Research Article | Vol. 13, No. 10 / 1 Oct 2022 / Biomedical Optics Express 5447
Biomedical Optics EXPRESS

RobustPPG: camera-based robust heart rate estimation using motion cancellation

AKASH KUMAR MAITY,[1,3,*] JIAN WANG,[2,3,4] ASHUTOSH SABHARWAL,[1,5] AND SHREE K. NAYAR[2,6]

[1]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA
[2]NYC Research Lab, Snap Inc., New York, NY 10036, USA
[3]Authors contributed equally
[4]jwang4@snap.com
[5]ashu@rice.edu
[6]snayar@snap.com
[*]akm8@rice.edu

Abstract: Camera-based heart rate measurement is becoming an attractive option as a non-contact modality for continuous remote health and engagement monitoring. However, reliable heart rate extraction from camera-based measurement is challenging in realistic scenarios, especially when the subject is moving. In this work, we develop a motion-robust algorithm, labeled RobustPPG, for extracting photoplethysmography signals (PPG) from face video and estimating the heart rate. Our key innovation is to explicitly model and generate motion distortions due to the movements of the person's face. We use inverse rendering to obtain the 3D shape and albedo of the face and environment lighting from video frames and then render the human face for each frame. The rendered face is similar to the original face but does not contain the heart rate signal; facial movements alone cause pixel intensity variation in the generated video frames. Finally, we use the generated motion distortion to filter the motion-induced measurements. We demonstrate that our approach performs better than the state-of-the-art methods in extracting a clean blood volume signal with over 2 dB signal quality improvement and 30% improvement in RMSE of estimated heart rate in intense motion scenarios.

1. Introduction

Vital signs provide a way to gauge the physical well-being of a person, especially in ambulatory care settings. Apart from health monitoring, measuring vital signs has multiple applications in human-computer interaction and driver monitoring systems. Research in remote vital signs monitoring has made significant progress over the last few years, mainly due to being a non-contact modality. Camera-based heart rate estimation utilizes photoplethysmography (PPG), where tiny blood volume changes underneath the skin surface due to the heart's pumping action is captured by a camera sensor [1]. However, there are significant challenges in extracting the heart rate from a camera-based remote photoplethysmography (rPPG) signal. The rPPG signal from the camera has extremely low signal strength and is corrupted by sensor noise and motion artifacts. In many cases, even small movements of the subject contribute to large artifacts in the camera measurements, which affects the accuracy of the estimated vital signs [2]. There have been significant advances in developing robust algorithms [3–9] for extracting vital signs like heart rate and breathing rate in low SNR regimes. However, most of these methods fail to deal with motion artifacts arising from intense movements, limiting their use to only in-lab settings.

In an ideal scenario, when the user remains stationary, and the light and the camera's position remain fixed, the skin pixel intensity variation consists of the clean artifacts-free rPPG signal. However, during any user's movements, the change in orientation of the skin surface results in unwanted motion contamination of the rPPG signal. The key approach in this work is to exploit a

#465143 https://doi.org/10.1364/BOE.465143
Journal © 2022 Received 27 May 2022; revised 3 Aug 2022; accepted 15 Aug 2022; published 26 Sep 2022

Biomedical Optics EXPRESS

light reflection model to relate facial movements to the motion artifacts in the measured rPPG signal. To the best of our knowledge, we are the first to explicitly generate motion artifacts based on image rendering. We use the generated motion artifacts, see Fig. 1, to filter the motion-induced rPPG signals using a bi-directional long short-term memory (Bi-LSTM) network.

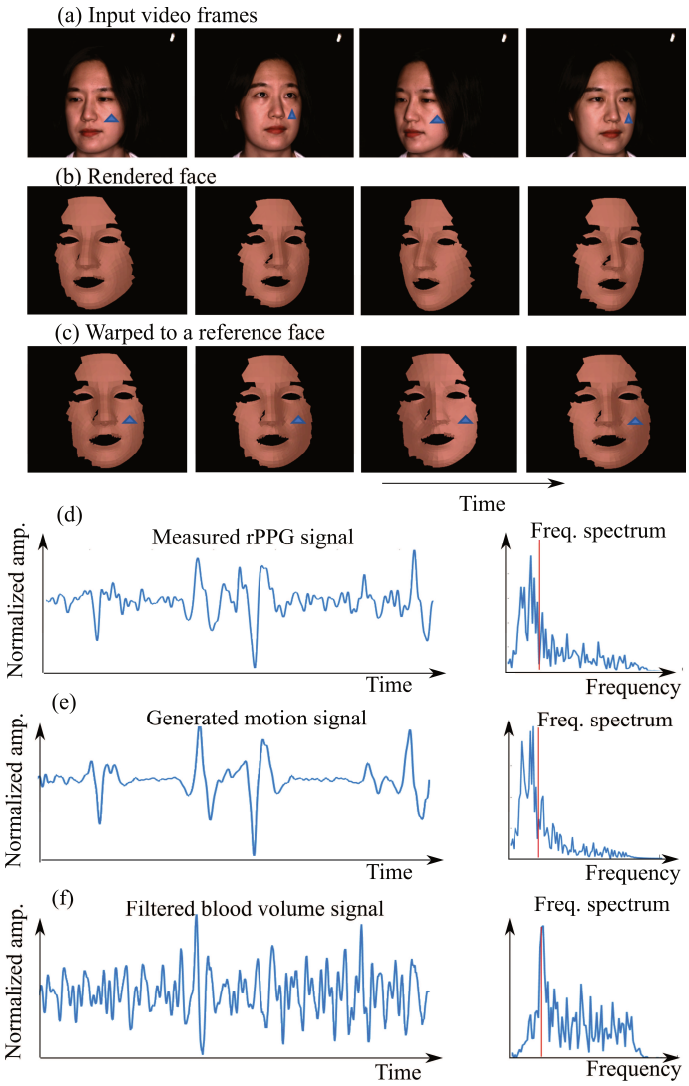

Fig. 1. A representative figure demonstrating our approach. We use a light reflection model to render the human face from the video. We use estimated 3D face shape and lighting information from video frames to render the human face and use the rendered face to cancel out motion distortions to extract a clean rPPG signal. (a) Input video frames of a subject. (b) Rendered face based on estimated 3D shape and lighting. (c) We register consecutive frames with respect to a reference frame. Note the intensity variation for different orientations of the face, *especially the left and right cheeks in the second and third frames here*. (d) Measured pixel intensity variation from a tracked point on the cheek from (a). (e) Motion signal from the corresponding pixel from (c). We then use the generated motion signal to cancel out motion distortion, resulting in a clean blood volume signal in (f). The red line in the frequency spectrum denotes the actual heart rate.

Biomedical Optics EXPRESS

To generate motion distortions, we need to know the face geometry and the scene lighting. On the one hand, one may use a depth camera to acquire the geometry of the human face during recordings. Similarly, the knowledge of scene lighting requires a pre-calibration setup in the same environment. However, obtaining prior knowledge of the face geometry and the scene lighting in practical scenarios might be infeasible. In this work, we use proxy measurements to build our prior knowledge. First, we obtain the 3D geometry of the face by using a 3D FaceMesh model, which gives an approximate geometry of the face at each instant. Secondly, we use a sequence of frames to estimate the scene illumination based on an approximate 3D face geometry. Next, we use image rendering-generated motion signals to filter the rPPG signals, leaving out a clean heart rate signal. Overall, the main contributions in the development of RobustPPG are:

1. We develop a framework that uses a 3D face model and scene illumination, estimated by inverse rendering, to explicitly model motion artifacts in camera-based rPPG signal. We use the generated motion signal to filter the motion distortions in the rPPG signals using a bi-directional long short-term memory (Bi-LSTM) network, resulting in a clean signal.

2. We show that our approach consistently outperforms existing state-of-the-art methods in the extracted rPPG signal quality and the estimated heart rate accuracy. Our method RobustPPG improves over 2 dB in signal quality in complex motion scenarios over the state-of-the-art methods. RobustPPG also improves heart rate estimation by 33% over the second-best method for intense movement scenarios.

3. We use an extended photometric stereo setup to validate the pipeline. The FaceMesh generated surface normals deviate on an average of 13° from the ground truth surface normals generated using photometric stereo. We show that even with an approximate face geommetry estimate, the normalized root mean squared error between the estimated motion signal using FaceMesh and the ground truth motion signals is less than 10%. The FaceMesh generated 3D facial geometry achieves near-optimal performance in terms of rPPG signal extraction.

We organize the rest of the paper as follows. First, we briefly describe the relevant prior work in Section 2. Next, we describe our proposed method for generating motion distortions for rPPG signal filtering in Section 3. We present our experimental results on datasets and comparisons against some state-of-the-art methods in Section 4. Finally, we discuss the limitations of our approach and scope for future work in Section 5.

2. Background and Key Challenges

For a camera-based heart rate estimation, we first record the video of a person's face. The temporal pixel intensity variations in the video capture the changing blood volume signal, also called the remote-PPG or rPPG signal. However, the rPPG signal has a very low signal strength. Therefore, the signal of interest is mainly dominated by camera sensor noise and large distortions that arise from a person's movements. The main objective is to develop a robust algorithm to recover the rPPG signal from the skin pixel intensity fluctuations in the video. The heart rate is then estimated from the recovered rPPG signal.

