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 research, we synchronously acquire and publicly share a multimodal data set that contains 20 sets of ECG and BCG signals as well as uncompressed, rPPG-dedicated videos from ten subjects in a stable state and large-motion state (MS) without and with large face and body movements in a sitting position. We demonstrate through experimental comparisons that the proposed multimodal HR sensor is more robust and accurate than the state-of-the-art single-modal HR sensor solely with rPPG- or BCG-based methods. The mean absolute error (MAE) of HR estimation is 7.13 BPM lower than the BCG algorithm and 3.12 BPM lower than the model-based plane-orthogonal-to-skin (POS) algorithm in the MS.**

***Index Terms*— Ballistocardiogram (BCG), deep facial ROI tracker (DFT), heart rate (HR) estimation, Kalman filter (KF), motion disturbances, multimodal heart rate sensor, remote photoplethysmography (rPPG).**

I. Introduction

HEART rate (HR), HR variability, and respiratory rate (RR) are important physiological indices [1], [2] of a person's health and mental state. The more accurate and timely we can estimate the HR and HRV, the more we can fully utilize these physiological indices to help estimate the mental state. By utilizing these physiological indices from emotional arousal, the polygraph was first invented by James McKenzie [2]. Usually, modern polygraphs record breathing patterns, cardiovascular activities, and electrodermal responses through direct person-to-device contact, which is inconvenient for use in real applications. In addition, the cost of polygraph equipment is high. Therefore, there is an increasing need to develop a modern polygraph system with contactless physiological sensors that are cost-effective and extremely easy to use. Several contactless or quasi-contactless physiological sensors, such as the RGB camera for remote photoplethysmography (rPPG) sensing and a fiber-optic cushion for ballistocardiogram (BCG) detection, can be expeditiously applied to computer-aided police interrogation for the best balance of smart public security construction and personal privacy protection. Fig. 1 shows the typical HR waveform from an electrocardiogram (ECG) signal, PPG, and BCG signal. There is a high correlation between the different peak-to-peak intervals of ECG, PPG, and BCG signals, i.e., RR interval, PP interval, and JJ interval. The RR interval is often used as a reference standard [3] for evaluating the PP interval from PPG and the JJ interval from BCG in performing HR and HRV calculations. In actual scenes, the suspect sitting face-to-face during police interrogation is completely free to various types of body movements such that the target physiological signal acquired from these sensors will be dramatically affected by motion artifacts. Therefore, accurately extracting the target physiological signals from the complex motion artifacts is a challenging problem in the real process of police interrogation. Sadek

BCG cushion is a noninvasive HR measurement method that measures the body movements produced by the blood

Manuscript received January 8, 2021; accepted February 8, 2021. Date of publication February 19, 2021; date of current version March 5, 2021. This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 19dz1200500 and Grant 19411951507, in part by the National Natural Science Foundation of China under Grant 61271320 and Grant 61871206, and in part by the Shanghai Jiao Tong University Cross Research Fund for Translational Medicine under Grant ZH2018ZDA19. The Associate Editor coordinating the review process was Dr. Anirban Mukherjee. (Yiming Liu, Binjie Qin, and Rong Li are co-first authors.) *(Corresponding authors: Binjie Qin; Min Liu.)*

Yiming Liu, Binjie Qin, and Anqi Huang are with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: bjqin@sjtu.edu.cn).

Rong Li and Min Liu are with the Criminal Investigation Department, Shanghai Public Security Bureau, Shanghai 200025, China (e-mail: minliu110@126.com).

Xintong Li and Haifeng Liu are with ECData Information Technology Company Ltd., Shanghai 200127, China.

Yisong Lv is with the School of Continuing Education, Shanghai Jiao Tong University, Shanghai 200240, China.

Digital Object Identifier 10.1109/TIM.2021.3060572

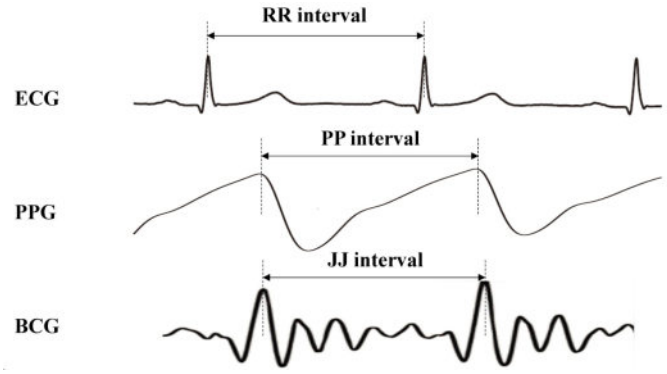


Fig. 1. Examples of typical ECG, PPG, and BCG physiological signals with high correlation among their peak-to-peak intervals.

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 biomonitoring technique for detecting cardiorespiratory activities by monitoring the illuminance variation in diffuse reflection [5] in the face image of video sequences. The periodic variation in the illuminance is caused by a change in the amount of hemoglobin molecules and proteins in blood vessels to directly reflect the HR information. Different from traditional PPG, where a sensor is deployed close to the tissue, rPPG needs to robustly and accurately select an effective facial region of interest (ROI) from large face and head movements in real time by image recognition technology. Due to the above-mentioned different operating principles, the BCG is thus sensitive to body movements, whereas the rPPG is the most susceptible to the face and head movements during face-to-face conversation in a sitting position.

For BCG, Sadek and Biswas's [4] "maximal overlap discrete wavelet transform" (MODWT)-based BCG algorithm has been proved to achieve more accurate HR estimation than fast Fourier transform, cepstrum, and autocorrelation-based methods. However, in the motion state (MS), the amplitude of the force signal caused by body movement far outweighs the pulsatile signal. HR is rarely detected by the traditional model-based method. To solve this problem, machine-learning-based methods have been proposed. For example, Alivar *et al.* [6] proposed a two-stage algorithm using a sequential detection algorithm to determine whether the BCG signal frame is corrupted by motion and building a parametric model (autoregressive model, Wiener smoother estimator) of the BCG signal to reconstruct the motion-corrupted signal. To separate the target HR signal from the large-motion artifacts in a single-channel BCG sensor, the fusion of multiple channel BCG sensors was implemented to benefit the extraction of sensitive pulsatile signals from large-motion artifacts. For instance, Brüser *et al.* [7] used a Bayesian fusion approach on multiple BCG signal sources to achieve a more accurate HR estimation. Inspired by the above works of BCG and rPPG with their differential sensitivities to different motion disturbances, the proposed contactless rPPG sensor explores the fusion of BCG signal measurements for accurate and timely HR estimation.

To the best of our knowledge, this is the first study to propose a multimodal quasi-contactless HR sensor by fusing optical-fiber-based BCG into motion-robust video-based rPPG, which is implemented by proposing a landmark-based deep facial ROI tracker (DFT) and pose constrained Kalman filter (KF) to enhance the performance of rPPG with further motion correction of BCG fusion for accurately estimating the HR from various motion disturbances. This DFT tracker gets the full benefit of large-data-driven deep learning strategy that can easily access to large amount of annotated video data for human-face tracking in the research field of computer vision. Compared with the limitation of small amount of annotated physical data for purely data-driven rPPG methods [8]–[10], the proposed method combines data-driven DFT with model-driven rPPG method [11] to ensure generalization and robustness of the facial ROI tracker for accurately estimating HR from head-motion-disturbed rPPG signals. The contributions of this article are summarized as follows.

