The approximation in (8) suggests that Blind Source Separation (BSS) techniques might be ideal candidates for de-mixing $\mathbf{C}(t)$ into different sources for pulse retrieval. The general procedure of BSS-based rPPG methods can be expressed as:

$$\mathbf{Y}(t) = \mathbf{W} \cdot \mathbf{C}(t), \tag{9}$$

where $\mathbf{Y}(t)$ denotes the factorized source-signals consisting of the pulse and noise; $\mathbf{W}$ denotes the de-mixing matrix that can either be estimated by PCA [2] or ICA [3], i.e., the sorting problem in ICA was further solved by a constrained-ICA based approach introduced by [18]. The essential difference between PCA and ICA is their assumptions w.r.t. the relationship between $i(t)$, $s(t)$ and $p(t)$, i.e., the source-signals are either uncorrelated or independent. The BSS operation is followed by selecting the most periodic signal from $\mathbf{Y}(t)$ as the pulse. As a consequence, these methods cannot deal with the cases in which the motion is also periodic, as typically occurs when the subject is exercising in a fitness setting.

Moreover, PCA and ICA have different limitations when estimating $\mathbf{W}$: (i) PCA uses the covariance of RGB-signals to estimate $\mathbf{W}$ (i.e., eigenvectors), which requires the variation in the amplitude of pulse and noise to be sufficiently different to determine the eigenvector directions; (ii) ICA assumes that the components in $\mathbf{Y}(t)$ are statistically independent and non-Gaussian for deriving $\mathbf{W}$, which requires $\mathbf{C}(t)$ to be a long signal to enable a statistical measurement. It may make the separation even harder, since different frequency components (e.g., respiration and Mayer-wave) may be included as well. Furthermore, the procedure of BSS in estimating an exact $\mathbf{W}$ is completely blind (i.e., a black box), which is not tractable for algorithm development.

Most importantly, BSS techniques are statistical and computational solutions for general signal-processing problems, which do not exploit the unique and characteristic skin reflection properties that can be used to solve the rPPG-specific problem. Especially illustrative in this respect is the ICA-based approach which normalizes the standard deviation (i.e., AC-components) of RGB-signals upfront thus ignoring the fact that the PPG-signal induces different yet known relative amplitudes in the individual RGB channels.

B. Model-based methods (PBV/CHROM)

In contrast to the BSS-based methods that impose no assumption on the colors associated with the source-signals, the model-based methods [4], [5] use knowledge of the color vectors of the different components to control the de-mixing. Therefore, these methods have one step in common: eliminating the dependency of $\mathbf{C}(t)$ on the average skin reflection color (i.e., DC-level), including the light source color and intrinsic skin color. This can be done by the temporal normalization[2]: dividing RGB-signals by their temporal mean, which does not harm the AC components. In (8), the temporal mean is considered as the large steady component over a time interval:

$$\overline{\mathbf{C}(t)} \approx \mathbf{u_c} \cdot I_0 \cdot c_0, \tag{10}$$

[2]An alternative to the temporal normalization is to take the logarithm, which for small variations as the PPG-signal has practically the same effect.

which is used to uniquely define a (diagonal) normalization matrix $\mathbf{N}$:

$$\mathbf{N} \cdot \overline{\mathbf{C}(t)} = \mathbf{N} \cdot \mathbf{u_c} \cdot I_0 \cdot c_0 = \mathbf{1}, \tag{11}$$

where $\mathbf{N}$ is used to temporally normalize $\mathbf{C}(t)$ as:

$$\begin{aligned}\mathbf{C_n}(t) =& \mathbf{N} \cdot \mathbf{C}(t) \\ =& \mathbf{N} \cdot \mathbf{u_c} \cdot I_0 \cdot c_0 + \mathbf{N} \cdot \mathbf{u_c} \cdot I_0 \cdot c_0 \cdot i(t) + \\ & \mathbf{N} \cdot \mathbf{u_s} \cdot I_0 \cdot s(t) + \mathbf{N} \cdot \mathbf{u_p} \cdot I_0 \cdot p(t) \\ =& \overbrace{\mathbf{1} \cdot \big(1 + i(t)\big)}^{\text{Intensity}} + \overbrace{\mathbf{N} \cdot \mathbf{u_s} \cdot I_0 \cdot s(t)}^{\text{Specular}} + \overbrace{\mathbf{N} \cdot \mathbf{u_p} \cdot I_0 \cdot p(t)}^{\text{Pulse}}.\end{aligned} \tag{12}$$

There are a number of qualitative observations w.r.t. (12):

- **Intensity**: $\mathbf{1} \cdot (1 + i(t))$ denotes the light intensity variations along the direction of $\mathbf{1}$, which is the temporally normalized skin-tone direction. This is usually the largest component in $\mathbf{C}_n(t)$, i.e., the largest distortion (e.g., motion-induced intensity variations) is typically simultaneously and equally present in all three channels.
- **Specular**: $\mathbf{N} \cdot \mathbf{u_s} \cdot I_0 \cdot s(t)$ denotes the temporal variations along the direction of the scaled specular reflection. Under the white light condition, we have $\mathbf{N} \cdot \mathbf{u_s} \cdot I_0 = \mathbf{N} \cdot \mathbf{1} \cdot I_0$, i.e., it is scaled with the inverse of the skin-tone. Under the non-white light condition, $\mathbf{u_s}$ depends on both the light source spectrum and camera sensitivity, while $\mathbf{N}$ depends on the same variables but also on the skin properties (e.g., optical absorption of skin melanin).
- **Pulse**: $\mathbf{N} \cdot \mathbf{u_p} \cdot I_0 \cdot p(t)$ denotes the pulse-induced temporal color variations, i.e., the component of interest. $\mathbf{N} \cdot \mathbf{u_p} \cdot I_0$ is the pulse-induced color variation direction in the temporally normalized RGB space. It depends on the luminance spectrum and camera sensor but is largely skin-tone independent [5]. Over a wide range of lighting spectra and commonly used camera sensitivities, the G-channel has the largest pulsatile amplitude, followed by the B-channel and R-channel, respectively.

Based on (12), we perform a detailed analysis on CHROM and PBV separately to see how they use the physiological/optical properties of skin reflections to address the problem of signal de-mixing. Both methods use the DC-removed signals of $\mathbf{C_n}(t)$ for pulse extraction, which is defined as:

$$\begin{aligned}\tilde{\mathbf{C}}_\mathbf{n}(t) =& \mathbf{C_n}(t) - \mathbf{1} \\ =& \mathbf{1} \cdot i(t) + \mathbf{N} \cdot \mathbf{u_s} \cdot I_0 \cdot s(t) + \mathbf{N} \cdot \mathbf{u_p} \cdot I_0 \cdot p(t),\end{aligned} \tag{13}$$

where $\tilde{\mathbf{C}}_\mathbf{n}(t)$ denotes the (zero-mean) color variation signals.

*1) **PBV**:* It chooses to directly retrieve the pulse from the pulsatile component by restricting all color variations to the pulsatile direction. It does so by projecting $\tilde{\mathbf{C}}_\mathbf{n}(t)$ onto a single direction $\mathbf{z}$ to create an estimate $\hat{p}(t)$ that is proportional to $p(t)$:

$$\hat{p}(t) = \tilde{\mathbf{C}}_\mathbf{n}^\top(t) \cdot \mathbf{z} = k \cdot p(t), \tag{14}$$

where $\mathbf{z}$ denotes the 3×1 projection vector containing the combining weights; k denotes the proportionality factor ($k \neq 0$). Next, it assumes that $p(t)$ (and therefore $\hat{p}(t)$) is uncorrelated with the other signal sources:

$$\mathbb{E}\{p(t) \cdot i(t)\} = \mathbb{E}\{p(t) \cdot s(t)\} \approx 0. \tag{15}$$

Now considering the expected value $\mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\hat{p}(t)\}$, we have:

$$\begin{aligned}\mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\hat{p}(t)\} &= \mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\tilde{\mathbf{C}}_\mathbf{n}^\top(t)\}\cdot\mathbf{z}\\ &= k\cdot\mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot p(t)\}\\ &\approx k\cdot\mathbf{N}\cdot\mathbf{u_p}\cdot I_0\cdot\mathbb{E}\{p(t)\cdot p(t)\}.\end{aligned}\tag{16}$$