In practical scenarios, natural movements like talking or face tilting introduce significant unwanted distortions in the measured raw pixel intensity variation. Generally, a bandpass filter ($[0.5 - 5]$Hz) is suitable to filter out signals unrelated to the heart rate. However, for scenarios where the dominant frequency of motion distortion lies close to the heart rate, it is challenging to extract the rPPG signal and heart rate reliably from the pixel intensity variation alone. In such cases, any information about the motion signal may allow one to extract a very clean rPPG signal through a noise-cancellation process. However, obtaining an exact knowledge of the motion distortion is challenging due to the following reasons:

Biomedical Optics EXPRESS

1. Motion distortion depends on local surface orientation. To improve motion robustness in a camera-based system, the very first step is to use an automatic face detector and tracker [10,11] on the recorded video. The human face is not a rigid body because different facial regions move separately under natural movements. For example, when a person is talking or smiling, the cheek regions move differently than the forehead region. Secondly, even if each pixel on the face is tracked perfectly within frames, the tracked pixel intensity still contains motion distortions due to changes in the orientation of the skin surface. In essence, the pixel intensity variation depends on the changes in surface orientation during facial movements. Therefore, even if we track the different regions of the face separately, the motion distortions would be different for different facial areas, as shown in Fig. 2(a). Hence, knowledge of 3D facial geometry is important in generating motion distortions at various points on the face. One simple way of acquiring 3D face geometry is to use additional sensors to collect data. For example, one can use a depth sensor like structured light [12,13] or photometric stereo [14] to acquire 3D geometry of the human face. However, the use of any additional sensors limits the practicality of such an approach.

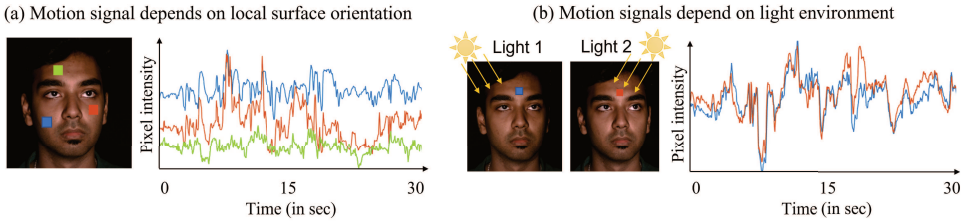

Fig. 2. Temporal motion signals variation due to a person's movements. (a) Motion signal depends on the location of the pixel and its local surface orientation. The pixel intensity variations for the left, right cheek, and forehead are different for facial movements. (b) Motion signal also depends on the environment light. For a point on the face, the same movements under different light directions cause different pixel intensity variations.

2. Motion distortion depends on the light environment. Lighting can change an object's appearance significantly. For a given point on the face, the pixel intensity variation due to movements will depend on the environment lighting, as shown in Fig. 2(b). Lighting estimation includes estimating the color and strength of rays of different directions at different locations on the face. Under some circumstances, one can make assumptions to make the task simpler. If the light source is far away from the object and the object's size is relatively small, it can be considered as distant lighting where each point on the object receives the same lighting (same light direction, intensity, and color). If the light source is nearby, it is near lighting [15]. Light fall-off should be considered. A scene point can receive light from many directions, but for a point with Lambertian reflectance, all the light rays can be linearly combined as one effective lighting. If the scene is illuminated by one distant light source, a mirror sphere [16] and a diffuse sphere [17] can be used to calibrate its direction and intensity, respectively. For effective lighting, a diffuse sphere or a rotated diffuse checkerboard [18] is enough. For more complicated lighting, one may use a mirror sphere to capture the high dynamic range whole environment lighting [19,20]. However, the use of a diffuse sphere or checkerboard might be infeasible in practical scenarios due to the hardware constraints. In the absence of additional hardware, human faces can be used as a light probe where shading [21], highlights and shadows [22,23] can be exploited, and can also be used to calibrate near lighting [15]. Even then, the rPPG signal fluctuations in the pixel intensity variations may affect the accuracy of the light estimation.

To bypass the challenges of generating reference motion signals, one of the most common approaches to achieving motion-robustness is to utilize some known properties of the unwanted distortions, the rPPG signal of interest, or both. Typically, a color camera captures images in red,

Biomedical Optics EXPRESS

green, and blue channels. The raw pixel intensity fluctuation in each channel contains different signal strengths of both the signal of interest and the distortions. The authors in [3,24] assume that the rPPG signal and the distortions arising from a person's movements are uncorrelated or independent of each other. Hence, blind source separation (BSS) techniques like ICA and PCA are used to extract a clean rPPG signal from the observed RGB channels in a face video. Methods like 2SR [25], CHROM [26], PBV [27] and POS [28] assume prior knowledge of the strength of blood pulsation in each color channel and use this prior information to weigh the observed raw pixel intensity fluctuations in the color channels. More recent approaches [29,30] utilize *explicitly* generated reference signals to filter out unwanted distortions in the pixel intensity fluctuations, leaving out a clean rPPG signal. In [29], the authors note that the location of a facial point changes according to the facial movements of a person, and hence the temporal change in tracked pixel coordinates can be used as a surrogate measure of motion distortions. The approach in [30] uses *distraction regions* from inverse attention mask to generate distortion signals. These distraction regions consist of subject hair or the background that do not contain any physiological signals. Although the recent approaches show significant robustness improvement in extracting the rPPG signal, analysis of the origin of motion distortions remains lacking.

3. Proposed Method: RobustPPG

In this work, we use **inverse rendering** to explicitly generate the motion distortions from a video. First, we estimate the geometry and albedo of human faces and the lighting. Next, we render the human face for each frame as the reference motion distortion since the rendered faces do not include rPPG signals. We then use the reference motion signals to cancel distortions in the pixel intensity variation. We name this approach RobustPPG, as illustrated in Fig. 3, and introduce our motion-signal model in this section. We then describe our rPPG extraction algorithm that consists of two main subparts - i) generating reference motion signals by face rendering and ii) using the reference signals to filter the motion distorted rPPG signals.

For the rest of the paper, we use notations as the following, where "a" and "b" as examples, a as a scalar, $\mathbf{a}$ as a vector, $\mathbf{A}$ as a matrix, $\mathbf{Ab}$ as matrix-vector product, $\mathbf{AB}$ as matrix-matrix product, $\mathbf{a} \cdot \mathbf{b}$ as the inner product of the two vectors with the same length, $*$ as a product between scalars, vectors and matrices (it can be omitted but it is sometimes used for separation; *if it is with the inner product, it has lower priority to avoid parentheses*), $\mathbf{A} \odot \mathbf{B}$ as element-wise product between two matrices or vectors, and $\mathbf{A}^\top$ as transpose of matrix $\mathbf{A}$.

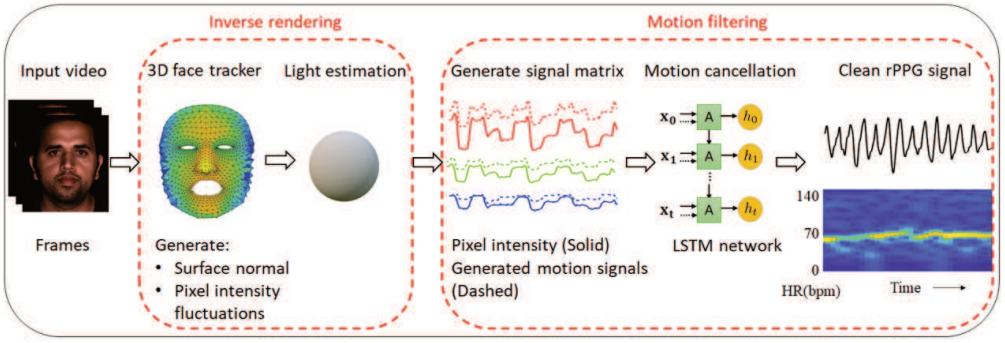

Fig. 3. Flowchart of RobustPPG. We use **inverse rendering** to explicitly generate motion distortions with high accuracy. First, we use FaceMesh, a 3D face tracker, to get 3D face geometry per frame. Next, we estimate light direction using the 3D geometry of the human face and generate accurate motion distortions at each triangle location. We then simultaneously use the motion signals and the corrupted raw pixel intensity fluctuations in a Bi-LSTM architecture to obtain clean filtered rPPG signals.

Biomedical Optics EXPRESS

3.1. Motion signal model

According to the Dichromatic Reflection Model (DRM) [31], the RGB pixel intensity at any 3D location $\mathbf{r}$ on the surface and at time t can be described as a sum of diffuse and specular components,

$$\mathbf{i}(\mathbf{r}, t) = \mathbf{i}_{\text{diffuse}}(\mathbf{r}, t) + \mathbf{i}_{\text{specular}}(\mathbf{r}, t), \tag{1}$$

where $\mathbf{i}_{\text{diffuse}}(\mathbf{r}, t)$ and $\mathbf{i}_{\text{specular}}(\mathbf{r}, t)$ represent the diffuse component and the specular component, respectively, and both $\in \mathbb{R}^{3\times 1}$. The diffuse component models subsurface light interaction with the human skin and tissue, and can be modeled as follows,

$$\mathbf{i}_{\text{diffuse}}(\mathbf{r}, t) = i_0(t) * (\mathbf{c} + \mathbf{e} * p(t)) * \mathbf{n}(\mathbf{r}, t) \cdot \mathbf{l}(\mathbf{r}, t), \tag{2}$$

$$\mathbf{i}_{\text{diffuse}}(\mathbf{r}, t) = \mathbf{c}_{\text{cam}} \odot \mathbf{c}_{\text{light}} \odot (\mathbf{c}_{\text{skin}}(\mathbf{r}) * \alpha + \mathbf{e}_{\text{ppg}} * p(t)) * l * \mathbf{n}(\mathbf{r}, t) \cdot \mathbf{l}(\mathbf{r}, t), \tag{3}$$

where all vectors $\in \mathbb{R}^{3\times 1}$; $\mathbf{c}_{\text{cam}}$, $\mathbf{c}_{\text{light}}$ and $\mathbf{c}_{\text{skin}}(\mathbf{r})$ are color response of the camera, color of the light and color of the skin, respectively, and all are unit vectors; the pulsatile blood volume signal is given by $p(t)$, and $\mathbf{e}_{\text{ppg}}$ denotes the strength of the pulsation in the color channels; α is albedo of the skin; l is light intensity; $\mathbf{n}(\mathbf{r}, t)$ and $\mathbf{l}(\mathbf{r}, t)$ are the time-varying unit vectors of surface normal and light source direction at a point located at $\mathbf{r}$ at time t, respectively.

