Remote Heart Rate Measurement From Near-Infrared Videos Based on Joint Blind Source Separation With Delay-Coordinate Transformation

Juan Cheng, *Member, IEEE*, Ping Wang, Rencheng Song, *Member, IEEE*, Yu Liu, *Member, IEEE*, Chang Li, *Member, IEEE*, Yong Liu, *Member, IEEE*, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Noncontact and low-cost heart rate (HR) measurement based on imaging photoplethysmography (iPPG) technology is commonly desired for health care monitoring. However, the usually employed red–green–blue (RGB) cameras are sensitive to illumination variations and cannot work under dark situations. In this study, we propose a novel framework of applying joint blind source separation with delay-coordinate transformation (DCT-JBSS) to evaluate HR from a single-channel near-infrared (NIR) camera in dark situation. First, three facial regions of interest (ROIs) are determined by face detection technique and a single-channel signal is constructed through a frame-by-frame pixel averaging within each ROI. Second, each single-channel signal is transformed into time-delayed multichannel signal through DCT and then treated as a separate ROI signal set. Third, the three ROI signal sets are simultaneously processed by JBSS to derive the underlying shared HR source component vector (SCV), which is usually ordered the first and has the highest correlation across each signal set. Finally, the fast Fourier transform (FFT) is applied to the HR SCV and the corresponding dominant frequency (within the range from 0.7 to 2.5 Hz) with the highest signal-to-noise ratio (SNR) is determined as the target HR frequency. The proposed framework, as well as several other typical iPPG methods, is validated on public DROZY and MR-NIRP databases. The proposed method achieves the best performance, providing a probable way to widen the application of remote and continuous HR measurement during night conditions.**

Manuscript received September 5, 2020; revised October 30, 2020; accepted November 18, 2020. Date of publication November 27, 2020; date of current version December 24, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61922075, Grant 61701160, and Grant 41901350; in part by the Provincial Natural Science Foundation of Anhui under Grant 1808085QF186 and Grant 2008085QF285; and in part by the Fundamental Research Funds for the Central Universities under Grant JZ2020HGPA0111 and Grant JZ2019HGBZ0151. The Associate Editor coordinating the review process was Anirban Mukherjee. *(Corresponding authors: Rencheng Song; Xun Chen.)*

Juan Cheng, Ping Wang, Rencheng Song, Yu Liu, and Chang Li are with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: chengjuan@hfut.edu.cn; wangping123@mail.hfut.edu.cn; rcsong@hfut.edu.cn; yuliu@hfut.edu.cn; changli@hfut.edu.cn).

Yong Liu is with the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore 639798 (e-mail: stephenliu@ntu.edu.sg).

Xun Chen is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230001, China, and also with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230026, China (e-mail: xunchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TIM.2020.3041083

***Index Terms*— Delay-coordinate transformation (DCT), independent vector analysis (IVA), joint blind source separation (JBSS), near-infrared (NIR) video, noncontact heart rate (HR).**

I. Introduction

HEART rate (HR) is an important vital sign of human body, which can reflect the physiological and mental status of humans. HR measurement is highly desired for many applications in the fields of continuous health monitoring and driver's status monitoring that include night situations. Conventional contact HR measurements require patients to wear adhesive gel patches or finger clips, which may cause skin allergy or discomfort. In contrast, noncontact HR measurement provides a convenient way to estimate HR for cases where direct contact with the skin has to be prevented (e.g., neonates and subjects with skin damage) or prolonged monitoring is desired (e.g., surveillance and fitness) [1], [2]. Recently, researchers have paid growing attention to noncontact HR measurement techniques. The imaging photoplethysmography (iPPG) is such a kind of video-based HR monitoring method, which detects pulsatile information caused by the cardiac activity from invisible facial color changes within the exposed skin from a distance.

The potentials of iPPG are promising. However, the iPPG pulsatile signal is quite weak and can be easily contaminated by noise interference, typically as illumination variations and motion artifacts [3], [4]. A series of studies have been proposed to improve the performance of HR measurement during realistic situations. Among them, the blind source separation (BSS)-based methods are widely utilized. As indicated in [5], the skin color signal can be treated as a combination of time-varying intensity signal, varying specular reflection signal, and pulsatile signal. Therefore, the pulsatile signal can be demixed by BSS techniques [e.g., independent component analysis (ICA)] under given statistical assumptions.

Usually, the iPPG technique adopts the red–green–blue (RGB) cameras that can be effective during the ambient light situation. However, the RGB cameras will be inaccurate or even powerless during dark and night situations [6]. The study of Aarts *et al*. [1] demonstrated that the RGB cameras could well monitor the HR of new infants in the Neonatal Intensive

Care Unit (NICU). However, low ambient light level and infant motions would prevent the successful HR measurement. An alternative scheme is to employ a near-infrared (NIR) camera with an active illumination to remotely measure HR under dark situations while significantly reducing the illumination variations [7]. Unlike the three-color-channel signal provided by RGB cameras, only one single-channel signal is derived from NIR cameras. As we know, the number of the extracted sources by BSS is not larger than that of the observation channels. In other words, insufficient channel of observations will influence the effect of demixing. Consequently, the BSS technique cannot be directly applied to the single-channel NIR signal. Zhao *et al.* [8] proposed to convert the single-channel signal from NIR images into the multichannel signal through delay-coordinate transformation (DCT), and then, ICA-based BSS was followed to evaluate the HR in the presence of motion artifacts. The principle of DCT is that the reconstructed state space is equivalent to that of the original single-channel signal consisting of all the dynamic variables. Then, the transformed multichannel signal containing the pulsatile information in a mixed form can be demixed by the BSS method.

On the other hand, it is well known that the HR information is commonly shared on different skin regions when the subject keeps relatively static. Conventional BSS methods, such as ICA, usually work with a single signal set consisting of channels from a single region of interest (ROI) (RGB channels) or multiple ROIs. Some studies [9]–[11] have demonstrated that extracting the iPPG signal from multiple ROIs simultaneously based on BSS techniques can improve the performance of HR measurement. For instance, Favilla *et al*. demonstrated that the HR could be extracted by applying the BSS technique to multiple green-channel signals extracted from multiple ROIs. The mean absolute error (MAE) was reduced about 3.81 ms for normal-to-normal intervals when using ICA preprocessing, compared to that of directly using the single green-channel signal followed by the bandpass filter [9].

However, on the one hand, the ICA-based BSS methods have a problem of permutation indeterminacy [12]. On the other hand, they can only handle a single signal set. In fact, different facial ROIs all contain pulsatile information although the qualities of them are different. The joint BSS (JBSS), instead of BSS, can extract the underlying sources within each signal set while keeping a consistent order of the extracted sources across multiple signal sets. Therefore, JBSS can solve the permutation indeterminacy problem while making full use of the correlation of sources across the multiple sets. [13]–[15].

Inspired by the success of DCT and JBSS separately, we propose a novel framework of applying the JBSS technique to multiple ROI signal sets to measure HR from a single NIR camera during dark situations. In order to make the single-channel NIR signal suitable for the input of the JBSS, the original single-channel NIR signal will first be constructed into the multiple-channel signal via the DCT technique and then treated as a separate ROI signal set. Under the assumption that the state space of the generated multichannel signal set is equivalent to that of the original single-channel signal including all the dynamic variables, each multichannel signal set contains the pulsatile information. Meanwhile, the JBSS provides a powerful capability to demix the shared source signal across multiple signal sets with the same modality or even distinct modalities [16]. Since NIR cameras, with an active illumination, can significantly reduce the illumination variations and suitable for darkness usage, the most common sources existing in all NIR signal sets derived from different ROIs are considered to contain the pulsatile information under relatively static situations. In this case, the underlying pulsatile sources can be separated by JBSS from each signal set, which usually has the largest correlation among all the source components.

The main contributions of this article are as follows. To the best of our knowledge, it is the first time that the framework of DCT-JBSS is proposed to remotely estimate HR for NIR videos. With the help of DCT, the single-channel signal can be expanded to a multichannel signal set, which contains the pulsatile information in a mixed form. Besides, through JBSS instead of BSS, multiple ROIs can be simultaneously utilized to provide the underlying common pulsatile information across each multichannel signal set, which improves the performance of HR measurement during dark and relatively static situations. The performance of the proposed DCT-JBSS framework, as well as other typical iPPG methods, has been evaluated and compared on publicly available DROZY and MR-NIRP databases. The best performances have been achieved by our proposed DCT-JBSS.

The remainder of this article is organized as follows. Section II introduces some existing work closely related to our proposed approach. Section III describes the details of our methods. Section IV introduces the experimental setup, results and discussions. Section V concludes our study.

II. Related Work

The iPPG methods based on RGB videos have achieved good performance of HR measurement during bright and stable illumination environments, even in the presence of motion artifacts. An increasing number of studies, based on realistic optical models and advanced signal processing techniques, have been conducted to remotely measure the PPG signals from facial videos [5], [17]–[20]. As for motion artifact elimination, the methods can be divided into BSS- and motion-based ones according to the recently proposed mathematical models in [5]. The progress has been summarized in several relevant review articles [4], [21], [22].