At this point, PBV assumes a prior-known blood volume pulse vector $\mathbf{u_{pbv}}$ (3×1 column vector) that satisfies:

$$\mathbf{u_{pbv}} = \mathbf{N}\cdot\mathbf{u_p}\cdot I_0.\tag{17}$$

Thus (16) can be rewritten as:

$$\mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\tilde{\mathbf{C}}_\mathbf{n}^\top(t)\}\cdot\mathbf{z} = \mathbf{u_{pbv}}\cdot k\cdot\mathbb{E}\{p^2(t)\},\tag{18}$$

and the projection vector $\mathbf{z}$ can be derived by:

$$\mathbf{z} = \mathbb{E}\{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\tilde{\mathbf{C}}_\mathbf{n}^\top(t)\}^{-1}\cdot\mathbf{u_{pbv}}\cdot k\cdot\mathbb{E}\{p^2(t)\}.\tag{19}$$

Instead of the ensemble averages, PBV uses a 3×3 temporal covariance matrix:

$$\tilde{\mathbf{\Sigma}} = \{\overline{\tilde{\mathbf{C}}_\mathbf{n}(t)\cdot\tilde{\mathbf{C}}_\mathbf{n}^\top(t)}\},\tag{20}$$

where $\overline{\{\cdot\}}$ denotes the temporal averaging operator for deriving the covariance in the time domain, and takes a k such that $\mathbf{z}$ has unit length. Finally, combining (19) and (20), the projection vector $\mathbf{z}$ is estimated by:

$$\mathbf{z}\propto\tilde{\mathbf{\Sigma}}^{-1}\cdot\mathbf{u_{pbv}},\tag{21}$$

which is used to derive the pulse-signal in (14). Since $\tilde{\mathbf{\Sigma}}$ is estimated from the video content, the key-point of PBV is in defining the blood volume pulse vector $\mathbf{u_{pbv}}$[3].

PBV has a clear advantage: when the assumption of (17) holds, the estimated projection-axis is optimal for pulse retrieval. However, it has two limitations. Firstly, the solution in (21) does not exist when $\mathbf{rank}(\tilde{\mathbf{\Sigma}})<3$, i.e., $\tilde{\mathbf{\Sigma}}$ cannot be inverted. In a singular or near-singular case, the obtained $\mathbf{z}$ is noise driven and, for $\mathbf{rank}(\tilde{\mathbf{\Sigma}})=1$, any $\mathbf{z}$ is a valid solution. This is the typical case in the model when $i(t)=s(t)=0$, i.e., the skin is measured in perfect conditions that are distortion-free. Secondly, it requires accurate knowledge of the blood volume pulse vector for correct noise suppression, i.e., the projection brings quality drops when $\mathbf{u_{pbv}}\neq\mathbf{N}\cdot I_0\cdot\mathbf{u_p}$. The outcome of the algorithm is sensitive to a particular parameter setting of PBV, and PBV in turn is defined by (and thus restricted to) a particular recording setup, depending on the light spectrum and camera sensor.

*2) **CHROM**:* Different from the straightforward one-step solution of PBV, CHROM chooses to introduce flexibility when estimating the projection direction and reduce the sensitivity to the prior knowledge used for pulse extraction. It first reduces the dimensionality of the de-mixing task by eliminating the specular component. This is achieved by only considering the chrominance-signals, which we shall describe as a projection of $\tilde{\mathbf{C}}_\mathbf{n}(t)$ onto the plane orthogonal to the specular variation direction. In order to allow correct functioning regardless the color of the illumination, the method assumes a standardized skin-tone vector, which enables automatic white-balancing of the images. Accordingly, we define the standardized skin-tone vector and the associated mapping matrix as:

$$\mathbf{M}^{-1}\cdot\mathbf{u_{skin}} = \mathbf{1}.\tag{22}$$

where $\mathbf{u_{skin}}$ denotes the 3×1 average skin-tone vector under white light (obtained from a large scale experiment in [4]); $\mathbf{M}$ denotes the diagonal mapping matrix, which is used to map $\tilde{\mathbf{C}}_\mathbf{n}(t)$ as:

$$\begin{aligned}\mathbf{M}\cdot\tilde{\mathbf{C}}_\mathbf{n}(t) =& \mathbf{M}\cdot\mathbf{1}\cdot i(t)+\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_s}\cdot I_0\cdot s(t)\\ &+\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_p}\cdot I_0\cdot p(t),\end{aligned}\tag{23}$$

where the temporally normalized skin color is mapped to the assumed standardized skin-tone under white light. The specular reflection vector $\mathbf{N}\cdot\mathbf{u_s}\cdot I_0$ is approximately mapped to the direction of white light:

$$\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_s}\cdot I_0\approx\kappa\cdot\mathbf{1}.\tag{24}$$

where κ is a proportionality factor. The next step of CHROM is projecting $\mathbf{M}\cdot\tilde{\mathbf{C}}_\mathbf{n}(t)$ onto a plane orthogonal to $\mathbf{1}$ to be independent of the specular variations (after the skin-tone correction):

$$\begin{aligned}\mathbf{S}(t) &= \mathbf{P_c}\cdot\mathbf{M}\cdot\tilde{\mathbf{C}}_\mathbf{n}(t)\\ &\approx \mathbf{P_c}\cdot\mathbf{M}\cdot\mathbf{1}\cdot i(t)+\mathbf{P_c}\cdot\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_p}\cdot I_0\cdot p(t),\end{aligned}\tag{25}$$

subject to

$$\mathbf{P_c}\cdot\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_s}\cdot I_0\approx\kappa\cdot\mathbf{P_c}\cdot\mathbf{1}=\mathbf{0},\tag{26}$$

where $\mathbf{P_c}$ is the 2×3 initial projection matrix used by CHROM that consists of projection-axes in rows, which defines a plane in the temporally normalized RGB space. Note that $\mathbf{P_c}\cdot\mathbf{M}$ is the resulting projection matrix used by CHROM[4]. Such a projection matrix has an attractive property: it creates two projected-signals in $\mathbf{S}(t)$, where the motion-induced/pulse-induced variations appear in in-phase/anti-phase. The reason for this phenomenon has not been explained in [4], but we will show it later.

The in-phase/anti-phase property in $\mathbf{S}(t)$ allows a simple way to create an estimate $\hat{p}(t)$ to approximate $p(t)$, namely "alpha-tuning" [4]:

$$\hat{p}(t) = S_1(t)-\alpha\cdot S_2(t)\ \text{with}\ \alpha=\frac{\sigma(S_1)}{\sigma(S_2)},\tag{27}$$

where $\sigma(\cdot)$ denotes the standard deviation operator; S_i denotes the i-th projected-signal. When the pulsatile components dominate, S_1 and S_2 are anti-phase and thus add up in a constructive way, i.e., $\hat{p}(t)\approx 2\cdot S_1(t)\propto p(t)$. When motion-induced disturbances dominate, (27) cancels the in-phase motion components to approximate $p(t)$. Only when the strengths of motion-induced and pulse-induced components balance, $\hat{p}(t)$ is a sub-optimal estimate of $p(t)$.

[3]As specified in [5], $\mathbf{u_{pbv}}$ used by PBV is measured as $[0.33, 0.77, 0.53]^\top$ for RGB channels, based on the condition of a halogen lamp and the optical RGB-filters of an UI-2220SE-C camera.

[4]As specified in [4], the initial projection matrix used by CHROM is $\mathbf{P_c}=\begin{pmatrix}1 & -1 & 0\\ 0.5 & 0.5 & -1\end{pmatrix}$. $\mathbf{P_c}$ is mapped to the assumed standardized skin-tone vector $\mathbf{u_{skin}}=[0.77, 0.51, 0.38]^\top$ obtained from a large scale experiment, resulting in a new projection matrix $\mathbf{P_c}\cdot\mathbf{M}\approx\begin{pmatrix}3 & -2 & 0\\ 1.5 & 1 & -1.5\end{pmatrix}$ that is eventually used by CHROM.

The strength of CHROM is that it has a certain robustness to non-white illuminations. However, it requires (24) to hold, i.e., the specular component $\mathbf{N}\cdot\mathbf{u_s}\cdot I_0$ measured in real video content must be compensated by the assumed standardized skin-tone vector $\mathbf{u_{skin}}$. Otherwise, it will exhibit specular residuals in the projected-signals, typically when $\mathbf{M}\cdot\mathbf{N}\cdot\mathbf{u_s}\cdot I_0 \neq \kappa\cdot\mathbf{1}$.