Here, we make the following assumptions (as illustrated in Fig. 4):

1. The light source is a point source located far away from the human face and the position of the light source and the camera remains constant at all times. Hence, the light source direction is parallel and independent of position and time at all locations. The intensity of the light source l also remains constant with time.

2. A diffuse Lambertian object can entirely model the human face, and the specularities, if present, are sparse and hence can be ignored. In addition, even if multiple point sources are present in the scene, all the sources can be modeled as one single effective point source under the assumption of a Lambertian surface.

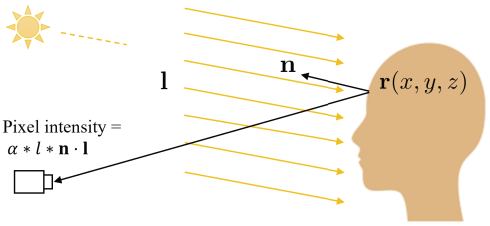

Fig. 4. We assume that (1) light source is located far away from the human face and then each point on the face receives the same lighting direction and intensity, and (2) human face has Lambertian reflectance.

Under these assumptions, substituting Eq. (3) in Eq. (1) and combining terms, we get,

$$\mathbf{i}(\mathbf{r}, t) = \mathbf{c} * \mathbf{n}(\mathbf{r}, t) \cdot \mathbf{l} + \mathbf{e} * p(t) \odot (\mathbf{c} * \mathbf{n}(\mathbf{r}, t) \cdot \mathbf{l}), \tag{4}$$

where $\mathbf{e} = \mathbf{e}_{\text{ppg}}/(\alpha * \mathbf{c}_{\text{skin}}(\mathbf{r}))$, $\mathbf{c} = \mathbf{c}_{\text{cam}} \odot \mathbf{c}_{\text{light}} \odot \mathbf{c}_{\text{skin}}(\mathbf{r}) * \alpha * l$, and is skin's RGB intensity in the camera without the shading term. We will refer $\mathbf{c}$ as skin color in short in this paper. In Eq. (4), we note that the first term on the right side of the equation is dominant and contributes to motion distortions in the measured rPPG signal $\mathbf{i}(\mathbf{r}, t)$.

Biomedical Optics EXPRESS

3.2. Generating motion signal

In order to extract our signal of interest $p(t)$ in Eq. (4), our approach is to generate the motion distortions first. Generating the reference motion distortion requires knowledge of three parameters: i) the surface normal direction $\mathbf{n}(\mathbf{r}, t)$, ii) direction of effective light source $\mathbf{l}$, and iii) the average skin color $\mathbf{c}$ which remains constant for a person over time.

3.2.1. 3D face modelling

In an ideal scenario, we can get surface normal directions of the human face using a depth-enabled camera, which might be infeasible in a practical scenario. A more practical way is to use 3D Morphable Models (3DMM) for face fitting, which outputs the approximate 3D facial geometry for each frame in the video. In this work, we use FaceMesh [32,33] for face tracking and fitting in each frame of a video. One may use other methods like [11,34–36]. First, the face is detected and tracked across each frame for a given video. Then facial landmarks are detected in each frame. Next, 3DMM is used for face fitting and generating 3D face geometry and texture. For each frame, the fitting process generates dense triangular meshes, the location of which is used to estimate surface normal at each triangle centroid in the mesh, as shown in Fig. 5. We compute an average of pixel intensities inside each triangle locally. Hence for each video, we have surface normal measurements $\mathbf{N} \in \mathbb{R}^{K\times T\times 3}$, where K is the number of triangles in each frame, T is the total number of frames in the video, the third dimension represents the x, y and z component of the surface normals. Similarly, we have intensity measurements $\mathbf{I} \in \mathbb{R}^{K\times T\times 3}$, where the third dimension contains the pixel intensities in red, green, and blue channels.

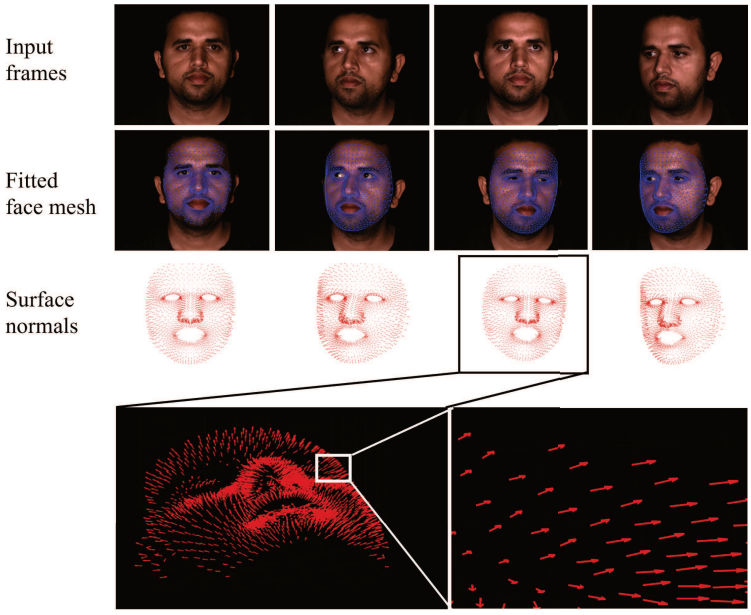

Fig. 5. For given frames (top row), the fitted face mesh is shown (second row). We compute surface normals (third row) at each triangle centroid from their 3D location.

3.2.2. Lighting estimation

Next, we need to estimate the effective light source direction $\mathbf{l}$ from the given video. For the estimation process, we use measurement from all the triangular meshes on the face (excluding highlights, lips and hair region) for a sequence of frames. Writing Eq. (4) in matrix multiplication

Biomedical Optics EXPRESS

form by ignoring the rPPG part, we have

$$\mathbf{I} = \mathbf{N} * \mathbf{l} * \mathbf{c}^{\top}, \tag{5}$$

where $\mathbf{I} \in \mathbb{R}^{K \times T_w \times 3}$ is the pixel intensity measurement, $\mathbf{N} \in \mathbb{R}^{K \times T_w \times 3}$ is the surface normal directions, and T_w denotes the number of frames. The effective light source direction $\mathbf{l} \in \mathbb{R}^{3 \times 1}$ and average skin color (assuming skin color at different locations is similar) $\mathbf{c} \in \mathbb{R}^{3 \times 1}$ need to be estimated. Typically, $p(t)$ in Eq. (4) is considered to be a zero mean quasi-periodic signal. Hence we consider a time window larger than a period of $p(t)$, so that the parameter $\mathbf{c}$ in Eq. (5) is a close approximate of the average skin color. Using Eq. (5), $\mathbf{l} * \mathbf{c}^{\top} \in \mathbb{R}^{3 \times 3}$ is obtained by solving a least square problem in an over-determined system. Here the matrix $\mathbf{l} * \mathbf{c}^{\top}$ is a low-rank matrix in simple lighting environment. Hence we compute Singular Value Decomposition (SVD) on the estimated matrix $\mathbf{l} * \mathbf{c}^{\top}$ resulting in an estimated light source direction $\widehat{\mathbf{l}}$ and average skin color $\widehat{\mathbf{c}}$. In essence, we only estimate $\mathbf{l}$ for calculating light estimation accuracy; we do not need to separately estimate $\mathbf{l}$ and $\mathbf{c}$ for motion signal generation because we only need $\mathbf{l} * \mathbf{c}^{\top}$ for generating motion signals.

3.2.3. Generating signal matrix

Once we estimate the effective lighting direction $\widehat{\mathbf{l}}$ and the average skin color $\widehat{\mathbf{c}}$, we generate the motion signal $\mathbf{m}(\mathbf{r}, t)$ in each triangle location $\mathbf{r}$ as follows (illustrated in Fig. 6),

$$\mathbf{m}(\mathbf{r}, t) = \widehat{\mathbf{c}} * \mathbf{n}(\mathbf{r}, t) \cdot \widehat{\mathbf{l}} = (\mathbf{n}(\mathbf{r}, t)^{\top} * \widehat{\mathbf{l}} * \widehat{\mathbf{c}}^{\top})^{\top}. \tag{6}$$

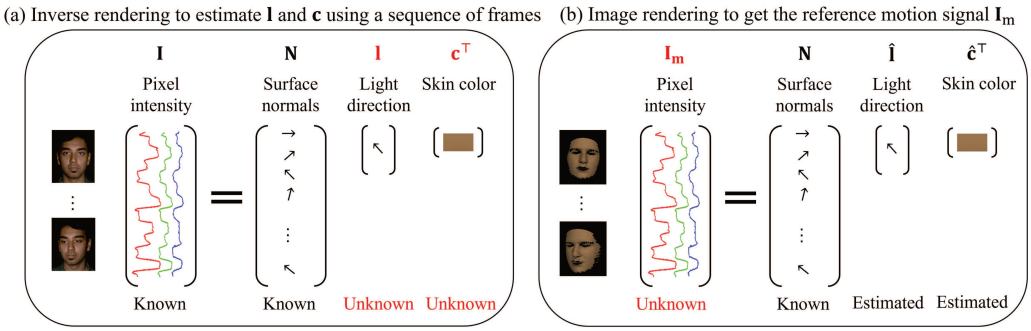

Fig. 6. Image rendering for motion signal generation. We estimate skin color and effective lighting direction using a sequence of frames while lighting does not change (a), and render face for each frame, where the rendered faces have motion-caused pixel intensity variation but do not have rPPG signal. Here, $\mathbf{c}$ = (skin albedo (scalar) * skin color (unit vector)) $\odot$ (light strength (scalar) * light color (unit vector)).