Since NIR cameras, with active illuminations, are much less sensitive to illumination variations than RGB ones, they are usually adopted to measure HR during the situations with varying illumination or full darkness. For instance, Jeanne *et al*. [23] demonstrated that NIR cameras are feasible to estimate HR under highly dynamic light conditions, including dark situations. The results obtained by their system show high accuracy (root-mean-square error (RMSE) less than 1 beat per minute (bpm) under disco-light situations) and a

correlation score above 0.99 when compared with a reference measurement method. Van Gastel *et al*. [24] demonstrated that respiration rate (RR) can be successfully detected with a camera in both visible and dark situations by using the close similarity between pulse and respiration-induced color variations of the skin. The MAE for guided breathing scenarios is 1.74 and 2.27 bpm in visible light and infrared, respectively. Nowara *et al.* [25] proposed a novel denoising algorithm (SparsePPG) based on robust principal components analysis (RPCA) and sparse frequency spectrum estimation to measure HR during in-car situations. They have found that it is possible to accurately measure HR using NIR cameras with an active 940-nm illumination, in both controlled and varying light situations. However, the signal-to-noise ratio (SNR) is significantly decreased in NIR compared to RGB in controlled lighting conditions. Kado *et al*. [26] proposed to measure HR using simultaneously recorded RGB and NIR face videos, based on the key idea of automatically selecting suitable face patches for HR estimation in both spatial and spectral domains. The simultaneous usage of RGB and NIR enabled robust HR estimation under various illumination conditions. It can be discovered from these studies that usually, a single-channel NIR signal is processed to measure HR in case of illumination variations without motion artifacts.

For more realistic situations, many algorithms adopted for RGB videos can also be employed for NIR videos. For instance, Verkruysse *et al*. found that fast Fourier transform (FFT) combined with bandpass filters to the green channel of the facial RGB videos could achieve the best HR measurement compared to that of the red or blue channel [27]. Chen *et al.* [28] applied a similar method to the single NIR channel to obtain a robust HR measurement. In order to resist the influence of the changing illumination when capturing RGB videos, Chen *et al.* [29] applied ensemble empirical mode decomposition (EEMD) to the green channel for separating the real pulsatile signal from the environmental illumination noise. Analogously, Zhang *et al*. [30] used an empirical mode decomposition (EMD) technique to derive HR measurements under the driving situation with complex illumination variations from NIR videos. An alternative way to measure HR from NIR videos is according to the skin or motion magnification framework [31]. He *et al*. [32] employed an NIR camera in conjunction with Eulerian video magnification (EVM) method to measure HR in dark conditions. Van Gastel *et al.* [24] verified the feasibility of motion robust pulse detection in NIR based on the PBV method, which was shown a good HR measurement performance for rPPG in visible light. In order to make the PBV method workable, three NIR cameras were constructed by the three monochrome cameras with different optical filters, with the center wavelengths of 675, 800, and 842 nm. It should be mentioned that although motion-based methods could also achieve a successful HR measurement performance from a single ROI for NIR videos, they usually need to employ at least two NIR cameras with different wavelengths [24]. In this article, we utilize only one single NIR camera, and we mainly focus on the BSS-based methods.

A. Conventional BSS

As known, the pulsatile information together with specular and diffuse reflections can be modeled as an optical model [5], [33]. Based on this model, the skin color channel signal can be treated as a linear combination of the time-varying intensity signal, the varying specular reflection signal, and the pulsatile signal [5]. Therefore, the pulsatile signal can be demixed and derived by applying the BSS technique (e.g., ICA) to RGB channels under certain statistical assumptions. Poh *et al*. [34] proposed an ICA-based BSS algorithm to extract the HR component using a single facial ROI during subtle-motion conditions. The RMSE corresponding to motion situations was sharply reduced, demonstrating the feasibility of BSS for HR measurement. Sun *et al*. [35] introduced an artifact-elimination method consisting of planar motion compensation and BSS. Their BSS mainly referred to the single-channel ICA (SCICA), which isolated multiple components using only the temporal information inherent in a single-channel signal. SCICA assumes that a set of the observed data points from a single-channel signal is a linear combination of unknown and statistically independent sources. Similarly, Zhao *et al*. [8] extracted both HR and RR from NIR videos by combining DCT with ICA (also called SCICA). The multichannel signal was generated from the single-channel signal via DCT and the desired pulsatile component was demixed and separated via BSS.

The abovementioned studies mainly focus on choosing one single facial ROI to derive HR. In fact, different facial ROIs all contain pulsatile information although the qualities of them are different. Therefore, several studies aim to improve the performance of HR estimation by simultaneously analyzing multiple different ROIs. Lam and Kuno [36] assumed that the extraction of HR from multiple facial ROIs could be treated as a linear BSS problem. According to the skin appearance model that describes how illumination variations and cardiac activity affect the appearance of the skin over time, HR can be well estimated by randomly selecting pairs of green-channel traces and majority voting. Wei *et al*. [11] proposed to measure HR by applying a second-order BSS to the six-channel RGB signals that yielded from dual facial ROIs. This method can suppress the respiratory motion artifacts for robust HR measurement. Favilla *et al*. [9] selected three different areas of facial skin, including the forehead, the left cheek, and the right cheek as ROIs, and obtained a three-green-channel signal. The detrended three-channel signal was fed into ICA to derive three independent components (ICs). By applying FFT to the ICs, the pulsatile component was identified as the one with the dominant frequency falling into the range from 0.75 to 2.0 Hz. The results prove that the proposed method can effectively improve the HR assessment from the iPPG signal. All these studies propose to utilize multiple ROIs and the conventional BSS technique. On the one hand, the conventional BSS-based methods, such as ICA, have a problem of permutation indeterminacy [37]. On the other hand, due to the fact that conventional BSS techniques can only handle one single signal set, the underlying shared pulse information across multiple ROIs cannot be well explored.

B. Joint BSS

With the increasing availability of multiple signal sets, various JBSS methods have been proposed to simultaneously accommodate them. Chen *et al*. [16] provided a thorough overview of representative JBSS methods for realistic neurophysiological applications from multiset and multimodal perspectives. They highlighted the benefits of the JBSS methods for neurophysiological data analysis. JBSS tries to extract the underlying sources within each signal set while keeping a consistent order of the extracted sources across multiple signal sets. Thereby, JBSS methods are excellent options for more accurate HR measurement. For instance, Guo *et al*. [13] first introduced the JBSS method into iPPG fields. They applied the independent vector analysis (IVA) to jointly analyze the multiple signal sets derived from multiple facial ROIs. Preliminary experimental results revealed an improved performance of HR measurement compared with that of the ICA-based BSS method. Later, Qi *et al.* [14] utilized the JBSS approach, including IVA and multiset canonical correlation analysis (MCCA) for HR measurement by exploring correlations among multiple RGB facial ROIs. The experimental results on a large public database showed that the JBSS method outperformed previous BSS methods. In order to eliminate illumination variations, Cheng *et al*. [15] proposed to apply a JBSS-EEMD framework to facial and background ROIs for extracting the underlying shared illumination variation sources, which were then removed from the facial ROI to reconstruct the clean facial ROI signal set. By this means, a robust noncontact HR measurement can be realized even during dynamically changing illumination variation situations.

To the best of our knowledge, there are a few studies aiming to evaluate HR by simultaneously adopting multiple facial ROIs from NIR videos, and JBSS methods have not yet been employed to measure HR for NIR videos. In this article, we propose a framework of JBSS with DCT technology, termed DCT-JBSS, to extract the pulsatile signal from multiple facial ROIs for NIR videos on public databases. The single NIR channel signal will be transformed into a multichannel signal set through the DCT technology, the state space of which is assumed equivalent to that of the original single-channel signal, including different dynamic variables. Besides, since NIR cameras are insensitive to illumination variations and suitable for HR measurement under totally dark situations, the most common source existing in all facial ROIs contains the pulsatile information under relatively static situations. Consequently, the JBSS technique is a preferred option to derive the pulsatile source from multiple ROI signal sets with multichannels generated by DCT for accurate HR measurement.

It is worth mentioning that we mainly focus on the BSS-based methods to measure HR in this article. There are many other advanced iPPG methods, such as model-based methods and deep-learning-based methods for RGB-based HR measurement for realistic situations. The progress has been summarized in several relevant reviews or articles [4], [5], [17], [18], [38]–[40].

III. Method

The flowchart of the proposed DCT-JBSS framework is shown in Fig. 1. First, three facial ROIs are identified through a face detection algorithm for each frame, and the corresponding pixel averaging of each ROI is calculated and concatenated frame by frame to generate the single NIR channel signal. Second, the DCT technique is applied to each single-NIR-channel signal for deriving multiple ROI signal sets. Third, the underlying pulsatile sources commonly existing in all the ROI signal sets will be extracted by the IVA-based JBSS algorithm and further processed by FFT to obtain the corresponding power spectral density (PSD). The dominant frequency with the highest SNR falling in the normal HR range of interest (ROI) will be selected as the target HR frequency. We elaborate the detailed main steps in the following.