C. Data-driven method (2SR)

The recently developed 2SR method [6] creates a subject-dependent skin-color space and tracks the hue-change over time to measure the pulse, where the instantaneous hue is determined based on the statistical distribution of the skin-pixels in the image domain. The notion of using the hue as a fundamental parameter for pulse extraction is supported by the analysis of using different color-spaces to measure pulse in [19]. Since the hue drives the measurement, the method is inherently suppressing all intensity variations at an early stage. In this sense, 2SR is akin to the approach introduced in the next section which defines a projection plane orthogonal to $\mathbf{1}$ in the temporally normalized RGB space for pulse extraction. However, the subspace-axes constructed by 2SR are completely data-driven without physiological considerations. In practice, this implies performance issues when the spatial measurements are unreliable as may occur, i.e., if the skin-mask is noisy or poorly chosen.

IV. POS ALGORITHM

So far, we have shown how different rPPG methods relate to the model. Based on the understanding, we are also able to design new algorithms targeting certain applications or effects. We illustrate this by considering how to introduce the main feature of 2SR into an algorithm based on model (12).

A. Analysis

Since the main feature of 2SR is to consider the hue-change (i.e., to disregard the intensity), its counterpart in (12) is to first eliminate the intensity variations in the direction of $\mathbf{1}$. Therefore, we project $\mathbf{C_n}(t)$ onto the plane orthogonal to $\mathbf{1}$, which is expressed as:

$$\begin{aligned}\mathbf{S}(t) &= \mathbf{P_p}\cdot\mathbf{C_n}(t)\\ &\approx \mathbf{P_p}\cdot\mathbf{N}\cdot\mathbf{u_s}\cdot I_0\cdot s(t) + \mathbf{P_p}\cdot\mathbf{N}\cdot\mathbf{u_p}\cdot I_0\cdot p(t),\end{aligned} \tag{28}$$

subject to

$$\begin{cases}\mathbf{P_p}\cdot\mathbf{1} = (0 \quad 0)^\top,\\ \mathbf{P_{p,1}}\cdot\mathbf{P_{p,2}}^\top = 0,\end{cases} \tag{29}$$

where $\mathbf{P_p}$ denotes a 2×3 projection matrix; $\mathbf{P_{p,i}}$ denotes the i-th row/projection-axis of $\mathbf{P_p}$, which in our definition is assumed to be orthogonal to each other, as the non-orthogonal axes always results in a separate component on the other direction and thus exhibits redundancy. In this case, $\mathbf{P_p}$ defines a plane orthogonal to $\mathbf{1}$ in the temporally normalized RGB space, which is in fact *a plane orthogonal to the temporally normalized skin-tone*, i.e., it is different from the projection plane defined by CHROM[5].

[5]The projection plane in CHROM (i.e., $\mathbf{P_c}\cdot\mathbf{M}$ in (25)) is orthogonal to the specular variation direction by assuming a standardized skin-tone vector. In our case, $\mathbf{P_p}$ in (28) is orthogonal to the intensity variation direction.

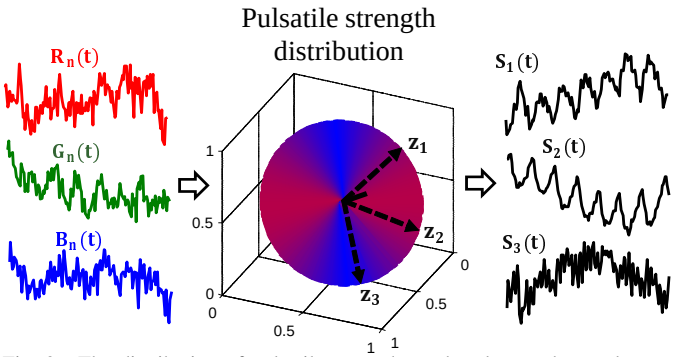

Fig. 2. The distribution of pulsatile strength on the plane orthogonal to $\mathbf{1}$, where the pulsatile strength is the absolute pulsatility value. The projection plane consists of 360 (discrete) projection-axis $\mathbf{z}$ sampled with 1° difference, where the red/blue color denotes the regions with stronger/weaker pulsatile strength. We exemplify three projection-axes on the plane: $\mathbf{z_1}=[-2,1,1]^\top$, $\mathbf{z_2}=[1,-2,1]^\top$ and $\mathbf{z_3}=[1,1,-2]^\top$, which have the pulsatilities -0.64, 0.68 and -0.04 according to (31). We project a temporally normalized RGB signal $\mathbf{C_n}(t)=[R_n(t),G_n(t),B_n(t)]^\top$, measured from the skin in a video, onto $\mathbf{z_1}$, $\mathbf{z_2}$, $\mathbf{z_3}$ and obtain $S_1(t)$, $S_2(t)$, $S_3(t)$.

Conceptually, there are two advantages for $\mathbf{P_p}$: (i) the (motion-induced) intensity variations are usually larger distortions that influence all three channels simultaneously; and (ii) it does not require exact knowledge of $\mathbf{N}\cdot\mathbf{u_s}\cdot I_0$ and $\mathbf{N}\cdot\mathbf{u_p}\cdot I_0$ to define the main distortion direction at the moment. Although the normal vector of the projection plane has been determined (as $\mathbf{1}$), the actual projection-axes in $\mathbf{P_p}$ are not defined yet. One might consider to define a $\mathbf{P_p}$ that can further project $\mathbf{C_n}(t)$ onto the direction orthogonal to the specular distortion on the plane. However, this is not a feasible option, since $\mathbf{N}\cdot\mathbf{u_s}\cdot I_0$ and $\mathbf{N}\cdot\mathbf{u_p}\cdot I_0$ may not be orthogonal to each other, typically $\mathbf{N}\cdot\mathbf{u_s}\cdot I_0$ is not well-defined due to different motion-types. In contrast, $\mathbf{N}\cdot\mathbf{u_p}\cdot I_0$ is relatively stable when the light source and camera filter are fixed during the measurement [5]. Thus we prefer to use $\mathbf{N}\cdot\mathbf{u_p}\cdot I_0$ to define $\mathbf{P_p}$, exploiting the physiological property of PPG-absorption.

According to [5], the blood pulsation has different relative PPG-contributions in RGB channels, which can be expressed as a blood volume pulse vector $\mathbf{u_{pbv}}$:

$$u_{pbv}(c) = \frac{\int_{400}^{700} H_c(\lambda)\cdot\frac{I(\lambda)}{I_h(\lambda)}\cdot PPG(\lambda)\,d\lambda}{\int_{400}^{700} H_c(\lambda)\cdot\frac{I(\lambda)}{I_h(\lambda)}\cdot\rho_{\mathbf{s}}(\lambda)\,d\lambda}, \tag{30}$$

where $u_{pbv}(c)$ denotes the pulsatile strength (i.e., scalar) of the c-th color channel of the camera sampled at the wavelength $\lambda\in[400,700]$ nm; $H_c(\lambda)$ denotes the response of c-th color channel of the camera; $I(\lambda)$ and $I_h(\lambda)$ denote the spectral compositions of the given light source and the halogen lamp used for measuring the absolute PPG-amplitude $PPG(\lambda)$; $\rho_s(\lambda)$ denotes the skin-reflection spectra.

To fully understand how the projection-axes in $\mathbf{P_p}$ affects the quality of the projected-signals $\mathbf{S}(t)$, we investigate the pulsatility of the projection direction on the plane using $\mathbf{u_{pbv}}$. Assuming one projection-axis in $\mathbf{P_p}$ as $\mathbf{z}$, the pulsatility on the direction of $\mathbf{z}$ is:

$$p = \mathbf{u_{pbv}}^\top\cdot\mathbf{z}, \tag{31}$$

where $\mathbf{z}$ denotes the 3×1 projection vector; $\mathbf{u_{pbv}}$ denotes the 3×1 blood volume pulse vector given by (30); p denotes the

pulsatility on the direction of $\mathbf{z}$, i.e., a scalar that can either be positive or negative. The pulsatile strength is the absolute value of p, which reflects the amplitude of pulsatile variations (AC-level) on the direction of $\mathbf{z}$.

Fig. 2 shows the distribution of the pulsatile strength on the plane orthogonal to $\mathbf{1}$ as a function of $\mathbf{z}$. From this figure, we can see that the projection direction is highly related to the pulsatility (and thus the pulsatile strength) that determines the signal quality, i.e., different $\mathbf{z}$ may give very different projected-signals. For example, $\mathbf{z_1}$ and $\mathbf{z_2}$ show negative and positive pulsatilities giving the anti-phase signals $S_1(t)$ and $S_2(t)$; $\mathbf{z_3}$ shows a much lower pulsatile strength giving the noisy signal $S_3(t)$. This implies that the projection-axis on the plane cannot be arbitrarily selected, but should depend on physiological reasoning.