Therefore, for each triangle on the face, we have six signals: three measured RGB pixel intensities $(i_{red}(t), i_{green}(t), i_{blue}(t))$ and three RGB motion signal that we synthetically generate $(m_{red}(t), m_{green}(t), m_{blue}(t))$. According to our proposed motion model, we rewrite Eq. (4) as a function of temporal motion distortions $\mathbf{m}(t)$ as follows,

$$\mathbf{i}(\mathbf{r}, t) = \mathbf{m}(\mathbf{r}, t) + \mathbf{e} * p(t) \odot \mathbf{m}(\mathbf{r}, t), \tag{7}$$

where $\mathbf{i}(\mathbf{r}, t)$ is the motion contaminated rPPG signal, $\mathbf{m}(\mathbf{r}, t)$ is the motion signal distortion, and $p(t)$ is the clean rPPG signal. Using the generated motion signals, we construct a signal feature matrix $\mathbf{S_r}$ by composing $\mathbf{S_r} = [i_{red}(t), i_{green}(t), i_{blue}(t), m_{red}(t), m_{green}(t), m_{blue}(t)]^{\top}$, where each $\mathbf{S_r} \in \mathbb{R}^{6 \times t}$.

Biomedical Optics EXPRESS

3.3. Motion cancellation of rPPG signals

The objective of the filtering process is to estimate a clean PPG signal with the help of estimated reference motion corruption signals. In other words, extract a clean PPG signal from the signal matrix $\mathbf{S_r}$. Previously, blind signal separation techniques were used to separate different signal sources, in this case, two sources being the rPPG signal and unwanted motion signals. In [29], the authors proposed a weighting vector based on a discriminative signature to extract the rPPG signal. Recently, recurrent neural networks like long short-term memory (LSTM) have been used on PPG signals as a signal denoising technique [30,37] and signal quality assessment [38]. In this work, we use a bi-directional long short-term memory (Bi-LSTM) network to filter the motion-distorted pixel intensity signal with the help of generated motion signals. The matrix $\mathbf{S_r}$ containing the pixel intensities and the reference motion signals are used as input to the Bi-LSTM network. The contact-based pulse-oximeter waveforms serve as ground truth labels for training the network. The objective of the Bi-LSTM network is to use the generated motion signals to filter out distortions in the observed pixel intensity signals and obtain a clean rPPG signal.

For the architecture, we use a 3-layer Bi-LSTM network with 30 hidden units. We split each of the signals into 4 seconds window with an overlap of 2 seconds, and the segments are then used as input to the network. We train the architecture with Adam optimizer, which optimized the mean squared error (MSE) loss between the predicted rPPG signal and the target ground-truth waveforms. The overall flowchart for our proposed approach is demonstrated in Fig. 3.

3.4. Other implementation details

We have assumed that a Lambertian surface can completely model the human face. However, oil or sweat makes the skin specular, violating the assumption. As a preprocessing step, we remove triangle meshes that correspond to any facial hair. We remove the triangles by a thresholding step, where we ignore meshes with extremely low pixel values that may indicate the presence of facial hair. Secondly, we also disregard pixels from the eye and the lip regions since the eyes and lips do not contain any heart rate information. For light estimation, we use a weighting mask to assign less importance to triangles on non-Lambertian surfaces, such as facial hair or specular regions. We use pixel color transformation to chromaticity space to obtain the weighting mask.

4. Experimental Evaluation of RobustPPG

We perform an extensive set of experiments with two main objectives - i) check the accuracy of intermediate parameter estimation such as light direction and ii) quantify the quality of rPPG signal extraction compared with other state-of-the-art methods.

4.1. FaceMesh Validation

In our proposed approach, the accuracy of the estimated face geometry obtained from the FaceMesh face tracker determines the accuracy of the estimated lighting, both of which affect the quality of synthetic motion signal generation. Therefore, the main goal of this section is to use a photometric stereo setup to validate FaceMesh. We design the experiments in this section to i) quantify the accuracy of the face geometry estimated using FaceMesh, ii) quantify the error in estimated lighting direction using FaceMesh, iii) evaluate the quality of motion signal generated using both mannequin and human subjects, and iv) evaluate the influence of inaccuracy in motion generation on the rPPG signal estimation.

4.1.1. Photometric stereo setup

To validate FaceMesh, we resort to photometric stereo to obtain ground truth 3D face geometry during movements. Generally, a photometric stereo setup is used to accurately estimate an object's 3D geometry. In a typical photometric stereo setup, the object is illuminated by three or

Biomedical Optics EXPRESS

more lights kept at different positions. Then, the surface normal is estimated from the camera images, given the prior knowledge of the lighting environment. It has been shown [39] that photometric stereo can obtain a highly detailed structure of the human face.

We use the photometric stereo setup to obtain continuous face orientation changes during facial movements. In our experimental setup, as shown in Fig. 7, we use three LEDs at three corners of a triangular geometry with the camera at the center of the triangle. The camera, which is kept at a distance of d = 1.8 m, records continuous movements of a subject at 60 fps. The LEDs are synchronized with the camera, such that three consecutive frames of the camera corresponds to the subject illuminated from three LEDs at different positions. We assume face is static during three consecutive frames. The images corresponding to a single LED position are effectively captured at $60/3 = 20$ fps.

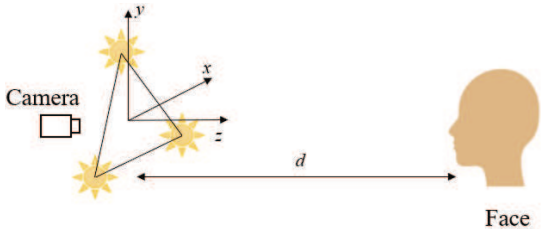

Fig. 7. Experimental setup for validating FaceMesh using photometric stereo. Photometric stereo can get ground truth surface normal. Facemesh can get an approximate. We want to know how accurate the surface normal by FaceMesh is, and what the influence on the PPG estimation based on the slight inaccuracy in the surface normal.

We capture videos of both mannequins and human subjects under various head movements. The advantage of using a mannequin is that it is devoid of any physiological signals; hence any pixel variation on the face is solely due to the movements of the mannequins. We capture four videos of two different mannequins under different head rotational movements, with each video corresponding to 30 seconds duration. However, realistic scenarios consist of complex facial movements that are far more complicated than simple rotational and translation movements and cannot be achieved with a mannequin alone. Therefore, we collect our dataset on human subjects with realistic facial and intense head movements. We further elaborate on the methodology of human experiments later in section 4.2.1.

Light calibration. We use a diffuse checkerboard to calibrate the light strength and direction of each LED like [18]. The image formation for a diffuse object is pixel intensity = albedo * light strength * (surface normal · light direction). In photometric stereo, we change light direction to calculate surface normal. In our light calibration, we change surface normal to get light information. More specifically, we do the geometric calibration for the camera first. Then, we turn on one LED. Next, we put the diffuse checkerboard at the reference location where head will be placed and rotate it while the camera captures the images. From each image, we can calculate the pixel values of the white grids and the surface normal of the checkerboard (because the camera is calibrated and the dimension of the checkerboard is known). By rotating the checkerboard at least three times, we can calibrate light strength and light direction of the LED. We repeat this process for other LEDs. The light calibration process gives us lighting matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$ where each vector corresponds to each light.

4.2. 3D Face geometry estimation

We use the captured images in the photometric stereo setup to obtain the ground truth of 3D face geometry. We use FaceMesh to track a subject's face for each frame in a video and average the pixel intensities within each triangle in the mesh for further analysis. For each triangle location m,

Biomedical Optics EXPRESS

the averaged pixel intensities from three consecutive frames (three consecutive frames correspond to three different LED positions) are then used to obtain an estimate of surface normals $N_{m;t}$ at time instant t, for a pre-calibrated lighting matrix $\mathbf{U}$. An example of the estimated face geometry obtained from FaceMesh N_{fm}, as well as photometric stereo N_{ps}, is shown in Fig. 8 for both mannequins and human subjects. We compute the angular error between the surface normals N_{fm} and N_{ps} at each triangle and time instant. From the error map in Fig. 8, we observe that the error is maximum in the nose region. The mean angular error computed across the triangle location (excluding the eyes, nose and mouth region) and time is 13.8° for mannequin videos and 18.37° for the human dataset. The larger angular error for real human datasets is mainly attributed to cases where some of our assumptions fail, such as the presence of specular regions and uneven albedo distribution. The facial movements will result in surface normal variation over time. To observe the effects of motion on surface normal variation, we compute the angular variation of both $N_{fm,t}$ and $N_{ps,t}$ at time instants t with respect to surface normals at $t = 0$, *i.e.* $N_{fm,0}$ and $N_{ps,0}$, respectively. We show the temporal angular variation from one triangle on the forehead of a mannequin during rotational movements and a human subject during a talking scenario in Fig. 8. There is a good agreement between the surface normal variation estimated from the two methods. Despite the large mean angular error of FaceMesh (Fig. 8 top row), FaceMesh is able to capture the temporal surface normal variation accurately, with the mean angular variation difference being less than 5°.

