Graph-Based Denoising for Respiration and Heart Rate Estimation During Sleep in Thermal Video

Cheng Yang, Member, IEEE, Menghan Hu, Member, IEEE, Guangtao Zhai, Senior Member, IEEE, and Xiao-Ping Zhang, Fellow, IEEE

Abstract—Quality sleep is a basic human need for well-being, yet sleep deprivation has been a long-term global problem. A common type of sleep deprivation is obstructive sleep apnea, where people repeatedly stop breathing during sleep with subsequent abnormal vital signs, namely, respiration rate and heart rate. While tremendous effort has been made for vital signs monitoring systems during sleep, existing works still lack portability for bulky and intrusive systems and reliability for consumer-level, nonintrusive systems. To bridge the gap between practicability and accuracy and facilitate Internet of Things for smart healthcare, in this article, we propose a vital signs estimation system during sleep via a thermal camera. The system first captures thermal image sequences of a sleeping subject and then processes the facial regions within the thermal images for vital signs signal extraction. Specifically, leveraging on the inherent graph structure among subregions of the facial area, we propose a graph-based, spatial–temporal signal denoising scheme. Experimental results show that the graph-based denoising scheme in our system effectively reduces the noise level introduced by cameras and subjects, and our proposed system outperforms state-of-the-art nonintrusive vital signs monitoring systems. Since the algorithm components in our system have relatively low time complexity and no model training is required, our system can be deployed efficiently at the edge devices in a smart home setting. The extracted vital signs can then be used for sleep abnormality detection and disease screening.

Index Terms—Graph signal processing (GSP), sleep monitoring, vital signs estimation.

I. INTRODUCTION

QUALITY sleep is required for every individual to have sufficient physiological and mental recovery and well-being [1], [2]. As the ratio of elderly population to the total population becomes over 11% since 2013 [3] and the health awareness [4], [5] keeps improving, quality sleep becomes more significant. However, sleep deprivation has been a long-term health problem among the world population. Sleep deprivation is generally caused by sleep disorder [6]. One of its common syndromes is obstructive sleep apnea (OSA) [7], [8], where people repeatedly stop breathing during sleep with subsequent abnormal vital signs or sleep quality indicators, namely, respiration rate and heart rate [9], [10]. Specifically, OSA mainly consists of hypopnea (overly-shallow breathing within a unit period) and apnea (completely no breathing within a unit period), both of which directly affects the respiration rate and pattern. On the other hand, the heart rate and heartbeat pattern become abnormal due to these above repetitive blood oxygen supply reduction [6]. Therefore, tracking or monitoring the vital signs during sleep can evaluate a subject’s sleep quality [11].

The crux of sleep monitoring is vital signs tracking, either intrusively or nonintrusively. Accuracy-wise, the most reliable way to track the vital signs is to adopt an intrusive sleep monitoring system [12]. The systems are widely used for periodical evaluation for patients with sleep disorder in the sleep medicine community [12]. A conventional sleep monitoring system consists of multiple stripes with vital signs sensors attached to the patient’s skin, a facial mask for breathing volume tracking, and multiple EEG sensors attached to the patient’s head [12]. Although comprehensive and accurate signals of the patient are acquired, the patients generally have uncomfortable sleep experience.

As a tradeoff between bulky equipment with high accuracy and cost-effective systems with suboptimal accuracy, the research on portable intrusive sensors and completely nonintrusive systems has emerged for over a decade. These sensors and systems allow home or even on-the-go usage [13]. In particular, smartphones, watches, and wristbands can track the respiration rate, heart rate, blood pressure, or even sleep stages [14]. On the other hand, various systems adopt cameras to acquire video frames of people’s head or upper body region for estimation and tracking of vital signs. However, the accuracy of these devices is generally not clinically validated or far less accurate than gold standard [13].
In addition, most existing camera-based nonintrusive vital signs monitoring systems require visible light for cameras being able to capture acceptable video frames. This restricts these systems to be used only during day time or environments with artificial light, which possibly brings privacy issues.

We review related works on: 1) intrusive and nonintrusive vital signs tracking systems with nonvision-based sensors and 2) nonintrusive vital signs tracking systems using video capture devices in detail as follows.

The most recent medicine-level sleep monitoring systems can be found in [15] and [16]. In particular, a full nocturnal polysonmograpy tracking device has the function of the monitoring of the heart beat pattern, lung and brain activities, respiration pattern, arm and leg movements, and blood oxygen levels during sleep [12]. Apparently, such type of system makes users uncomfortable while breathing and making body movements during sleep, which may introduce unrealistic sleep monitoring results. For the systems that focus on part of the vital signs tracking, Chu et al. [17] used a strain sensor to be attached to the ribcage and abdomen to measure both the respiration rate and volume. The data are then transmitted via a wireless Bluetooth module. Therefore, there is no bulky set of wires surrounding the patient. However, the double-sided adhesive required for the strain sensor to be attached to the patient’s skin still rises the concern of the comfort. Similarly, Farhad et al. [18] used a spirometer attached to the ribcage of a subject to track the respiration rate and heart rate of a subject. However, it is less robust against noise than the strain-type sensors in terms of signal fidelity.

There is an increasing need for nonintrusive systems due to the fact that there is no need to wear or take off the devices. Some of the existing contact-less works are based on radio-frequency (RF) [19] and acoustic [20] sensing techniques. In particular, RF-based works focus on the detection of the thorax wall vibration via radar-modulation techniques [19], [21] of the wireless-wave signals for estimation of the respiration rate and heart rate. See [22] and [23] for review of the state-of-the-art RF identification (RFID) and WiFi sensor-based systems, respectively. Specifically, Batchu et al. [24] used a beam-steering Doppler sensor for both heart rate and respiration rate estimation. However, the Doppler sensor-based system is vulnerable to even small body movements during sleep. Similarly, Schellenberger et al. [25] used a set of radars for the classification of the subject’s movement during sleep. However, this work is limited to the movement classification only and does not perform any vital signs estimation. Liu et al. [26] used a single WiFi device for simultaneous respiration rate and heart rate estimation. However, as indicated in [23], [26] is sensitive to the movement, location, and orientation of the subject. The sensitivity issue is more severe for heart rate estimation via [26] due to the fact that heartbeat is overlapped with and covered by the chest and abdomen respiration movements. On the other hand, Xu et al. [27], Ha et al. [28], and Kroschel and Luik [29] leveraged higher frequency sensing techniques to extract the vital sign information from the thorax wall vibration. Specifically, Xu et al. [27] and Ha et al. [28] utilized the millimeter-wave (mmWave) emitted by an mmWave sensing probe for heart rate estimation. Kroschel and Luik [29] adopted the laser signal emitted by a vibrometer for estimation of the heart beat. However, similar to [26], the movement, location, and orientation of the experimental subject can easily affect the performance of [27]–[29], and the systems in [27]–[29] require customized high-end Doppler active radios, which in general are not readily available on the market and therefore incur prohibitive cost [22]. Acoustic sensing techniques have also been applied to vital signs monitoring [20]. However, the acoustic-based systems generally have limited measurement range and are susceptible to environment audio noise [20].