A. ROI Detection and Tracking

Reliable ROI detection and tracking are important for robust RGB-video-based HR estimation, which is also important to HR estimation for NIR videos. As demonstrated in [41], the forehead and both cheeks are optimal ROIs. However, since the DROZY database was designed to simultaneously capture facial videos and polysomnography (PSG) signals, including electrocardiogram (ECG), electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) using PSG electrodes, some of the PSG electrodes are placed on one side (usually the right side) of the cheek. Besides, the captured face of the DROZY database accounted for a small proportion of the whole image, which meant that the area of one side of cheek was quite small. In order to cover a certain number of pixels within an ROI, a whole cheek ROI, including the nose part, was selected. In addition, as stated in [42] and [43], the chin ROI is rich in capillaries that will generate strong signal strength, which is also demonstrated to have a relative lower SNR than that of the forehead or cheek ROI [41]. Therefore, three ROIs, including the forehead ROI, the whole cheek ROI, and the chin ROI, are determined.

The well-known Viola–Jones face detection algorithm [44] is adopted to detect the facial rectangle, with the length l and the width d. Fig. 2 shows the localizations of both facial ROI subregions and background ROI. The ROI of the cheek area is a rectangle, with the size of $0.5d \times 0.2l$, and the center of the rectangle is the same as that of the facial rectangle. The ROIs of forehead and chin are both 10×10 pixel squares, and the center of squares is located upstraight and downstraight that of the facial rectangle, respectively. In order to compare with the following Seg-ICA [45], a background ROI should also be determined, shown as the blue rectangle in Fig. 2, with the size of $0.2d \times 0.8l$. Then, the Kanade–Lucas–Tomasi (KLT) algorithm [46], a standard facial tracker, is employed to detect the corners inside the ROIs for subsequent video frames. This ensures an accurate and fast ROI detection, which also helps to compensate for the interference of motion artifacts. Afterward, the pixels within each ROI are averaged frame by frame and the single-channel signal is formed. In total, three single-channel time sequences corresponding to

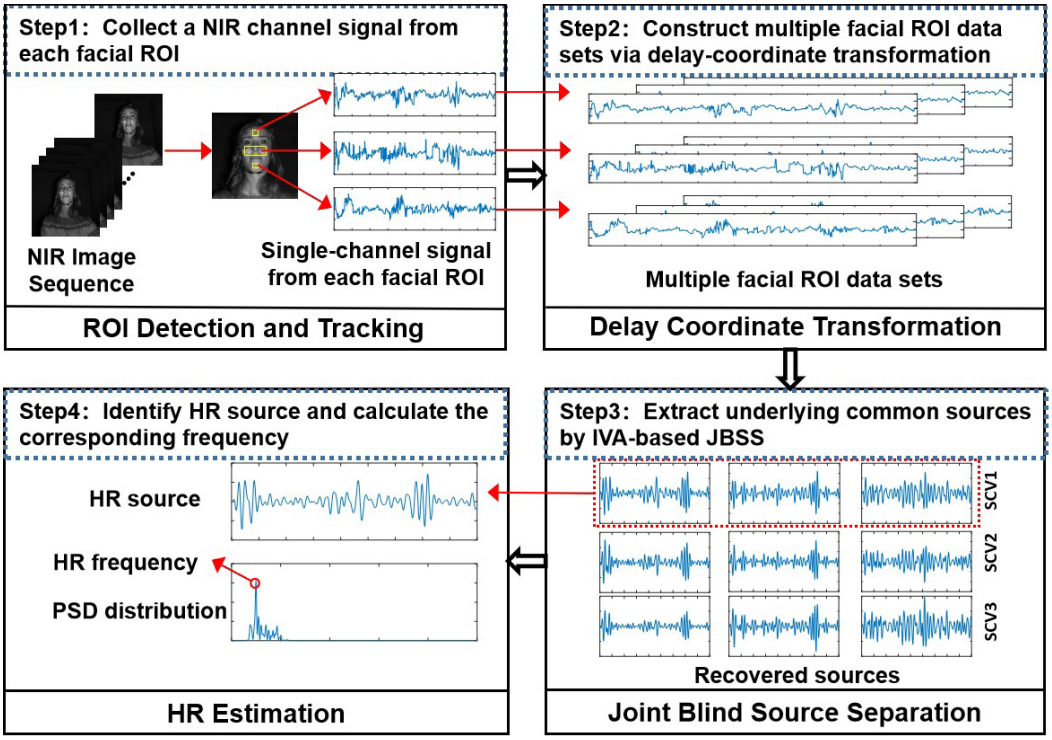

Fig. 1. Overview of the proposed DCT-JBSS framework to measure HR for NIR videos.

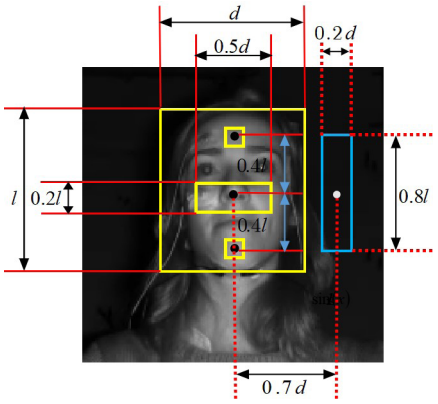

Fig. 2. Illustration of determined facial ROIs and background ROI.

facial ROIs are generated and a single-channel time sequence corresponding to background ROI is also generated.

Before using DCT, the generated three single-channel time sequences are first to be normalized to have zero means and unit variances [34]. Then, a detrending filter [47] is utilized to eliminate the slow linear or more complex trends. The smoothness parameter was experimentally set as 1000 to get a good performance.

B. Delay-Coordinate Transformation

The cardiovascular system is a nonlinear and dynamic in nature. The dynamics of the system are similar to those of other deterministic systems showing chaotic properties, which have irregular periodicity as well as an exquisite sensitivity to the initial conditions [48]. DCT, usually referring to time-delay embedding theorem, is proposed to describe nonlinear dynamics of a deterministic system showing chaotic properties or similar dynamics of a cardiovascular autonomic system. James and Lowe [49] demonstrated that an appropriate embedding matrix, constructed out of a series of delay vectors from the measured EEG signal, contained the information of artifact components, seizure components, theta band, and so on, which can later be deconstructed by applying ICA to this embedding matrix. Aston *et al.* [50] proposed a U.S. patent that periodic (periodic, pseudoperiodic, or approximate-periodic) data, such as physiological data, can be analyzed by obtaining a vector of delay coordinates for each one of a plurality of samples of the periodic data in a time window.

With the help of the sliding window technique, several sequences beginning with different time stamps are employed to form an embedding matrix, the state space of which is equivalent to that of the original (unobservable) signal, containing all the dynamic variables. The single-channel signal derived from the kth facial ROI is marked as $A^{[k]} = [a_1^{[k]}, a_2^{[k]}, \ldots, a_n^{[k]}, \ldots, a_N^{[k]}]$, where $a_n^{[k]}$ is the mean pixel value within the k-th facial ROI for the nth frame and N is the total number of the single NIR signal. The lag/delay version through DCT is denoted as

$$A^{[k]}[n] \equiv \left[a_n^{[k]}, a_{n+\tau}^{[k]}, a_{n+2\tau}^{[k]} \ldots, a_{n+(m-1)\tau}^{[k]}\right] \tag{1}$$

where m is the embedding dimension and τ is the time delay. The value of $(Q+m-1)\tau$ should be integer and is not larger than $N-1$. If τ is set as one sample point, $(Q+m-1)$ equals to $N-1$. Consequently, the embedding matrix can be

constructed by a number of consecutive delay vectors as

$$X^{[k]} = \begin{bmatrix} a_1^{[k]} & a_{1+\tau}^{[k]} & \cdots & a_{1+Q\tau}^{[k]} \\ a_{1+\tau}^{[k]} & a_{1+2\tau}^{[k]} & \cdots & a_{1+(Q+1)\tau}^{[k]} \\ \vdots & \vdots & & \vdots \\ a_{1+(m-1)\tau}^{[k]} & a_{1+m\tau}^{[k]} & \cdots & a_{1+(Q+m-1)\tau}^{[k]} \end{bmatrix} \quad (2)$$

which contains the pulsatile information in a mixed form that can be demixed by BSS technique, and the underlying HR source can be uncovered.

In practice, the number of delay vectors Q is determined by the length of the observed signal N. With the help of DCT, the multichannel signal of the kth facial ROI, treated as a separate signal set, can be constructed, marked as $X^{[k]}$.

According to the mathematical model of iPPG, the skin color signal can be treated as a linear combination of the time-varying intensity signal, the varying specular reflection signal, and the pulsatile signal [5]. Most of the previous BSS-based studies demonstrated that the three-channel signal is enough to extract the HR source. Zhao *et al.* [8] already proved that when embedding the single green-channel to 3-D embedding matrix, the HR can be well evaluated. Besides, Buzug and Pfister [51] demonstrated that since the reconstruction is an embedding, i.e., there is a topological mapping from the original phase space to the embedding space, the embedding dimension has to be $m \geq 2n + 1$ (n is the dimension of the flow in the original space). In this study, n is equal to 1 (the single-channel NIR signal), and m can set to 3. In addition, we evaluated the performance metrics varying with different dimensions of m. Experimental results showed that when m was set as 3, the performance was really good. Thereby, m was set as 3 in this study. Besides, the frame rate of the proposed NIR videos from the public database is not larger than 30 frames per second (fps). To ensure the performance of HR measurement, each time point (without downsampling) should be included so that τ was set as 1.