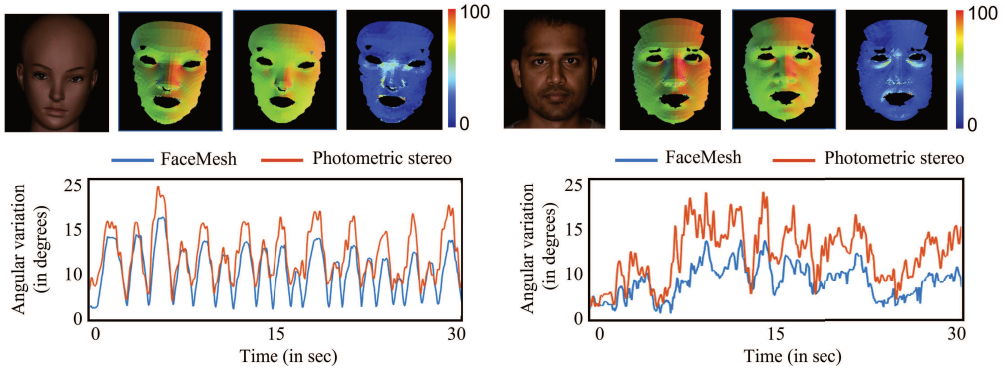

Fig. 8. Examples of face geometry estimated using FaceMesh vs. photometric stereo (PS) for a mannequin (left) and a human face (right) on a single frame (top row, 2nd and 3rd sub-figures in each group). The error map in degrees indicates unreliable estimates in nose, mouth and eye regions (top row, 4th sub-figure in each group). One example of the temporal angular variation of surface normal using FaceMesh and photometric stereo with respect to the first frame of the video (bottom row) is shown. The FaceMesh estimated surface normal during facial movements closely follows the surface normal variation using photometric stereo.

4.3. *Light environment estimation accuracy*

Next, we use the estimated 3D geometry from FaceMesh to obtain an estimate of the lighting matrix $\widehat{\mathbf{U}}$. The average error on four mannequin videos is 4.56° between the estimated lighting direction and the actual lighting direction for all the three LEDs used in our setup. Although we show that we can accurately estimate the lighting in the case of point sources, our approach can be extended to various lighting conditions. Furthermore, as demonstrated later, our approach can be used to estimate the rPPG signal accurately in uncontrolled settings as well.

Biomedical Optics EXPRESS

4.4. Motion signal generation

Given the knowledge of accurate 3D geometry of the face for every frame, we should be able to generate motion signals accurately. To validate our hypothesis, we use the generated 3D face geometry to simulate the motion signal corresponding to the pixel intensity from a single lighting position, as shown in Fig. 6. We show one example of a synthetic motion signal generated from photometric stereo and FaceMesh and the actual pixel intensity signal from a single triangle in Fig. 9. Next, we filter the motion signals using a bandpass filter ($[0.5 - 5]$Hz) since the human heart rate belongs in this frequency range. To evaluate the quality of estimated surface normals, we calculate two metrics - i) normalized root mean square error (NRMSE) and ii) normalized cross-correlation (NCC) between estimated motion signals and the actual pixel intensity from the mannequin videos, reported in Table 1. We observe that the generated motion signals using face geometry N_{ps} obtained from photometric stereo provided the best estimate of the actual motion signal, with a small NRMSE error and high correlation coefficient. FaceMesh still offers a good estimate of the reference motion signal with the NRMSE being less than 10%, although the performance is inferior to that obtained using photometric stereo. The slightly inferior performance of FaceMesh is expected since FaceMesh provides only an approximate estimate of the surface geometry of the human face. If the motion signals generated are very accurate, the rPPG signal, after a filtering process, ideally should not contain any motion distortions. We discuss the detail of our rPPG filtering technique in the next section. As a sanity check, we use our filtering technique to estimate residuals in mannequin videos using both rendered motion signals from FaceMesh and photometric stereo. A mannequin is devoid of any physiological signal; hence the residual signal after filtering should not contain any strong signals. The last row of Fig. 9 shows the residual signal obtained after the filtering step. Rightly, the filtering step successfully gets rid of any motion signal, as evident from the residual signal spectrogram. Additionally, we observe that the difference in spectrogram between the two methods is negligible, which validates that the performance of FaceMesh is good enough to generate motion signals with comparable accuracy.

Table 1. Accuracy of generated motion distortions by using 3D face geometry obtained from photometric stereo and FaceMesh.

Metric	Photometric stereo	FaceMesh
NRMSE (%)	5.6	8.03
NCC	0.9954	0.9930

4.5. rPPG signal estimation: FaceMesh vs. photometric stereo

Any error in the motion signal generated by FaceMesh may result in quality deterioration in the extracted rPPG signal in the case of human subjects. To quantify the performance deterioration due to slight inaccuracy in the generated motion signal, we extract the rPPG signal using surface geometry from both photometric stereo and FaceMesh. We observe that using a motion signal generated by photometric stereo results in an insignificant improvement of 0.15 dB ($p>0.05$) over FaceMesh with respect to the average signal-to-noise ratio of the estimated rPPG signal. Hence, the motion signals generated using FaceMesh achieve near optimal performance in terms of rPPG signal extraction.

Biomedical Optics EXPRESS

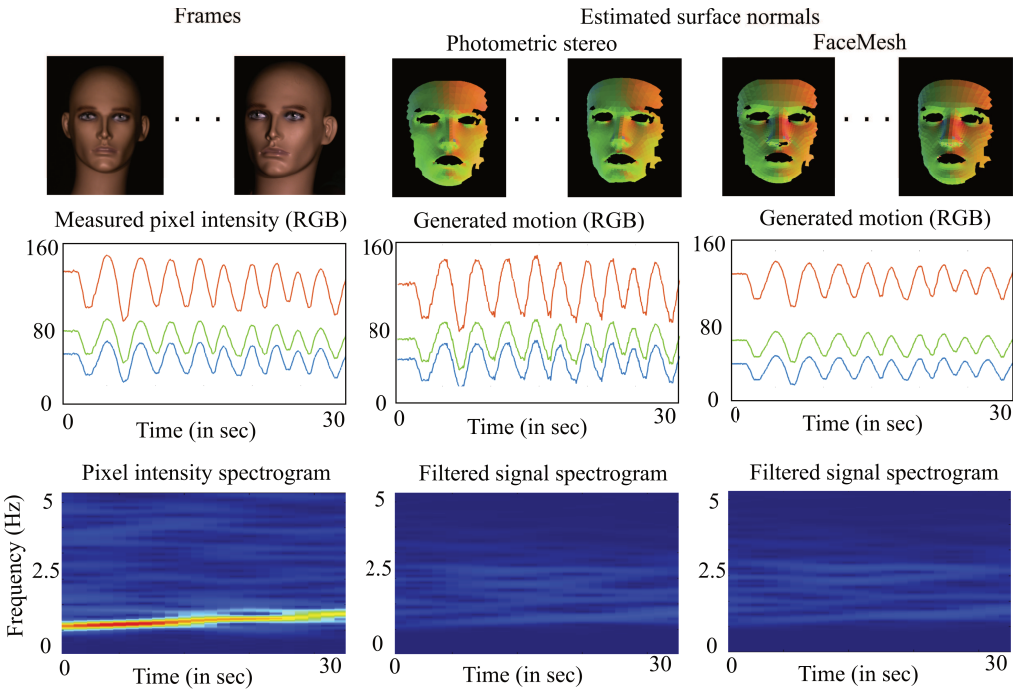

Fig. 9. An example of motion signal generated by photometric stereo and FaceMesh. Both methods' synthetic motion signals match the pixel intensity variations from a single triangle on the mannequin (second row). The pixel intensity has a strong signal due to the movements of the mannequin, as evident from the spectrogram. However, the filtered residual signal's spectrogram is devoid of any strong component due to motion distortion in the pixel intensity variation.

5. PPG Signal Estimation

Next, we discuss the details of our dataset to evaluate our proposed approach and compare the performance against some existing state-of-the-art methods.

5.1. Dataset

Our approach is mainly geared toward a reliable estimate of PPG signals under complicated motion scenarios. To validate our approach, we use a publicly available PURE dataset [40,41], which is one of the few datasets to contain complicated motion scenarios. The PURE dataset contains face videos of ten participants with synchronized physiological data under six motion conditions, including rotation of the head and talking. The duration of the video is around one minute, and the ground truth PPG waveform is provided by a contact-based pulse-oximeter.

To validate our approach on intense motion scenarios, we create a separate dataset, which consists of videos from 12 subjects (9 male, 3 female) captured in our experimental setup, as shown in Fig. 7. For each subject, we collected two motion scenarios, each of duration 2 minutes in the calibrated lighting environment. For each motion scenario, 3 videos are captured corresponding to 3 lighting positions, resulting in 6 videos per subject. Our total dataset consists of 72 videos. The first motion scenario consisted of fast head rotational movements, and the second was a 2 minutes clip of the subjects talking in front of the camera with natural expressions and movements. A pulse-oximeter (CMS50D+) was used on the finger to record the PPG signal synchronously which serves as the ground truth PPG signal. These dataset collections were

Biomedical Optics EXPRESS

approved by the Rice University Institutional Review Board (Protocol number: 14-45E, Approval Date: 3/04/2014). We name the dataset as RICE-motion dataset [42].

We also collected several toy videos in indoor and outdoor settings to evaluate our proposed approach in practical lighting scenarios. The videos are each one minute, captured by a hand-held smartphone camera. In addition, we record the user's heart rate before and after the video recording using a standard pulse-oximeter for ground truth heart rate.

5.2. Training and validating the motion cancellation network

We use our generated signal matrices $\mathbf{S_r}$ as described in Section 3.2.3 as input features to our Bi-LSTM architecture. The trained model learns the function that relates the pixel intensity variation with the motion distortions and the physiological signal of interest. Since the size of the real datasets used in this work is comparatively small, we generate a synthetic dataset by generating various motion distortions in the rPPG signals for training the network.

Synthetic motion dataset: We use Eq. (7) to generate synthetic signal matrices. The clean PPG signal $p(t)$ is generated using a parametric model as in [43], where heart rate is used as the key parameter in generating a clean PPG signal waveform. The heart rate is randomly chosen from a uniform distribution from 30 bpm to 240 bpm, and a clean PPG signal of 30 seconds duration is generated corresponding to the heart rate. We use a random Brownian noise generator to generate the reference motion signal $\mathbf{m}(t)$. Finally, we add random white noise to the pixel intensity variation that simulates the modeling error and the camera sensor noise. The parameter $\mathbf{e}_{\text{ppg}} = [0.18, 0.78, 0.60]$ is kept constant for each sample in the synthetic dataset. We generate 400 samples of motion corrupted signals, each of 30 seconds duration. The signal matrix $\mathbf{S_r}$ is then generated using the synthetic RGB pixel intensities $\mathbf{i}(t)$ and the motion signals $\mathbf{m}(t)$.