As opposed to the above nonvideo-based systems, the most widely researched nonintrusive vital signs tracking field is the video-based systems, including the ones utilizing the color, depth, infrared, thermal, and multimodal cameras. See [30] for representative noncontact video-based systems. We review the related works on video-based systems in detail next.

1) Color Camera: One of the most cited color-camera-based heart rate estimation system is proposed in [31]. The system takes average of the facial region spatially and adopts independent component analysis (ICA) to extract the heart beat signal. However, this work was only tested in upright sitting position. Two variants of the color-camera-based heart rate estimation system in [31] are proposed in [32] and [33], respectively. Specifically, Wu et al. [32] adopted a series of classical signal processing tools to extract the heart rate by enlarging the subtle, temporal change of the pixel intensities within the facial region of a subject. Likewise, Balakrishnan et al. [33] used feature point tracking and principal component analysis (PCA) to extract the heart beat signal within the forehead and mouth region of a subject. However, neither [32] nor [33] is robust against abrupt head movement. Wang et al. [34] first assumed multiple candidate soft knowledge spaces among the red, green, and blue color channels and then selected the best estimate. Mehta and Sharma [35] and Yin et al. [36] applied different signal decomposition methods, i.e., variational mode decomposition (VMD) and ensemble empirical mode decomposition (EEMD), respectively, for heart rate estimation. A deep learning-based heart rate and breath rate estimation system is proposed in [37]. The network architecture consists of a skin-reflection-based motion model and a face-appearance model for robust against various ambient lighting conditions and large head motion. Wang et al. [38] adopted a heuristic cross-point counting (CPC) approach [39] to estimate the respiration rate and a recurrent neural network for respiration pattern classification. However, none of the above color-camera-based methods is privacy preserved and they fail to estimate the heart rate or respiration rate when the light is off, i.e., complete darkness that is common during sleep.

2) Infrared Camera: As in the color-camera-based systems, the infrared and near-infrared-camera-based systems also extract the vital sign signals from the change of the temporal pixel intensities within a time period. In contrast to
the color cameras, these systems can be used in complete darkness. Zhang et al. [40] used near-infrared camera and a series of classical filtering processes for heart rate estimation and classification of the subject’s drowsiness levels. Similarly, Nowara et al. [41] used a sparse-based denoising method and other classical filtering methods for heart rate and heart beat rhythm estimation for drive state classification. Wang et al. [42] extended the techniques in [34] in near-infrared video frames and jointly denoises and recover the heart rate. However, similar to the color-camera-based approaches, the infrared and near-infrared camera-based approaches are not privacy preserved.

3) Depth Camera: Depth camera-based systems have emerged thanks to their ability of operation in complete darkness and privacy protection. The core process of depth-camera-based approaches is to extract the temporal change of the distance between the subject’s cloth or skin and the depth camera for respiration rate and heart rate estimation. Specifically, Wang et al. [43] used depth camera for classification of various respiration patterns. This system tracks the subject’s facial region and is robust against various facial cover, including surgical masks. Yang et al. [44] used depth camera for estimating heart beat patterns. This system consists of a block-wise graph-based joint temporal denoising/bit-depth enhancement module that helps to restore the vital signs signal and tracks the subject’s head movement and is robust against various headings, including front, side-way, and back. Yang et al. [45] used depth camera to track breathing patterns during sleep and detect abnormal breathing events. The system consists of a block-wise graph-based temporal denoising module that is similar to [44]. However, the graph-based temporal filtering modules in [44] and [45] suffer from the expensive computation cost due to the repetitive pixel-wise graph construction, and all the above depth-camera-based approaches suffer from the inherent, relatively low resolution and noisy observations acquired from the depth camera. Besides, the above approaches may still fail to work when major head motion is present.

4) Thermal Camera: Similar to depth camera-based systems, there also exist thermal-camera-based approaches that protect users’ privacy and robust in various lighting conditions. The privacy-preserving property is mainly due to the fact that the facial details decrease as the camera imaging wavelength increases [46]. Since thermal cameras generally have the longest imaging wavelengths compared to the color, infrared and depth cameras, the captured thermal images retain the least amount of facial details. These systems rely on a periodical change of the skin temperature instead of heart beat-induced subtle motions. In particular, Cosar et al. [47] used thermal videos and a series of signal filtering tools for heart rate and respiration rate estimation. Pereira et al. [48] adopted a three-combined estimators for respiration rate and breath-to-breath interval estimation. Kim et al. [49], Calopa [50], and Garbey et al. [51] used similar approaches and adopted infrared thermal camera to locate the carotid artery and track its periodic intensity change for heart rate estimation. However, similar to the depth camera-based systems, current consumer-level thermal cameras suffer from inherent noticeable acquisition noise and low resolution, making it difficult to extract vital signs accurately.

5) Multimodal Camera: For multimodal camera, Chen et al. [52] used the face detection result from the color video to estimate the respiration rate from the thermal video. However, this method requires repetitive dynamic time warping (DTW) to select the most reliable signal segments for vital signs extraction, which is impractical in near real-time vital signs tracking applications. Hu et al. [39] proposed a noncontact vital signs estimation system and creates a data set that consists of infrared and thermal videos for estimating both heart rate and respiration rate. However, the initial facial-region-detection phase requires time-consuming neural network training, and the vital signs were estimated by a heuristic CPC approach that is sensitive to the parameter setting. Hu et al. [53] proposed a similar system that is based on peak detection (PD) with the combination of color and thermal videos. Negishi et al. [54] used color videos for heart rate estimation and thermal for respiration rate estimation to screen subjects that potentially have seasonal influenza. However, Negishi et al. [54] adopted a computationally expensive multiple signal classification (MSC) algorithm for vital signs rate extraction. It can also be easily seen that [52]–[54] have drawback of privacy issues due to the facial-region detection module. See Table I for the list of the above video-based systems.

To bridge the gap between practicability and accuracy and facilitate Internet of Things for smart healthcare, in this article, we propose a nonintrusive vital signs estimation scheme via a graph-signal-processing-based thermal camera system for people during sleep. The system consists of a video capture module and a signal processing module. In a nutshell, we first capture thermal video frames that contain the facial region of interest (ROI) of a sleeping subject and then employ a series of signal processing tools for dimensionality reduction of the video signal and vital signs extraction. Specifically, leveraging on the inherent graph structure among subregions of the subject’s facial area, we propose a novel, graph-based, spatial–temporal signal denoising scheme that effectively reduces the noise level within the video signal. Taking the denoised video signal as the input, the system estimates the respiration rate and heart rate simultaneously. We summarize the contributions of this article as follows.

1) We propose an end-to-end, nonintrusive vital signs estimation system, where people can get near real-time vital signs reading during sleep.

2) Most existing works that adopt heuristic trend-removal process to reduce the body motion before vital signs signal extraction, which may introduce signal bias. We solve this problem by simply removing the trend-removal process, extracting 1-D temporal signal from each subregion of the ROI, and directly applying PCA.

