Motion-Robust Multimodal Heart Rate Estimation Using BCG Fused Remote-PPG With Deep Facial ROI Tracker and Pose Constrained Kalman Filter

Yiming Liu, Binjie Qin, *Member, IEEE*, Rong Li, Xintong Li,
Anqi Huang, Haifeng Liu, Yisong Lv, and Min Liu

Abstract—**The heart rate (HR) signal is so weak in remote photoplethysmography (rPPG) and ballistocardiogram (BCG) that HR estimation is very sensitive to face and body motion disturbance caused by spontaneous head and body movements as well as facial expressions of subjects in conversation. This article proposed a novel multimodal quasi-contactless HR sensor to ensure the robustness and accuracy of HR estimation under extreme facial poses, large-motion disturbances, and multiple faces in a video for computer-aided police interrogation. Specifically, we propose a novel landmark-based approach for a deep facial region of interest (ROI) tracker and face pose constrained Kalman filter to continuously and robustly track target facial ROIs for estimating HR from face and head motion disturbances in rPPG. This motion-disturbed rPPG signal is further fused with a minimally disturbed BCG signal by the face and head movements via a bank of notch filters with a recursive weighting scheme to obtain the dominant HR frequency for final accurate HR estimation. To facilitate reproducible HR estimation researc(o)1.(l)-(31.(fi(wi2 u.(ncia)7(c).(ilita)7(t)0.(e)-2.7(r)17.(e).(p)-3.(r)17.(o)7(duc).(ible)-2**

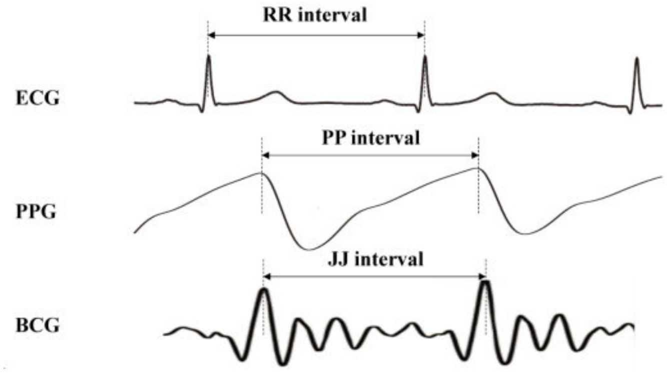


Fig. 1. Examples of typical ECG, PPG, and BCG physiological signals with high correlation among their peak-to-peak intervals.

that is ejected and moved during each cardiac cycle [4]. The BCG signal is a combination of cardiac activities, respiratory activities, and body movements such that the BCG signal can simultaneously reveal a person's HR and RR. In addition to the BCG technique, the rPPG-based method is an emerging branch of PPG, which is a simple and low-cost video-based biomonmotion disturbances as well as different movement types introduced in the data set.

This article is organized as follows. We review the related works on contactless HR estimation in Section II. The design of DFT and pose constrained ROI landmark KF for rPPG as well as motion-artifact correction via BCG fusion are presented for HR estimation in Section III. An illustration of the experimental results is presented in Section IV. Conclusions and discussions on our work are presented in Section V.

II. Related Works

HR estimation based on rPPG needs to track several areas of facial skin, such as the forehead and cheek regions, which are denoted as ROIs, to obtain high-quality rPPG signals [12]. Subsequently, the light intensities of the spatially averaged pixel values in the facial ROIs are filtered to recover the rPPG signal. However, when the suspect's face, head, and body move, it is more difficult to extract HR signals from the contactless rPPG compared to contact PPG measurement methods for the following reasons: 1) it is difficult to ensure that face ROIs are always correctly identified and tracked in rPPG measurement and 2) in rPPG measurement, the relative position and orientation between the camera and moving facial tissue change frequently with the distance being largely varied such that the radiant flux on the ROI and its camera response, as well as disturbances from light sources, are largely varied to introduce serious motion artifacts in the distorted rPPG signal. To solve these two motion-caused problems, more intelligent facial ROI trackers and motion-artifact suppression for rPPG are worth studying in this article.

To achieve an intelligent facial ROI tracker, face detection [13] must first be implemented to determine where the target face is located when there are occasionally several faces or face occlusions in the video sequence. After face detection, the whole facial region should be continuously tracked via object tracking algorithms. Among these object tracking methods, Kanade–Lucas–Tomasi (KLT) [14] based on sparse optical flow vectors from good features (such as corners) in two subsequent frames of a video can achieve fast face tracking after the manual definition of the target face in an environment where the brightness of the object is assumed to remain invariant. Some rPPG works [15]–[17] used only KLT to track a person's face. However, due to optical flow equations relying on the first-order Taylor expansion and easily breaking down when large motions occurred between sequential frames, KLT tracking accuracy on unsolved challenges inherent in the optical flow technique [18], such as large face movement and partial occlusion cases as well as handling textureless facial areas, is not ideal for estimating motion-robust HR for rPPG [16], [17].

By introducing powerful multicue and multidimensional features, including both handcrafted and deep neural network features, discriminative correlation filtering (DCF) algorithms [19], [20] have been proved to achieve more accurate tracking but are somewhat more computationally expensive than others. Therefore, an efficient convolution operator (ECO) algorithm [21] for object tracking was proposed with a compact generative model and factorized convolution operator to cluster historical frames and employ dimension reduction to reduce memory and time complexity. Some researchers have demonstrated that ECO tracking accuracy and real-time performance are superior to previous object tracking algorithms, which is then an important motivation of this work for integrating an ECO-based face tracker into facial ROI tracker design for motion-robust rPPG.

After the face tracker obtains the data matrix of the face, the desired facial ROIs containing the high-quality rPPG signal should be continuously and accurately identified or tracked. Traditional methods generally use face segmentation [22], [23] and face alignment [24] to achieve facial ROI tracking. Usually, the entire face generated by face segmentation is denoted as ROI. This method is simple in principle and fast in implementation. Its core idea is to define an "explicit skin cluster" classifier that expressly defines the boundaries of the skin cluster in color space [22]. However, when illumination is locally uneven and the background color is close to the skin color, the ROI tracked by this segmentation method is usually incoherent and contains noisy background areas. Pursche *et al.* [25] used a CNN to select ROI and compared the effectiveness of network based on a different number of training samples. The ROIs calculated by CNN lead to significantly better and faster results compared to ROIs from classical approaches. For face alignment, Kazemi and Sullivan [24] used an ensemble of regression trees (ERT) to estimate the landmark positions of faces. The facial ROI was then tracked from the landmark coordinates. The ERT optimized the sum of square error loss and naturally handled missing or partially labeled data. It achieved face alignment in milliseconds with high-quality predictions. To guarantee face alignment accuracy in extreme face pose or occlusion situations, a recently proposed practical facial landmark detector (PFLD) [26] was designed with a dual network structure to implement a backbone network for predicting landmark position and used an auxiliary network for face pose determination for regularizing face landmark localization in the backbone network. However, when the face has large movement, the landmark localization by the alignment-based method still has errors and introduces abrupt shifts in the facial ROI. The Kalman filter (KF) [17] is assumed to be capable of modeling head motion for correction of landmark coordinate errors of the landmarks generated by PFLD. Therefore, we are inspired to conduct a deep study on this facial ROI tracking problem in large face movement disturbances.