C. Joint Blind Source Separation

The common pulsatile sources existing in all ROI signal sets will be extracted by utilizing JBSS. JBSS aims to separate the underlying sources commonly existing in each signal set while keeping a consistent order of the extracted sources across multiple signal sets. We first describe the extraction of pulsatile sources via JBSS.

Given K signal sets (here $K \geq 2$), with each containing P (here, P is equal to m) channels and $Q+1$ samples, the kth signal set $X^{[k]}$ can be expressed by its corresponding column vectors as

$$X^{[k]} = \left[x_{(1)}^{[k]}, x_{(2)}^{[k]}, \ldots, x_{(q)}^{[k]}, \ldots, x_{(Q+1)}^{[k]}\right], \quad 1 \leq k \leq K \quad (3)$$

where $x_{(q)}^{[k]}$ is the qth realization of the column vector $X^{[k]}$ with the size of $P \times 1$. Each signal set is treated as a linear mixture of L underlying independent sources

$$X^{[k]} = \mathrm{B}^{[k]}\mathrm{S}^{[k]}, \quad 1 \leq k \leq K \quad (4)$$

where $\mathrm{B}^{[k]}$'s are mixing matrices and $\mathrm{S}^{[k]}$'s are underlying source matrices. $\mathrm{S}^{[k]}$ can be expressed by its corresponding vectors $\mathrm{S}^{[k]} = [s_1^{[k]}, s_2^{[k]}, \ldots, s_L^{[k]}]^T$, where the superscript T denotes the transpose operation.

In the JBSS framework, source component vector (SCV) has been defined across multiple signal sets [52], The lth SCV is a random vector independent of all other SCVs and the components within each SCV are dependent. JBSS aims to identify the aforementioned SCVs by finding the mixing matrices $\mathrm{B}^{[k]}$'s or the demixing matrices $W^{[k]}$'s and the corresponding source vector estimates $y^{[k]} = W^{[k]}X^{[k]}$. The estimate of the lth SCV is given as $y_l = [y_l^{[1]}, y_l^{[2]}, \ldots, y_l^{[k]}, \ldots, y_l^{[K]}]^T$. Here, $y_l^{[k]}$ is the estimation of the lth component in the kth signal set given by $y_l^{[k]} = (w_l^{[k]})^T X^{[k]}$, where $(w_l^{[k]})^T$ is the lth row of $W^{[k]}$.

The goal of IVA is to identify the L independent SCV from K signal sets, which can be achieved by minimizing the mutual Information I_{IVA} among the estimated SCVs as

$$\begin{aligned} I_{IVA} &\triangleq I[y_1; y_2; \ldots; y_L] = \sum_{l=1}^{L} H[y_l] - H[y_1, y_2, \ldots, y_L] \\ &= \sum_{l=1}^{L} H[y_l] - H[W^{[1]}X^{[1]}, \ldots, W^{[K]}X^{[K]}] \\ &= \sum_{l=1}^{L}\left(\sum_{k=1}^{K} H[y_l^{[k]}] - I[y_l]\right) - \sum_{k=1}^{K} \log|\det(W^{[k]})| - C_1 \end{aligned} \quad (5)$$

where $H[\cdot]$ indicates the entropy and C_1 is the constant term $H[X^{[1]}, X^{[2]}, \ldots, X^{[K]}]$. The final representation shows that minimizing the cost function of IVA is equivalent to simultaneously minimizing the entropy of all components and maximizing the mutual information within each estimated SCV. Consequently, each estimated SCV is independent of all other estimated SCVs, while the components within each SCV are dependent on each other. IVA can ultimately solve permutation ambiguity when applying BSS techniques to multiple signal sets.

The most widely used specific distributions of IVA are Laplace distribution and Gaussian distribution, with the corresponding implementation algorithms called IVA-L and IVA-G. IVA-L assumes that each SCV follows a multivariate Laplace distribution (i.e., isotropic and without second-order correlation), whereas IVA-G exploits the linear dependencies across multiple signal sets by supposing that each SCV follows a multivariate Gaussian distribution. For most neurophysiological applications, a second-order dependence across signal sets may be optimal [16]. In this article, the IVA-G was employed to implement the JBSS framework. Due to the fact that the most dependent information among all the three facial ROIs is the pulsatile information, it is anticipated that the sources included in the first order of SCV should be selected as the HR source.

D. HR Estimation

After JBSS, the first order of recovered SCVs containing three sources (with each corresponding to one signal set), termed SCV1, will be bandpass filtered [fourth-order Butterworth filter, with the lower cutoff at 0.70 Hz (42 bpm) and the upper cutoff at 2.5 Hz (150 bpm)] and treated as the HR

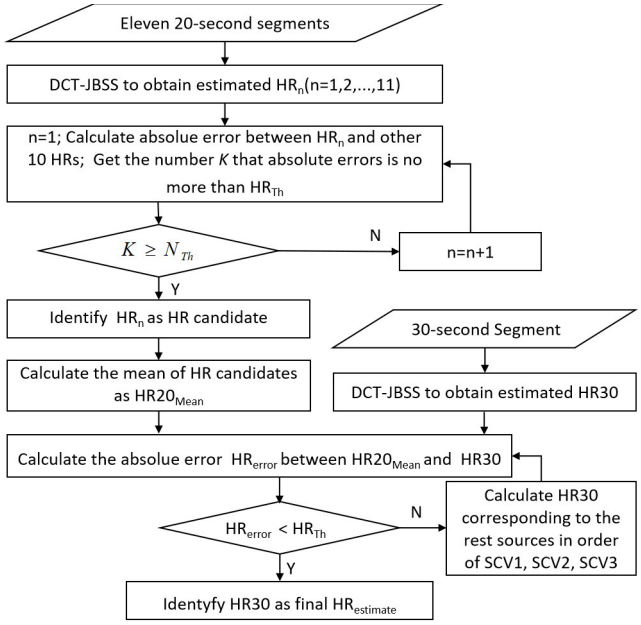

Fig. 3. Flowchart of the DCT-JBSS-M algorithm.

source candidates. Then, the PSD distribution of each HR source candidate is calculated with FFT, where the dominant frequency is determined as $f_{\max}$. The final HR measurement is calculated as $60 \times f_{\max}$ bpm.

When dealing with the final HR measurements, Niu *et al*. [53] proposed to model the temporal relationship of neighboring HR rhythms via HR distribution to improve the performance of the succeeding HR estimations. In our study, there may exist common noises in different ROIs due to contaminated motion artifacts, and the pulsatile signal may appear in other sources with lower SNR in SCV1 or other orders of SCVs rather than SCV1. To solve this problem, we develop a modified DCT-JBSS framework, termed DCT-JBSS-M to modify the HR measurements. The HR continuity from all the subtime-duration (20 s) segments is utilized to get rid of the HR outliers of the 30-s duration segment. The DCT-JBSS-M framework includes two main steps. The first step is to derive the most probable HR of the 30-s segment, represented by the mean HR value ($\mathrm{HR20_{Mean}}$) of all the HR candidates derived from 20-s duration segments. The second step is to select the most appropriate HR from all the recovered SCVs according to $\mathrm{HR20_{Mean}}$. The detailed flowchart is shown in Fig. 3.

Specifically, during Step One, a 30-s duration segment is first divided into 11 20-s duration consecutive segments with 1-s sliding step, which means that there is 19-s overlapping between every two adjacent segments. Therefore, the HRs of these 11 video segments have a certain continuity and small variations, due to the fact that there is a refractory period for an excitable cardiac muscle membrane to be ready for a second stimulus. For each 20-s duration segment, the original DCT-JBSS approach is adopted to measure the HR. Then, a threshold of HR variation during 1 s $\mathrm{HR_{Th}}$ is set to identify HR candidates. HR_n will be identified as the HR candidate if the number (K) of the absolute error (not larger than $\mathrm{HR_{Th}}$) between HR_n derived from current nth 20-s duration segment and the other ten 20-s duration HRs is not smaller than N_{Th}. After all the HR candidates are determined, the mean of these HR candidates $\mathrm{HR20_{Mean}}$ is calculated.

During Step Two, HR30 is calculated from the current 30-s duration segment by DCT-JBSS and may be modified based on the previously obtained $\mathrm{HR20_{Mean}}$. If the absolute error HR_{Error} between HR30 and its $\mathrm{HR20_{Mean}}$ is not larger than the threshold $\mathrm{HR_{Th}}$, the original HR30 will be retained. Otherwise, HR30 corresponding to SCV2 and even SCV3 will be calculated and also compared with $\mathrm{HR20_{Mean}}$, until the absolute error is not larger than $\mathrm{HR_{Th}}$. In this case, the new HR30 will be determined as the target HR. However, if no HR candidate is derived or none of HR30 can satisfy the criterion, the original HR30 derived by DCT-JBSS will be retained. In this article, $\mathrm{HR_{Th}}$ was set as 7 bpm, whereas N_{Th} was set as 5.