3) Since the 1-D temporal signal of each subregion inherently correlates with each other, both spatially and temporally, where one can effectively define such correlation via a graph structure, our graph signal processing (GSP)-based spatial–temporal denoising scheme
TABLE I

SYSTEMS WITH VIDEO-BASED SENSORS

<table>
<thead>
<tr>
<th>method</th>
<th>camera</th>
<th>vital signs</th>
<th>core differentiators</th>
<th>disadvantages</th>
<th>privacy-preserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>[31]</td>
<td>color</td>
<td>heart rate</td>
<td>first camera-based heart rate estimator</td>
<td>limited to upright sitting position</td>
<td></td>
</tr>
<tr>
<td>[32]</td>
<td>heart</td>
<td>heart rate</td>
<td>pixel intensity magnification</td>
<td>vulnerable to abrupt movement</td>
<td></td>
</tr>
<tr>
<td>[33]</td>
<td>heart</td>
<td>heart rate</td>
<td>heart beat-induced head motion tracking</td>
<td>vulnerable to abrupt movement</td>
<td></td>
</tr>
<tr>
<td>[34]</td>
<td>heart</td>
<td>soft knowledge</td>
<td>spaces among color channels</td>
<td>not robust to change of lighting conditions</td>
<td></td>
</tr>
<tr>
<td>[37]</td>
<td>heart</td>
<td>respiration rate</td>
<td>end-to-end deep nets</td>
<td>not robust to poor lighting conditions</td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td>infrared</td>
<td>heart rate</td>
<td>feature extraction for drowsiness levels</td>
<td>not movement-robust</td>
<td></td>
</tr>
<tr>
<td>[41]</td>
<td>heart</td>
<td>heart beat rhythm</td>
<td>sparse-based denoising drive monitoring</td>
<td>vulnerable to abrupt movement</td>
<td></td>
</tr>
<tr>
<td>[42]</td>
<td>heart</td>
<td>heart rate</td>
<td>joint denoising and heart rate estimation</td>
<td>require prior signal of head motion</td>
<td></td>
</tr>
<tr>
<td>[43]</td>
<td>depth</td>
<td>heart rate</td>
<td>facial-cover-robust abnormal breathing classifier</td>
<td>vulnerable to abrupt movement, inherent low resolution</td>
<td></td>
</tr>
<tr>
<td>[44]</td>
<td>heart</td>
<td>heart rate</td>
<td>robust to various head facing</td>
<td>not robust to major head motion, inherent low resolution</td>
<td></td>
</tr>
<tr>
<td>[45]</td>
<td>response</td>
<td>heart rate</td>
<td>detection of abnormal breathing events</td>
<td>inherent low resolution</td>
<td></td>
</tr>
<tr>
<td>[47]</td>
<td>thermal</td>
<td>heart rate</td>
<td>robot-assisted camera facing</td>
<td>computationally expensive algorithm</td>
<td></td>
</tr>
<tr>
<td>[48]</td>
<td>response</td>
<td>heart rate</td>
<td>multiple combined estimators</td>
<td>inherent low resolution</td>
<td></td>
</tr>
<tr>
<td>[49],[50],[51]</td>
<td>heart rate</td>
<td>carotid artery detection</td>
<td>limited head facing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[52]</td>
<td>color+</td>
<td>respiration rate</td>
<td>dynamic time warping for selective signal segments</td>
<td>computationally expensive</td>
<td></td>
</tr>
<tr>
<td>[39]</td>
<td>thermal</td>
<td>heart rate</td>
<td>adaptive filtering</td>
<td>face detector requires neural nets training</td>
<td></td>
</tr>
<tr>
<td>[53]</td>
<td>color+</td>
<td>heart rate</td>
<td>robust face detector</td>
<td>not robust to poor lighting conditions</td>
<td></td>
</tr>
<tr>
<td>[54]</td>
<td>color+</td>
<td>heart rate</td>
<td>screening of seasonal influenza</td>
<td>computationally expensive</td>
<td></td>
</tr>
<tr>
<td>ours</td>
<td>thermal</td>
<td>heart rate</td>
<td>graph-based spatial-temporal filtering</td>
<td>-</td>
<td>✓</td>
</tr>
</tbody>
</table>

effectively reduces the noise level within the video signal.

4) The algorithm components in our system have relatively low time complexity and no model training is required. Therefore, our system can be efficiently adopted at the edge devices in a smart home setting, where the extracted vital signs can then be used for abnormal sleep event detection and disease screening.

The remainder of this article is organized as follows. We overview our proposed system in Section II. We explain our system components in detail in Section III. We show our experimental results on respiration rate and heart rate estimation in Section IV, and conclude this article in Section V.

II. PROPOSED SYSTEM

Our overall system consists of two parts, i.e., a video capture module and a signal processing module. The video capture module consists of a thermal camera with the resolution 640 × 480 at 30 frames/s that captures thermal videos [39]. As in [39], the camera is posed in a slightly look-down position so that the camera view captures the facial region of a sleeping subject. The signal processing module first detects the head/facial region using a Chamfer matching method [55], [56] in the thermal video frame and then tracks the above ROI using an off-the-shelf 2-D tracker in the subsequent thermal video frames. Unlike the existing methods that require feature selection [57] and tracking within the facial region [39], our method only requires a pyramid of binary head templates for computationally efficient Chamfer matching. The entire thermal video frame is then thresholded by fixed parameters so that only the relatively high temperature regions of interest are left in each frame. Next, for each subregion, we average the thermal image within the facial ROI to get dimensionality-reduced 1-D temporal signals. Here, we propose a novel graph-based, spatial–temporal 1-D signal filtering scheme, so that each channel (by block) is filtered with noise-level effectively reduced. We subsequently adopt a Butterworth filter for further noise-level reduction. We now apply PCA to the filtered multichannel 1-D signals and automatically pick the most periodic PCA component via fast Fourier transform (FFT) as the extracted vital signs signal. Finally, we designate the amplitude of the vital signs based on the periodicity of the selected signal. Our system components are summarized...
in Fig. 1. The signal processing steps are elaborated in the following sections.

III. VITAL SIGNS ESTIMATION DURING SLEEP

The main idea of our system is to detect the facial region in the thermal video frames in a privacy-preserved way and then track the facial ROI in the thermal video frames. We describe the signal processing module in detail next.