The PPG- and BCG-based robust HR measurement algorithms with motion-artifact suppression can be divided into three categories: the blind source separation (BSS)-based algorithm, and the model-based and deep-learning-based algorithms. BSS refers to extracting a source signal from a mixed signal without knowing the mixing process in advance. Among BSS algorithms, independent component analysis (ICA) is a commonly used method. Some ICA-based methods are applied to rPPG to estimate HR, and the accuracy of experiments proves their feasibility. However, it assumes that the distribution of different signals is statistically independent and non-Gaussian [27]. To calculate the decomposition matrix, sufficiently long signal data is necessary for data analysis.

Therefore, it cannot guarantee real-time and high-accuracy performance for real applications.

The model-based algorithm uses prior knowledge of different color components to achieve cardiac signal separation. De Haan and Jeanne [28] proposed the chrominance-based (CHROM) method, which needs a constant "skin-tone" vector under white light to help suppress the effect of motion disturbance. This constant vector was experimentally determined and is not invariant in different experimental environments. Therefore, the accuracy of estimating HR in different experimental environments varies greatly. Afterward, the blood-volume pulse vector-based (PBV) method [29] was proposed to improve motion robustness. The PBV method utilized the blood volume change signature to distinguish pulse-induced color changes from motion artifacts. The covariance matrix of the color data matrix should be calculated in the PBV method. Then, the matrix is inverted for subsequent calculation. However, if the matrix is not invertible, the algorithm cannot complete the extraction of the cardiac signal. Later, Wang *et al.* [11] compared the previous BSS-based and model-based algorithms and proposed a new model-based rPPG algorithm called POS, which outperformed other algorithms via experimental comparison in recent review work [5].

Most deep learning methods [8]–[10] are inherently data-driven and supervised such that they depend on various large labeled data sets to accommodate the diversity of data sets acquired from different video devices and the large variation in various head motions and lighting conditions. For example, deep skin segmentation [10] via nonskin and skin classification requires considerable human effort and training data to implement skin labeling and annotation for HR estimation. The learned mapping from these training data sets to the desired skin segmentation prediction is achieved by setting large parameters of deep neural networks to minimize the specific distance measure (or loss function) between the ground-truth label and the deep model's predicted segmentation. This learned mapping over training examples is thus very dependent on the trained data set and labeling such that it is insensitive and ineffective to the newly acquired data sets with their specific skin properties, challenging light conditions, and unexpected large-motion disturbances. Therefore, there may be some tradeoff between motion robustness and measurement accuracy for newly acquired data sets from real scenarios. Other distortion artifacts, such as the artifacts caused by video compression, can be referred to in [30]. A detailed comparison and review of the rPPG algorithm can be seen in the newly published review papers [5], [31], [32]

Many existing methods have reported their performance using private databases that only consist of videos and gold-standard signals, such as ECG or PPG. The MAHNOB-HCI database [33] was first used for remote HR estimation. Face videos, audio signals, eye gaze, and peripheral/central nervous system physiological signals, including HR with small head movement and facial expression variation under laboratory illumination, were recorded. Stricker *et al.* [34] released the PURE database consisting of 60 videos from ten subjects, in which all the subjects were asked to perform six kinds of movements, such as talking or head rotation. Reference data were captured in parallel using a finger clip pulse oximeter. Hsu *et al.* [35] released the PFF database, consisting of 104 videos of ten subjects, in which only small head movements and illumination variations were involved and ground-truth results were recorded using the MIO Alpha II wrist wearable device. These two databases are limited by the number of subjects and recording scenarios, making them unsuitable for building a real-world HR estimator. Soleymani *et al.* [33] built a large-scale multimodal HR database (named VIPL-HR), which consists of 2378 visible face videos and 752 NIR face videos from 107 subjects. Three different recording devices (RGB-D camera, smartphone, and web camera) were used for video recording, and the PBV signal of the fingertip oximeter served as the ground truth.

III. Materials and Methods

The proposed multimodal quasi-contactless HR sensor fuses two different physiological sensors to estimate HR. Specifically, a DARMA optic fiber-based BCG sensor with a sampling rate of 50 Hz and an FLIR BLACKFLY BFS-U3-19S4 RGB camera are used to build our multimodal HR sensor and acquire the corresponding multimodal data set. The duration of each sample is 30 s, and ten volunteers are involved in the acquisition of these multimodal data.

For a better explanation, we generally divide motion disturbances into two cases in the acquisition of these multimodal data sets: the SS without obvious large body and head movement and the MS with the subject's body moving and the head varying yaw angle being exceeding 30° or varying coordinates exceeding 30 pixels. Specifically, in the MS, the subjects played a game called "Mafia" [36], in which the "mafia" has to cheat the "detectives" and "ordinary citizens" and vote on a player to eliminate in each round. The players communicate and function in a way that resembles real interrogation situations. In the discussion part of the game, every "mafia" member should make a statement. When they lied or were queried, large emotional fluctuations may emerge, which led to large body motions and large variations in head movements and facial expressions. The video sequences capturing the face image in 25 frames/s with a resolution of 640 × 480 were recorded by a camera. The experimental results were compared with ground-truth HR acquired by Heal Force's three lead PC-80B ECG Monitor. The damaged segments in the ECG signal (for example, due to body movement or equipment motion) were manually commented and deleted to ensure the correctness of the reference value. In the experiment, the reference ECG signals that were eliminated did not exceed 5% of the total signal. The whole framework of the motion-robust quasi-contactless HR sensor is schematically shown in Fig. 2.