1) To our best knowledge, this is the first work to propose a BCG-rPPG multimodal HR sensor (called DFT-KF-plane-orthogonal-to-skin (POS) multimodal HR sensor) by exploiting the differential sensitivities of BCG and rPPG to body movements and head and face movements in a sitting position. This multimodal sensor can be smoothly run from a stable state (SS) to a large-MS for accurately estimating HR from various motion disturbances for computer-aided police interrogation.
2) We propose a new facial ROI tracker via integration of data-driven and model-based methods to accurately track the target facial ROIs for model-based rPPG. Specifically, a DFT is implemented by integrating a deep-learning-based face tracker and its landmark alignment, which is further compensated for landmark position errors via a face pose constrained KF of moving landmarks for correcting abrupt motion disturbances of facial ROIs to obtain a motion-robust rPPG signal.
3) We further exploit a bank of FIR notch filters to estimate the instantaneous frequency of BCG and rPPG signals in a real-time manner. The specific outputs of the filter bank are fused in a recursive weighting scheme to obtain the dominant HR frequency for final accurate HR estimation from the two sources of input signals. Specifically, we map the output-to-input ratios of all notch filters into adaptive weights via an exponential function and perform a weighted summation of the dominant frequencies corresponding to all notch filters to obtain the final HR of two signals.
4) We synchronously build a multimodal HR database that consists of 20 sets of data from ten subjects in stable and large MSs without and with large face and head movements as well as body movements in the sitting position. To the best of our knowledge, this is the first data set that synchronously records ECG, BCG original signals, and rPPG-dedicated video sequences for HR estimation with different amplitudes and time scales of

motion 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

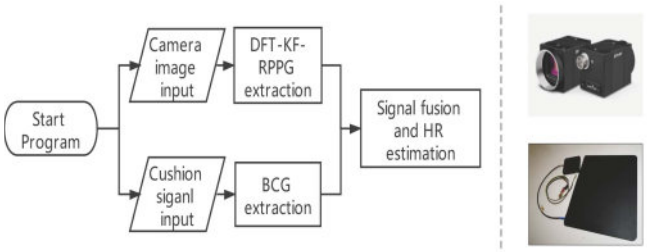


Fig. 2. Framework of a multimodal motion-robust quasi-contactless HR sensor.

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 the advantage of C-COT's target detection scores predicted as a continuous function enabling accurate subgrid localization, ECO constructs a filter as a linear combination of basis filters and introduces a factorized convolution operator for a continuous T-periodic multichannel convolution filter to jointly learn the basis filter and coefficients in matrix factorization optimization. This factorization strategy reduces the number of parameters in the deep facial tracker, which is further combined with a compact generative model of the training sample distribution to drastically reduce computational complexity while enhancing sample diversity. The state-of-the-art facial tracking performance is further assured by a simple yet effective deep model update strategy that reduces overfitting to recent facial object samples.

TABLE I

BACKBONE NET CONFIGURATION. EACH LINE REPRESENTS A SEQUENCE OF IDENTICAL LAYERS, REPEATING n TIMES. ALL LAYERS IN THE SAME SEQUENCE HAVE THE SAME NUMBER c OF OUTPUT CHANNELS. THE FIRST LAYER OF EACH SEQUENCE HAS A STRIDE s. THE EXPANSION FACTOR t IS ALWAYS APPLIED TO THE INPUT SIZE

Input	operator	t	c	n	s
$112^2 \times 3$	Conv3 × 3	-	64	1	2
$56^2 \times 64$	Depthwise Conv3 × 3	-	64	1	1
$56^2 \times 64$	Bottleneck	2	64	5	2
$28^2 \times 64$	Bottleneck	2	128	1	2
$14^2 \times 128$	Bottleneck	4	128	6	1
$14^2 \times 128$	Bottleneck	2	16	1	1
(S1)$14^2 \times 16$	Conv3 × 3	-	32	1	2
(S2)$7^2 \times 32$	Conv7 × 7	-	128	1	1
(S3)$1^2 \times 128$	-	-	128	1	-
(S1,S2,S3)	Full Connection	-	196	1	-

2) DFT With Pose Constrained KF: After the deep face tracker successfully finds the matrix containing the target face in the sequential input images, the facial matrix is passed into the deep landmark detector called PFLD [26] for facial ROI generation. Specifically, the facial landmarks are first predicted by a backbone CNN-based PFLD network, where the multiscale features from the global structures of the human face, such as symmetry and spatial relationship between eyes, mouth, and nose, are directly extracted via several convolutional layers and used for landmark prediction. To overcome the bottleneck in terms of processing speed and model size in the backbone network, a simple/small MobileNet architecture built on depthwise separable convolution filters is adopted to replace the traditional convolution operation in PFLD to significantly reduce the computational load and accelerate the speed of the backbone network. To further reduce the number of feature maps without accuracy degradation, the backbone network is compressed by adjusting the width parameter of the network. Furthermore, a proper auxiliary network takes an identical standard frontal face and its 11 landmarks as references of each face and its corresponding landmarks to estimate the face rotation matrix and then computes three Euler angles, including yaw, pitch, and roll angles, to make the landmark localization stable and robust. Table I provides the configuration of the backbone network.

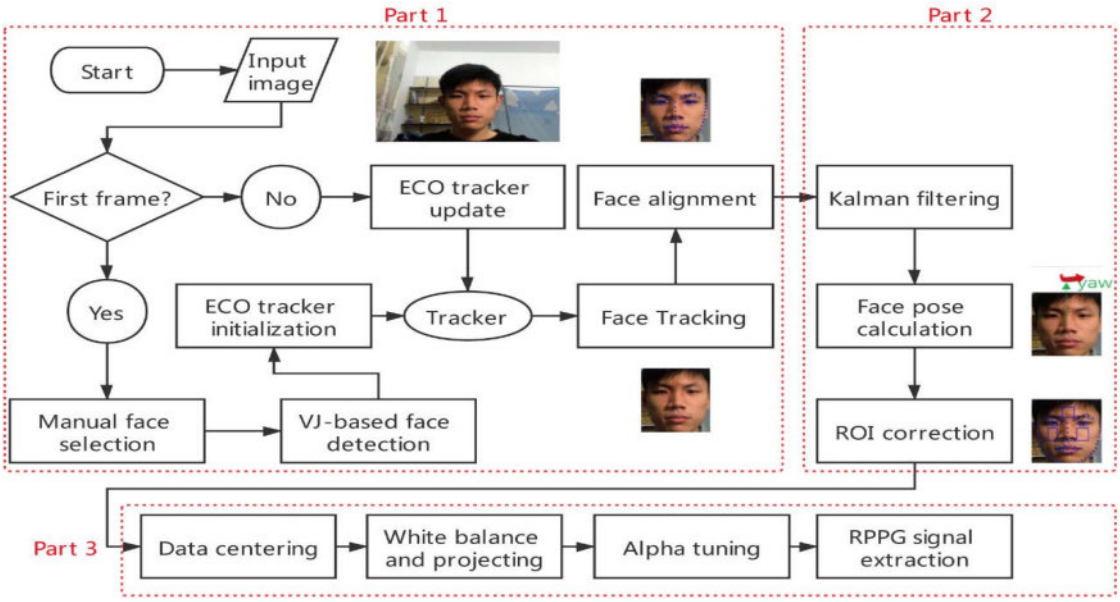


Fig. 3. Data processing flowchart for the proposed motion-robust rPPG with a DFT and face pose constrained KF.