A. ROI Detection and Tracking

There may exist multiple irrelevant objects with relatively high temperature within the scene, which results in multiple irrelevant regions consisting of high pixel intensities in the thermal video frames. To reliably detect the facial region, we first apply scale-invariant Chamfer matching [55] with a pyramid of binary head templates, which returns the potential head locations. Next, we extract a circular region around each detected location and fit each region with a hemisphere of binary head templates, which returns the potential head locations. Next, we extract a circular region around each detected location and fit each region with a hemisphere to locate the probable head position. Then, for simplicity, we designate a square area $S = \{p_i\}_{i=1}^M$ with $M$ blocks of size $K \times K$ pixels centered inside the hemisphere as our ROI for tracking using an off-the-shelf 2-D spatial–temporal tracker [59]. Specifically, let $x_{t+1}$ be the head location in frame $t+1$, which is the center of the square area $S$, then, $x_{t+1}$ is predicted based on maximization of the following confidence map:

$$x_{t+1} = \arg \max_{x \in S(x)} m_{t+1}(x)$$

(1)

where $m_{t+1}(\cdot)$ involves fast Fourier convolution between a spatial convex model and a low-pass temporal-filtered abrupt appearance model, see [59] for details. We apply the above spatial–temporal tracker [59] for the entire image sequence, which results in a sequence of head locations $\{x_i\}_{i=1}^T$. We then obtain the pixel-intensity signal of each of the $M$ image blocks by spatial averaging, i.e.,

$$(y_i)_{t} = \frac{1}{K^2} \sum_{a=1}^K \sum_{b=1}^K \sum_{c=1}^T [p(a, b, c)]_t$$

(2)

where $T = 3$ for three color channels of the thermal videos. This leads to an $M$-channel length-$N$ 1-D temporal signal $Y = [y_1, \ldots, y_M] \in \mathbb{R}^{N \times M}$, which is used for spatial–temporal trajectory filtering to be described next.

B. Spatial–Temporal Trajectory Filtering

As shown in Fig. 2, the underlying true signal, i.e., the periodically changed skin/air temperature signal due to respiration/heart beat within the facial area is corrupted by the video acquisition noise and quantization noise. One can take block-wise spatially averaging of the pixel intensities to reduce the computational load. Apparently, the pixel-wise error within the captured video frames gets propagated [60] to the averaged signal. The above error propagation effect gets more significant if one further takes moving average operation to the block-wise spatially averaged signal. On the other hand, denoising of the thermal video frames [44], [45] in pixel level is computationally expensive due to high dimensionality. For the tradeoff between error propagation [60] and the computational cost, we resort to divide the ROI into multiple subregions, take the average for dimensionality-reduction block-wisely, and then denoise the spatially averaged signal. Given a thermal video frame sequence with $N$ frames, we first block-wise average the pixels within the facial ROI, which results in an $M$-channel ($M$ blocks within the facial ROI) length-$N$ 1-D signals. Since each subregion has a 1-D temporal signal

![Fig. 2: Proposed graph-filtering scheme (in blue) that reduces the noise level within the pixel intensities and suffices the tradeoff between computational load and error propagation. The vertical arrows in gray denote the processes that introduce error propagation.](image-url)
that has explicit time instants, and different subregions have explicit spatial correlations given the center pixel locations of each subregion, the averaged video signal has an inherent graph structure that can effectively describe the correlation among the signal observations both spatially and temporally. The above spatial–temporal covariance among the observations can be estimated using existing sparse inverse covariance estimation methods, e.g., a graphical lasso [61]. However, the lasso-based methods require repetitive update of the entries of the target covariance matrix row- and column-wisely, which can be computationally expensive for large number of observations. Therefore, we apply a novel graph-based spatial temporal trajectory filtering scheme for noise reduction within the preprocessed 1-D signals, which we explain in detail next. Furthermore, unlike [44] and [45] that constructed the graph without any graph update mechanism, our following proposed graph-based spatial–temporal filtering scheme iteratively restores the pixel intensities signal and updates the graph metric via the restored pixel intensities signal, which help to finally recover the underlying noise-free vital signs signal.

1) Graph-Based Signal Denoising: The foundation of our proposed graph-based filtering scheme is based on the following GSP tools, including the basic structure of a graph, graph Laplacian regularization (GLR), and graph total variation (GTV). Specifically, given a set of observations, one can define an undirected graph \( G = [V, E, A] \) with: 1) a set of vertices \( V \) that correspond to the observations and 2) a set of edges \( E \) that correspond to the levels of similarities among the vertices \( V \). With a Gaussian kernel, each edge within \( E \) for vertices \( i \) and \( j \) is given by the following nonnegative edge weights:

\[
    w_{ij} = \begin{cases} 
    \exp\left(-\left(\mathbf{f}_i - \mathbf{f}_j\right)^\top \mathbf{M} \left(\mathbf{f}_i - \mathbf{f}_j\right)\right), & \text{if } i \neq j \\
    0, & \text{otherwise (o.w.)} 
    \end{cases}
\]  

all of which define an adjacency matrix \( A \). Here, \( \mathbf{M} \in \mathbb{R}^{3 \times 3} \) is a symmetric, positive-definite matrix used to define a Mahalanobis feature distance \( (\mathbf{f}_i - \mathbf{f}_j)^\top \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j) \) [62]. \( w_{ij} \) measures the similarity among features of two samples, \( \mathbf{f}_i \in \mathbb{R}^3 \) denotes the sample feature, i.e., the horizontal and vertical pixel locations and pixel intensity of an observation. Now, we define a diagonal, degree matrix \( D \) with diagonal entries \( D_{ij} = \sum_j A_{ij} \). We can finally define a combinatorial graph Laplacian \( \mathbf{L} = \mathbf{D} - \mathbf{A} \), which has been widely adopted to solve inverse problems in imaging [63]–[67]. Given \( \mathbf{L} \) is a symmetric, positive semidefinite matrix proved via the Gershgorin circle theorem (GCT) [68], we can apply eigendecomposition \( \mathbf{L} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top \), where each column entry of \( \mathbf{U} \) denotes the eigenvector and each diagonal entry \( \lambda_k \) of \( \mathbf{A} \) denotes the corresponding eigenvalue.

As shown in Fig. 3, given observed, \( M \)-channel length-\( N \) 1-D temporal signal \( \mathbf{Y} \in \mathbb{R}^{N \times M} \), we first vectorize \( \mathbf{Y} \) as \( \text{vec}(\mathbf{Y}) \in \mathbb{R}^{NM} \), and then aim to recover a signal \( \text{vec}(\mathbf{V}) \in \mathbb{R}^{NM} \) that is an (ideally) noise-free \( M \)-channel length-\( N \) 1-D temporal signal. Starting from a maximum a posteriori estimation formulation, we have

\[
    \text{vec}(\mathbf{V})_{\text{MAP}} = \arg \max_{\text{vec}(\mathbf{V})} f(\text{vec}(\mathbf{Y})|\text{vec}(\mathbf{V}))p(\text{vec}(\mathbf{V})).
\]  

Assuming a Gaussian Markov random fields (GMRF) model [64], we can now express the likelihood function \( f \) and the prior distribution \( p \) explicitly

\[
    f(\text{vec}(\mathbf{Y})|\text{vec}(\mathbf{V})) = \exp\left(-\|\text{vec}(\mathbf{Y} - \mathbf{V})\|_2^2\right) \\
    p(\text{vec}(\mathbf{V})) \propto \exp\left(-\text{vec}(\mathbf{V})^\top \mathbf{L} \text{vec}(\mathbf{V})\right).
\]  