A. Motion-Robust rPPG via DFT and Face Pose Constrained KF

In general, rPPG is very dependent on facial ROI selection for HR estimation. With the facial ROI selection strategy mentioned in Section II, we used the face-alignment-based ROI algorithm. The proposed motion-robust rPPG is shown

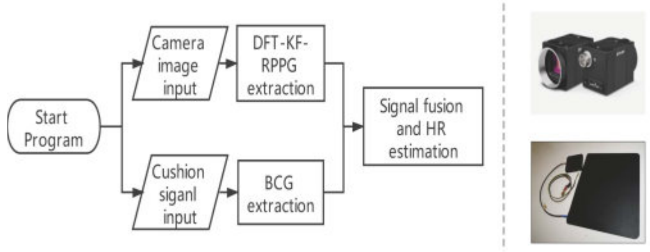


Fig. 2. Framework of a multimodal motion-robust quasi-contactless HR sensor.

in Fig. 3. The proposed rPPG sensor consists of three important submodules that are listed as follows.

1) A DFT is implemented via the integration of two subsequent modules as follows. A deep face tracker selects and tracks the target face image matrix via the face detector and the state-of-the-art ECO tracker, and the sequential face image matrix in the video is then aligned with the recently introduced PFLD facial landmark detector network [26] to achieve facial ROI tracking.
2) This robust DFT further corrects the sequential face landmarks' error using Kalman filtering of landmark coordinates and prior constraints of face pose information.
3) The POS method [11] is utilized to extract the pulsatile signal from the target pixels in the facial ROIs.

1) Deep Face Tracker: Our deep face tracker is built on the classical face detector and the recently introduced object tracker. Specifically, an image matrix containing the target face is selected semiautomatically. Then, a classical face detector via the OpenCV Haar classifier based on the Viola–Jones algorithm [13] is utilized to detect whether the selected area contains faces. In the real process of police interrogation, nontarget faces may be captured during video recording. This may cause an incorrect selection of the target facial ROI. Traditional face detection methods have no way to determine which face belongs to the concerned suspect/witness. To utilize the close correlation of target faces in sequential video frames to deal with nontarget face disturbances, we believe that correlation filtering-based tracking algorithms can continuously track target faces at rapid speeds with high accuracy.

Considering the robustness in situations with large head rotation and real-time requirements of the desired facial tracker, we utilize the online ECO [21] tracker once we acquire the target face central coordinates of the "face rectangle" and its length and width in the initial frame by the Viola–Jones algorithm. We input the original image as well as the central coordinates, length and width of the head region into the ECO tracker, which will have a high response to the target face and a low response to the background in the next few frames.

As a discriminative correlation filter-based tracking method, the ECO tracker adopts a continuous convolution operator tracker (C-COT) as the baseline to extract multiresolution facial feature maps by performing convolutions in the continuous domain without the need for explicit resampling. With

TABLE I

BACKBONE [illegible] CONFIGURATION. EACH LINE REPRESENTS A SEQUENCE OF IDENT[illegible] LAYERS, REPEATING n TIMES. ALL LAYERS IN THE SAME S[illegible]CE HAVE THE SAME NUMBER c OF OUTPUT CHAN[illegible] FIRST LAYER OF EACH SEQUENCE HAS A STR[illegible]XPANSION FACTOR t IS ALWAYS APPLIED [illegible] TO THE INPUT SIZE

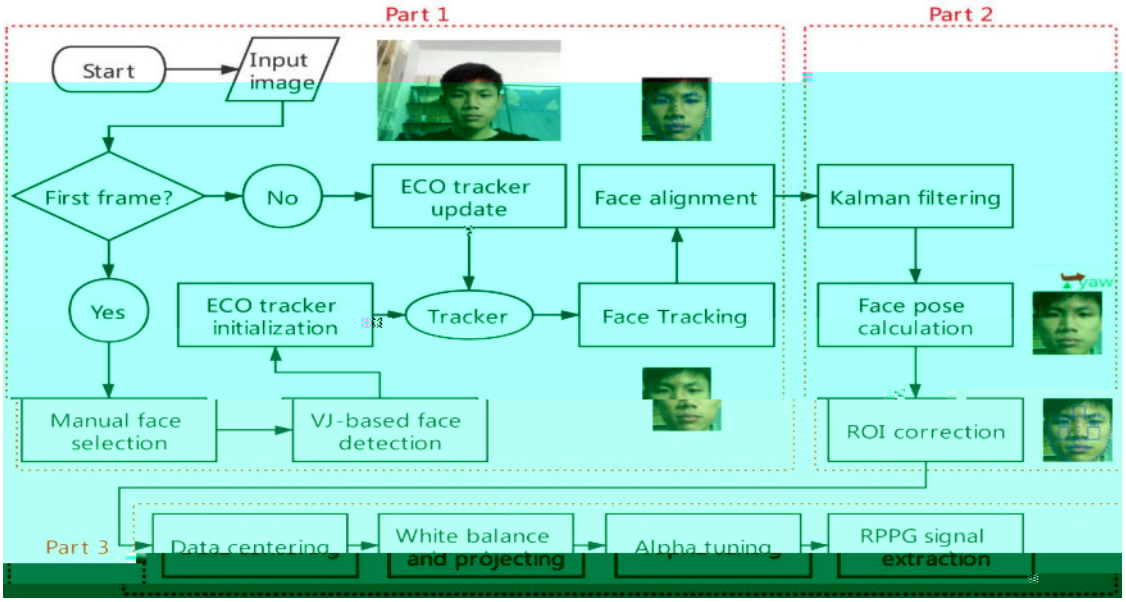
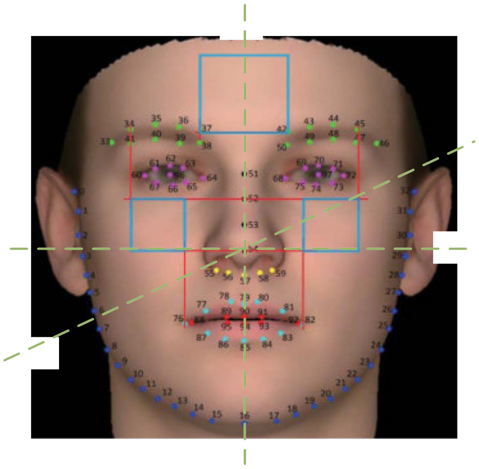


Fig. 3. Data processing flowchart for the proposed motion-robust rPPG with a DFT and face pose constrained KF.