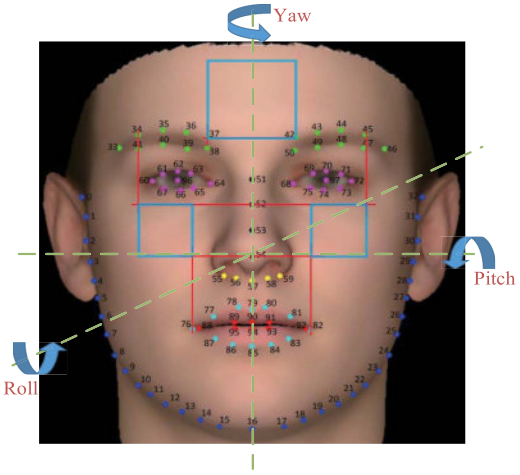


Fig. 4. Facial ROI identification by face landmarks and three degrees of freedom (roll, pitch, and yaw) of the face in 3-D space.

With the abovementioned face landmark detection and alignment, we can continuously and consistently obtain 98 landmark points on each target face image from the video sequence to determine the facial ROIs. Here, two pieces of cheek skin under the eyes and one piece of forehead skin are chosen as the target facial ROIs because they have proved to contain good rPPG SNR signal quality [37]. As shown in the blue box in Fig. 4, we use numbers 34, 76, 82, and 45 landmarks to determine the horizon-axis endpoint coordinate of the cheek skin ROI rectangle. We use numbers 52 and 54 landmarks to determine the vertical-axis endpoint coordinates of the cheek skin ROI rectangle. Number 37 and 42 landmarks are the endpoints of the bottom edge for the forehead ROI square. The square's side length is set as the coordinate distance between the numbers 37 and 42 landmarks on the horizon axis.

Although the advanced face alignment method in DFT has become quite accurate at predicting the facial landmark locations, they do not simultaneously assure the robustness and accuracy of their predicted locations from sudden large head motion and various lighting conditions as well as large occlusion in real applications. Although we use facial landmarks to help locate the ROIs, there are still some measurement errors caused by landmark fluctuation. For a robust estimation method from a series of noisy measurements, KF [17], [38] is a highly efficient recursive filter that can estimate the state of a dynamic system from a series of incomplete and noisy measurements. A typical example of the KF is to predict the coordinates and velocity of an object's position from a limited set of noisy observations of its position. We thus apply a KF as a face motion estimation model to correct the facial landmarks (No. 34, 37, 42, 45, 52, 54, 76, and 82) for facial ROI correction in video sequences.

Specifically, KF is used to predict the state at time t by the state at time $t-1$. For real-time landmark tracking, the state of specific landmark positions $\mathbf{x}_k$ can be assumed to have continuous first- and second-order derivatives, denoted by $\dot{\mathbf{x}}_k$ and $\ddot{\mathbf{x}}_k$, respectively, so that the state of position and velocity vectors varies between frames in terms of changes in velocity and acceleration as follows:

$$\begin{cases} \mathbf{x}_k = \mathbf{x}_{k-1} + h\dot{\mathbf{x}}_{k-1} + \dfrac{1}{2}h^2\ddot{\mathbf{x}}_{k-1} \\ \dot{\mathbf{x}}_k = \dot{\mathbf{x}}_{k-1} + h\ddot{\mathbf{x}}_{k-1}. \end{cases} \tag{1}$$

In implementing the motion-robust rPPG sensor, the positions of eight landmarks related to ROI tracking are subject to the following KF. The task of KF is to estimate the values of state vectors $\mathbf{S}_k = [\mathbf{x}_k, \dot{\mathbf{x}}_k]$ given observation of vector $\mathbf{M}_k$. We assume that the random tracking process to be estimated

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 $\mathbf{F}-\mathbf{1}$ (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

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

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

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

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:

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

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:

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

where

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

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:

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

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

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

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

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

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:

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

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 $\mathbf{e}_k$:

$$\begin{aligned}\mathbf{P}_{k+1}^- &= E\left[\mathbf{e}_{k+1}^-\mathbf{e}_{k+1}^{-T}\right] = E[(\mathbf{A}\mathbf{e}_k + \mathbf{w}_k)(\mathbf{A}\mathbf{e}_k + \mathbf{w}_k)^T] \\ &= \mathbf{A}\mathbf{P}_k\mathbf{A}^T + \mathbf{Q}_k. \end{aligned} \tag{14}$$

Having these required quantities at time t_{k+1}, we now can assimilate the measurement $\mathbf{m}_{k+1}$ for its optimal use in updating the estimate $\hat{\mathbf{s}}_{k+1}$. This recursive Kalman filtering with its pertinent equations and the sequence of computational

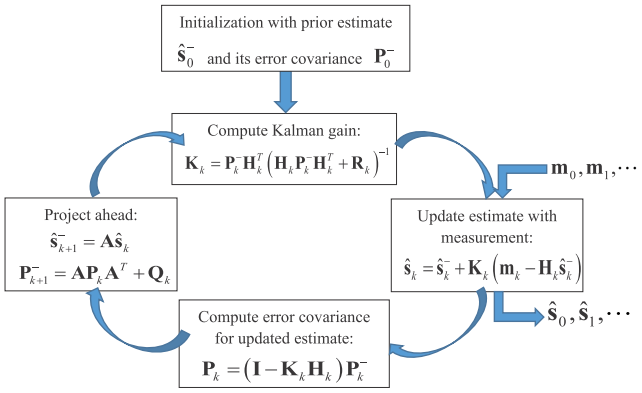


Fig. 5. Facial landmark KF framework.

steps are shown as follows. As shown in Fig. 5. As shown in Fig. 5. The positions of the current landmarks are predicted by the MS of corrected landmarks in the last iteration. The Kalman gain is updated through the covariance matrix in the last iteration. Then, the Kalman gain updates the covariance matrix and calculates the MS of corrected landmarks based on input measured landmarks and predicted landmarks.

However, the KF models motion on the image coordinate system. A large head rotation in the 3-D world coordinate system will affect the prediction performance of the KF. In addition, head rotation gives rise to incorrect facial ROI tracking, and a face pose constrained KF is proposed as follows. As shown in Fig. 4, three XYZ fixed angles, including roll ϕ, pitch ψ, and yaw θ angles [39], can describe face pose intuitively in which they use three separate angles to decompose a rotation into three different rotations around the fixed XYZ reference frame. In general, the initial pose of the world coordinate system is defined as the face oriented to the camera ($[\phi,\ \psi,\ \theta] = [0,\ 0,\ 0]$). We observe that the Kalman-filtered partial ROI is occluded when the yaw angle of the face pose is large. If this yaw angle is larger (or smaller) than 45° (or −45°), the right (or left) ROI on the cheek is excluded for HR calculation.