Subsequently, we can formally express our optimal denoising problem [69] as

\[
    \min_{\text{vec}(\mathbf{V})} \|\text{vec}(\mathbf{Y} - \mathbf{V})\|_2^2 + \mu \text{vec}(\mathbf{V})^\top \mathbf{L} \text{vec}(\mathbf{V})
\]  

where \( \mu \) is a weighting parameter. The optimal denoising [70] problem in (6) has two quadratic terms and this quadratic program (QP) has a closed-form solution \( \text{vec}(\mathbf{V}^*) = (\mathbf{I} + \mu \mathbf{L})^{-1} \text{vec}(\mathbf{Y}) \). Specifically, with a combinatorial graph Laplacian \( \mathbf{L} \), the second term

\[
    \text{vec}(\mathbf{V})^\top \mathbf{L} \text{vec}(\mathbf{V}) = \frac{1}{2} \sum_{i,j \in E} w_{ij} \Delta \text{vec}(\mathbf{V})_{ij}^2 = \sum_{k=1}^{N} \lambda_k \text{vec}(\tilde{\mathbf{V}})^2
\]

is called a GLR that is based on a graph signal piecewise smoothness (PWS) prior [63], where \( \Delta \text{vec}(\mathbf{V})_{ij} = \text{vec}(\mathbf{V})_{ij} - \text{vec}(\mathbf{V})_{ij} \), \( \text{vec}(\mathbf{V}) = \mathbf{U}^\top \text{vec}(\mathbf{V}) \) denotes the graph Fourier transform coefficients of \( \text{vec}(\mathbf{V}) \). Since \( \mathbf{L} \succeq 0 \), (7) is lower bounded by 0. Minimizing (7) aims to reduce the energy of signal \( \text{vec}(\mathbf{V}) \) that corresponds to the high graph frequencies. Similarly, another popular graph signal total variation prior [71], [72] is given by

\[
    \|\text{vec}(\mathbf{V})\|_{\text{GTV}} = \sum_{i,j \in E} w_{ij} |\Delta \text{vec}(\mathbf{V})_{ij}|. 
\]  

Directly replacing the GLR term in (6) with the GTV in (8) will not give us a closed-form solution. As in [71], the GTV \( l_1 \)-norm term is transformed to a quadratic term by first defining a new adjacency matrix \( \mathbf{S} \) with an edge weight

\[
    s_{ij} = \max \left| \frac{w_{ij}}{|\Delta \text{vec}(\mathbf{V})_{ij}|}, \alpha \right|
\]

where \( |\Delta \text{vec}(\mathbf{V})_{ij}| \) is an estimated signal and \( \alpha \) is a heuristic chosen parameter that avoids numerical instability when \( |\Delta \text{vec}(\mathbf{V})_{ij}| \) is close to 0. Since

\[
    s_{ij} \Delta \text{vec}(\mathbf{V})_{ij}^2 = \frac{w_{ij}}{|\Delta \text{vec}(\mathbf{V})_{ij}|} \Delta \text{vec}(\mathbf{V})_{ij}^2 \approx w_{ij} |\Delta \text{vec}(\mathbf{V})_{ij}|
\]

we define a new adjacency matrix \( \mathbf{S} \). And then

\[
    s_{ij} = \frac{w_{ij}}{|\Delta \text{vec}(\mathbf{V})_{ij}|} |\Delta \text{vec}(\mathbf{V})_{ij}| \]
the GTV $l_1$-norm term can now be transformed to a quadratic form. To do this, we define an $l_1$-Laplacian matrix $L_\nu$ as
\[
L_\nu = \text{diag}(S1) - S
\] (11)
where every entry within a column vector $1 \in \mathbb{R}^N$ is equal to 1. Apart from the PWS-prior based GLR optimal denoising problem in (6), we now have a GTV-prior based GLR optimal denoising problem
\[
\min_{\text{vec}(V)} \|\text{vec}(V - Y)\|^2_2 + \mu \|\text{vec}(V)^\top L_\nu \text{vec}(V)\|_* + \beta \|\text{vec}(V)\|_1
\] s.t. $M > 0, \text{Tr}(M) \leq C$ (12)
For both (6) and (12), since $L \succeq 0, L_\nu \succeq 0$, and thus $I + \mu L \succ 0, I + \mu L_\nu \succ 0$. By defining a condition number $\gamma = \lambda_{\max}/\lambda_{\min}$ where $\lambda$’s denote the eigenvalues of the matrix $I + \mu L$ and $I + \mu L_\nu$, and setting an upper bound for $\mu \leq (\gamma - 1)/(2d)$ [73], where $d$ denotes the maximum degree of the vertices in a given graph $G$, one can numerically guarantee that both solution $\text{vec}(V^*) = (I + \mu L)^{-1}\text{vec}(Y)$ for (6) and solution $\text{vec}(V'^*) = (I + \mu L_\nu)^{-1}\text{vec}(Y)$ for (12) are stable [73]. Similar to the PWS-prior based GLR term, the GTV-prior based GLR term is also lower bounded by 0 [74].

2) Graph-Based Signal Denoising With Optimal Feature Metric: Directly solving (6) or (12) with $M = I$ (I denotes an identity matrix with a proper dimension) in (3) may still result in suboptimal solutions, since the feature distance in (3) is not optimized by simply fixing $M = I$. In this article, we propose to alternately optimize the Mahalanobis feature distance via updating graph metric matrix $M$ in (3) and optimize the target signal $\text{vec}(V)$, with a low-rank prior [75]. Therefore, the overall objective function is given by
\[
\min_{\text{vec}(V)} \|\text{vec}(Y - V)\|^2_2 + \alpha \|\text{vec}(L_\nu V)\|_2 \|\text{vec}(V)\|_* + \beta \|\text{vec}(V)\|_1
\] s.t. $M > 0, \text{Tr}(M) \leq C$ (13)
where $\alpha$ and $\beta$ denote weighting parameters for the PWS/GTV term and the nuclear-norm term, respectively. $C$ denotes a constant, $\text{Tr}(-)$ denotes the trace of a matrix, i.e., the sum of the diagonals, and $\|\cdot\|_*$ denotes the nuclear norm of a matrix [75], i.e., the sum of singular values of a matrix. Since $M$ is positive definite, $\|\cdot\|_*$ is equivalent to $\text{Tr}(-)$.