for landmarks can be modeled in the following state equation:

$$\begin{aligned}\mathbf{s}_k &= \begin{bmatrix} x_k \\ \dot{x}_k \\ y_k \\ \dot{y}_k \end{bmatrix} = \begin{bmatrix} 1 & h & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & h \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ \dot{x}_{k-1} \\ y_{k-1} \\ \dot{y}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{h^2}{2} & 0 \\ h & 0 \\ 0 & \frac{h^2}{2} \\ 0 & h \end{bmatrix} \begin{bmatrix} \ddot{x}_{k-1} \\ \ddot{y}_{k-1} \end{bmatrix} \\ &\quad + \mathbf{w}_k \\ &= \mathbf{A}\mathbf{s}_{k-1} + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k. \end{aligned} \tag{2}$$

The landmark observation (measurement) of the tracking process is assumed to occur at discrete points in time in accordance with the following measurement equation:

$$\mathbf{m}_k = \mathbf{H}_k\mathbf{s}_k + \mathbf{v}_k \tag{3}$$

where $\mathbf{s}_k$ is the predicted facial landmark state vector $[x_k, \dot{x}_k, y_k, \dot{y}_k]$ in frame F, $\mathbf{s}_{k-1}$ is the existing facial landmark state vector for frame $\mathbf{F-1}$ (current frame), and $\mathbf{u}_k$ denotes the acceleration vector of the landmark in frame $\mathbf{F}$, which is always ignored. $\mathbf{w}_k$ and $\mathbf{v}_k$ are assumed to be a white sequence and are known as the process noise and measurement noise, respectively, of the landmark in frame $\mathbf{F}$; $\mathbf{A}$ is the usual state transition matrix reflecting the effect of the previous state on the current state. The matrix $\mathbf{B}$ is the optional control input, which is always ignored with the acceleration vector $\mathbf{u}_k$ of the landmark. The matrix $\mathbf{H}$ in the measurement (3) gives a noiseless connection between the landmark state $\mathbf{s}$ and the measurement $\mathbf{m}$ in the current frame of the video sequence. The covariance matrices for $\mathbf{w}_k$ and $\mathbf{v}_k$ are given by

$$E\big[\mathbf{w}_k\mathbf{w}_i^T\big] = \begin{cases} \mathbf{Q}_k, & i = k \\ \mathbf{0}, & i \neq k \end{cases} \tag{4}$$

$$E\big[\mathbf{v}_k\mathbf{v}_i^T\big] = \begin{cases} \mathbf{R}_k, & i = k \\ \mathbf{0}, & i \neq k \end{cases} \tag{5}$$

$$E\big[\mathbf{w}_k\mathbf{v}_i^T\big] = \mathbf{0}, \quad \text{for all } k \text{ and } i \tag{6}$$

where $\mathbf{0}$ denotes a matrix with zero elements. The respective covariance matrices, $\mathbf{Q}_k$ and $\mathbf{R}_k$, are assumed to be known.

By initializing KF filtering at some point t_k, we have a prior landmark position estimate denoted as $\hat{\mathbf{s}}_k^-$ and the corresponding error $\hat{\mathbf{e}}_k^- = \mathbf{s}_k - \hat{\mathbf{s}}_k^-$ having its prior covariance matrix $\mathbf{P}_k^- = E[\hat{\mathbf{e}}_k^-\hat{\mathbf{e}}_k^{-T}] = E[(\mathbf{s}_k - \hat{\mathbf{s}}_k^-)(\mathbf{s}_k - \hat{\mathbf{s}}_k^-)^T]$. With the prior estimate $\hat{\mathbf{s}}_k^-$, we now use the measurement $\mathbf{m}_k$ to improve the prior estimate. To this end, we adopt the following update recursion:

$$\hat{\mathbf{s}}_k = \hat{\mathbf{s}}_k^- + \mathbf{K}_k\big(\mathbf{m}_k - \mathbf{H}_k\hat{\mathbf{s}}_k^-\big) \tag{7}$$

where the updated (posterior) estimate is equal to the prior estimate plus a correction term, which is proportional to the error in predicting the newly arrived observation vector and its prediction based on the prior estimate. Matrix $\mathbf{K}_k$, known as the Kalman gain, controls the amount of correction, and its value is determined to minimize the following mean square error $J(\mathbf{K}_k)$ derived from the trace of posteriori error covariance matrix associated with the updated estimate since the trace is the sum of the mean square errors in the estimates of all the elements of the state vector:

$$J(\mathbf{K}_k) \equiv E\big[\mathbf{e}_k^T\mathbf{e}_k\big] = \text{trace}\{\mathbf{P}_k\} \tag{8}$$

where

$$\mathbf{P}_k = E\big[\mathbf{e}_k\mathbf{e}_k^T\big] = E[(\mathbf{s}_k - \hat{\mathbf{s}}_k)(\mathbf{s}_k - \hat{\mathbf{s}}_k)^T]. \tag{9}$$

After substituting (4) into (8) and then substituting the resulting expression for $\hat{\mathbf{s}}_k$ into (9) as well as using $\mathbf{P}_k^- = E[(\mathbf{s}_k - \hat{\mathbf{s}}_k^-)(\mathbf{s}_k - \hat{\mathbf{s}}_k^-)^T]$ as a prior estimation error, which is uncorrelated with the current measurement error $\mathbf{v}_k$, we obtain the following result:

$$\begin{aligned}\mathbf{P}_k &= E\big\{\big[\big(\mathbf{s}_k - \hat{\mathbf{s}}_k^-\big) - \mathbf{K}_k\big(\mathbf{H}_k\mathbf{s}_k + \mathbf{v}_k - \mathbf{H}_k\hat{\mathbf{s}}_k^-\big)\big] \\ &\qquad \times \big[\big(\mathbf{s}_k - \hat{\mathbf{s}}_k^-\big) - \mathbf{K}_k\big(\mathbf{H}_k\mathbf{s}_k + \mathbf{v}_k - \mathbf{H}_k\hat{\mathbf{s}}_k^-\big)\big]^T\big\} \\ &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\ &= \mathbf{P}_k^- - \mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^- - \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T + \mathbf{K}_k\big(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\big)\mathbf{K}_k^T. \end{aligned} \tag{10}$$

We proceed to differentiate the trace of $\mathbf{P}_k$ with respect to $\mathbf{K}_k$ and note that the trace of $\mathbf{P}_k^-\mathbf{H}_k^T\mathbf{K}_k^T$ is equal to the trace of its transpose $\mathbf{K}_k\mathbf{H}_k\mathbf{P}_k^-$. The derivative result is

$$\frac{d(\text{trace } \mathbf{P}_k)}{d\mathbf{K}_k} = -2\big(\mathbf{H}_k\mathbf{P}_k^-\big)^T + 2\mathbf{K}_k\big(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\big). \tag{11}$$