IV. Experiments

A. Experimental Setup

To verify the performance of HR measurement based on our proposed method, two public databases, the ULg Multimodality Drowsiness Database (also called DROZY) and MERL-Rice NIR Pulse Data set (MR-NIRP), are employed [54]. DROZY provides multiple modalities of data (contact HR reference and NIR videos) to tackle the design of drowsiness monitoring systems and related experiments. Fourteen healthy subjects (11 females and 3 males), aged 22.7 $\pm$ 2.3 [mean $\pm$ standard deviation (SD)] years old, participated in the data collection. Each subject took three psychomotor vigilance tests (PVTs) over two consecutive days, under conditions of increasing sleep deprivation induced by acute and prolonged waking. The PVTs were all performed in a quiet, isolated laboratory environment, and the room lights were turned off for PVT2 and PVT3. The database contains fully synchronized raw data, including PSG signal, and NIR facial videos. The NIR facial videos were recorded in the MP4 format (compressed videos) using the Microsoft Kinect v2 sensor, with 830-nm active infrared illuminance. The pixel resolution of NIR videos was 512×424. ECG was included in the PSG signal and sampled at 512 Hz, which was the HR reference in our study. There exist some motions such as body swerve and mouth coverage with hands due to drowsiness or sleep. The total number of the NIR videos was 36, and 20 of them were recorded at 30 fps, whereas the rest 16 ones were recorded at 15 fps. However, due to the blurred ECG signal, video 2-1 was excluded, resulting in 19 videos with 30 fps. Each NIR video has a duration of about 10 min, and the first 570 s were utilized for HR analysis. The processing window was set as 30 s, and 19 nonoverlapping segments for each video were obtained. Consequently, 361 video segments with 30 fps were derived, whereas 298 video segments with 15 fps were generated.

MR-NIRP database [25] contains both RGB videos and narrowband NIR videos. Eight healthy subjects (two females and six males), aged 20–40 years old, with varying skin tones (four Indians, three Caucasians, and one Asian), participated

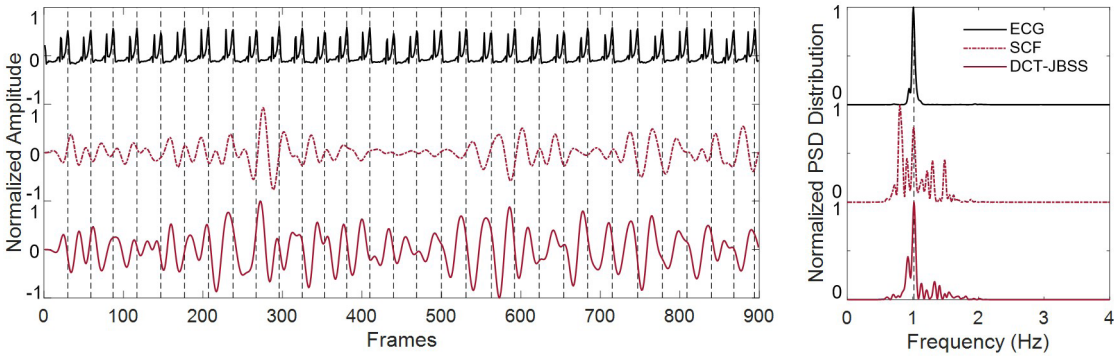

Fig. 4. Taking the fourth segment of the Video 13-2 for instance, the illustration of both the temporal and spectral (PSD) characteristics of the target HR source, the HR reference waveform, and the filtered single-channel signal. (a) Temporal characteristic. (b) Spectral characteristic.

in the data collection. The videos were recorded in an indoor environment, and all subjects were asked to sit still but allowed for natural head motion. The monochrome camera, Point Grey Grasshopper GS3-U3-41C6NIR-C, fitted with a narrowband 940-nm bandpass filter with a 10-nm passband, was used to record NIR images. The raw 10-bit images were recorded with 640×640 resolution at 30 fps, and each video lasted about 3 min. A Contec CMS50D+ finger pulse oximeter was used to obtain a reference PPG waveform recorded at 60 fps. The processing window was also set as 30 s without overlapping, resulting in 51 segments in total.

B. HR Estimation Results and Discussion

To demonstrate the feasibility of our proposed method for remote HR measurement for NIR videos, the same five quality metrics as [55] were employed. Specifically, the MAE $\mathrm{HR}_{\mathrm{mae}}$, the SD $\mathrm{HR}_{\mathrm{sd}}$, the RMSE $\mathrm{HR}_{\mathrm{rmse}}$, the mean error rate percentage $\mathrm{HR}_{\mathrm{mer}}$, and the Pearson's correlation coefficient (CC) r. The SD $\mathrm{HR}_{\mathrm{sd}}$ is defined as

$$\mathrm{HR}_{\mathrm{sd}} = \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mathrm{HR}_e^{(i)} - \overline{\mathrm{HR}_e})^2} \right) \tag{6}$$

where $\mathrm{HR}_e = \mathrm{HR}_{\mathrm{predict}} - \mathrm{HR}_{\mathrm{label}}$ is the error of HR and $\overline{\mathrm{HR}_e}$ indicates the mean value of HR_e. Besides, the percentage of the HR error HR_e is less than 3 or 5 bpm is employed in %, termed PTE3 or PTE5.

On the one hand, although motion-based methods could also achieve a satisfactory performance of HR measurement for NIR videos [24], they usually need at least two IR cameras with different wavelengths. In this article, we only adopted a single NIR camera and excluded these excellent motion-based methods for comparison. On the other hand, we adopted the same public MR-NIRP database to evaluate our proposed DCT-JBSS method, and the baseline SparsePPG method [25] was not reimplemented, but the same performance metrics were compared and discussed. Considering all these factors, five typical iPPG methods, most related to the BSS techniques or the single-channel signal processing strategies, namely, SCF (single channel filtering) [27], [28], EEMD [30], [56], SCICA [8], MRICA (multiregion ICA) [9], and Seg-ICA (segmented ICA) [45], were also implemented on the DROZY and MR-NIRP databases for comparison. For MRICA, the selected three ROIs were the same as our proposed approach. For SCF, EEMD, SCICA, and Seg-ICA, the single-channel signal was generated from the spatial averaging within all the three facial ROIs. For our proposed DCT-JBSS method, the results without and with modification were both recorded.

In our experiment, three orders of SCVs in total were recovered from the three ROI signal sets. Each order of SCV contained three sources, with each corresponding to one signal set. Taking the fourth segment of Video 13-2 for instance, the CC between every two sources from the same order is calculated. The minimum CC of SCV1 is 0.36, whereas the maximum CC can reach 0.73. The minimum CC of SCV2 is only 0.01, whereas the maximum CC is 0.29. As for SCV3, the minimum CC is 0.03, whereas the maximum CC is 0.05. It is proved that the commonly shared source among all the three facial ROIs orders the first and contains the pulsatile information. Besides, through the DCT-JBSS, both the temporal and spectral (PSD) characteristics of the target HR source match those of the HR reference waveform better than the original single-channel signal. The results are shown in Fig. 4, with Fig. 4(a) and (b) corresponding to the temporal and spectral characteristics, respectively. The peaks of the target HR source recovered by DCT-JBSS match well and are more prominent than those recovered by the SCF method. The dominant frequencies of the three sources in SCV1 are 1.23, 0.81, and 1.02 Hz, with corresponding SNRs -6.50, -5.81, and -1.58 dB, respectively. By the proposed DCT-JBSS, the target frequency having the largest SNR is determined as 1.02 Hz, whereas the dominant frequency of the HR reference signal is 1.01 Hz. However, when adopting the SCF method, the target frequency is only 0.79 Hz, less accurate compared with that of the HR reference signal.

For all the 361 segments from DROZY with 30 fps, the percentage of which order the SCV was selected to serve as the target HR source was calculated. The percentage of the target HR source selected from SCV1 is 82.27% (297/361), the percentage of the target HR source selected from SCV2 is 14.14% (51/361), and the percentage of the target HR source selected

TABLE I

PERFORMANCE COMPARISON ON THE DROZY DATABASE USING DIFFERENT iPPG METHODS (30 fps)

	HR_{mae} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer} (%)	r	$PTE5$ (%)
SCF	7.57	9.90	12.45	11.97	0.39	60.39
EEMD	8.39	9.70	12.81	14.54	0.32	53.46
SCICA	7.75	9.82	12.50	12.49	0.52	58.17
MRICA	6.27	8.30	10.39	10.26	0.64	62.88
Seg-ICA	6.85	10.02	10.95	7.32	0.68	56.23
DCT-JBSS	5.46	7.29	9.10	8.73	0.72	67.31
DCT-JBSS-M	3.25	4.06	5.20	5.20	0.90	81.16

TABLE II

PERFORMANCE COMPARISON ON THE MR-NIRP DATABASE USING DIFFERENT iPPG METHODS

	HR_{mae} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer} (%)	r	$PTE5$ (%)
SCF	2.29	3.89	4.48	3.37	0.88	90.20
EEMD	2.41	3.69	4.38	3.58	0.89	86.27
SCICA	2.71	5.25	5.86	3.97	0.81	84.31
MRICA	2.71	4.35	5.08	3.88	0.84	88.24
Seg-ICA	3.86	3.84	5.41	5.56	0.84	68.63
DCT-JBSS	1.75	1.89	2.56	2.53	0.95	90.20

from SCV3 is only 3.60% (13/361). These results have also proved our assumption that the most relevant pulsatile sources can be extracted from the first order of SCV recovered by our proposed DCT-JBSS approach.