Although $\mathbf{u_{pbv}}$ remains stable when the recording setup is fixed during the measurement, the light source and camera filter are usually unknown in a video and can vary in different setups, which renders it difficult to use a fixed off-line $\mathbf{u_{pbv}}$ for accurate on-line measurement. Inspired by CHROM [4], we use *the knowledge of the blood volume pulse to define a rough projection region on the plane orthogonal to the temporally normalized skin-tone direction, and refine an exact projection direction on the plane by real-time tuning*.

Therefore, the key-point is in defining two projection-axes on the plane that can bound a most likely pulsatile region (e.g., the red regions on the plane of Fig. 2), where larger pulsatilities can be found within the boundaries by tuning, i.e., by $S_1(t) + S_2(t)$. Based on our requirement, the projected-signals in (28) can be expressed in such a general form:

$$\mathbf{S}(t) = \begin{pmatrix} S_1(t) \\ S_2(t) \end{pmatrix} \text{ with } \begin{cases} S_1(t) = d_1(t) - d_2(t), \\ S_2(t) = d_1(t) + d_2(t) - 2d_3(t), \end{cases} \tag{32}$$

where $\mathbf{D}(t) = [d_1(t), d_2(t), d_3(t)]^\top$ is a vector having the same entries as $\mathbf{C_n}(t)$ but differently ordered. The entries in $\mathbf{D}(t)$ are ordered according to the *decreasing pulsatile strength* of RGB channels in $\mathbf{C_n}(t)$, i.e., the descending channel-ranking based on $\mathbf{u_{pbv}}$. The projection-axes defined in (32) are orthogonal to each other and also to $\mathbf{1}$. Most importantly, both projection-axes exhibit *positive pulsatilities* and thus generate *in-phase* pulse-signals.

Taking a single light source (e.g., the fluorescent lamp) as the example, the skin pulsatility is usually the largest in the G-channel, followed by the B-channel and R-channel. Based on such a channel-ranking and our requirements in (32), the projection-axes in (28) can be defined as:

$$\mathbf{P_p} = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{pmatrix}, \tag{33}$$

which in fact combines temporally normalized RGB-signals as: $S_1(t) = G_n(t) - B_n(t)$ and $S_2(t) = G_n(t) + B_n(t) - 2R_n(t)$.

The last step is to tune an exact projection direction within the bounded region of (32), where the specular and pulsatile components in (12) can further be separated. Before tuning, it needs to be shown that the specular distortion and pulse are physically separable on the projection plane. To this end, we

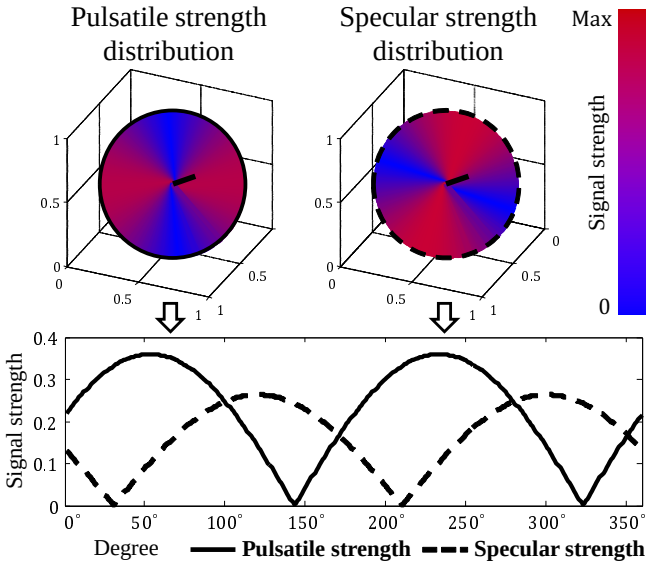

Fig. 3. The distribution of pulsatile and specular strengths on the plane orthogonal to $\mathbf{1}$, where the red/blue color denotes the regions with stronger/weaker signal strength. Here we apply (i) $\mathbf{u_{skin}} = [0.77, 0.51, 0.38]^\top$ of CHROM to approximately define the specular vector as $[0.37, 0.56, 0.75]^\top$, based on (22) and (24); and (ii) $\mathbf{u_{pbv}}$ of PBV to define the pulsatile vector as $[0.33, 0.77, 0.53]^\top$. These two (color variation) vectors are used to estimate the strengths of both components on the projection plane, which is similar to the procedure in Fig. 2. Considering the degree of the projection-axis as a variable, the distribution of pulsatile and specular strengths on the plane can be compared by strength curves.

compare the distribution of the signal strength (i.e., "strength" means the amplitude of the signal variation) between these two components on the plane orthogonal to $\mathbf{1}$ (see Fig. 3). Based on the assumed $\mathbf{u_{skin}}$ and $\mathbf{u_{pbv}}$, it shows that the specular and pulse components have almost opposite distributions on the plane, i.e., their strength curves have a clear phase shift.

When the distributions of $\mathbf{N} \cdot \mathbf{u_s} \cdot I_0$ and $\mathbf{N} \cdot \mathbf{u_p} \cdot I_0$ on the projection plane are sufficiently different, the specular and pulsatile components are algorithmically separable. In this sense, our tuning depends on the hypothesis that specular and pulse have different relative strengths in the temporally normalized RGB channels. Such hypothesis seems to be true under normal conditions: the hemoglobin and melanin contents in human skin tissues lead to specific chromophore concentrations. The skin color (including dark skin) under white light looks more reddish and less bluish, i.e., nobody has an inborn blue face. Thus the specular vector (e.g., the inverse of the skin-tone after temporal normalization) is unlikely to coincide with the blood volume pulse vector. However, it remains questionable whether such hypothesis also holds for extreme luminance conditions, since the lighting spectrum affects the relative contributions of both the specular and pulsatile components to RGB channels. We will verify this hypothesis in the experimental section by using the benchmark videos recorded in various illumination conditions with different lighting spectra.

Assuming for the moment that such hypothesis is true, the region bounded by the two projection-axes in (32) has large pulsatile strength and low specular strength, and thus the projected-signals have (i) in-phase pulsatile components, and (ii) anti-phase specular components. Similar to CHROM, we leave the task of finding an exact projection direction to

the alpha-tuning [4], which can be expressed as:

$$h(t) = S_1(t) + \alpha \cdot S_2(t) \text{ with } \alpha = \frac{\sigma(S_1)}{\sigma(S_2)}, \qquad (34)$$

where $\sigma(\cdot)$ denotes the standard deviation operator. Note that the sign in (34) is different from the sign in (27) of CHROM (i.e., $+$ instead of $-$), as it depends on the pulsatility of two projected-signals[6]. The alpha-tuning used by [4] has an attractive property: (i) when the pulsatile variation dominates $\mathbf{S}(t)$, $S_1(t)$ and $S_2(t)$ appear in in-phase. Adding two in-phase signals together will boost the resulting-signal strength, i.e., the value of α is non-critical at this point; (ii) when the specular variation dominates $\mathbf{S}(t)$, $S_1(t)$ and $S_2(t)$ appear in anti-phase. The α can pull/push the specular variation strength of one signal to the same level as the other one, i.e., $\sigma(S_1) = \sigma(\alpha \cdot S_2)$. Adding two anti-phase signals with the same amplitude will cancel out the specular distortion. However, its performance becomes sub-optimal when the pulsatile strength and specular strength are very close to each other, i.e., α is driven by a mixture of both and is thus not well-defined. This drawback will be discussed together with CHROM in Section IV.C.

Assuming that $h(t)$ is estimated from short video intervals in a sliding window (with length l), we can derive a long-term pulse-signal $H(\tau)$ by overlap-adding the partial segments $h(t)$ (after making them zero-mean) as in [4]. Consequently, $H(\tau)$ is the final output pulse-signal that can be used for further analysis, such as the pulse-rate estimation. To be more specific, the setting of l depends on the camera frame-rate, which should include at least one cardiac cycle for processing. On top of that, a short l is preferred, as it can quickly adapt the alpha-tuning to suppress instantaneous distortions in a short interval and also avoid the influence of low frequency components like respiration. The overlap-adding length must be smaller than l. In our case, the window slides 1 frame for the overlap-adding (i.e., overlap-adding length is thus $l-1$), which includes more measurements.