3) Algorithm Development: When solving $\text{vec}(V)$, the third term does not have optimization variables, and thus (13) has a closed-form solution $\text{vec}(V^*) = (I + \alpha L)^{-1}\text{vec}(Y)$. When solving the metric matrix $M$, the first term does not have optimization variables, and thus we iteratively apply gradient descent and positive-definite-cone projection. Specifically, we define $Q(M) = \alpha \text{vec}(V)^\top L(M)\text{vec}(V) + \beta \|\text{vec}(V)\|_1$. The above gradient descent and projection steps are given by
\[
M_t = \text{Proj}(M_{t-1} - \gamma_t \nabla Q(M_{t-1}))
\] (14)
where $\text{Proj}(-)$ denotes the PD-cone projection, and $\gamma_t$ is a step size that is initialized heuristically, increased by a small amount if the gradient descent yields a smaller objective and decreased by half otherwise, the gradient w.r.t. $M_{m,n}$ is given by
\[
\frac{\partial Q(M)}{\partial M_{m,n}} = \begin{cases} -\alpha \sum_{i,j} \Delta(f_{ij})_n \Delta(f_{ij})_m w_{ij} \Delta \text{vec}(V)_{ij}^2, & \text{if } m \neq n. \\ -\alpha \sum_{i,j} \Delta(f_{ij})_m^2 w_{ij} \Delta \text{vec}(V)_{ij}^2, & \text{o.w.} \end{cases}
\] (15)
where $\Delta f_{ij} = f_i - f_j$. To perform PD-cone projection, we first find the eigendecomposition of
\[
M_{t-1} - \gamma_t \nabla Q(M_{t-1}) = \sum_i \lambda_i w_i w_i^\top
\] (16)
where $w_i$ denotes the eigenvector that corresponds to the eigenvalue $\lambda_i$. We then define the projection
\[
M_t = \sum_i \max(\lambda_i, 0) w_i w_i^\top
\] (17)
followed by a scaling:
\[
M_t^* = CM_t/\text{Tr}(M_t), \quad \text{if } \text{Tr}(M_t) > C.
\] (18)
We solve (13) until convergence. In summary, we call our above proposed graph-based spatial–temporal filtering schemes: 1) multichannel ($M$-channel signal) PWS (based on a combinatorial graph Laplacian $L$ in (7)) with a nuclear-norm term (MPT) and 2) multichannel GTV (based on an $l_1$-Laplacian matrix $L_\nu$ in (11)) with a nuclear-norm term (MGT), respectively.

Our objective in (13) is different from a recent GSP-based vital signs estimation work in [76], where an additional low-rank prior is considered to promote the underlying low-rank structure among the features. We show Lemma 1 and its proof that, by conditioning $\beta$, our proposed formulation can lead to a smaller objective than [76], and thus improved denoising results.

Lemma 1: If $\beta \leq \alpha \sum_{n\neq m} \sum_{i,j} \Delta(f_{ij})_m \Delta(f_{ij})_n - \Delta(f_{ij})_m w_{ij} \Delta \text{vec}(V)_{ij}^2$, then the Mahalanobis feature distance $\Delta(f_{ij})_m^2 M \Delta(f_{ij})_n$ has a larger upper bound by solving $M$ via (13) than solving $M$ without the nuclear-norm term, which potentially produce smaller edge weights in (3), and thus leads to a smaller objective.

Proof: Since
\[
\beta \leq \alpha \sum_{n\neq m} \sum_{i,j} \Delta(f_{ij})_m \Delta(f_{ij})_n - \Delta(f_{ij})_m w_{ij} \Delta \text{vec}(V)_{ij}^2
\] (19)
we have
\[
\alpha \sum_{i,j} \Delta(f_{ij})_m^2 w_{ij} \Delta \text{vec}(V)_{ij}^2 - \beta \\
\geq \alpha \sum_{n\neq m} \sum_{i,j} |\Delta(f_{ij})_m \Delta(f_{ij})_n| w_{ij} \Delta \text{vec}(V)_{ij}^2
\] (20)
i.e.,
\[
-\gamma_t \frac{\partial Q(M)}{\partial M_{m,n}} \geq -\gamma_t |\sum_{n\neq m} \frac{\partial Q(M)}{\partial M_{m,n}}|
\] (21)
and thus $-\gamma_t \nabla Q(M_{t-1}) \succeq 0$ by GCT [68]. Since $M_{t-1} > 0$, we have $M_t > 0, M_t^* > 0$. Let the eigendecomposition of $M_t^*$ be $M_t^* = U \Lambda U^\top$, then
\[
\Delta(f_{ij})_m \Delta(f_{ij})_n \leq \lambda_{\max} \Delta(f_{ij})_m^2.
\] (22)
If we optimize $\mathbf{M}$ without the nuclear-norm term in (13), let the corresponding most recent metric matrix be $\mathbf{M}^*$, then, $\mathbf{M}_t = \mathbf{U} \mathbf{A} \mathbf{U}^\top$, and
\[
\Delta (\mathbf{f}_{ij})^\top \mathbf{M}_t^* \Delta (\mathbf{f}_{ij}) \leq \hat{\lambda}_{\text{max}} \Delta (\mathbf{f}_{ij})^2.
\] (23)
Since $\text{Tr}(\mathbf{M}_t) < \text{Tr}(\mathbf{M}_t)$, and thus $\hat{\lambda}_{\text{max}} < \lambda_{\text{max}}$. Since the Mahalanobis feature distance optimized by (13) has a larger upper bound than solving $\mathbf{M}$ without the nuclear-norm term, we conclude that (13) potentially produce smaller edge weights, and thus lead to a smaller objective than [76].

The core idea of the above iterative graph-based optimal denoising/graph metric learning is to leverage on its the spectral properties of GSP, i.e., useful information is concentrated in low frequencies. As discussed in Section I, graph based has been used in existing nonintrusive vital signs estimation systems, such as [44] and [45]. Our proposed graph-based scheme differs from [44] and [45] in the sense that we utilize the inherent graph structure within the video signal, together with a graph update mechanism, and effectively reduce the noise level within the pixel intensities and have better tradeoff between computational load and error propagation.

C. Signal Post-Processing for Vital Signs Monitoring

The noise level within graph-filtered signal $\mathbf{Y}^*$ is effectively reduced. However, potential inherent noise may still exist within $\mathbf{Y}^*$. Therefore, we further apply Butterworth filtering, PCA, and FFT for post-processing of the signal and vital signs monitoring. We explain each of these three steps in detail next.

1) Butterworth Filtering: We first apply a second-order Butterworth filter to each of the $M$ channels in $\mathbf{Y}^* \in \mathbb{R}^{N \times M}$. For fair comparison and to discriminate the respiration rate and the heart rate, we follow the same settings of the frequency band in [39], and set the frequency band $[0.167, 0.667]$ Hz for breathing rate estimation and $[0.667, 1.667]$ Hz for heart rate estimation.

2) PCA Signal Decomposition: Now, we apply PCA to the Butterworth filtered, $M$-channel length-$N$ 1-D temporal signal $\mathbf{Y}^* = [\mathbf{y}_1, \ldots, \mathbf{y}_M], \mathbf{y}_i \in \mathbb{R}^{N}$, with $K$ components. Specifically, PCA first computes the average of $\mathbf{y}_i$’s
\[
\bar{\mathbf{Y}}^* = [\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_M], \quad \bar{\mathbf{y}}_i = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i
\] (24)
and then computes the following covariance matrix:
\[
\Sigma_{\mathbf{Y}^*} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{y}_i - \bar{\mathbf{Y}}^*) (\mathbf{y}_i - \bar{\mathbf{Y}}^*)^\top.
\] (25)
Next, PCA solves the following eigen-pair problem:
\[
\Sigma_{\mathbf{Y}^*} \mathbf{E} = \mathbf{E} \mathbf{A}
\] (26)
where $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_M]$ denotes the $M$-column eigenvectors that correspond to the $M$-channel length-$N$ 1-D temporal signal $\mathbf{Y}^*$, and $\mathbf{A} = \text{diag}(\lambda_1, \ldots, \lambda_M)$ denotes a diagonal matrix with diagonals being the eigenvalues corresponding to the eigenvectors $\mathbf{e}_i$’s in $\mathbf{E}$. Finally, PCA projects $\mathbf{Y}^*$ onto $\mathbf{e}_i$’s to get the projected 1-D temporal signal, i.e., PCA component
\[
\text{proj}_{\mathbf{e}_i} \mathbf{Y}^* = \mathbf{Y}^* \cdot \mathbf{e}_i.
\] (27)
Note that the PCA component $\text{proj}_{\mathbf{e}_i} \mathbf{Y}^*$ is perpendicular to another PCA component $\text{proj}_{\mathbf{e}_j} \mathbf{Y}^*$, where $i, j \in \{1, \ldots, M\}$.