Meanwhile, the performance comparison on the DROZY database by applying different iPPG methods for NIR-videos with 30 fps is shown in Table I. It can be seen from Table I that our proposed DCT-JBSS achieves the best performance in terms of all the six quality metrics compared with other five iPPG methods. The MAE HR_{mae} is 5.46 bpm, the SD of HR errors HR_{sd} is 7.29 bpm, the RMSE HR_{rmse} is 9.10 bpm, the mean error rate percentage HR_{mer} is 8.73%, the Pearson's CC r is 0.72, and the percentage of HR error within 5 bpm PTE5 is 67.31%. It can also be seen from Table I that the performance of HR measurement based on SCICA is comparable to that based on SCF. The performance of HR measurement based on MRICA is somewhat better than that based on SCF and SCICA. However, the performance of HR measurement can be significantly improved when combining JBSS with DCT. The metric comparison between the SCICA and the MRICA demonstrates that the quality of the pulsatile information underlying the multiple facial ROIs is better than that of the multichannel signal constructed by applying DCT to the single-channel signal from one single facial ROI. This is also the direct motivation to propose our DCT-JBSS approach. It can be seen from Table I that the proposed DCT-JBSS-M can achieve a further improvement than the original DCT-JBSS. The MAE HR_{mae} is only 3.25 bpm, the SD HR_{sd} is 4.06 bpm, the RMSE HR_{rmse} is 5.20 bpm, the mean error rate percentage HR_{mer} is 5.20%, the Pearson's CC r increases to 0.90, and the PTE5 achieves an improvement of 13.85%, which indicated that when there exist severe motions, our proposed DCT-JBSS methods will also be challenged. The performance can be improved by the DCT-JBSS-M method, considering the property that the pulse cannot dramatically change within a very short time.

The performance comparison on the other MR-NIRP database using all the abovementioned iPPG methods is shown in Table II. It can also be concluded that our proposed DCT-JBSS achieves the best performance in terms of all quality metrics. Since the videos of MR-NIRP database were acquired during indoor environment when the subjects kept still, the motion interference was much slighter than that in the DROZY database. PTE5 derived by DCT-JBSS on the MR-NIPR database is 22.89% higher than that on the DROZY database. Besides, since the largest HR error is not larger than 7 bpm, the advanced DCT-JBSS-M will not work and will be omitted from this database. It can also be found from Table II that PTE5 derived by the SCF is the same as that by DCT-JBSS, whereas PTE5 derived by EEMD, SCICA, or MRICA is at most 5.89% lower, which demonstrates that these iPPG methods work well under stationary situations. We all calculated PTE3 (the HR error within 3 bpm) for all the methods, and our DCT-JBSS can achieve PTE3 with 86.27%, while SCF, EEMD, SCICA, and MRICA can achieve 80.39%, 76.47%, 78.43%, and 78.43%, respectively. Both PTE5 and PTE3 derived by Seg-ICA are much lower than the other five methods, which shows that the Seg-ICA cannot tackle the motion artifacts. This can be explained that the employment of the background ROI is more suitable for illumination variation elimination [45]. Similar conclusions can also be drawn on this database that under relatively static situations, our proposed DCT-JBSS method can reduce the RMSE HR_{rmse} to 2.56 bpm, whereas the SparsePPG method is 1.06 bpm. The percentage of the HR error within 5 bpm PTE5 is 90.20%, and PTE6 is 92.16%, whereas PTE6 of the SparsePPG method is 95.18% [25]. Until now, the SparsePPG method has achieved the best performance of HR measurement on the MR-NIRP database. The comparison between DCT-JBSS and SparsePPG demonstrates that the performance of our DCT-JBSS is quite competitive.

Fig. 5 shows the Bland–Altman analysis [57] between each iPPG method (except Seg-ICA) and the HR reference for DROZY database with 30 fps. From Fig. 5(a)–(d), the estimated HRs based on SCF, EEMD, SCICA, and MRICA are compared to their corresponding HR references, whereas Fig. 5(e) and (f) shows the agreement between the proposed DCT-JBSS and the HR references, as well as between the modified DCT-JBSS-M approach and the HR references. Fig. 5 shows that by SCF, EEMD, SCICA, and MRICA methods, the corresponding 1.96 times SD is at least 20.0 bpm, and the maximum can reach 24.7 bpm. However, by our proposed DCT-JBSS, the 1.96 times SD can decrease to 16.9 bpm. Besides, by the modified version, the 1.96 times SD can further decrease to 9.9 bpm. The smallest mean bias 1.3 bpm can be achieved by the DCT-JBSS-M approach.

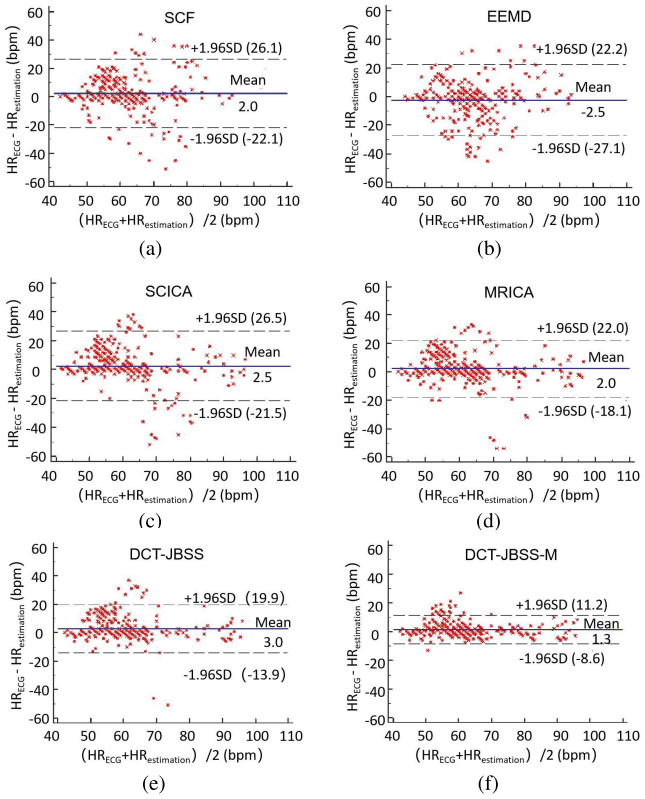

Fig. 5. Bland–Altman plots analyzing the agreement of HR estimation between ECG and each iPPG method (30 fps). (a) SCF. (b) EEMD. (c) SCICA. (d) MRICA. (e) DCT-JBSS. (f) DCT-JBSS-M.

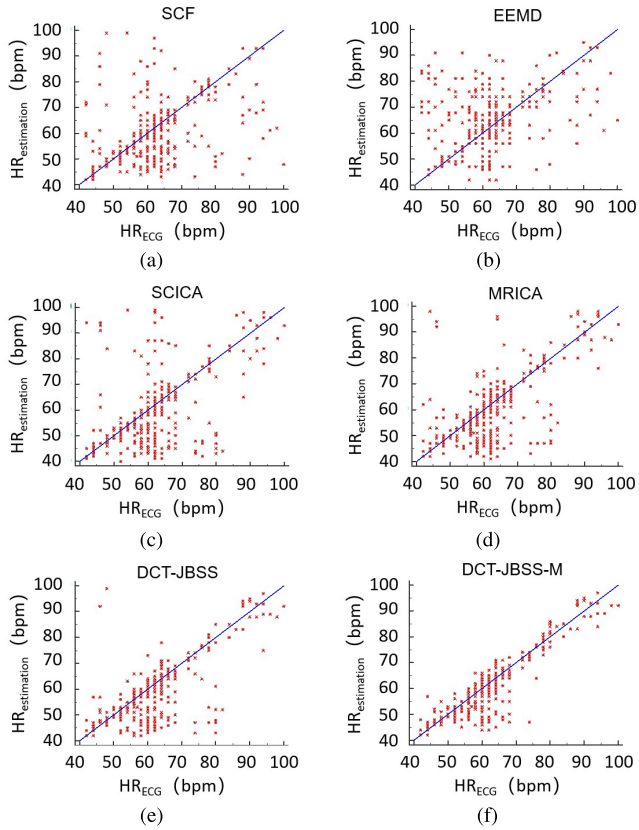

Fig. 6. Scatter plots for HR measurements on the DROZY database using different iPPG methods (30 fps). (a) SCF. (b) EEMD. (c) SCICA. (d) MRICA. (e) DCT-JBSS. (f) DCT-JBSS-M.

Fig. 6 shows the scatter plots for HR measurements on the DROZY database with 30 fps using different iPPG methods (except Seg-ICA), with Fig. 6(a)–(f) corresponding to SCF, EEMD, SCICA, MRICA, the proposed DCT-JBSS, and the DCT-JBSS-M, respectively. It can be seen from Fig. 6 that a lot of inaccurate HR estimations have been obtained by the first five aforementioned methods. The accuracy is improved by our proposed DCT-JBSS method, which can be seen in Fig. 6(e) that the HR measurements are much more concentrated around the baseline $y = x$, showing a stronger correlation. When adopting the modified version, the inaccurate outliers can further be rectified and the correlation can be enhanced. Table I and Figs. 5 and 6 show the significant improvement of our proposed DCT-JBSS and the modified version DCT-JBSS-M.