Specifically, the camera is fixed during video recording in police interrogation, and the transformation matrix mapping world system coordinate to the camera system coordinate is constant. We simply assume that the world coordinate system completely coincides with the camera coordinate system, such that the transformation matrix from the world coordinate system to the camera coordinate system is $\mathbf{I}$. Therefore, the change in rotation angle is directly related to the change in facial landmarks in the world coordinate system. The landmarks are assumed to be viewed from a sufficiently large distance, such that the video camera system satisfies weak perspective projection conditions [40], [41]. The correspondence between the 2-D face landmark $\mathbf{l}_i' = [x_i', y_i']^T$ under the image coordinate system and the 3-D face landmark $\mathbf{L}_i = [X_i; Y_i; Z_i]^T$ under the world coordinate system is

$$\mathbf{l}_i' = \mathbf{A}[\mathbf{R} \mid \mathbf{T}]\mathbf{I}\begin{bmatrix}\mathbf{L_i}\\ 1\end{bmatrix} = \begin{bmatrix}\alpha & 0 & 0\\ 0 & \alpha & 0\end{bmatrix}\begin{bmatrix}r_1 & r_2 & r_3 & t_1\\ r_4 & r_5 & r_6 & t_2\\ r_7 & r_8 & r_9 & t_3\end{bmatrix}\begin{bmatrix}X_i\\ Y_i\\ Z_i\\ 1\end{bmatrix} \tag{15}$$

where $\mathbf{R}$ is the rotation matrix represented by roll–pitch–yaw angles [39], $\mathbf{T} = [t_1, t_2, t_3]^T$ is the translation vector, $\alpha = f/t_3$, and f is the focal length of the camera.

We assume that all facial landmarks have identical translation vectors. Let $\bar{X} = (\sum_{i=1}^{N} X_i)/N$, and we define $X_i^{(n)} = X - \bar{X}$, where superscript (n) represents the first letter of 'normalization'. Similarly, we define $Y_i^{(n)}$, $Z_i^{(n)}$, $x_i'^{(n)}$ and $y_i'^{(n)}$. We therefore normalize all the landmark coordinates, and the vector $\mathbf{T}$ can be eliminated in (14)

$$\begin{bmatrix}x_i'^{(n)}\\ y_i'^{(n)}\end{bmatrix} = \begin{bmatrix}\alpha & 0 & 0\\ 0 & \alpha & 0\end{bmatrix}\begin{bmatrix}r_1 & r_2 & r_3 & 0\\ r_4 & r_5 & r_6 & 0\\ r_7 & r_8 & r_9 & 0\end{bmatrix}\begin{bmatrix}X_i^{(n)}\\ Y_i^{(n)}\\ Z_i^{(n)}\\ 1\end{bmatrix}. \tag{16}$$

We augment (15) to all landmark points, which means

$$\begin{bmatrix}\mathbf{x}'^{(n)}\\ \mathbf{y}'^{(n)}\end{bmatrix} = \begin{bmatrix}x_1'^{(n)} & x_2'^{(n)} & \cdots & x_N'^{(n)}\\ y_1'^{(n)} & y_2'^{(n)} & \cdots & x_N'^{(n)}\end{bmatrix}$$

and

$$\mathbf{C} = \begin{bmatrix}\mathbf{X}^{(n)}\\ \mathbf{Y}^{(n)}\\ \mathbf{Z}^{(n)}\end{bmatrix} = \begin{bmatrix}X_1^{(n)} & X_2^{(n)} & \cdots & X_N^{(n)}\\ Y_1^{(n)} & Y_2^{(n)} & \cdots & Y_N^{(n)}\\ Z_1^{(n)} & Z_2^{(n)} & \cdots & Z_N^{(n)}\end{bmatrix}.$$

Later, we define

$$\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3] = [\alpha, 0, 0]\mathbf{R} = [\alpha r_1, \alpha r_2, \alpha r_3]$$
$$\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3] = [0, \alpha, 0]\mathbf{R} = [\alpha r_4, \alpha r_5, \alpha r_6].$$

Thus, we can write

$$\begin{aligned}\boldsymbol{\lambda}\mathbf{C} &= \mathbf{x}'^{(n)}\\ \boldsymbol{\lambda}^T &= (\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{x}'^{(n)T}.\end{aligned} \tag{17}$$

Note that the matrix $\mathbf{C}\mathbf{C}^T$ is nonsingular when all of the points L_i are not coplanar, and similarly

$$\boldsymbol{\gamma}^T = (\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{y}'^{(n)T}. \tag{18}$$

The first two rows of $\mathbf{R}$ are obtained as $[(\lambda_1^2/\alpha), (\lambda_2^2/\alpha), (\lambda_3^2/\alpha)]$ and $[(\gamma_1^2/\alpha), (\gamma_2^2/\alpha), (\gamma_3^2/\alpha)]$. The third row is then obtained as the cross product of these two rows. The yaw angle θ is thus computed by its relationship to the corresponding rotation matrix element $\mathbf{R}$ that is represented by roll–pitch–yaw angles [39].

3) rPPG Extraction Based on POS: Taking the face image at time t, the RGB values in the ROI are spatially averaged to generate temporal signals $\mathbf{x}(t)$. According to the POS model built by Wang *et al.* [11], it can be expressed as follows:

$$\mathbf{x}(t) \approx \mathbf{u}_c \cdot I_0 \cdot c_0 \cdot (1 + i(t)) + \mathbf{u}_s \cdot I_0 \cdot s(t) + \mathbf{u}_p \cdot I_0 \cdot p(t) \tag{19}$$

where I_0 represents the stationary parts of luminance intensity. $\mathbf{u}_c$, $\mathbf{u}_s$, and $\mathbf{u}_p$ are 3×1 vectors. $\mathbf{u}_c$ denotes the unit color vector of the skin reflection. $\mathbf{u}_s$ denotes the unit color vector of the light spectrum. $\mathbf{u}_p$ denotes the relative pulsatile strengths in RGB channels. $i(t)$, $s(t)$, and $p(t)$ are zero-mean signals. $i(t)$ denotes the varying parts of luminance intensity, which is related to light source illumination variation. $s(t)$ denotes the varying parts of specular reflections. It is related to body

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 eliminate the effect of camera quantization noise.

Subsequently, a 2×3 projection matrix $\mathbf{P}_{\mathrm{p}} = \left(\begin{smallmatrix} 0 & 1 & -1 \\ -2 & 1 & 1 \end{smallmatrix}\right)$ is identified to project $\bar{\mathbf{x}}(t)$ on two axes of the color space plane, which is orthogonal to the vector $\mathbf{1} = (1\ 1\ 1)^{\mathrm{T}}$. Through this projection, the largest motion-induced distortion of light intensity variation along the direction $\mathbf{1}$ is eliminated from all three camera channels

$$\begin{aligned} \tilde{\mathbf{x}}(t) &= \mathbf{P}_p \cdot \mathbf{N} \cdot \bar{\mathbf{x}}(t) \\ &= \mathbf{P}_p \cdot \mathbf{N} \cdot \mathbf{u}_s \cdot I_0 \cdot s(t) + \mathbf{P}_p \cdot \mathbf{N} \cdot \mathbf{u}_p \cdot I_0 \cdot p(t) \\ &= \begin{bmatrix} S_1(t) \\ S_2(t) \end{bmatrix} \end{aligned} \quad (21)$$

where $S_1(t)$ and $S_2(t)$ denote the two projected components. According to Wang *et al.*'s experiment [11], motion disturbances in two decomposed components are antiphase. Alpha tuning can fuse two components to suppress motion disturbances. It can be expressed as

$$p(t) = S_1(t) + \alpha \cdot S_2(t) \qquad \text{with } \alpha = \frac{\sigma(S_1)}{\sigma(S_2)} \quad (22)$$

where $\sigma(\cdot)$ denotes the standard deviation operator. When motion disturbances dominate term $\tilde{\mathbf{x}}$, the standard deviation of each projected signal represents the motion disturbance amplitude on each axis. Coefficient α can push the motion disturbance strength of two projected signals into the same level, i.e., $\sigma(S_1) = \sigma(\alpha \cdot S2)$; adding two antiphase signals with the same amplitude will suppress the motion disturbances.