3) Signal Selection and Vital Signs Monitoring: We designate the most periodic PCA component as the extracted vital sign signal. To find the most periodic PCA component, we follow [33], apply FFT to each PCA component $\text{proj}_{\mathbf{e}_i} \mathbf{Y}^*$, and calculate the proportion of the amplitude of the single-sided power that corresponds to the dominant frequency and its $l$-harmonics:
\[
p = \frac{P_t}{\sum_{i=1}^{P_t}} \in [0, 1]
\] (28)
where $P_t$ denotes the largest amplitude of the single-sided power, and $P_t, i > 1$ denotes the $i$-th-harmonic of $P_t$. The PCA component with the largest proportion $p$ is then designated as the extracted vital sign signal. Finally, we estimate the respiration rate and heart rate via averaging the dominant frequency $f_1$ that corresponds to $P_t$ with a length-$m$ FFT window on the selected PCA component and then multiply it by 60, which results in the respiration rate/heart rate estimation $v = \tilde{f}_1 \times 60$ breaths/beats per minute, respectively.

IV. EXPERIMENTAL RESULTS

In this section, we evaluate our system using the video data set in [39] for estimation of both respiration rate and heart rate. We do this by first comparing with state-of-the-art video-based vital signs estimation systems, and then comparing the variants of our system via an ablation study, which we discuss in detail next.

A. Experimental Settings

We adopt a publicly available thermal video data set [39] that consists of 28 trials and 12 subjects (two females and ten males, from 21 to 38 years old) from Shanghai Jiao Tong University, sleeping with various body movements and head positions with respect to the thermal video camera (MAG62, Magnity Electronics Company Ltd., Shanghai, China). See the first column of Tables III and IV for the detail head positions, including front facing, slight head rotation, and frequent head rotation. Also, see Fig. 4 for head region of sample thermal images in grayscale. Each trial consists of a 35–60 s thermal video. The thermal camera has the resolution of $640 \times 480$ and the thermal imaging wavelength is 7.5–14 $\mu$m, which belongs to the long-wave bands according to [46] and the captured thermal image retains little amount of the facial details [46] compared to RGB and infrared cameras, i.e., being privacy preserved. The subject was positioned 1 to 3 m away from the thermal camera. The illumination intensity of the video recording environment varied between 0 and 3 Lux. We show the head region of the sample thermal images for each of the 12 subjects in the first two rows of Fig. 4. We also show the head region of the sample thermal images for two
TABLE II
PARAMETER SETTINGS

<table>
<thead>
<tr>
<th>params.</th>
<th>usage</th>
<th>settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>GLR weighting</td>
<td>$10^{2}$</td>
</tr>
<tr>
<td>β</td>
<td>trace-weighting</td>
<td>$1 - 10^4$, change by $10^2$</td>
</tr>
<tr>
<td>γ</td>
<td>metric step size</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>$P^*$</td>
<td>PCA components</td>
<td>5</td>
</tr>
<tr>
<td>l</td>
<td>number of harmonics</td>
<td>2</td>
</tr>
<tr>
<td>m</td>
<td>fft window size</td>
<td>2-30 sec</td>
</tr>
<tr>
<td>C</td>
<td>trace constraint</td>
<td>3</td>
</tr>
</tbody>
</table>

Fig. 4. Head region of sample thermal grayscale images. First two rows: sample images for 12 tested subjects; last two rows: sample images for two of the 28 thermal video sequences, showing various head positions with respect to the thermal video camera due to different body postures/sleep positions.

of the 28 thermal video sequences in the last two rows of Fig. 4. One can see various head positions with respect to the thermal video camera due to different body postures/sleep positions.

The parameter settings required for the experiments are summarized in Table II. Specifically, the trace-weighting parameter $\beta$ and FFT window size vary within a heuristically set range and are finally designated when the averaged amplitude of the power spectrum peak is the largest. All experiments were conducted in MATLAB R2020a.

B. Comparing With State of the Art

We compare our system (two proposed schemes, namely, MPT and MGT, respectively) with the following six state-of-the-art video-based systems: two blind source separation-based approaches, including VMD [35] and EEMD [36], two signal signature detection-based approaches, including PD [53] and CPC [39], and two signal similarity measurement-based approaches, including DTW [52] and MSC [54]. All above competing schemes were implemented by ourselves, and the parameters were set following the experimental settings in the respective papers. For the parameters that are not specified in the respective papers, we set them empirically. As described in Section I, the nonvideo RF-based contact-less sensing techniques, including WiFi-based, RFID-based, mmWave-based, and laser-based respiration rate and heart rate estimation techniques, are sensitive to the movement, location, and orientation of the subject [23]. The sensitivity issue is more severe for heart rate estimation via these RF-based contact-less works due to the fact that heartbeat is overlapped with and covered by the chest and abdomen respiration movements. In addition, mmWave-based and laser-based methods require customized high-end Doppler active radios, which in general are not readily available on the market and therefore incur prohibitive cost [22] and are difficult for experimental use. The nonvideo acoustic-based systems generally have limited measurement range and are susceptible to environment audio noise [20]. On the other hand, RF-based, acoustic-based and video-based works handle different types of signal noise, making it difficult to have a fair comparison. Specifically, RF-based works deal with the high-frequency environment noise [19], acoustic-based works deal with the environment audio noise [20], whereas the video-based works deal with the image noise. Therefore, we focus on performance comparison of the aforementioned video-based works.
Specifically, for the PCA phase in our system, we perform PCA with a $P$-component decomposition, where $P = 5$ across all experiments. We can clearly see from Fig. 5 that PCA component 2 for respiration signal of Trial 3 is the most periodic one among all components since the corresponding power spectrum has a largest peak among all PCA components.
components. Similar to the respiration rate estimation, as shown in Fig. 6, PCA component 5 for heart beat signal of Trial 21 is the most periodic one among all components. Fig. 7 shows the selected PCA components as respiration/heart beat signal and estimated respiration/heart rate, respectively.

We conducted a Kolmogorov–Smirnov (KS) statistical significance test [77] to measure the similarity of the distributions between the estimated respiration/heart rates and the ground truth, with the null hypothesis being that the observations are from the same distribution. First, as shown in the KS-test p-values in Table III, at a 5% significance level, 7 of the 8 pairs of distributions in respiration rate estimation, including VMD [35], EEMD [36], PD [53], CPC [39], MSC [54], and the proposed MPT and MGT, were found to be similar, and the result of DTW [52] was found not to be similar to the ground truth. Similarly, as shown in the KS-test p-values in IV, at a 5% significance level, 6 of the 8 pairs of distributions in heart rate estimation, including VMD [35], EEMD [36], CPC [39], MSC [54], and the proposed MPT and MGT, were found to be similar, and the result of PD [53] and DTW [52] was found not to be similar to the ground truth.