We standardize the signals by subtracting the temporal mean and then dividing by the temporal standard deviation of signals in each time window. The signals are then filtered using a bandpass filter ($[0.5 - 5]$Hz) to reject any out-of-band distortions. We then combine the real and synthetic datasets for training our motion cancellation network. In this work, we adopt the subject-specific training and test via cross-validation approach. For a particular subject, we train the network on combined simulated and real data from all the remaining subjects for all motion scenarios in the dataset. We then use region-specific signal matrices $\mathbf{S_r}$ for the corresponding subject to extract rPPG signals from each triangle location $\mathbf{r}$. We then obtain the overall rPPG signal by spatially averaging the model-predicted rPPG signal from all the triangle locations on the face. The cross-validation approach is repeated for all the subjects in the dataset, and the average error across all the subjects is reported.

5.3. Performance comparison

In order to validate the effectiveness of our approach, we utilize our rendered motion signal in both Bi-LSTM architecture (RobustPPG) and a discriminative filtering technique introduced in [29] and name the variant DRPPG (Discriminative RobustPPG). Note that both RobustPPG and DRPPG belong to the same family where motion signals are generated by inverse rendering. We compare our two approaches with three prior state-of-the-art methods, namely, DIS-M which uses the coordinate location of face tracker [29], POS [28] and CHROM [26]. We also implement a convolution attention network (CAN-MTTS) [44] to extract rPPG signal from phone videos. We keep the length and overlap of the time windows fixed for all methods for fair comparison.

To evaluate the quality of the estimated PPG signal, we compute the signal-to-noise ratio (SNR) of the rPPG signal extracted using different approaches. Furthermore, we also compute heart rate based on the extracted rPPG signal. Instead of using the entire duration of video recording to estimate the average heart rate, we use short overlapping windows of 5 seconds and 1 second overlap to compute instantaneous heart rate. We then compute the average root mean square error (RMSE) between the heart rate estimates and the ground truth for all subjects.

Biomedical Optics EXPRESS

5.4. Results and discussion

We show examples of the generated motion distortions from one triangle located on the forehead of subjects from both PURE and RICE-Motion dataset in Fig. 10. For the PURE dataset, we show results of a subject, but we use a different subject's face to show the movements, due to the dataset terms and rules for privacy concerns. The same applies for Fig. 11. For both datasets, we observe that the generated motion distortion has a good correlation with the overall trend in the pixel intensity fluctuations recorded by the camera, even for complicated movement of the subjects.

The average SNR values across all subjects are reported in Table 2 for all six types of movements for the PURE dataset. Here, we make some key observations. First, for stationary or simple movements like slow translation and rotations, the SNR for all the methods remains fairly consistent. However, in complex facial movements like in the talking scenario, the SNR decreases for all the methods. Secondly, RobustPPG performs consistently better than the other methods in all motion scenarios, with the maximum improvement in the talking scenario. The consistently better performance of RobustPPG may be attributed to the accurate generation of reference motion distortions for complicated movements. We show the spectrogram of the blood volume signal estimated from some of the methods in Fig. 11 against the ground truth spectrogram for both PURE and RICE-motion dataset. We observe that the heart rate signal in the spectrogram estimated using RobustPPG is much cleaner than the other methods. The clinical metrics associated with heart rate variability (HRV) can be computed from the extracted heart rate trend. We also report the average RMSE error across different motion scenarios in Table 2. The results show that in simple motion scenarios, almost all methods have similar accuracy in terms of predicting the heart rate with CHROM being the best of them. However, for complicated motions like head rotations and talking, RobustPPG clearly better predicts the heart rate than all the other methods. The improvement in error for RobustPPG is more than 1 bpm above the DIS-M (second-best prior method) for the talking scenario. Therefore, our approach can provide a more reliable measure of HRV in addition to the average heart rate.

Table 2. Signal-to-noise ratio and heart rate RMSE on PURE dataset which contains different motion scenarios. Our methods show improvement in the talking scenario.

	Signal-to-noise ratio (dB)					Heart Rate RMSE (bpm)				
Movement	RobustPPG (Ours)	DRPPG (Ours)	DIS-M	POS	CHROM	RobustPPG (Ours)	DRPPG (Ours)	DIS-M	POS	CHROM
Static	6.34	5.14	5.13	7.76	**7.85**	1.56	1.92	1.92	1.41	**1.37**
Slow trans.	7.56	5.73	4.97	**7.77**	7.69	1.65	1.87	2.85	**1.62**	1.63
Fast trans.	**7.64**	6.92	5.99	6.26	6.75	**1.82**	2.01	2.15	2.13	2.04
Slow rot.	**7.72**	4.65	4.20	7.18	7.71	1.71	1.81	1.85	1.78	**1.69**
Fast rot.	**7.54**	6.69	5.91	5.36	6.58	**1.72**	1.85	1.99	2.54	1.90
Talking	**3.35**	1.87	1.29	0.98	1.06	**2.38**	2.99	3.27	5.43	3.55

We report the SNR values as well as the RMSE heart rate on our RICE-Motion dataset, which is more challenging than PURE, in Table 3. Although in the RICE-Motion dataset, the face occupies a larger portion of the frame as compared to the PURE dataset, the subjects are asked to perform more vigorous head and facial movements for the RICE-Motion dataset. Hence, the motion contamination is stronger in the RICE-Motion dataset compared to the PURE dataset. In the case of heavy rotational movements, we observe around 3 dB improvement in SNR of the estimated PPG signal using RobustPPG over the DIS-M approach (second-best prior method). Our approach is successful in capturing large motion distortions and hence can effectively cancel out the distortions in the pixel intensity fluctuations. We also observe an improvement of over 1 bpm in the RMSE of the estimated heart rate, with an improvement of

Biomedical Optics EXPRESS

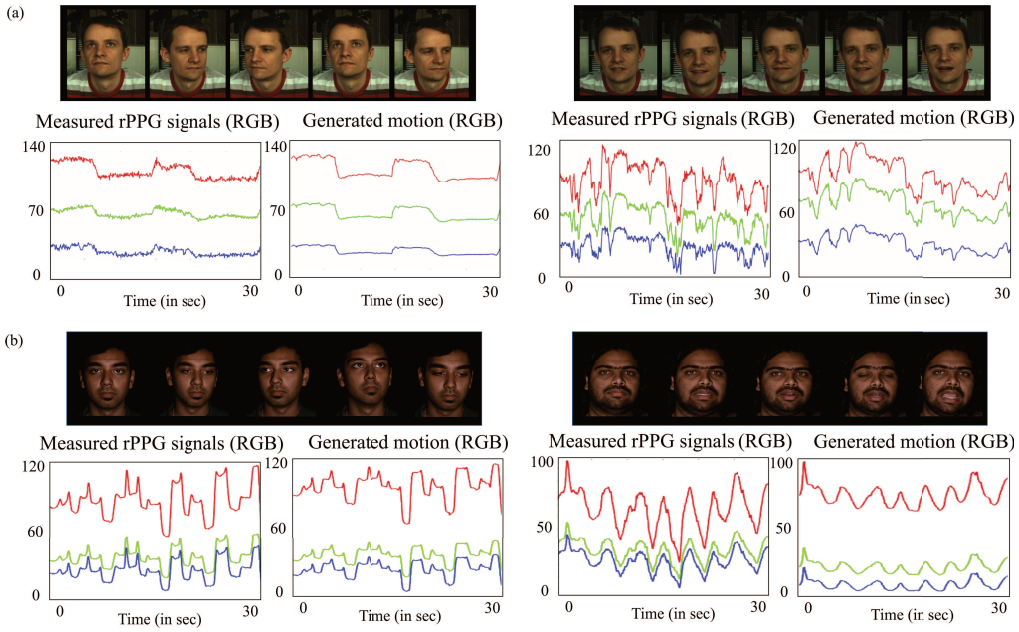

Fig. 10. Examples of measured raw rPPG signals and generated motion signals from two dataset: (a) PURE dataset and (b) RICE-motion dataset. We observe good agreement between the pixel intensity variation from a single triangle (left column) vs. generated motion distortions (right column) for different motion scenarios such as rotational movments and talking.

over 28% using RobustPPG over DIS-M. For the talking scenario, RobustPPG obtains around 1.5 bpm improvement in estimated heart rate over DIS-M and an SNR improvement of around 2.5 dB. For small movements, the x, y coordinates of the face-tracker can serve as a decent approximation of the motion distortions in the observed pixel intensity variations. However, RobustPPG is significantly effective for large motions in canceling out the motion distortions, as evident from the relatively cleaner signal spectrograms. Additionally, we observe that the DRPPG method is the second best method for extracting the rPPG signal in both datasets. The use of accurate reference motion signals helps in achieving improved performance over the other state-of-the-art methods. However, the Bi-LSTM architecture shows better filtering capability over the discriminative signature-based filtering [29], as evident from the superior performance of RobustPPG over DRPPG.

Table 3. Signal-to-noise ratio and heart rate RMSE on RICE-motion dataset which contains fast movements. RobustPPG performs better than the other methods.