We set the derivative equal to zero and obtain the following optimal Kalman gain:

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T\big(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\big)^{-1}. \tag{12}$$

The posterior error covariance matrices $\mathbf{P}_k$ for the optimal estimate are now computed and related to the prior error covariance matrix $\mathbf{P}_k^-$ by substituting the optimal gain expression into (10) as follows:

$$\begin{aligned}\mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{P}_k^-\mathbf{H}_k^T\big(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\big)^{-1}\mathbf{H}_k\mathbf{P}_k^- \\ &= \mathbf{P}_k^- - \mathbf{K}_k\big(\mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k\big)\mathbf{K}_k^T \\ &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^-. \end{aligned} \tag{13}$$

Note that we need prior estimate $\hat{\mathbf{s}}_k^-$ and covariance matrix $\hat{\mathbf{P}}_k^-$ to assimilate the measurement $\mathbf{m}_k$ for the updated estimate $\hat{\mathbf{s}}_k$ by the use of (7), and we can expect such a similar need at the next iteration to make optimal use of the next measurement $\mathbf{m}_{k+1}$. The updated $\hat{\mathbf{s}}_k$ is projected forward as $\hat{\mathbf{s}}_{k+1}^- = \mathbf{A}\hat{\mathbf{s}}_k$ via the transition matrix while ignoring the contribution of $\mathbf{w}_k$ due to (4).

The prior error covariance matrix $\mathbf{P}_{k+1}^-$ associated with $\hat{\mathbf{s}}_{k+1}^-$ is obtained by transforming the prior error $\mathbf{e}_{k+1}^- = \mathbf{s}_{k+1} - \hat{\mathbf{s}}_{k+1}^- = (\mathbf{A}\mathbf{s}_k + \mathbf{w}_k) - \mathbf{A}\hat{\mathbf{s}}_k = \mathbf{A}\mathbf{e}_k + \mathbf{w}_k$, that is, we can write the expression for $\mathbf{P}_{k+1}^-$ as follows by considering that $\mathbf{w}_k$ is the process noise for the previous state and has zero cross correlation with $\mathbf{e}_k$:

$\mathbf{P}_{k+1}^- = E\big[\mathbf{e}_{k+1}^-\mathbf{e}_{k+1}^{-T}\big] = E[(\mathbf{A}\mathbf{e}_k + \mathbf{w}_k)($sAeö

of its tranf..7-.03 -.0 TD(k)TjF 1 Tf.2 0 0 32.2(its)-33 TD-.. 0e1updat

motion, which influences the geometric structure between the light source, skin surface, and camera. $p(t)$ denotes the cardiac pulse signal that we are interested in.

A 3×3 normalization matrix $\mathbf{N}$ with constraint $\mathbf{N} \cdot \mathbf{u}_c \cdot I_0 \cdot c_0 = \mathbf{1}$ is used to temporally normalize $\mathbf{x}(t)$ as

$$\bar{\mathbf{x}}(t) = \mathbf{N} \cdot \mathbf{x}(t) \approx \mathbf{1} \cdot (1 + i(t)) + \mathbf{N} \cdot \mathbf{u}_s \cdot I_0 \cdot s(t) + \mathbf{N} \cdot \mathbf{u}_p \cdot I_0 \cdot p(t). \quad (20)$$

This temporal normalization can simply be implemented by dividing its samples by their mean over a temporal interval, i.e., $\bar{\mathbf{x}}(t) = \mathbf{x}(t)/\mu(\mathbf{x})$, where $\mu(\mathbf{x})$ can be a running average centered around a specific image or an average of an overlap-add processing interval that includes the specific image. In either case, the temporal normalization is preferably taken over a number of images such that the interval contains at least a pulse period. This temporal normalization can

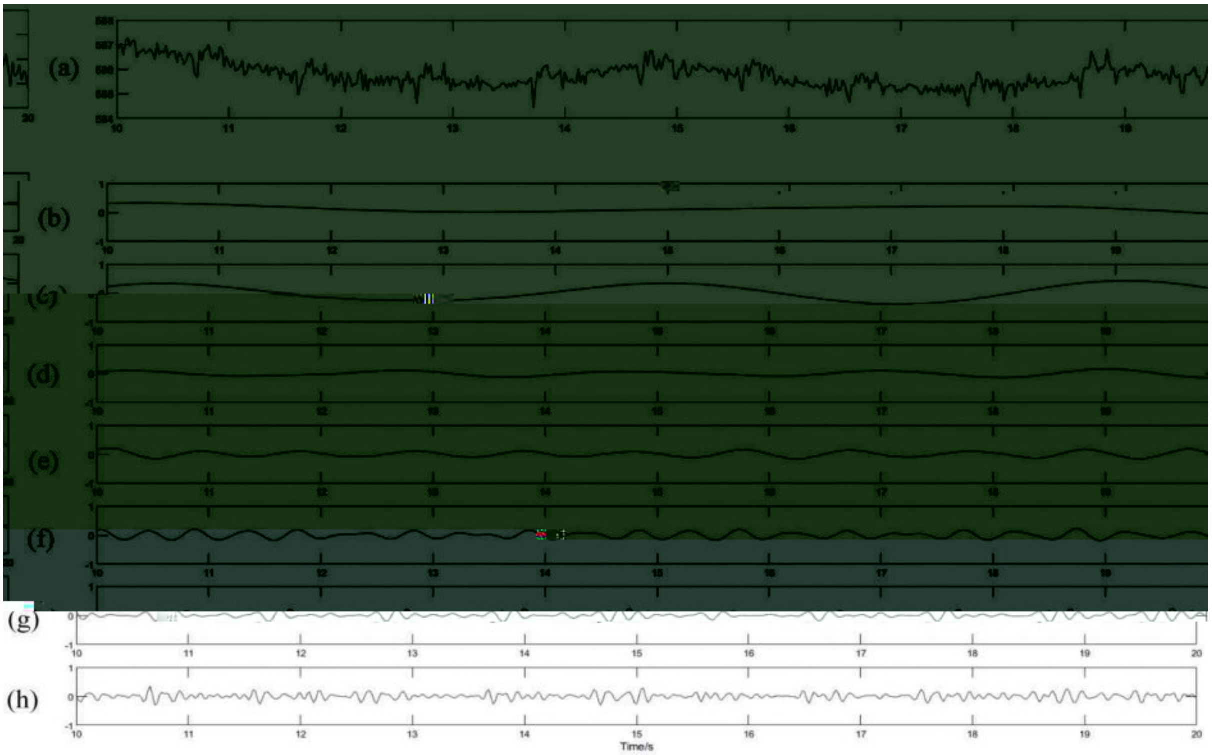


Fig. 6. BCG signal decomposition: (a) the original BCG signal; (b) the 8th level harmonic; (c) the 7th level harmonic; (d) the 6th level harmonic; (e) the 5th level harmonic (**selected component**); (f) the 4th level harmonic; (g) the 3rd level harmonic; (h) the 2nd level harmonic.