1) Comparison Between DCT-JBSS and DCT-JBSS-M: Fig. 7 shows the HR error distributions derived from both the proposed DCT-JBSS and the modified version DCT-JBSS-M on the DROZY database with 30 fps. It can be seen that by the proposed DCT-JBSS approach, the HR errors range from -37 to 15 bpm, whereas the modified version DCT-JBSS-M can shrink the range from -27 to 9 bpm. Besides, by the proposed DCT-JBSS, the percentage of the absolute HR error within 5 bpm is 67.31% (243/361), which can be improved to 81.16% (293/361) when adopting DCT-JBSS-M.

Furthermore, for cases that the absolute HR errors were less than 5 bpm (81.16%, 293/361), the selected orders of SCV treated as the target HR source were analyzed. The analysis demonstrates that among these 293 segments, only ten out of them were not selected from SCV1. This proves the assumption of our proposed DCT-JBSS that the first order of SCV is prone to contain pulsatile information for NIR videos during relatively static situations. The conclusion can hold that the performance of HR measurement on the MR-NIRP database is much better than that on the DROZY database. The probable reason is that the movement interference involuntary caused by drowsiness on DROZY is more distinct than that under static situations on MR-NIRP. When evaluated on the MR-NIRP database, the maximum HR error was not larger than 7 bpm, where the advanced DCT-JBSS-M would not be employed.

Fig. 8 shows the HR measurements for DROZY database with 30 fps based on the DCT-JBSS-M approach, where each of the 19 segments corresponds to a single video. It can be seen from Fig. 8 that the overall HR errors of each video are different. Specifically, the HR errors of Video 6-1 (HR_{mae} 6.63 bpm), 9-3 (HR_{mae} 5.84 bpm), and 11-2 (HR_{mae} 5.11 bpm) are significantly larger than the overall error of all the videos (HR_{mae} 3.25 bpm) when adopting DCT-JBSS-M. We further observed these three videos and found that in some clips, the head movements of the subject were large or the ROI was obscured by his/her hand. It indicates that our proposed approach will be challenged by strong motion artifacts or other complex situations, which will be the direction of our future work.

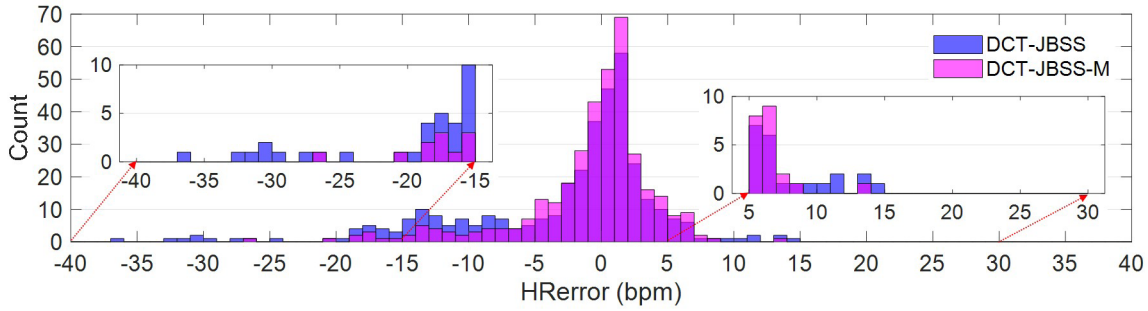

Fig. 7. HR error distributions derived from both DCT-JBSS and DCT-JBSS-M on the DROZY database with 30 fps.

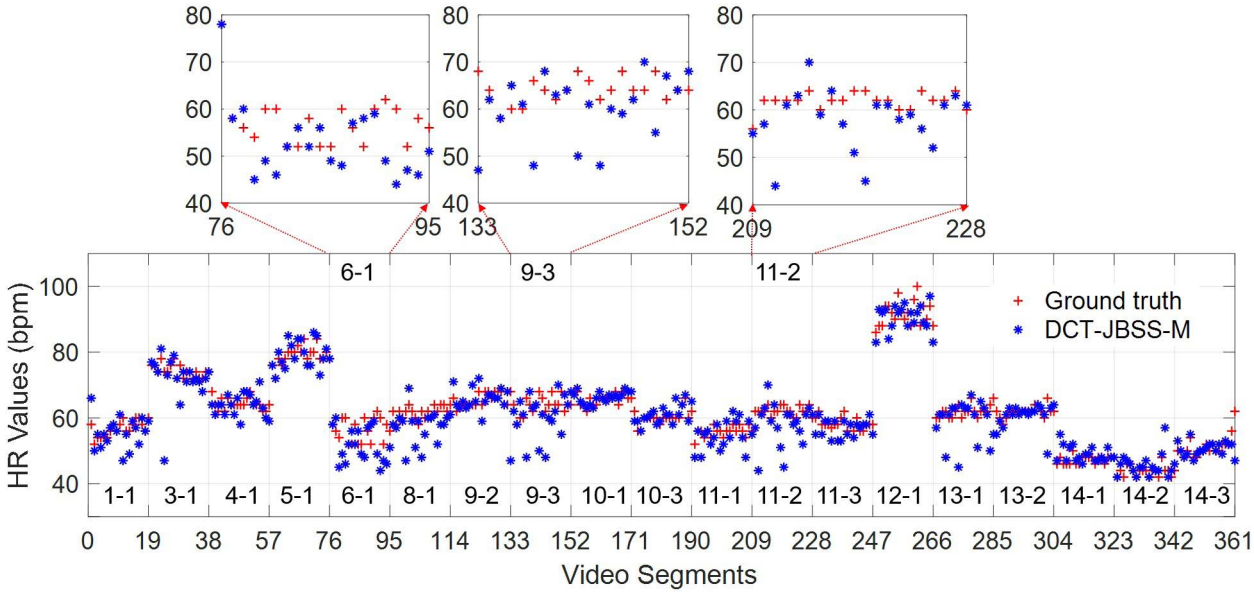

Fig. 8. HR measurements based on DCT-JBSS-M.

2) Influence of the Frame Rate: Besides the adopted 19 NIR videos from DROZY recorded at 30 fps, there are 16 NIR videos from DROZY recorded at 15 fps and can be divided into totally 298 segments. In the experiment, we also evaluated the performance of HR measurement for 15-fps NIR videos from the DROZY database by applying the proposed DCT-JBSS and DCT-JBSS-M methods. The performance was also compared with the aforementioned five methods. The comparison results are shown in Table III, demonstrating that our proposed DCT-JBSS-M also achieves the best performance in terms of all quality metrics. The MAE HR_{mae} is 5.27 bpm, the SD HR_{sd} is 6.60 bpm, the RMSE HR_{rmse} is 8.44 bpm, the mean error rate percentage HR_{mer} is 7.12%, and the Pearson's CC r is 0.67. When comparing Table I with Table III, the overall performance decreases a bit for NIR videos with 15 fps. The probable reason is that signal length has an influence on the quality of the independent sources extracted from BSS methods. We believe that when the frame rate of NIR videos is much higher, the reduced frame rate will have little impact on the performance of HR measurement. Such a conclusion is in accordance with the findings in [12] and [58] that when the frame rate decreased from 120 to 30 fps, little observable difference was observed in terms of MAE or error distributions.

TABLE III
PERFORMANCE COMPARISON ON THE DROZY DATABASE USING DIFFERENT iPPG METHODS (15 fps)

	HR_{mae} (bpm)	HR_{sd} (bpm)	HR_{rmse} (bpm)	HR_{mer} (%)	r	$PTE5$ (%)
SCF	9.39	9.78	13.54	12.51	0.36	48.66
EEMD	9.28	8.79	12.78	13.04	0.25	43.62
SCICA	10.19	9.87	14.17	13.73	0.44	44.27
MRICA	10.28	9.68	14.11	13.68	0.39	49.36
Seg-ICA	9.76	9.44	13.56	13.31	0.45	45.30
DCT-JBSS	7.66	8.22	11.23	10.30	0.53	62.75
DCT-JBSS-M	5.27	6.60	8.44	7.12	0.67	73.88

It should be noticed that when there are common rigid motions that affect the dominant frequency of HR after performing DCT-JBSS, the motion-elimination procedure should be added before DCT-JBSS. This is in accordance with the finding that the spatial input of the (J)BSS techniques is advantageous and the outcome of the (J)BSS techniques is strongly dependent on the input quality [37]. Previous study based on DIScriminative signature-based extraction (DIS) [59]

method has made a great contribution to rPPG measurement using NIR cameras, with the aim of widening the application scope to dark situations while having a comparative performance in presence of motions. A positive assumption can be made that the combination of DIS and DCT-JBSS will promote the performance of HR measurement from multispectral and multiROI NIR videos. Besides, with the help of characterizing the motion trajectory via tracking facial landmarks, the motion can first be extracted and eliminated from facial ROIs to improve the performance of HR measurement. As for illumination variations, the background ROI can be utilized to extract the underlying illumination variation source, such as Seg-ICA [45] or JBSS-EEMD algorithms, and then eliminate this interference source from facial ROIs. This is the future direction of our study under realistic situations.

V. Conclusion

In this article, we proposed a novel noncontact HR measurement method based on JBSS with DCT for NIR videos. First, a separate multichannel signal set was generated by applying DCT to the single-channel signal from each facial ROI. Second, with the help of JBSS that is designed to handle multiple signal sets simultaneously, the underlying pulsatile information commonly shared among different facial ROIs was extracted. The combination of JBSS and DCT can significantly improve the performance of HR measurement for NIR videos under relatively static situations. Other five typical iPPG methods, including SCF, EEMD, SCICA, MRICA, and Seg-ICA, were also employed for comparison on two publicly available DROZY and MR-NIRP databases. The experimental results demonstrated the feasibility of our proposed DCT-JBSS framework. This study will widen the iPPG application during dark situations. In the future, the work focusing on HR measurements in a more challenging situation, including both motion artifacts and illumination variations, such as in-car environment, will be implemented.