	Signal-to-noise ratio (dB)					Heart Rate RMSE (bpm)				
Movement	RobustPPG (Ours)	DRPPG (Ours)	DIS-M	POS	CHROM	RobustPPG (Ours)	DRPPG (Ours)	DIS-M	POS	CHROM
Head mov.	**7.03**	4.23	3.89	3.22	1.31	**2.74**	3.43	3.87	4.66	5.75
Talking	**2.05**	0.08	-0.31	-1.42	-2.29	**3.11**	4.63	4.81	9.26	10.91

Although we have assumed a single distant light source in our signal model, our model also holds true for multiple distant light sources. From the principle of superposition and the assumption that the human face is a Lambertian surface, multiple light sources can be effectively represented by a single distant light source. Our model does not account for near-lighting

Biomedical Optics EXPRESS

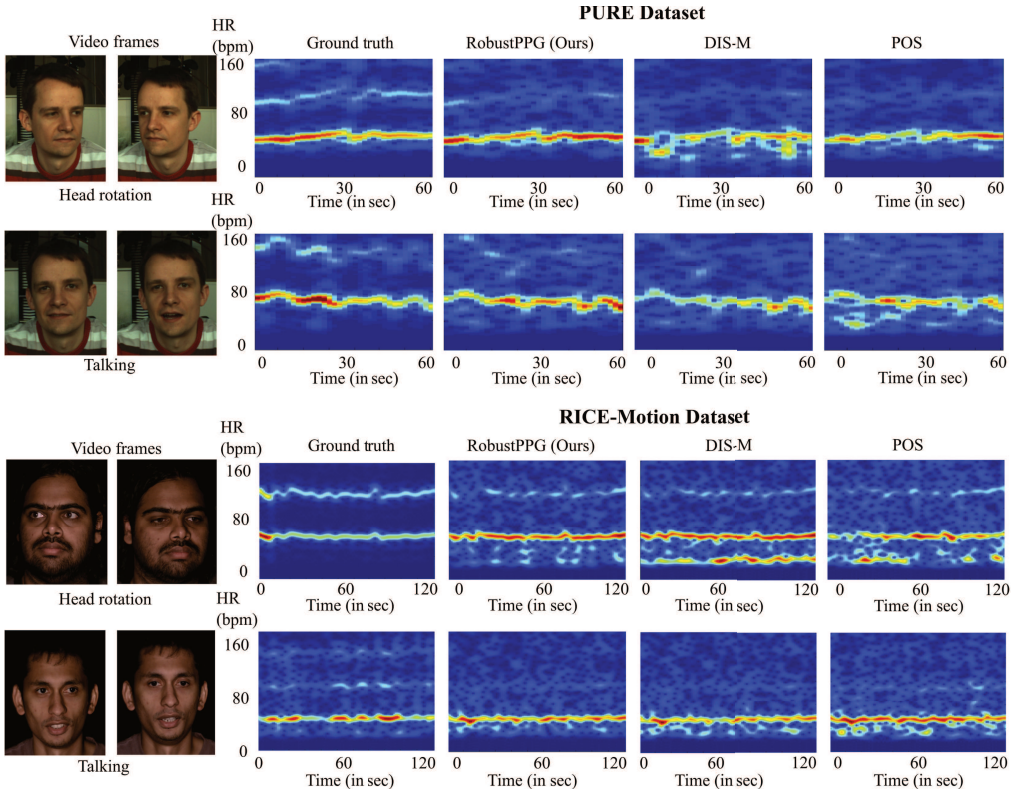

Fig. 11. Estimated spectrograms from our method RobustPPG, DIS-M and POS compared against the ground truth. The bright trace denotes the heart-rate signal. The heart-rate signal estimated from RobustPPG is comparatively cleaner and contains less motion distortions as compared to the other methods.

scenarios, and the performance of our approach may degrade. To this end, we evaluate our approach in out-of-lab settings. To this end, we test our method on phone videos in natural lighting such as outdoor settings (sunlight) and indoor room lighting. We observe that our method still performs better than the state-of-the-art methods in extracting the heart-rate signal, shown in Fig. 12. The estimated spectrogram extracted using RobustPPG has strong signal component present in the expected heart rate bandwidth (as indicated by the dotted box) and contains less unwanted distortion than the other methods, as evident from the relatively cleaner spectrograms. RobustPPG achieves an average improvement of over 1.5 dB over the second-best prior method and demonstrates good motion robustness in indoor and outdoor settings. Additionally, nearby light sources or arbitrary illumination can also be modeled, which may obtain even better performance, and we leave this direction for our future work.

A pre-processing step in our algorithm removes triangle meshes that correspond to facial hair. Since cheek regions contain strong PPG signals, removing triangle meshes from the cheek due to the presence of facial hair, say beard, may result in degradation of performance compared to subjects with no beard. In order to quantify the degradation, we test our approach RobustPPG on the subjects from the RICE-Motion dataset. Only three participants with olive skin tone have facial hair (beard). To make the comparison fair, we compare the accuracy for bearded vs. non-bearded olive skin-tone subjects. On the limited RICE-Motion dataset, RobustPPG achieves

Biomedical Optics EXPRESS

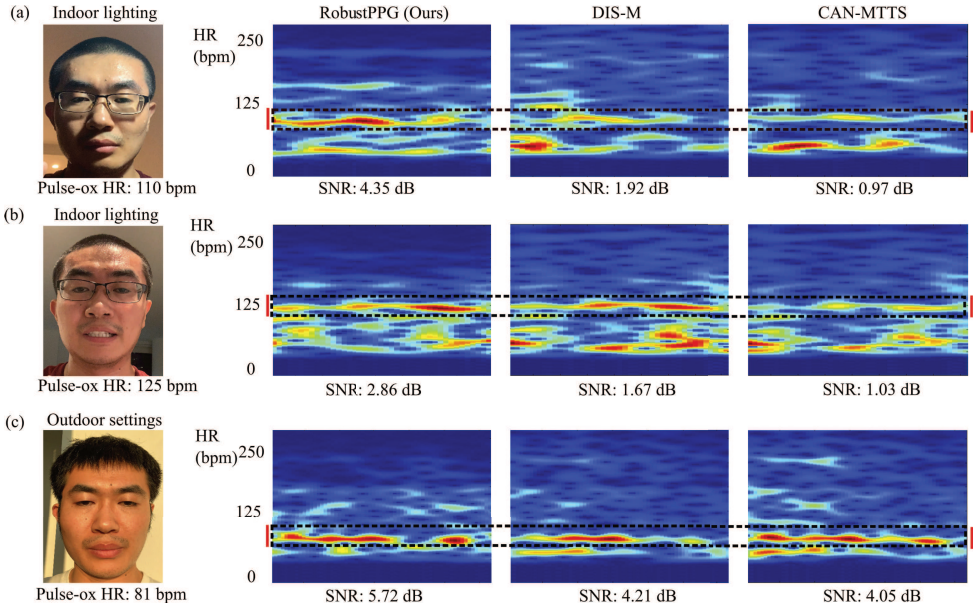

Fig. 12. Phone camera-based rPPG signal estimation in the wild. We show the spectrograms of extracted rPPG signals using RobustPPG (our method), DIS-M, and Convolutional attention network (CAN-MTTS) for three videos with different lighting scenarios - (a) room lighting, (b) gym lighting, and (c) sunlight, and under natural movements. The dotted box represents the frequency bandwidth of the ground truth heart rate. RobustPPG achieves the highest quality in signal extraction even under uncontrolled lighting environments over the other methods.

an average heart rate RMSE of 1.92 bpm and SNR of 5.16 dB for subjects with no facial hair vs. 2.31 bpm heart rate RMSE and SNR of 4.75 dB for subjects with facial hair.

Generally, skin color is a dominant factor that affects the quality of the extracted rPPG signals from a video. Darker skin tone absorbs more light, resulting in lower intensity levels as captured by a camera; thereby, reliable extraction of rPPG signals becomes more challenging. In the RICE-Motion dataset, we observe an improved performance with an average heart-rate RMSE of 2.33 bpm and SNR of 5.89 dB for fair-skin-toned subjects over olive-skin-toned subjects, with an average heart-rate RMSE of 2.94 bpm and 4.96 dB SNR. On the other hand, compared with the DIS-M method, the improvement in performance using our method is slightly more for olive-skin-toned subjects. Compared to DIS-M, RobustPPG achieves a 0.75 bpm reduction in RMSE heart rate and an increase in signal quality by 2 dB SNR for fair skin-toned subjects vs. 1.34 bpm reduction in RMSE heart rate and an SNR improvement by 2.8 dB SNR for olive skin-toned subjects. This demonstrates that our method outperforms the other state-of-the-art methods in recovering weaker rPPG signals in the presence of complicated movements.

Finally, for training the Bi-LSTM network, we use the pulse-oximeter on the finger as the training label. Using finger PPG signals as training labels is a common strategy in most learning-based methods [8,45]. However, we note that the signal from a pulse-oximeter obtained from a finger is slightly different from the facial pulse wave in two ways - i) the waveform shape (finger PPG has more features and higher harmonics compared to that from the face) and ii) a phase delay due to pulse transit time. We tackle the first point by low-pass filtering the finger pulse waveform so that the high-frequency harmonics are filtered out. The phase delay between waveforms is a more difficult problem to tackle. In [9], the authors show that using camera-based signals obtained from a motion-robust algorithm (POS, CHROM) as training labels

Biomedical Optics EXPRESS

for a CNN architecture achieves slightly higher accuracy since there is no phase delay. However, in extreme motion scenarios, using camera-based signals as training labels is highly unreliable as the signals are corrupted. Ear-based pulse-oximeter is another alternative to get ground truth labels; however, facial and head movements may affect the quality of the signals. Hence, we resort to the filtered finger-based pulse-oximeter waveforms to obtain high-quality ground truth labels. The performance of our approach validates the use of finger PPG waveforms similar to the approach in [8,45].