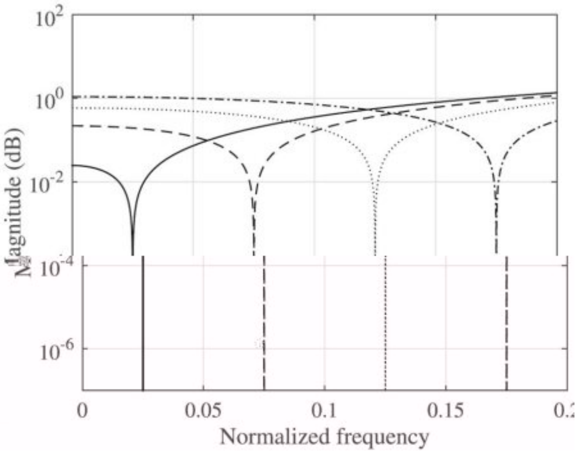


Fig. 7. Frequency response curve of length-3 FIR notch filters with different stopband frequencies f_i. The closer the dominant frequency of the input signal is to f_i, the greater the attenuation of the filtered signal.

of the notch filter, small output signals should be given more weight, whereas large output signals should be given less weight. Thus, we define $\mathbf{W}_i[m]$ for every frequency f_i

$$\mathbf{W}_i[m] = \exp\left(-\gamma \frac{1}{S} \sum_{j=1}^{S} \mathbf{R}[m, j]\mathbf{P}_i[m, j]\right) \tag{28}$$

where $\gamma = [\min_{i=1,\ldots,F}(\mathbf{R}[m, j]\mathbf{P}_i[m, j])]^{-1}$ and $\mathbf{R}[m, j]$ for $j = 1, \ldots, S$ are a set of weights related to the input signals. $\mathbf{R}$ are defined as the signal-to-output power ratios of the input signals for a notch filter centered on the target frequency. The signal-to-output ratios are computed and normalized to create a set of weights $\mathbf{R}$ for the S inputs as

$$\mathbf{R}[m, j] = \frac{\mathbf{U}[m, j]/\mathbf{O}[m, j]}{\sum_{j=1}^{S} \mathbf{U}[m, j]/\mathbf{O}[m, j]} \tag{29}$$

where $\mathbf{O}[m, j]$ is the mean squared value of the input

$$\begin{aligned}\mathbf{O}[m, j] &= \delta \mathbf{O}[m-1, j]\\ &\quad + (1-\delta) \sum_{k=1}^{freq} \mathbf{y}_f^2[(m-1) * freq + k, j]\end{aligned} \tag{30}$$

which is initialized to $\mathbf{O}[2, j] = U(\text{freq} + 1) + U(\text{freq} + 2) + \cdots + U(2*\text{freq})$ and $U(x) = (\mathbf{u}[x, j] - 2\mathbf{u}[x-1, j]\cos(2\pi f_1) + \mathbf{u}[x-2, j])^2$ and $\mathbf{y}_f$ are an output from a notch filter centered at the estimated frequency of the previous sample $(m-1)$

$$\begin{aligned}\mathbf{y}_f[n, j] &= \mathbf{u}[n, j] - 2\mathbf{u}[n-1, j]\cos(2\pi f[m-1])\\ &\quad + \mathbf{u}[n-2, j]\end{aligned} \tag{31}$$

where $f[m-1]$ is the previously estimated frequency (initialized to $f[2] = f_1$). The final frequency (HR) estimation of each second is then computed as the weighted sum of the notch frequencies of the filter bank

$$f[m] = \frac{\sum_{i=1}^{F} \mathbf{W}_i[m] f_i}{\sum_{i=1}^{F} \mathbf{W}_i[m]}. \tag{32}$$

IV. Experimental Results

To evaluate our multimodal sensor via comparison with other state-of-the-art methods, we first evaluate the effect of

DFT and corresponding motion-artifact suppression in the proposed method by replacing the traditional facial ROI tracking method (KLT + ERT) in the POS-based HR estimation framework with the proposed DFT and KF algorithms for experimental comparison. Thirty videos in SSs and motion disturbances are acquired and analyzed by the classical and proposed methods. Specifically, the state-of-the-art methods for comparison are as follows: MODWT-BCG [4], ICA [42], PBV [29], and POS [11] methods. The following metrics are used to evaluate the performances of facial ROI tracker and HR estimation.

1) *Mean Frame Rate (MFR):* The average number of video frames that the program can process in one second

$$\mathrm{MFR} = \frac{1}{N}\sum_{n=1}^{N} F(n) \tag{33}$$

where $F(n)$ represents the number of video frames which has been processed in the nth second.

2) *Tracker Quality (TQ):* We define the TQ metric as the ratio of the number of pixels n_{eff} in the valid facial areas to the number of pixels n_{all} in the overall ROIs. The valid facial areas are determined by manual marking. The TQ's value range should be [0, 1]. When the measured value is close to 1, it means that the face tracker has achieved high-precision tracking

$$\mathrm{TQ} = \frac{n_{\mathrm{eff}}}{n_{\mathrm{all}}}. \tag{34}$$

3) *Mean Absolute Error (MAE):* We use this metric to compare our method with other methods on the HR estimation accuracy and compare the effect of each module in the algorithm on the accuracy of the entire algorithm

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N} \left|\mathrm{HR}_{\mathrm{est}}^{n} - \mathrm{HR}_{g}^{n}\right| \tag{35}$$

where $\mathrm{HR}_{\mathrm{est}}^{n}$ is the estimation of HR and HR_{g}^{n} is the ground truth of HR.