B. Algorithm

In order to arrive at a fully specified algorithm, we assume that in most use-cases the channel-ranking in terms of pulsatility is relatively stable, i.e., the actual values in $\mathbf{u_{pbv}}$ may change, but their order cannot be easily altered. This is shown in Fig. 4 where the effects of luminance and skin-tone are qualitatively illustrated. Fig. 4 (a) exemplifies two strikingly different luminance spectra that are commonly used: incandescent lamp and fluorescent lamp. Since the spectrum of the incandescent light can be considered a low-pass filtered version of that of the fluorescent light, it is expected that the channel-ranking of $\mathbf{u_{pbv}}$ will not be very different in these two lighting conditions. Fig. 4 (b) shows the reflection spectra of different skin-tones. Since their shapes are rather

[6]Considering the blood volume pulse vector $\mathbf{u_{pbv}} = [0.33, 0.77, 0.53]^\top$, the pulsatilities of two projection-axes in CHROM $\begin{pmatrix} 3 & -2 & 0 \\ 1.5 & 1 & -1.5 \end{pmatrix}$ are -0.55 and 0.47 respectively, which indicates that the two projected-signals are anti-phase. So CHROM uses $S_1(t) - \alpha \cdot S_2(t)$ for alpha-tuning in (27). In contrast, POS directly finds two projection-axes giving in-phase signals, and thus its alpha-tuning is formulated as $S_1(t) + \alpha \cdot S_2(t)$ in (34).

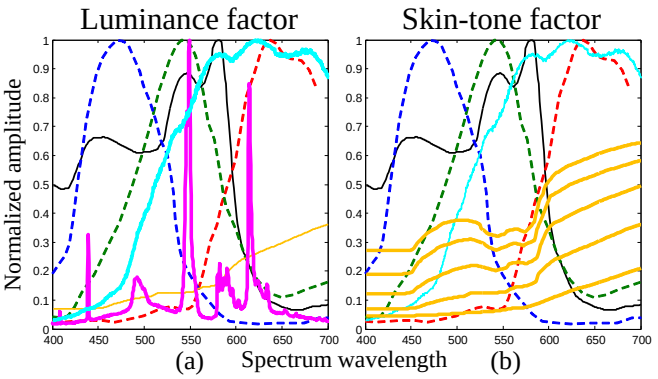

Fig. 4. The dashed red/green/blue curve denotes the response of R/G/B channel of the camera; black curve denotes the absolute PPG amplitude; yellow curve denotes the skin-reflection spectra; cyan/magenta curve denotes the spectrum of incandescent/fluorescent lamp. Since the PPG-amplitude is fixed and the order of RGB channel responses of a camera sensor will not be altered, we only investigate the luminance factor and skin-tone factor: (a) compares incandescent (cyan curve) and fluorescent (magenta curve) lighting conditions, and (b) compares different skin-tones (yellow curves).

constant, the channel-ranking of $\mathbf{u_{pbv}}$ is expected to be skin-tone independent [5]. For simplicity, we therefore take the incandescent lamp or fluorescent lamp as the typical light source to fix the projection-axes for benchmark, which is in fact the $\mathbf{P_p} = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{pmatrix}$ from (33).

The novelty of the newly proposed method is using the plane orthogonal to the skin-tone in the temporally normalized RGB space for pulse extraction. So we name it "Plane-Orthogonal-to-Skin" (POS), which is also the unique character distinguishing it from prior art. In order to highlight the fundamental/independent performance of POS, we keep its algorithm as clean and simple as possible, i.e., even the commonly used band-pass filtering is not used. The bare core algorithm of POS is shown in Algorithm 1, which can be implemented in a few lines of Matlab code.

C. Difference with model-based prior art

Since PBV, CHROM and POS are all approaches to de-mix (12) based on optical/physiological considerations, they share many properties. We will highlight their differences to forecast the difference in performance as discussed in Section VI.

Algorithm 1 Plane-Orthogonal-to-Skin (POS)

Input: A video sequence containing N frames

Initialize: $\mathbf{H} = \mathbf{zeros}(1, N)$, $l = 32$ (20 fps camera)
for $n = 1, 2, ..., N$ **do**
  $\mathbf{C}(n) = [R(n), G(n), B(n)]^\top \leftarrow$ spatial averaging
  if $m = n - l + 1 > 0$ **then**
    $\mathbf{C_n^i} = \dfrac{\mathbf{C_{m\to n}^i}}{\mu(\mathbf{C_{m\to n}^i})} \leftarrow$ temporal normalization
    $\mathbf{S} = \begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{pmatrix} \cdot \mathbf{C_n} \leftarrow$ projection
    $\mathbf{h} = \mathbf{S_1} + \dfrac{\sigma(\mathbf{S_1})}{\sigma(\mathbf{S_2})} \cdot \mathbf{S_2} \leftarrow$ tuning
    $\mathbf{H_{m\to n}} = \mathbf{H_{m\to n}} + (\mathbf{h} - \mu(\mathbf{h})) \leftarrow$ overlap-adding
  end if
end for

Output: The pulse-signal $\mathbf{H}$

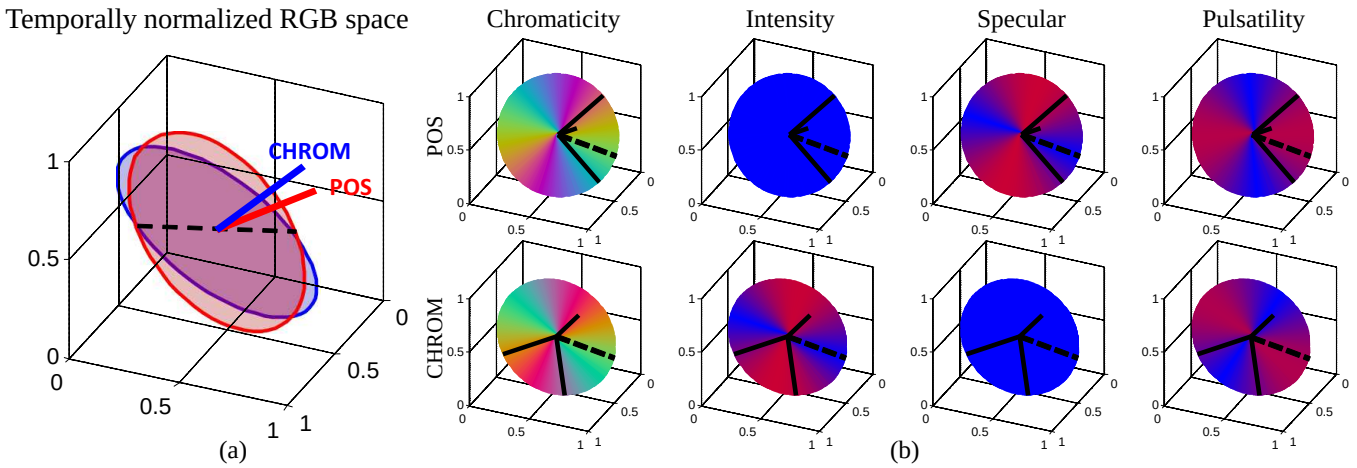

Fig. 5. (a) The projection planes of POS (red) and CHROM (blue) in the temporally normalized RGB space. Both projection planes intersect at the direction of $2G_n(t) - R_n(t) - B_n(t)$ (dashed black line), i.e., the direction for pulse extraction in CHROM when $\alpha = 1$. (b) The projection planes of POS and CHROM have different chromaticity distributions. Besides, they have different distributions of intensity, specular and pulsatile variations. Note that the solid black line denotes the principal normal vector and projection-axes in both methods.

1) PBV versus CHROM and POS: PBV is a one-step procedure to determine a single and optimum projection assuming that disturbances are present. It requires accurate knowledge of the blood volume pulse signature, which results in restrictions for the recording setup. Moreover, if the assumption, that large enough distortions are present, is invalid, the solution will become non-unique. Although CHROM and POS may have the problem in distinguishing between the pulsatile component and noise component when their amplitude-level are close to each other on respective planes, these two methods are less restrictive than PBV in terms of the amount of distortions, i.e., CHROM and POS perform well for both the stationary and motion situations when the alpha-tuning is either driven by pulse or large distortions, whereas PBV is particularly designed for the motion situation.