B. Fusion Between BCG and rPPG

Considering that the BCG signal is highly sensitive to body movements while the rPPG signal is extremely susceptible to face and head movements, we therefore design an adaptive weighting scheme to fuse these two signals from various motion disturbances for accurate and real-time HR estimation. Specifically, we use a bank of notch filters to detect the dominant frequency of these two input signals with the notch frequencies being evenly distributed in a certain frequency band. We can real-timely calculate the ratio of output to input signals for each filter at every second, which is then used to calculate adpative weights for the two signals at different frequencies. Furthermore, we cached the input signals and estimated HR 1 s, which will be multiplied with a forgetting factor the next. This iterative fusion strategy recursively uses the last-estimated HR and the current signals to estimate the current HR.

According to Sadek and Biswas's work [4], a bandpass filter with a frequency band of 0.45–4 Hz is implemented to filter out the noise beyond the target ranges of the HR frequency in the BCG signal. Then, the MODWT is adopted to process the filtered signal. The BCG signal decomposition can be seen in Fig. 6. Sadek and Biswas's work [4] claimed that the fifth harmonic of MODWT-based signal decomposition has the highest correlation to the cardiac cycle. This is exactly consistent with our experimental results. In Fig. 6(a), although the BCG signal is disturbed by noise, there is still a typical abruptly rising waveform before the J-peak in the BCG signal. Therefore, we can roughly judge the subject's cardiac cycle. Compared with each component of BCG decomposition, the fifth harmonic [see Fig. 6(e)] has the highest correlation to the cardiac cycle.

Thus, there are two filtered rPPG and BCG signals related to the pulse signal. We utilize a bank of length-3 FIR notch filters to process the filtered signals. The FIR notch filters' transfer function H is denoted as follows:

$$H(z) = 1 - 2z^{-1}\cos(2\pi f_i) + z^{-2} \quad (23)$$

where the discrete frequency is $f_i \in [f_1, \ldots, f_F]$ and F is denoted as the number of discrete frequencies. An example of a length-3 FIR filter is shown in Fig. 7.

We define $\mathbf{u}[n,\ j]$ as the input signal with $n = 3, \cdots$ and $j = 1, \ldots, S$, in which S is the number of input signals. In addition, $m = 1, 2, \ldots, T$ is the index of time on the second scale. At sample n and input signal j, the output of the filter is $\mathbf{y}_i[n, j]$

$$\mathbf{y}_i[n, j] = \mathbf{u}[n, j] - 2\mathbf{u}[n-1, j]\cos(2\pi f_i) + \mathbf{u}[n-2, j]. \quad (24)$$

The input $\mathbf{u}$ is filtered with all the filters of the bank. For each filter in each second, the output-to-input power is computed as

$$\mathbf{P}_i[m, j] = \frac{\mathbf{Y}_i[m, j]}{\mathbf{U}[m, j]} \quad (25)$$

with

$$\mathbf{Y}_i[m, j] = \delta \mathbf{Y}_i[m-1, j] + (1-\delta) \sum_{k=1}^{freq} \mathbf{y}_i^2[(m-1) * freq + k, j] \quad (26)$$

$$\mathbf{U}[m, j] = \delta \mathbf{u}[m-1, j] + (1-\delta) \sum_{k=1}^{freq} \mathbf{u}^2[(m-1) * freq + k, j] \quad (27)$$

where a forgetting factor range is $0 \leq \delta < 1$ and freq is the sampling frequency. To initialize, we set $\mathbf{Y}_i[2, j] = \mathbf{U}[2, j] = 0.5(\mathbf{u}^2[1, j] + \mathbf{u}^2[2, j] + \cdots + \mathbf{u}^2[2 * \mathrm{freq}, j])$.

Then, $\mathbf{P}_i$ is used to compute a set of weights $\mathbf{W}_i[m]$ such that the weighted sum of the notch frequencies can estimate the input dominant frequency. According to the characteristics

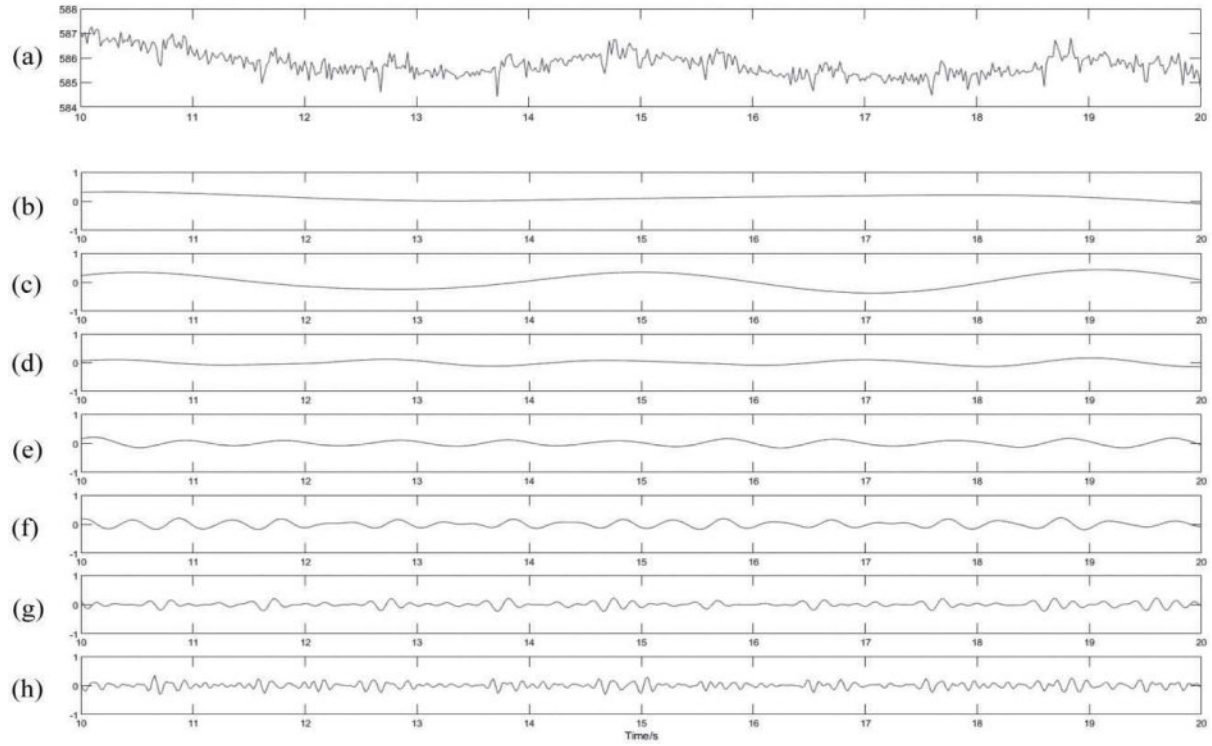


Fig. 6. BCG signal decomposition: (a) the original BCG signal; (b) the 8th level harmonic; (c) the 7th level harmonic; (d) the 6th level harmonic; (e) the 5th level harmonic (**selected component**); (f) the 4th level harmonic; (g) the 3rd level harmonic; (h) the 2nd level harmonic.