4) *Root-Mean-Square Error (RMSE):* We use RMSE to measure the difference between the reference HR and the HR calculated from the video. RMSE represents the sample standard deviation of the absolute difference between the reference value and the measured value, that is, the smaller the RMSE is, the more accurate the HR estimation is

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \left(\mathrm{HR}_{\mathrm{est}}^{n} - \mathrm{HR}_{g}^{n}\right)^{2}}. \tag{36}$$

5) *Pearson Correlation of HR:* The Pearson correlation r is applied to evaluate the correspondence of HR between the quasi-contactless signal and the ECG-reference

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}. \tag{37}$$

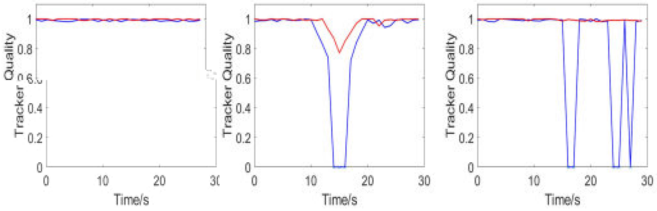
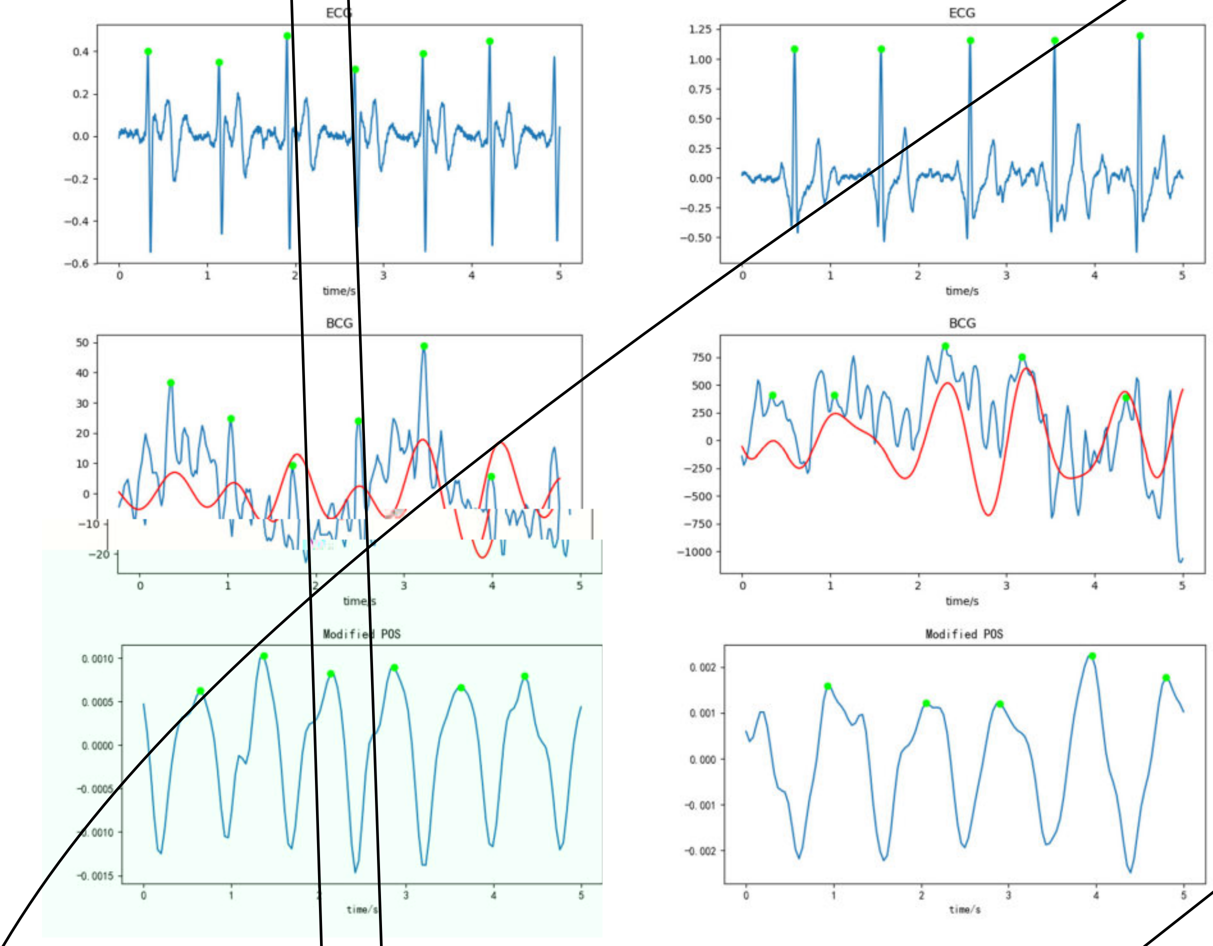


Fig. 8. Tracking quality in three cases of an SS, MS, and nontarget face entrance for the proposed DFT-KF and traditional KLT + ERT methods. The red lines for the DFT-KF method show better tracking quality than the blue lines for the KLT + ERT method.

As shown in Fig. 8(a), in the SS, there is no obvious difference between the proposed DFT-KF and traditional KLT + ERT methods in terms of tracking quality. Fig. 8(b) and (c) shows that the proposed method outperforms traditional methods in the two cases of motion and nontarget entrance disturbances.

We further evaluate different ROI tracking methods on five persons in terms of the mean TQ and MFR. Based on the above two metrics, the correctness and real-time performance of ROI selection can be evaluated. We collect 30-s-long data from the ECG, BCG, and rPPG sensors for each sensor. To ensure that the data are simultaneously collected at the same time in the log file, all the data collection programs were run on one computer. The log file saved the time node in each sampling for data alignment. 34283/F218.7(-)-2a785corro42nFFlaK3.8(w)1T.8(w)1



(a)

(b)

Fig. 9. Comparison of pulse signal extraction from different HR sensors in two states. The J-peak in the original BCG signal and the peak of the decomposed signal based on MODWT (red line) have evident correlations with the R-peak of the ECG signal in the SS. In the MS, there is no correlation between the J-peak of BCG and the R-peak of the ECG signal, whereas the P-peak of the rPPG signal from the proposed method approximately corresponds to the R-peak of the ECG signal. (a) Stable state. (b) MS.