References

[1] L. A. M. Aarts *et al.*, "Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit—A pilot study," *Early Hum. Develop.*, vol. 89, no. 12, pp. 943–948, 2013.

[2] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Robust heart rate from fitness videos," *Physiol. Meas.*, vol. 38, no. 6, p. 1023, 2017.

[3] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4264–4271.

[4] X. Chen, J. Cheng, R. Song, Y. Liu, R. Ward, and Z. J. Wang, "Video-based heart rate measurement: Recent advances and future prospects," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 3600–3615, Oct. 2019.

[5] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.

[6] Y. Cho, S. J. Julier, N. Marquardt, and N. Bianchi-Berthouze, "Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging," *Biomed. Opt. Exp.*, vol. 8, no. 10, pp. 4480–4503, 2017.

[7] S. B. Park, G. Kim, H. J. Baek, J. H. Han, and J. H. Kim, "Remote pulse rate measurement from near-infrared videos," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1271–1275, Aug. 2018.

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

[9] R. Favilla, V. C. Zuccala, and G. Coppini, "Heart rate and heart rate variability from single-channel video and ICA integration of multiple signals," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2398–2408, Nov. 2019.

[10] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3640–3648.

[11] B. Wei, X. He, C. Zhang, and X. Wu, "Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions," *Biomed. Eng. OnLine*, vol. 16, no. 1, p. 17, Dec. 2017.

[12] E. B. Blackford and J. R. Estepp, "Effects of frame rate and image resolution on pulse rate measured using multiple camera imaging photoplethysmography," *Proc. SPIE*, vol. 9417, Mar. 2015, Art. no. 94172D.

[13] Z. Guo, Z. J. Wang, and Z. Shen, "Physiological parameter monitoring of drivers based on video data and independent vector analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4374–4378.

[14] H. Qi, Z. Guo, X. Chen, Z. Shen, and Z. Jane Wang, "Video-based human heart rate measurement using joint blind source separation," *Biomed. Signal Process. Control*, vol. 31, pp. 309–320, Jan. 2017.

[15] J. Cheng, X. Chen, L. Xu, and Z. J. Wang, "Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 5, pp. 1422–1433, Sep. 2017.

[16] X. Chen, Z. J. Wang, and M. J. McKeown, "Joint blind source separation for neurophysiological data analysis: Multiset and multimodal methods," *IEEE Signal Process. Mag.*, vol. 33, no. 3, pp. 86–107, May 2016.

[17] A. Al-Naji, K. Gibson, S.-H. Lee, and J. Chahl, "Monitoring of cardiorespiratory signal: Principles of remote measurements and review of methods," *IEEE Access*, vol. 5, pp. 15776–15790, 2017.

[18] D. J. McDuff, J. R. Estepp, A. M. Piasecki, and E. B. Blackford, "A survey of remote optical photoplethysmographic imaging methods," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 6398–6404.

[19] J. Przybyło, E. Kańtoch, M. Jabłoński, and P. Augustyniak, "Distant measurement of plethysmographic signal in various lighting conditions using configurable frame-rate camera," *Metrol. Meas. Syst.*, vol. 23, no. 4, pp. 579–592, Dec. 2016.

[20] S. A. Siddiqui, Y. Zhang, Z. Feng, and A. Kos, "A pulse rate estimation algorithm using PPG and smartphone camera," *J. Med. Syst.*, vol. 40, no. 5, p. 126, May 2016.

[21] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 463–477, Mar. 2016.

[22] A. Al-Naji and J. Chahl, "Simultaneous tracking of cardiorespiratory signals for multiple persons using a machine vision system with noise artifact removal," *IEEE J. Translational Eng. Health Med.*, vol. 5, pp. 1–10, 2017.

[23] V. Jeanne, M. Asselman, B. den Brinker, and M. Bulut, "Camera-based heart rate monitoring in highly dynamic light conditions," in *Proc. Int. Conf. Connected Vehicles Expo (ICCVE)*, Dec. 2013, pp. 798–799.

[24] M. van Gastel, S. Stuijk, and G. de Haan, "Motion robust remote-PPG in infrared," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1425–1433, May 2015.

[25] E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1272–1281.

[26] S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, "Remote heart rate measurement from RGB-NIR video based on spatial and spectral face patch selection," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 5676–5680.

[27] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Opt. Exp.*, vol. 16, no. 26, pp. 21434–21445, 2008.

[28] J. Chen *et al.*, "RealSense = real heart rate: Illumination invariant heart rate estimation from videos," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

[29] D. Y. Chen *et al.*, "Image sensor-based heart rate evaluation from face reflectance using Hilbert–Huang transform," *IEEE Sensors J.*, vol. 15, no. 1, pp. 618–627, Jan. 2015.

[30] Q. Zhang, Y. Zhou, S. Song, G. Liang, and H. Ni, "Heart rate extraction based on near-infrared camera: Towards driver state monitoring," *IEEE Access*, vol. 6, pp. 33076–33087, 2018.

[31] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, Aug. 2012.

[32] X. He, R. Goubran, and F. Knoefel, "IR night vision video-based estimation of heart and respiration rates," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Mar. 2017, pp. 1–5.

[33] 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.

[34] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Exp.*, vol. 18, no. 10, pp. 10762–10774, 2010.

[35] Y. Sun, "Motion-compensated noncontact imaging photoplethysmography to monitor cardiorespiratory status during exercise," *J. Biomed. Opt.*, vol. 16, no. 7, Jul. 2011, Art. no. 077010.

[36] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3640–3648.

[37] S. Zaunseder, A. Trumpp, D. Wedekind, and H. Malberg, "Cardiovascular assessment by imaging photoplethysmography–a review," *Biomed. Eng./Biomedizinische Technik*, vol. 63, no. 5, pp. 617–634, 2018.

[38] X. Niu, S. Shan, H. Han, and X. Chen, "RhythmNet: End-to-End heart rate estimation from face via spatial-temporal representation," *IEEE Trans. Image Process.*, vol. 29, pp. 2409–2423, 2020.

[39] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 349–365.

[40] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," 2019, *arXiv:1905.02419*. [Online]. Available: http://arxiv.org/abs/1905.02419

[41] S. Kwon, J. Kim, D. Lee, and K. Park, "ROI analysis for remote photoplethysmography on facial video," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4938–4941.

[42] M. A. Hassan *et al.*, "Heart rate estimation using facial video: A review," *Biomed. Signal Process. Control*, vol. 38, pp. 346–360, Sep. 2017.

[43] S. Huynh, R. K. Balan, J. Ko, and Y. Lee, "VitaMon: Measuring heart rate variability using smartphone front camera," in *Proc. 17th Conf. Embedded Netw. Sensor Syst.*, 2019, pp. 1–14.

[44] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. I-511–I-518.

[45] J. Hu, Y. He, J. Liu, M. He, and W. Wang, "Illumination robust heart-rate extraction from single-wavelength infrared camera using spatial-channel expansion," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 3896–3899.

[46] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

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

[48] P. E. McSharry and G. D. Clifford, "A comparison of nonlinear noise reduction and independent component analysis using a realistic dynamical model of the electrocardiogram," *Proc. SPIE*, vol. 5467, May 2004, pp. 78–88.

[49] C. J. James and D. Lowe, "Extracting multisource brain activity from a single electromagnetic channel," *Artif. Intell. Med.*, vol. 28, no. 1, pp. 89–104, May 2003.

[50] M. C. Aston, Philip and M. Nandi, "Delay coordinate analysis of periodic data," U.S. Patent 9 940 741, Apr. 10, 2018.

[51] T. Buzug and G. Pfister, "Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global static and local dynamical behavior of strange attractors," *Phys. Rev. A, Gen. Phys.*, vol. 45, no. 10, p. 7073, 1992.

[52] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.* Berlin, Germany: Springer, 2006, pp. 165–172.

[53] X. Niu, H. Han, S. Shan, and X. Chen, "Continuous heart rate measurement from face: A robust rPPG approach with distribution learning," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 642–650.

[54] Q. Massoz, T. Langohr, C. Francois, and J. G. Verly, "The ULg multimodality drowsiness database (called DROZY) and examples of use," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–7.

[55] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, "Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7411–7421, Oct. 2020.

[56] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, Jan. 2009.

[57] J. M. Bland and D. G. Altman, "Agreement between methods of measurement with multiple observations per individual," *J. Biopharmaceutical Statist.*, vol. 17, no. 4, pp. 571–582, Jul. 2007.

[58] Y. Sun, S. Hu, V. Azorin-Peris, R. Kalawsky, and S. Greenwald, "Noncontact imaging photoplethysmography to effectively access pulse rate variability," *J. Biomed. Opt.*, vol. 18, no. 6, Oct. 2012, Art. no. 061205.

[59] W. Wang, A. C. den Brinker, and G. de Haan, "Discriminative signatures for remote-PPG," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 5, pp. 1462–1473, May 2020.