Therefore in comparison, CHROM and POS require less accurate knowledge of the blood volume pulse signature and are more tolerant to the amount of distortions, which can be considered as resorting to a sub-optimal and greedy algorithm.

2) CHROM versus POS: The essential difference between CHROM and POS is the order in which the distortions (intensity and specular) are eliminated and thus which distortion is used in the alpha-tuning. CHROM starts with the specular and uses the intensity for alpha-tuning, POS does it the other way round (see Fig. 5 (a)). The ordering difference in CHROM and POS implies symmetric performance issues in alpha-tuning, i.e., it may become sub-optimal when the amplitude of the intensity/specular variation is very close to that of the pulsatile variation on respective planes. Fig. 5 (b) further illustrates the difference between CHROM and POS by showing the strength distribution of intensity ($\mathbf{1}$), specular ($[0.37, 0.56, 0.75]^\top$), and pulsatility ($[0.33, 0.77, 0.53]^\top$) on their planes. The intensity/specular distortion is eliminated in the first step of POS/CHROM respectively, resulting in blue planes in Fig. 5 (b). The remaining components peak in almost orthogonal directions on each projection plane. This implies that both methods are essentially operable with the alpha-tuning. However, we argue that the cleanness (blueness) of their projections are different in practice: CHROM is expected to be more vulnerable due to a (over time) consistent difference between the assumed and actual directions of the specular distortion for each individual subject, while POS is expected to be more vulnerable to inhomogeneous illumination spectra.

We draw a brief conclusion on the comparison of three model-based methods: PBV requires accurate knowledge of the blood volume pulse direction. CHROM and POS use soft priors in blood volume pulsation (i.e., channel-ranking) to define a projection plane for alpha-tuning. Moreover, POS further softens the knowledge required by CHROM (i.e., standardized skin-tone) for defining the projection plane by using the data-driven approach, i.e., defining a plane orthogonal to the temporally normalized skin-tone direction. The differences in performance of the various methods (see Section VI) reflect this trade-off between exactness of upfront knowledge and greediness of the rPPG algorithm.

V. Experimental Setup

This section presents the experimental setup for the benchmarking. First, a large video dataset is introduced. Next, a evaluation metric is presented. Finally, a total of eight rPPG methods are adopted for comparison.

A. Benchmark dataset

A benchmark dataset containing 60 video sequences (with 147100 frames) has been built to evaluate the proposed rPPG method. The videos are recorded with a regular RGB camera[7] in an uncompressed bitmap format[8], 768×576 pixels, 8 bit depth, and 20 FPS. The ground-truth is either the contact-based PPG-signal sampled by a finger-based transmissive pulse oximetry[9] or the ECG-signal sampled by a polar chest belt[10] (in the fitness experiment). Both are synchronized with

[7]Global shutter RGB CCD camera USB UI-2230SE-C from IDS.

[8]The MAHNOB-HCI dataset created by [20] for affect recognition is unsuitable for the rPPG task, as the recorded videos are compressed in MPEG-4 format, i.e., subtle pulsatile information may be lost after compression or be polluted by compression artifacts.

[9]Model CMS50E from ContecMedical.

[10]Polar H3 heart-rate sensor.

Fig. 6. Snapshots of some recordings in the benchmark dataset. The frames in each row are exemplified from the corresponding category in the same order.

the video frames. Unless mentioned otherwise, the subject is illuminated by a frontal fluorescent lamp, and sits in front of the camera with the face visible.

To be as close as possible to practical use-cases, we perform recordings in four scenarios to include different challenges. This allows us to investigate these challenges independently, as described below (the bold number in brackets indicates the number of frames recorded for each challenge category):

• **Skin-tone (22500)** 15 subjects with various skin-tones are recorded and categorized into three skin-types based on the Fitzpatrick scale: 5 Western European subjects (*skin-type I-II*), 5 Eastern Asian subjects (*skin-type III*), and 5 Sub-Sahara Africa/Southern Asian subjects (*skin-type IV-V*).

• **Luminance (31500)** Luminance becomes a challenging factor when motion distortions appear [6]. Thus we define 3 basic motion-types, i.e., *stationary*, *rotation* (rigid motion) and *talking* (non-rigid motion), for a subject (skin-type III) to perform under 7 different luminance conditions including single/mixture of colored light sources, i.e., fluorescent lamp, red LED lamp, green LED lamp, blue LED lamp, red-green LED lamps, red-blue LED lamps, and green-blue LED lamps. Note that the colored LED lamps act as point sources.

• **Recovery after exercise (54000)** To evaluate the rPPG robustness to pulse-rate changes, a series of videos is recorded to analyze the pulse-rate recovery from a running exercise. In this category, 6 subjects (3 males and 3 females) in skin-types I-III participated in the recordings. Each subject performed 3 different levels of running (with different intensities) by adjusting speed and gradient of the treadmill: *low* (gradient=12°, speed=4-5 km/h), *medium* (gradient=14°, speed=5-6 km/h), and *high* (gradient=15°, speed=7-8 km/h). The duration of each running exercise is 3 minutes. After the exercise, the subject immediately sits in front of the camera for a recording.

• **Fitness exercise (39100)** The last experiment is to record subjects in fitness exercise for testing the rPPG robustness to vigorous body-motions. The body-motion due to the sporting exercise is much more significant and periodic [21] than the simulated head motions in the luminance category. In this category, 5 male subjects in skin-types I-V, illuminated by the ceiling fluorescent light, exercised on biking or stepping devices. The duration of recordings are between 2.5 minutes and 8.5 minutes. Since the finger-based PPG-sensor is vulnerable to body-motions due to exercise, the ECG-signals were recorded as the reference and post-processed to remove outliers.

Fig. 6 shows snapshots of some recordings in our benchmark dataset. All videos are pre-processed by a fast on-line learning based object tracker and OC-SVM classifier that have been used in [9] for localizing the face and selecting the skin-pixels. This study has been approved by the Internal Committee Biomedical Experiments of Philips Research, and informed consent has been obtained from each subject.

B. Evaluation metrics

The Signal-to-Noise-Ratio (SNR) metric used by [4] is adopted to assess the quality of the extracted pulse-signal. Similar to [4], the SNR of the pulse frequency is derived by the ratio between the energy around the first two harmonics and remaining parts in the frequency spectrum, where the location of the first two harmonics is determined by either the PPG-signal or ECG-signal. The SNR values, shown by each benchmarked method, are averaged over all videos under each challenge per category for statistical comparison.

C. Compared rPPG methods

The POS method proposed in this paper is intended as an algorithmic component in an rPPG monitoring-system. Thus we compare it as clean as possible with direct algorithmic alternatives. For a thorough evaluation, we benchmark it with seven commonly-used or state-of-the-art rPPG methods:

• **G (2007)** [1], the single wavelength method that is still popular as evidenced by recent researches [7], [10].

• **G-R (2008)** [22], a simple alternative to the single channel method by combining two channels.

• **PCA (2011)** [2], a Blind Source Separation (BSS) based method.

• **ICA (2011)** [3], the most famous BSS-based method that has been widely used.

• **CHROM (2013)** [4], the motion robust method based on the standardized skin-tone assumption.

• **PBV (2014)** [5], the motion robustness improved method using the blood volume pulse signature.

• **2SR (2016)** [6], a recent method exploiting the skin-pixel distribution in the image domain.

All these 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 POS strictly follows Algorithm 1 presented in this paper. Following the discussion at the end of Section IV.A, the sliding window length of POS is defined as $l = 32$ given a 20 fps camera, which measures cardiac activities in 1.6 s, i.e., it can capture at least one cardiac cycle of the measured signal in a broad pulse-rate range $[40, 240]$ beat per minute. The parameters in the benchmarked methods are set according to the original papers. For fair comparison, all parameters remained identical when processing different videos.

TABLE I
SNR PERFORMANCE INDICATORS PER RPPG METHOD AND CHALLENGE. THE RED AND BLUE COLORED NUMBERS IDENTIFY THE BEST AND SECOND-BEST RPPG METHODS IN EACH CHALLENGE.