RobustPPG has four subparts - i) face-tracking, ii) surface normal, pixel intensity fluctuation extraction and light estimation, iii) motion signal generation and iv) rPPG signal extraction at each triangle mesh using Bi-LSTM network, with the main computational bottleneck being in extracting pixel intensity from each triangle mesh from a frame. Overall, the whole running time of our algorithm on MATLAB is approximately 3.5 minutes for extracting heart rate trace for a 1-minute video at 30 fps on an Intel core i7 processor. There are some aspects where we can reduce the computation time. First, an optimized code for extracting pixel intensities for each triangle mesh may significantly reduce computation time. We can further speed up the computation using Python or C++ and/or GPUs. Secondly, we use an individual triangle mesh-based filtering approach to extract the rPPG signal for all the results reported in our paper. In this way, the Bi-LSTM network extracts PPG signal from each individual triangle mesh from a face. We can also have a unified filtering approach by performing a one-shot filtering process using the Bi-LSTM network instead of triangle-wise filtering. In a slight modification, we can rewrite Eq. (7) by adding up both sides of the equation across all the triangle locations $\mathbf{r}$. The filtering process can then be done on the summed-up intensity signals. This will further reduce the total computation time with a slight performance degradation. Third, we hope that a version of this algorithm could find its way to practical applications by optimizing the implementation architectures. For example, LSTM networks can be unrolled and mapped to feedforward architectures, making computations faster and more efficient. These unrolled algorithms [46] can also be mapped to efficient neural processors that are making their way into many commercial products.

6. Conclusion and future work

We present a novel algorithm called RobustPPG for camera-based rPPG signal extraction and heart rate estimation. We demonstrate that a 3D face tracker such as FaceMesh can be used to generate accurate motion distortions in the pixel intensity variation. Furthermore, using a Bi-LSTM network for signal filtering, we demonstrate better accuracy in rPPG signal extraction over state-of-the-art methods. We hope that this work will significantly push the limits on motion robustness for reliable heart rate estimation and can find its way into real-life applications.

In this work, we have only modeled the motion distortions arising from Lambertian modeling. The specular components can be taken into account to make the modeling more accurate. Secondly, we consider only distant lighting assumptions in our work. Near-lighting scenarios [47,48] call for modeling complexities that can be explored to better estimate the motion signals. Thirdly, we also consider the case when the camera is fixed. Movements of the camera will cause additional signal distortions in the rPPG signals, which may affect the accuracy of heart rate estimation in hand-held phone scenarios. These are all interesting avenues to explore and might serve as exciting directions for future work.

Funding. National Science Foundation (Expeditions Grant - NSF-1730574); Snap, Inc..

Disclosures. The authors declare no conflict of interest.

Data availability. The RICE-motion dataset presented in the paper is publicly available in the repository [42].

References

Biomedical Optics EXPRESS

1. W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," Opt. Express **16**(26), 21434 (2008).
2. Y. Sun, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," J. Biomed. Opt. **16**(7), 077010 (2011).
3. D. J. M. M.Z. Poh and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2011).
4. A. Pai, A. Veeraraghavan, and A. Sabharwal, "HRVCam: robust camera-based measurement of heart rate variability," J. Biomed. Opt. **26**(02), 1 (2021).
5. B. D. Holton, K. Mannapperuma, P. J. Lesniewski, and J. C. Thomas, "Signal recovery in imaging photoplethysmography," Physiol. Meas. **34**(11), 1499–1511 (2013).
6. M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," Biomed. Opt. Express **6**(5), 1565 (2015).
7. Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," IEEE Trans. Biomed. Eng. **63**(3), 463–477 (2016).
8. W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the european conference on computer vision (ECCV)*, (2018), pp. 349–365.
9. Q. Zhan, W. Wang, and G. de Haan, "Analysis of CNN-based remote-PPG to understand limitations and sensitivities," Biomed. Opt. Express **11**(3), 1268 (2020).
10. C. Tomasi and T. Kanade, "Detection and tracking of point features," Technical Report MU-CS-91-132, Carnegie Mellon University (1991).
11. Google, "MediaPipe face mesh," https://google.github.io/mediapipe/solutions/face_mesh.html. Accessed: 2022-08-27.
12. J. Geng, "Structured-light 3d surface imaging: a tutorial," Adv. Opt. Photonics **3**(2), 128–160 (2011).
13. V. Saragadam, J. Wang, M. Gupta, and S. Nayar, "Micro-baseline structured light," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), pp. 4049–4058.
14. A. Jones, G. Fyffe, X. Yu, W.-C. Ma, J. Busch, R. Ichikari, M. Bolas, and P. Debevec, "Head-mounted photometric stereo for performance capture," in *2011 Conference for Visual Media Production*, (IEEE, 2011), pp. 158–164.
15. C. Tsiotsios, A. J. Davison, and T.-K. Kim, "Near-lighting photometric stereo for unknown scene distance and medium attenuation," Image Vis. Comput. **57**, 44–57 (2017).
16. Y. Nie, Z. Song, M. Ji, and L. Zhu, "A novel calibration method for the photometric stereo system with non-isotropic led lamps," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, (2016), pp. 289–294.
17. D. Cho, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Semi-calibrated photometric stereo," IEEE Trans. Pattern Anal. Mach. Intell. **42**(1), 232–245 (2018).
18. J. Wang, Y. Matsushita, B. Shi, and A. C. Sankaranarayanan, "Photometric stereo with small angular variations," in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), pp. 3478–3486.
19. C.-H. Hung, T.-P. Wu, Y. Matsushita, L. Xu, J. Jia, and C.-K. Tang, "Photometric stereo in the wild," in *2015 IEEE Winter Conference on Applications of Computer Vision*, (IEEE, 2015), pp. 302–309.
20. P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *ACM SIGGRAPH 2008 classes*, (2008), pp. 1–10.
21. O. Aldrian and W. A. Smith, "Inverse rendering of faces with a 3d morphable model," IEEE Trans. Pattern Anal. Mach. Intell. **35**(5), 1080–1093 (2012).
22. R. Yi, C. Zhu, P. Tan, and S. Lin, "Faces as lighting probes via unsupervised deep highlight extraction," *Proceedings of the European Conference on computer vision (ECCV)*, (2018), pp. 317–333.
23. D. A. Calian, J.-F. Lalonde, P. Gotardo, T. Simon, I. Matthews, and K. Mitchell, "From faces to outdoor light probes," in *Computer Graphics Forum*, vol. 37 (Wiley Online Library, (2018), pp. 51–61.
24. D. J. M. M.Z. Poh and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," Opt. Express **18**(10), 10762 (2010).
25. W. Wang, S. Stuijk, and G. de Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," IEEE Trans. Biomed. Eng. **63**(9), 1974–1984 (2016).
26. G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013).
27. G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," Physiol. Meas. **35**(9), 1913–1926 (2014).
28. W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2017).
29. W. Wang, A. C. den Brinker, and G. de Haan, "Discriminative signatures for remote-PPG," IEEE Trans. Biomed. Eng. **67**(5), 1462–1473 (2020).
30. E. M. Nowara, D. McDuff, and A. Veeraraghavan, "The benefit of distraction: Denoising camera-based physiological measurements using inverse attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), pp. 4955–4964.
31. S. A. Shafer, "Using color to separate reflection components," Color Res. Appl. **10**(4), 210–218 (1985).

Biomedical Optics EXPRESS

32. Snap Inc. Lens Studio, Face mesh, https://docs.snap.com/lens-studio/references/guides/lens-features/tracking/face/face-effects/face-mesh/adding-a-face-mesh. Accessed: 2022-08-27.
33. Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, 2019).
34. Apple, Face tracking with ARKit, https://developer.apple.com/videos/play/tech-talks/601/. Accessed: 2022-08-27.
35. Meta Spark AR, The face mesh, https://sparkar.facebook.com/ar-studio/learn/articles/people-tracking/face-mesh/adding-a-face-mesh. Accessed: 2022-08-27.
36. E. Wood, T. Baltrusaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljevic, D. Wilde, S. Garbin, C. Raman, J. Shotton, T. Sharp, I. Stojiljkovic, T. Cashman, and J. Valentin, "3d face reconstruction with dense landmarks," in *Proceedings of the European Conference on computer vision (ECCV)*, 2022).
37. D. Botina-Monsalve, Y. Benezeth, R. Macwan, P. Pierrart, F. Parra, K. Nakamura, R. Gomez, and J. Miteran, "Long short-term memory deep-filter in remote photoplethysmography," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (IEEE, 2020).
38. H. Gao, X. Wu, C. Shi, Q. Gao, and J. Geng, "A LSTM-based realtime signal quality assessment for photoplethysmogram and remote photoplethysmogram," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (IEEE, 2021).
39. G. Vogiatzis and C. Hernández, "Self-calibrated, multi-spectral photometric stereo for 3d face capture," Int J Comput Vis **97**(1), 91–103 (2012).
40. R. Stricker, S. Muller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (IEEE, 2014).
41. R. Stricker, S. Mueller, and H. M. Gross, "PURE PPG dataset," Technische Universität Ilmenau, 2004 http://www.tu-ilmenau.de/neurob/data-sets/pulse. Accessed: 2022-08-27.
42. A. K. Maity, J. Wang, A. Sabharwal, and S. Nayar, "Robustppg," Version 1.0.0, Github, 2022, https://github.com/akashmaity/RobustPPG .
43. A. K. Maity, A. Veeraraghavan, and A. Sabharwal, "PPGMotion: Model-based detection of motion artifacts in photoplethysmography signals," Biomed. Signal Process. Control. **75**, 103632 (2022).
44. X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," arXiv preprint arXiv:2006.03790 (2020).
45. O. Perepelkina, M. Artemyev, M. Churikova, and M. Grinenko, "HeartTrack: Convolutional neural network for remote video-based heart rate monitoring," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (IEEE, 2020).
46. V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," IEEE Signal Process. Mag. **38**(2), 18–44 (2021).
47. J. Wang, X. Liang, Y. Matsushita, M. Chen, and B. Huang, "Gigapixel 3d camera," in Tech-report, (2015).
48. Z. Chen, Y. Ji, M. Zhou, S. B. Kang, and J. Yu, "3d face reconstruction using color photometric stereo with uncalibrated near point lights," in *2020 IEEE International Conference on Computational Photography (ICCP)*, (IEEE, 2020), pp. 1–12.