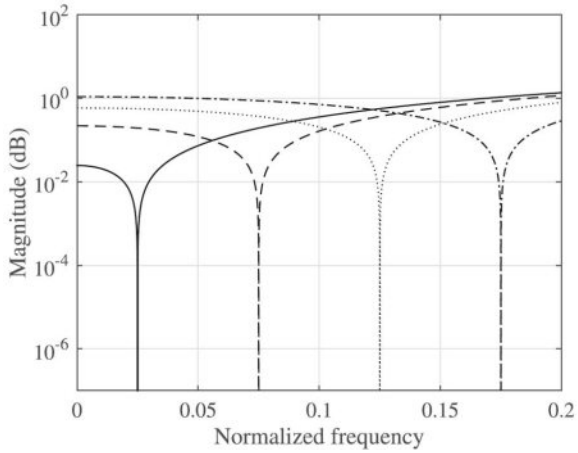


Fig. 7. Frequency response curve of length-3 FIR notch filters with different stopband frequencies f_i. The closer the dominant frequency of the input signal is to f_i, the greater the attenuation of the filtered signal.

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

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

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 $(m-1)$

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

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

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

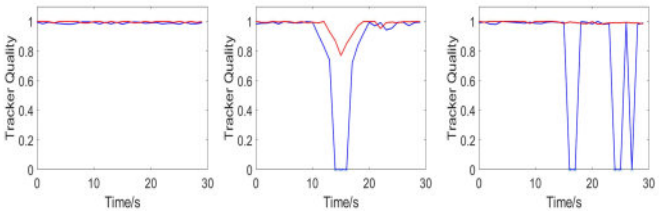
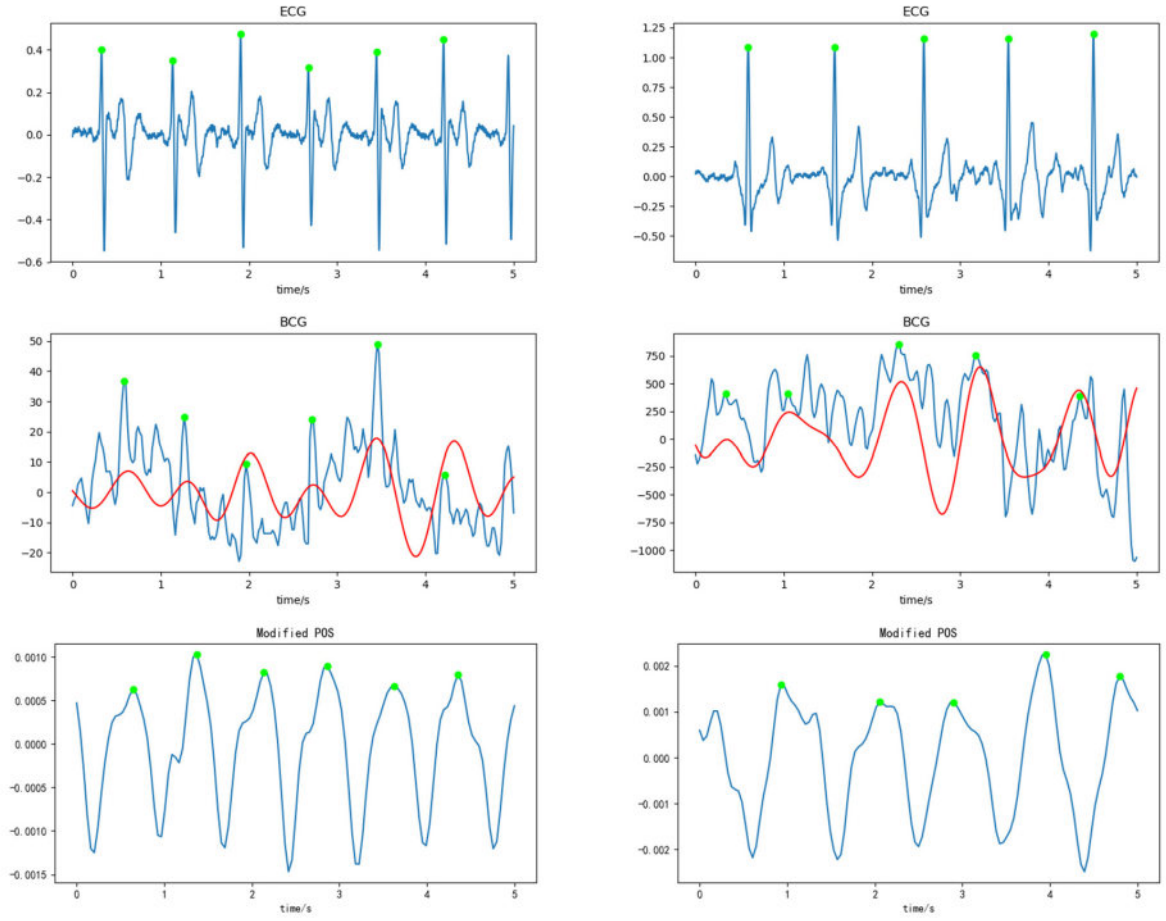


Fig. 8. Tracking quality in three cases of an SS, MS, and nontarget face entrance for the proposed DFT-KF and traditional KLT + ERT methods. The red lines for the DFT-KF method show better tracking quality than the blue lines for the KLT + ERT method.

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. As shown in Table I, KLT-ERT represents the traditional face alignment method. DFT and DFT-KF represent our proposed algorithm and the further improved algorithm, respectively. The proposed method with DFT and face pose constrained KF shows better superiority than the traditional facial ROI tracker, especially in the MS. In addition, the PFLD we adopted in DFT is an optimized model that is trained directly on CPU- and GPU-based computational frameworks without model inference and acceleration. The frame rates of the GPU-optimized DFT tracking methods are greater than those of the other methods in our experimental results.

The correlation between the ECG, BCG, and rPPG signals at different scales of motion disturbances is verified by experiments. The J-peak of BCG and the P-peak of PPG are supposed to appear later than the R-peak of ECG since the electrical activity precedes the mechanical activity. As shown in Fig. 9(a), in the SS, the J-peak in the original BCG signal has an evident correspondence with the peak of the fifth-harmonic component of MODWT. All the J-peaks of BCG appear slightly behind the R-peaks of reference-ECG. At the same time, the P-peak in the rPPG signal also has a similar exact correlation to the R-peak in the ECG signal. For the MS in Fig. 9(b), the motion disturbance overwhelms the characteristic waveform of BCG signal, the peak of the fifth-harmonic component of MODWT no longer has a strong correspondence with the J-peak. There is not an exact agreement between the J-peak of BCG and the R-peak of ECG signals. However, the P-peak of the rPPG signal can still approximately correspond to the R-peak in the ECG signal. Therefore, in the MS, it is suitable to estimate HR based on the rPPG signal estimated by our multimodal HR sensor.



(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
DFT-KT (GPU)	**19.5**		

To further compare the signal quality of the reference ECG-, MODWT-BCG-, and DFT-KF-based rPPG signals, we demonstrate the short-time Fourier transform (STFT) spectra of these signals in Fig. 10. The power spectrum of a clean PPG from a healthy subject should normally have peaks representing the HR with various harmonics of HR frequency. Since we applied a 0.5–4-Hz bandpass filter to the original signal, the spectrum should only retain one main HR component and three other harmonic components [43].