Category	Challenge	G(2007)	G-R(2008)	PCA(2011)	ICA(2011)	CHROM(2013)	PBV(2014)	2SR(2016)	POS
Skin-tone	Type I-II	2.67	**7.55**	5.85	6.51	6.47	5.57	7.44	**7.69**
	Type III	2.07	7.89	5.38	6.61	6.21	6.26	**7.90**	**8.04**
	Type IV-V	-0.49	6.40	2.25	4.56	5.43	4.04	**6.60**	**7.21**
Luminance	Stationary	8.10	10.14	8.70	**11.61**	9.42	6.57	**10.53**	10.51
	Rotation	0.81	3.34	1.46	4.04	3.63	**6.36**	6.16	**6.28**
	Talking	-0.62	3.75	0.46	3.11	3.99	4.01	**5.33**	**5.05**
Recovery	Low	-3.07	4.67	-0.60	1.78	2.66	1.95	**4.93**	**4.82**
	Medium	-3.19	4.97	-0.79	1.64	3.62	3.15	**5.26**	**5.21**
	High	-8.19	4.11	-6.51	-0.82	3.52	3.52	**4.84**	**4.74**
Fitness	Biking	-6.39	-3.38	-4.21	-5.40	**0.68**	0.57	-0.28	**0.78**
	Stepping	-12.59	-9.06	-11.41	-12.51	**-3.13**	**-2.85**	-4.50	-3.58
Overall	Average	-1.90	3.67	0.05	1.92	3.86	3.56	**4.93**	**5.16**

VI. RESULTS AND DISCUSSION

This section discusses the benchmark result. Table I summarizes the SNR values of the benchmarked rPPG methods obtained in each challenge per category, where the red/blue bold entry denotes the best/second-best result obtained by the corresponding method in each challenge. The overall category gives the performance averaged over all individual challenges for each method. Fig. 7 shows the qualitative comparison of spectrograms obtained in fitness challenges.

1) G: The single channel method G obtains on average the worst performance, which stems from the fact that no effort has been made to eliminate distortions by combining signals from different color sensors. This suggests that when multiple wavelength sensors are available in a regular RGB camera, it is better to profit from the statistics provided by all color channels, especially when the pulsatility in different channels is non-uniform. Even when one channel does not contain any pulsatile information, it can still be used as a noise-sensor to design a method that can be independent of such noise, i.e., signal de-noising. The poor performance of G is further confirmed in Fig. 7.

2) G-R: Surprisingly, G-R, a simple alternative to G by combining two channels, shows decent performance, i.e., it even outperforms the BSS-based methods that combine three color channels. In fact, G-R projects the temporally normalized RGB-signals onto the direction of $[-1, 1, 0]^\top$, which is also a projection-axis on the POS-plane that is orthogonal to $\mathbf{1}$ (i.e., independent of intensity variations). The essential difference with POS is that G-R does not exploit the B-channel, thus cannot further differentiate pulse from specular distortions. In non-fitness videos without significant body-motions, it can still profit from the physiological phenomenon that the G and R channels contain the maximal and minimal pulsatile amplitudes, i.e., such a combination maximizes the resulting pulsatility. However, it shows limitations in suppressing motion-induced specular distortions in fitness applications (see Fig. 7).

3) PCA/ICA: Comparing the BSS-based methods, ICA performs better than PCA in non-fitness challenges, i.e., especially in the stationary case of the luminance category, where ICA is the best. Both methods have clear quality drops when distortions appear, i.e., head motions in different luminance conditions or heavy breathing during the exercise recovery. In fitness challenges, PCA and ICA almost break down when significant and periodic body-motions appear, which is largely due to the de-mixing matrix estimation and target component selection. We also notice that ICA is more sensitive to large motion distortions than PCA in our fitness setup, which is in line with the findings in [4].

4) CHROM/PBV: Comparing the two existing model-based methods, CHROM performs slightly better than PBV in overall, especially in non-fitness challenges without strong distortions. The reason has been explained in Section III that the use of blood volume pulse signature in PBV brings a modest loss in signal quality when the subject is (nearly) stationary. In fitness challenges, the model-based methods (CHROM, PBV and POS) demonstrate significantly improved robustness as compared to the non-model based methods (G, G-R, PCA, ICA and 2SR). Among them, PBV reports on average the best performance in fitness, which is followed by CHROM and POS, although their differences are non-significant.

5) 2SR: The recently developed method 2SR gains the best position in the non-fitness comparison. Basically, 2SR replaces the spatial pixel averaging of skin-pixels by the Least-Mean-Square estimate of the skin-color space using spatial PCA. This may reduce the influence of outliers when skin-pixels dominate the measurement, but suffers performance degradations if the skin-mask is poorly defined, which typically occurs in fitness, i.e., some pixels may always contain a combination of skin and non-skin when the skin-region is moving at high-speed (due to motion blur) [6]. Fig. 7 shows that 2SR works properly in subject 3-4 where the skin-mask is relative clean, but fails with subject 6 (with dark skin-tone) where the skin-mask is seriously polluted. We have to note that the overall second best position gained by 2SR in the complete dataset is based on the preliminary condition that the skin-mask is well defined in the majority of videos.

6) POS: Table I shows that POS obtains the overall best performance. The comparison between three model-based methods in non-fitness challenges shows that the assumption

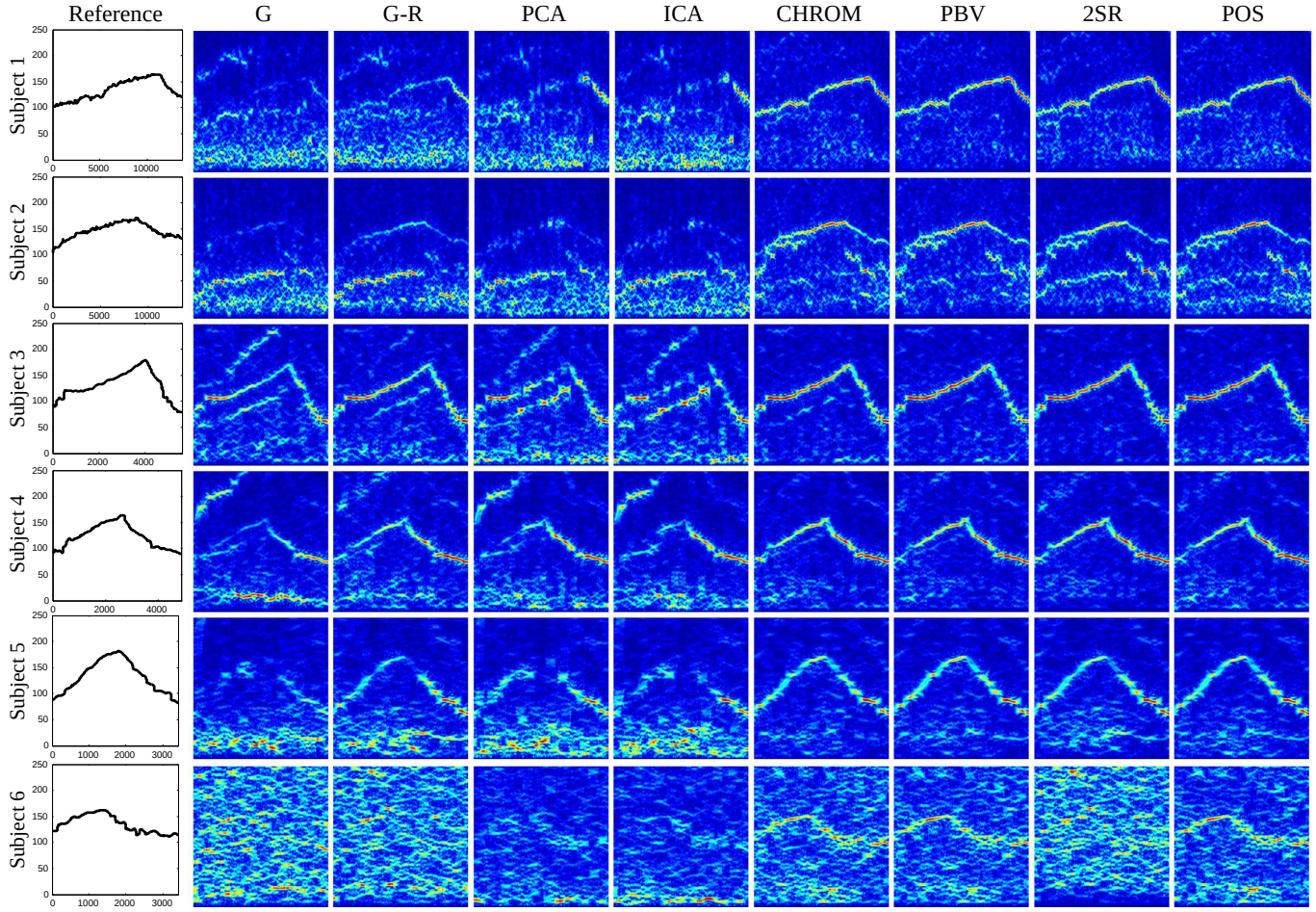

Fig. 7. The ECG reference signals and the spectrograms obtained by benchmarked rPPG methods on videos recorded in fitness exercise (i.e., subject 2 performs stepping exercise), where the x-axis and y-axis denote the frame number and frequency respectively.

made by POS, i.e., be independent of the intensity variations, is advantageous in simple use-cases without strong distortions. We also notice a quite similar performance of POS and 2SR in non-fitness challenges. This is likely explained by their shared feature of reducing the dimensionality of the de-mixing problem to the plane orthogonal to the (normalized) skin-tone direction. The main difference between POS and 2SR is in defining the orthonormal plane: the subspace-axes (e.g., the second and third principal components) of 2SR is purely determined by the image-based skin-pixel distributions, while the projection-axes of POS is built on physiological reasoning. Such a property is especially advantageous for POS in fitness challenges where the skin-mask is noisy (see subject 6 in Fig. 7).