TABLE II

Comparison of Different ROI Tracking Modules in SS and MS

Category	Mean Frame Rate	TQ (SS)	TQ (MS)
KLT-ERT	10.5	0.95	0.44
DFT (CPU)	5.4	0.97	0.82
DFT (GPU)	19.8		
DFT-KT (CPU)	5.3	0.97	0.92

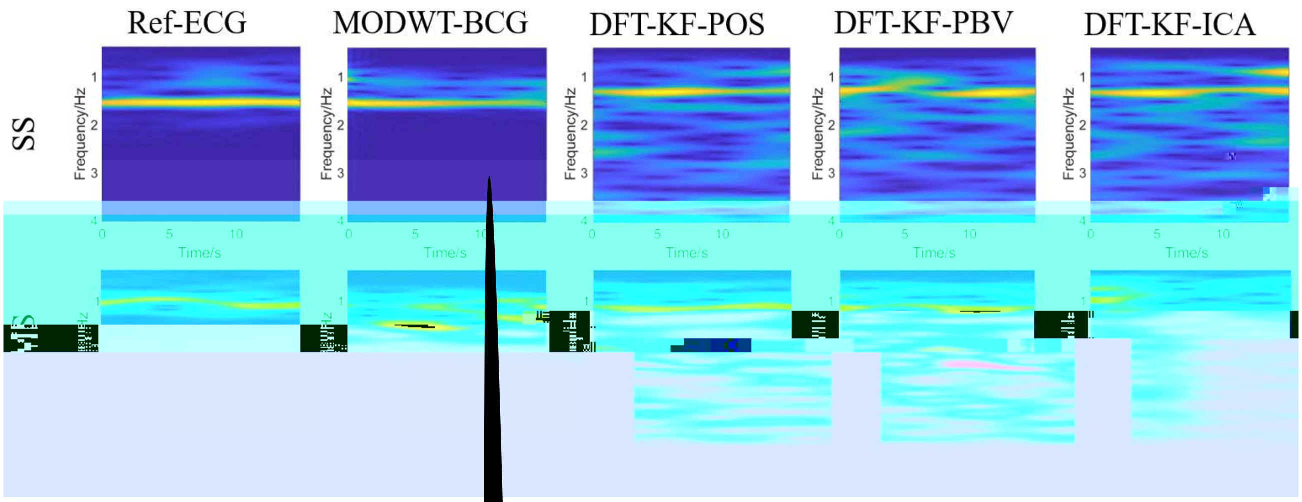


Fig. 10. Short-time Fourier transform spectra obtained by ECG-reference, MODWT-BCG, DFT-KF-POS, and two other rPPG algorithms based on DFT. The ground-truth ECG signal and the MODWT-BCG signal in the SS highlight the exact HR component (as one yellow stripe) having higher kurtosis and SNR with a clear contrast to the blue background. In both SS and MS, the STFT spectrum of the proposed DFT-KF-POS multimodal method contains more focused stipes that correspond to the several harmonics of HR frequency compared with the other two rPPG methods based on DFT and KF algorithms.

TABLE III
MAE OF DIFFERENT HR ESTIMATION METHODS IN THE SS AND MS MOTION

Method	MAE (SS)\bmp	MAE (MS)\bmp
	2.33	**1[illegible].3**

Finally, we conduct DFT-KT-ICA and BCG and DFT-KT-POS and BCG method comparisons. The Pearson correlation and Bland–Altman plots [44] are reported in Figs. 11 and 12, respectively. The RMSE of DFT-KT-POS and BCG is lower, and the correlation coefficient is higher than that of DFT-KF-ICA and BCG. The distance between limit lines (dotted line) and arithmetic mean of DFT-KT-POS and BCG is smaller. This means that DFT-KT-POS and BCG is more reliable in long-term HR estimation.

V. Conclusion

In this article, we propose a multimodal quasi-contactless HR sensor that can be used in computer-aided police interrogation by fusing optical-fiber-based BCG with video-based rPPG physiological signals via a microbending fiber-optic cushion sensor and RGB camera. We design a DFT via face alignment and object tracking technology, as well as a face pose constrained KF, to improve the robustness of the rPPG algorithm in extreme poses, motion disturbances, and multiplayer scenes. It can realize face tracking and correct selection of ROI in challenging situations, such as face occlusion, multiple faces, and large-angle rotation of the target face in real police interrogation.

The characteristics of these two multimodal signal types under different MSs were analyzed. In a relatively SS, the HR calculated based on the optical-fiber-based BCG sensor is more accurate than that calculated based on the video-based rPPG sensor. When the distortion of motion artifacts on the BCG signal is more intense, the video-based rPPG sensor produces more accurate HR estimation than the BCG sensor. The notch filters applied for two signal sources calculate the weights of different discrete frequencies. Simultaneously, the current HR estimation result is compensated by the consistent HR estimation in the past result. The multimodal HR sensor has higher accuracy than the method solely based on single-modal rPPG or BCG-based HR sensor.

More advanced rPPG-based contactless HR sensors with detail-preserving noise removal [45], [46], long-term face occlusion, as well as face and body shake resistance [47] will be developed in future work to be more robust and accurate to large-motion disturbances in various challenging conditions for calculating more useful physiological indices, such as respiration rate, HR variability, and blood pressure [45], in computer-aided police interrogation.

Acknowledgment

The authors would like to thank all the reviewers for their valuable comments on this article. They would also like to thank the volunteers from Shanghai Jiao Tong University for their efforts in creating the benchmark data set.

References

[1] K. Fox *et al.*, “Resting heart rate in cardiovascular disease,” *J. Amer. College Cardiol.*, vol. 50, no. 9, pp. 823–830, 2007.

[2] G. Duran, I. Tapiero, and G. A. Michael, “Resting heart rate: A physiological predicator of lie detection ability,” *Physiol. Behav.*, vol. 186, pp. 10–15, Mar. 2018.

[3] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, “Stress and heart rate variability: A meta-analysis and review of the literature,” *Psychiatry Investig.*, vol. 15, no. 3, p. 235, 2018.

[4] I. Sadek and J. Biswas, “Nonintrusive heart rate measurement using ballistocardiogram signals: A comparative study,” *Signal, Image Video Process.*, vol. 13, no. 3, pp. 475–482, Apr. 2019.

[5] 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.

[6] A. Alivar *et al.*, “Motion artifact detection and reduction in bed-based ballistocardiogram,” *IEEE Access*, vol. 7, pp. 13693–13703, 2019.

[7] C. Bruser, J. M. Kortelainen, S. Winter, M. Tenhunen, J. Parkka, and S. Leonhardt, “Improvement of force-sensor-based heart rate estimation using multichannel data fusion,” *IEEE J. Biomed. Health Informat.*, vol. 19, no. 1, pp. 227–235, Jan. 2015.

[8] X. Niu, S. Shan, H. Han, and X. Chen, “RhythmNet: End-to-end heart

[28] G. de Haan and V. Jeanne, “Robust pulse rate from chrominance-based rPPG,” *IEEE Trans. Biomed. Eng.*