Compared with the spectra of the standard ECG signal and BCG signal, the STFT spectra of the rPPG signal contain obvious bands of harmonic components. In the first row of Fig. 10 for the SS, BCG's HR component has higher kurtosis and SNR, appearing as a yellow stripe in the spectrogram with a clear contrast to the blue background. Among the rPPG-based algorithms, DFT-KF-POS shows a cleaner and more focused STFT spectrum with a high SNR, which manifests as a narrow and bright pulsatile stripe. The DFT-KF-ICA STFT spectrum shows one wider pulsatile stripe and one extra shorter stripe, which means that its HR is clearer than the DFT-KF-POS PPG signal. DFT-KF-PBV shows noisier and more diffusive spectra and performs worse than other methods. Its pulsatile stripe brightness is not high compared with the background, which indicates that the SNR of the signal is not good. In some time periods, the stripe is overwhelmed by noise to indicate its poor HR estimation performance.

In the second row of Fig. 10 for the motion disturbance state, the BCG signal and DFT-KF-ICA PPG signal are severely distorted, and the corresponding pulsatile stripe of the spectrum cannot be found. The SNR of the signals

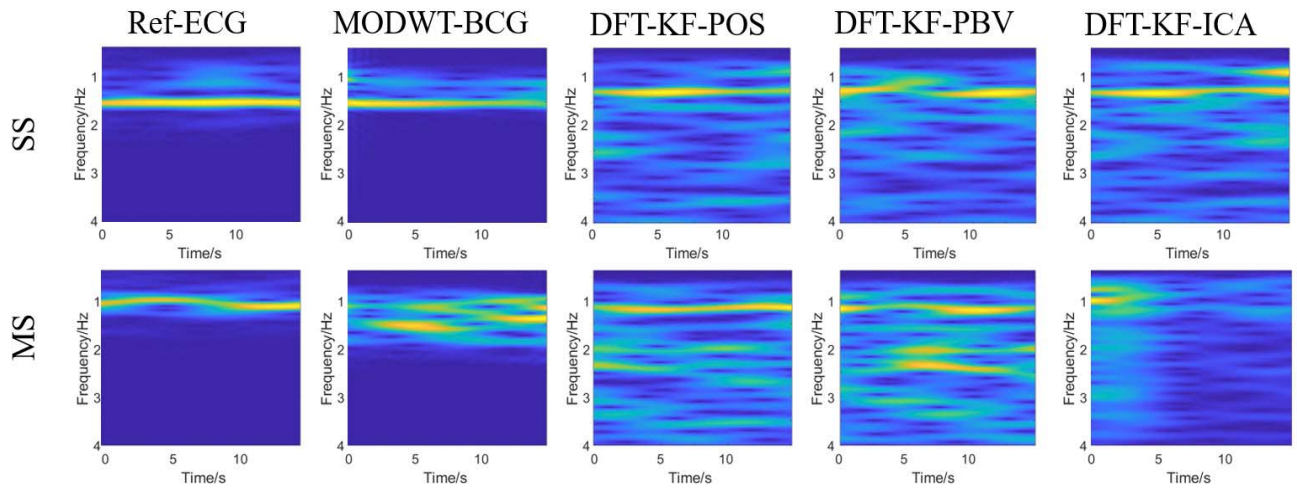


Fig. 10. Short-time Fourier transform spectra obtained by ECG-reference, MODWT-BCG, DFT-KF-POS, and two other rPPG algorithms based on DFT. The ground-truth ECG signal and the MODWT-BCG signal in the SS highlight the exact HR component (as one yellow stripe) having higher kurtosis and SNR with a clear 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
MODWT-BCG [4]	**2.33**	13.33
DFT-KF-ICA [42]	3.9	10.32
DFT-KF-PBV [29]	5.63	9.48
DFT-KF-POS [11]	5.47	9.32
DFT-KF-ICA & BCG	3.23	6.34
DFT-KF-POS & BCG	4.71	**6.20**

drops significantly. However, the pulsatile stripe of the spectrum in the other two rPPG signals is still visible. Among them, the pulsatile stripe of the proposed DFT-KF-POS is clearest from the extracted signal, which demonstrates the best HR estimation performance in motion disturbances.

Later, we conducted comparative experiments to evaluate the HR estimation performance of most state-of-the-art algorithms in the SS and MS. As shown in Table II, the whole MAE value is high because we completely and fairly analyze all estimations in whole time periods without deleting any gross errors in the HR estimation per second. Some inconsistent estimation results are significantly different from past calculations, which therefore raise the MAE values of the final statistical result. The accuracy of the POS-based methods is relatively excellent compared with other rPPG methods. In our experiments, the proposed rPPG with BCG fusion run in the SS performs slightly worse than the optical-fiber-based MODWT-BCG algorithm. This is because BCG has proved to produce accurate HR estimation when there is less or no motion disturbance. This may also be due to the possibility of introducing noise from the rPPG signal in BCG fusion.

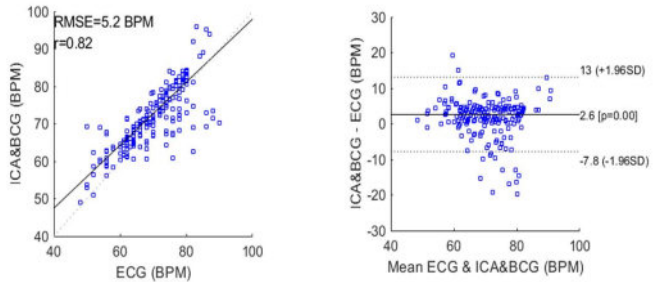


Fig. 11. Correlation and Bland–Altman plots of the DFT-KF-ICA and BCG HR estimation against ECG reference of eight users.

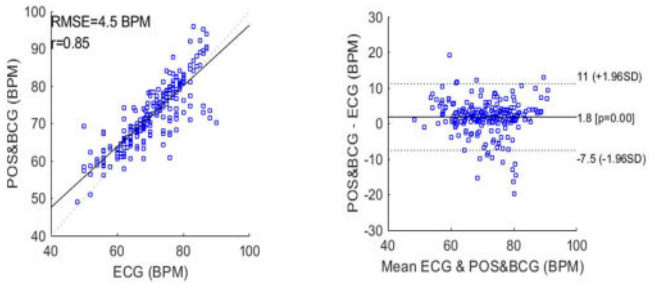


Fig. 12. Correlation and Bland–Altman plots of the DFT-KF-POS and BCG HR estimation against the ECG reference of eight users.

However, in the MS, the proposed motion-robust rPPG algorithm is superior to all the other algorithms, including the BCG and original POS algorithms, in terms of the MAE value (see Table II). The large face and body movements in the MSs introduce largely varying force signals that far outweigh the pulsatile signal for the BCG sensor and result in difficulties in face tracking and HR estimation for classical rPPG algorithms. However, by achieving the high performance of the KF-regularized DFT and motion-artifact correction via multimodal fusion, the proposed rPPG algorithm is robust to motion disturbance and provides consistent HR estimation in all HR measurements.

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 rate estimation from face via spatial-temporal representation," *IEEE Trans. Image Process.*, vol. 29, pp. 2409–2423, 2020.