As a final remark, we stress that this paper aims at increasing and improving the understanding to the algorithmic principles of rPPG, and providing more insights that may benefit the algorithm development in the future. Neither detailed algorithmic optimization nor dedicated signal processing were considered for attaining the highest accuracy in any of the discussed methods. Though the validity of the presented model might appear limited due to the assumption of a single light source, the benchmark shows that, in practice, such a limitation in the model-based methods is not as severe as might appear upfront. The validation also suggests that when developing a general purpose rPPG engine for a broad range of use-cases, one should consider to use the characteristic properties of rPPG to design a robust solution.

VII. Conclusion

A mathematical model for rPPG measurement is proposed, which is based on the optical and physiological considerations and assumption of a single light source with a constant spectrum. We use this model to understand the commonalities and differences between existing rPPG methods in pulse extraction. Our analysis shows that combining the model with different assumptions allows constructing various algorithms to extract the pulse-signal from a video, and further suggests an alternative method POS that resembles CHROM but alters the order in which the main expected color distortions are reduced using different priors. A large benchmark, involving various challenges, is executed on existing and newly-proposed rPPG methods to confirm our understanding.

Acknowledgment

The authors would like to thank Mr. Ger Kersten at Philips Research for creating the video recording system, and also the volunteers from Philips Research and Eindhoven University of Technology for their efforts in creating the benchmark dataset.

REFERENCES

[1] W. Verkruysse *et al.*, "Remote plethysmographic imaging using ambient light," *Opt. Exp.*, vol. 16, no. 26, pp. 21 434–21 445, Dec. 2008.

[2] M. Lewandowska *et al.*, "Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity," in *Proc. Federated Conf. Comput. Sci. Inform. Syst. (FedCSIS)*, Szczecin, Poland, Sept. 2011, pp. 405–410.

[3] M.-Z. Poh *et al.*, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[4] 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.

[5] 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, pp. 1913–1922, Oct. 2014.

[6] W. Wang *et al.*, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 9, pp. 1974–1984, Sept. 2016.

[7] X. Li *et al.*, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, June 2014, pp. 4264–4271.

[8] L. Tarassenko *et al.*, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, p. 807, May 2014.

[9] W. Wang *et al.*, "Exploiting spatial redundancy of image sensor for motion robust rPPG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

[10] M. Kumar *et al.*, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomed. Opt. Exp.*, vol. 6, no. 5, pp. 1565–1588, May 2015.

[11] S. Tulyakov *et al.*, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 2396–2404.

[12] A. R. Guazzi *et al.*, "Non-contact measurement of oxygen saturation with an RGB camera," *Biomed. Opt. Exp.*, vol. 6, no. 9, pp. 3320–3338, Sept. 2015.

[13] I. C. Jeong and J. Finkelstein, "Introducing contactless blood pressure assessment using a high speed video camera," *J. Med. Syst.*, vol. 40, no. 4, pp. 1–10, Apr. 2016.

[14] L. K. Mestha *et al.*, "Towards continuous monitoring of pulse rate in neonatal intensive care unit with a webcam," in *Proc. IEEE Conf. Eng. Med. Biol. Soc. (EMBS)*, Chicago, IL, USA, Aug. 2014, pp. 3817–3820.

[15] S. Fernando *et al.*, "Feasibility of contactless pulse rate monitoring of neonates using google glass," in *Proc. EAI Conf. Wireless Mobile Commun. Healthcare (Mobihealth)*, London, UK, Oct. 2015, pp. 198–201.

[16] J.-P. Couderc *et al.*, "Detection of atrial fibrillation using contactless facial video monitoring," *Heart Rhythm*, vol. 12, no. 1, pp. 195–201, Jan. 2015.

[17] D. McDuff *et al.*, "Remote measurement of cognitive stress via heart rate variability," in *Proc. IEEE Conf. Eng. Med. Biol. Soc. (EMBS)*, Chicago, IL, USA, Aug. 2014, pp. 2957–2960.

[18] G. R. Tsouri *et al.*, "Constrained independent component analysis approach to nonobtrusive pulse rate measurements," *J. Biomed. Opt.*, vol. 17, no. 7, p. 077011, July 2012.

[19] 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, p. 048002, Apr. 2015.

[20] M. Soleymani *et al.*, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[21] W. Wang *et al.*, "Quality metric for camera-based pulse rate monitoring in fitness exercise," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sept. 2016, pp. 2430–2434.

[22] M. Hülsbusch, "An image-based functional method for opto-electronic detection of skin perfusion," Ph.D. dissertation (in German), Dept. Elect. Eng., RWTH Aachen Univ., Aachen, Germany, 2008.

Wenjin Wang received the B.Sc. degree in Biomedical Engineering (in top class) from Northeastern University, Shenyang, China, in 2011 and the M.Sc. degree in Artificial Intelligence (with full scholarship) from University of Amsterdam, The Netherlands, in 2013. Currently, he is a Ph.D. candidate at Eindhoven University of Technology, The Netherlands, and cooperates with the Vital Signs Camera project at Philips Research Eindhoven.

Wenjin Wang works on problems in computer vision, i.e., remote photoplethysmography (rPPG).

Albertus C. den Brinker received the M.Sc. and PhD degrees from Eindhoven University of Technology in 1983 and 1989, respectively. From 1987 to 1999 he worked in the Signal Processing Group at the Department of Electrical Engineering, at Eindhoven University of Technology. In 1999, he joined the Digital Signal Processing Group at Philips Research Laboratories Eindhoven, being active in the field of Signal Processing of Audio and Speech. One of the activities concerned standardization of audio coders, especially standardization within MPEG. Major contributions were made to MPEG-4 Amendment 2 and MPEG Surround. Later research concerned systems and algorithms for classification and interpretation of audio signals. The current activities include signal processing, data analysis and data mining for healthcare applications.

Sander Stuijk received his M.Sc. (with honors) in 2002 and his Ph.D. in 2007 from the Eindhoven University of Technology. He is currently an assistant professor in the Department of Electrical Engineering at Eindhoven University of Technology. He is also a visiting researcher at Philips Research Eindhoven working on bio-signal processing algorithms and their embedded implementations. His research focuses on modelling methods and mapping techniques for the design and synthesis of predictable systems with a particular interest into bio-signals.

Gerard de Haan received BSc, MSc, and PhD degrees from Delft University of Technology in 1977, 1979 and 1992, respectively. He joined Philips Research in 1979 to lead research projects in the area of video processing/analysis. From 1988 till 2007, he has additionally taught post-academic courses for the Philips Centre for Technical Training at various locations in Europe, Asia and the US. In 2000, he was appointed "Fellow" in the Video Processing & Analysis group of Philips Research Eindhoven, and "Full-Professor" at Eindhoven University of Technology. He has a particular interest in algorithms for motion estimation, video format conversion, image sequence analysis and computer vision. His work in these areas has resulted in 3 books, 3 book chapters, 180 scientific papers and more than 180 patent applications, and various commercially available ICs. He received 5 Best Paper Awards, the Gilles Holst Award, the IEEE Chester Sall Award, bronze, silver and gold patent medals, while his work on motion received the EISA European Video Innovation Award, and the Wall Street Journal Business Innovation Award. Gerard de Haan serves in the program committees of various international conferences on image/video processing and analysis, and has been a "Guest-Editor" for special issues of Elsevier, IEEE, and Springer.