[9] F. Bousefsaf, A. Pruski, and C. Maaoui, "3D convolutional neural networks for remote pulse rate measurement and mapping from facial video," *Appl. Sci.*, vol. 9, no. 20, p. 4364, Oct. 2019.

[10] S. Chaichulee *et al.*, "Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning," *Physiological Meas.*, vol. 40, no. 11, Dec. 2019, Art. no. 115001.

[11] 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.

[12] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, and K.-W. Cheung, "Block-based adaptive ROI for remote photoplethysmography," *Multimedia Tools Appl.*, vol. 77, no. 6, pp. 6503–6529, Mar. 2018.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[14] B. D. Lucas *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 2. Vancouver, BC, Canada: International Society for Optics and Photonics, 1981, pp. 674–679.

[15] L. Feng, L.-M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 879–891, May 2015.

[16] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomed. Opt. Exp.*, vol. 6, no. 5, pp. 1565–1588, 2015.

[17] S. K. A. Prakash and C. S. Tucker, "Bounded Kalman filter method for motion-robust, non-contact heart rate estimation," *Biomed. Opt. Exp.*, vol. 9, no. 2, pp. 873–897, 2018.

[18] Z. Tu *et al.*, "A survey of variational and CNN-based optical flow techniques," *Signal Process., Image Commun.*, vol. 72, pp. 9–24, Mar. 2019.

[19] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.

[20] M. Fiaz, A. Mahmood, S. Javed, and S. K. Jung, "Handcrafted and deep trackers: Recent visual object tracking approaches and trends," *ACM Comput. Surv.*, vol. 52, no. 2, pp. 1–44, May 2019.

[21] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6638–6646.

[22] F. Gasparini and R. Schettini, "Skin segmentation using multiple thresholding," *Proc. SPIE*, vol. 6061, Jan. 2006, Art. no. 60610F.

[23] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognit. Lett.*, vol. 124, pp. 82–90, Jun. 2019.

[24] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[25] T. Pursche, R. Claus, B. Tibken, and R. Moller, "Using neural networks to enhance the quality of ROIs for video based remote heart rate measurement from human faces," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2019, pp. 1–5.

[26] X. Guo *et al.*, "PFLD: A practical facial landmark detector," 2019, *arXiv:1902.10859*. [Online]. Available: http://arxiv.org/abs/1902.10859

[27] 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.

[28] 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.
[29] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood vol. pulse, signature," *Physiological Meas.*, vol. 35, no. 9, p. 1913, 2014.
[30] C. Zhao, W. Chen, C.-L. Lin, and X. Wu, "Physiological signal preserving video compression for remote photoplethysmography," *IEEE Sensors J.*, vol. 19, no. 12, pp. 4537–4548, Jun. 2019.
[31] T. Tamura, "Current progress of photoplethysmography and SPO2 for health monitoring," *Biomed. Eng. Lett.*, vol. 9, no. 1, pp. 21–36, Feb. 2019.
[32] F. T. Z. Khanam, A. Al-Naji, and J. Chahl, "Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review," *Appl. Sci.*, vol. 9, no. 20, p. 4474, Oct. 2019.
[33] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.
[34] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. 23rd IEEE Int. Symp. Robot Hum. Interact. Commun.*, Aug. 2014, pp. 1056–1062.
[35] G.-S. Hsu, A. Ambikapathi, and M.-S. Chen, "Deep learning with time-frequency representation for pulse estimation from facial videos," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 383–389.
[36] M. Braverman, O. Etesami, and E. Mossel, "Mafia: A theoretical study of players and coalitions in a partial information environment," *Ann. Appl. Probab.*, vol. 18, no. 3, pp. 825–846, Jun. 2008.
[37] 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.
[38] R. G. Brown and P. Y. Hwang, *Introduction to Random Signals and Applied Kalman Filtering With MATLAB Exercises*, vol. 4. New York, NY, USA: Wiley, 2012.
[39] P. Corke, *Robotics, Vision and Control: Fundamental Algorithms In MATLAB Second, Completely Revised*, vol. 118. Cham, Switzerland: Springer, 2017,
[40] A. M. Bruckstein, R. J. Holt, A. N. Netravali, and T. S. Huang, "Optimum fiducials under weak perspective projection," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 223–244, 1999.
[41] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.
[42] Y. Li, D. Powers, and J. Peach, "Comparison of blind source separation algorithms," in *Proc. Adv. Neural Netw. Appl.*, 2000, pp. 18–21.
[43] C. Orphanidou, *Signal Quality Assessment in Physiological Monitoring: State of the Art and Practical Considerations*. Cham, Switzerland: Springer, 2017.
[44] J. Martin Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 327, no. 8476, pp. 307–310, Feb. 1986.
[45] D. Wang, X. Yang, X. Liu, J. Jing, and S. Fang, "Detail-preserving pulse wave extraction from facial videos using consume-level camera," *Biomed. Opt. Exp.*, vol. 11, no. 4, pp. 1876–1891, 2020.
[46] W. Zhao, Y. Lv, Q. Liu, and B. Qin, "Detail-preserving image denoising via adaptive clustering and progressive PCA thresholding," *IEEE Access*, vol. 6, pp. 6303–6315, 2018.
[47] H. Yue *et al.*, "Non-contact heart rate detection by combining empirical mode decomposition and permutation entropy under non-cooperative face shake," *Neurocomputing*, vol. 392, pp. 142–152, Jun. 2020.

Yiming Liu received the B.Eng. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2018, and the M.Sc. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2021.

His current research interests include medical image processing, machine learning, and physiological signal processing.

Binjie Qin (Member, IEEE) received the M.Sc. degree from the Nanjing University of Science and Technology, Nanjing, China, in 1999, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2002.

He was a Lecturer and an Associate Professor with the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. From 2012 to 2013, he was a Visiting Professor with the Department of Computer Science, University College London, London, U.K. He is currently an Associate Professor with the School of Biomedical Engineering, Shanghai Jiao Tong University. His current research interests include biomedical imaging, image processing, machine learning, computer vision, and biomedical instrumentation.

Rong Li currently works at the Shanghai Public Security Bureau, Shanghai, China. His current research interests include computer-aided interrogation, emotional arousal detection, and physiological and psychological monitoring.

Xintong Li received the B.Eng. degree in computer science and technology from Nanchang University, Nanchang, China, in 1996.

He currently works at ECData Information Technology Company Ltd., Shanghai, China. His current research interests include signal processing and algorithm development.

Anqi Huang received the B.Eng. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019.

Her current research interests mainly include super-resolution ultrasound imaging, acoustic characteristics of high-boiling-point phase-change nanodroplets, and the potential of nanoscale acoustic-sensitive particles in drug delivery.

Haifeng Liu received the B.Sc. degree in automation control from the Beijing University of Technology, Beijing, China, in 1997.

He currently works at ECData Information Technology Company Ltd., Shanghai, China. His current research interests include product innovation and information system design, development, and integration.

Yisong Lv received the M.Sc. degree in automatic instruments from the Kunming University of Science and Technology, Kunming, China, in 1999, and the Ph.D. degree in biomedical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2003.

His current research interests include image analysis, machine learning, and computer vision.

Min Liu currently works at the Shanghai Public Security Bureau, Shanghai, China. His current research interests include computer-aided interrogation, emotional arousal detection, and physiological and psychological monitoring.