We investigated the root-mean-square error (RMSE), $R^2$-value, and Bland–Altman plots [78] between the estimated respiration/heart rate and the ground truth. As shown in Tables III and IV, our proposed scheme MGT achieves the smallest RMSE, largest $R^2$-value, and smallest 95% confidence interval for the Bland–Altman plot for both respiration rate and heart rate estimation compared to the other competing schemes. Although various head positions were present in the captured thermal videos, all evaluated methods track the ROI successfully throughout each video sequence. For our proposed system, this is due to the deployment of a robust spatial–temporal ROI tracker [59]. When major head movements are present or the facing direction is not perfectly pointing toward the camera, the vital signs estimation may result in worse performance than that in an ideal setting, which can be seen in all evaluated methods. For example, the respiration rate estimation results of Trials 1, 12, 13, 14, 16, and 23 for six evaluated state-of-the-art methods have very poor agreement compared to other trials. Similar results can also be seen at the heart rate estimation of Trials 1, 8, 12, 13, 14, 17, 19, 24, and 25. This means that head movements can significantly affect the heat change of the air within the ROI region, which leads to poor quality of captured thermal videos that generally bring difficulty of signal filtering. The overall difference of the respiration rate estimation between all evaluated methods and the ground truth is smaller than that of the heart rate estimation, which is mainly due to the fact that the respiration has much more significant influence on the heat change of the air temperature within the ROI than the heart beat.

For competing schemes, VMD [35] and EEMD [36] focus on vital signs estimation with two state-of-the-art blind source separation techniques, but they do not perform well in vital signs estimation in thermal images, as shown in Tables III and IV. Specifically, both VMD [35] and EEMD [36] estimate the respiration rate and the heart rate with more fluctuation compared to the other experimented schemes. Although various head positions were present in the captured thermal videos, all evaluated methods track the ROI successfully throughout each video sequence. For our proposed system, this is due to the deployment of a robust spatial–temporal ROI tracker [59]. When major head movements are present or the facing direction is not perfectly pointing toward the camera, the vital signs estimation may result in worse performance than that in an ideal setting, which can be seen in all evaluated methods. For example, the respiration rate estimation results of Trials 1, 12, 13, 14, 16, and 23 for six evaluated state-of-the-art methods have very poor agreement compared to other trials. Similar results can also be seen at the heart rate estimation of Trials 1, 8, 12, 13, 14, 17, 19, 24, and 25. This means that head movements can significantly affect the heat change of the air within the ROI region, which leads to poor quality of captured thermal videos that generally bring difficulty of signal filtering. The overall difference of the respiration rate estimation between all evaluated methods and the ground truth is smaller than that of the heart rate estimation, which is mainly due to the fact that the respiration has much more significant influence on the heat change of the air temperature within the ROI than the heart beat.
Fig. 8. Bland–Altman plots (left) and correlation plots (right) of estimated respiration rate compared to state of the art. From top to bottom: VMD [35], EEMD [36], PD [53], CPC [39], DTW [52], and MSC [54].

Fig. 9. Bland–Altman plots (left) and correlation plots (right) of estimated heart rate compared to state of the art. From top to bottom: VMD [35], EEMD [36], PD [53], CPC [39], DTW [52], and MSC [54].

estimation is the smallest among all experimented schemes, as can be seen from Fig. 7. Furthermore, the GTV-based variant (namely, MGT) of our proposed system performs better than the PWS-based variant (namely, MPT) of our system, which shows that the GTV-prior can better reflect the correlation among the observations, both spatially and temporally.

C. Ablation Study

We now examine the contribution of our system components in the following ablation study. In particular, we compare our system schemes MGT and MPT with the following four variants of our schemes.

1) SC: A scheme with only a single channel, i.e., taking the average of the entire ROI.

2) MC: A scheme with multiple channels without our proposed graph-based spatial–temporal filtering component.

3) MP: A scheme with multiple channels with a PWS-prior-based graph filtering component, but without the trace constraint term.
Fig. 10. Estimated respiration signal and rate [(a) and (b), respectively] for Trial 3 (ground-truth respiration rate: 9 bpm) and heart beat signal and heart rate [(c) and (d)] for Trial 21 (ground-truth heart rate: 51 bpm) compared in an ablation study, including single channel, multichannel, MP, MG, and the proposed MPT and MGT.

### Table V

**Complexity Versus Performance.** *M* **Blocks. Length-N Time Series.** *E* **Denotes the Number of Trainable Parameters in the Head Detector in [39].** *Q ≪ N* **Denotes the Segment Length of the Signal for Denoising.** *k* **and *l* Denote the Number of Iterations of the Alternating Denoising/Metric Learning Process and That of Metric Learning, Respectively.**

<table>
<thead>
<tr>
<th>Method</th>
<th>time complexity</th>
<th>$R^2$ BR (Respiration)</th>
<th>$R^2$ HR (Heart Rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMD [35]</td>
<td>$O(N \log N)$</td>
<td>0.537</td>
<td>0.818</td>
</tr>
<tr>
<td>BEMD [36]</td>
<td>$O(N \log N)$</td>
<td>0.304</td>
<td>0.818</td>
</tr>
<tr>
<td>PD [53]</td>
<td>$O(M)$</td>
<td>0.705</td>
<td>0.533</td>
</tr>
<tr>
<td>CPC [39]</td>
<td>$O(M + E)$</td>
<td>0.817</td>
<td>0.920</td>
</tr>
<tr>
<td>DTW [52]</td>
<td>$O(MN)$</td>
<td>0.793</td>
<td>0.877</td>
</tr>
<tr>
<td>MSC [54]</td>
<td>$O(M^3N^3)$</td>
<td>0.846</td>
<td>0.739</td>
</tr>
<tr>
<td>SC</td>
<td>$O(1)$</td>
<td>0.562</td>
<td>0.637</td>
</tr>
<tr>
<td>MC</td>
<td>$O(M)$</td>
<td>0.731</td>
<td>0.887</td>
</tr>
<tr>
<td>MP</td>
<td>$O((MN/Q) (k(8 + \log 2)l + Q^3))$</td>
<td>0.827</td>
<td>0.985</td>
</tr>
<tr>
<td>MG</td>
<td>$O((MN/Q) (k(8l + Q^3)))$</td>
<td>0.871</td>
<td>0.918</td>
</tr>
<tr>
<td>MPT</td>
<td>$O((MN/Q) (k(8l + Q^3)))$</td>
<td>0.915</td>
<td>0.964</td>
</tr>
</tbody>
</table>

4) **MG:** A scheme with multiple channels with a GTV-prior-based graph filtering component, but without the trace constraint term.

Similar to our above comparison with the state of the art, as shown in Tables III and IV, our two schemes MPT and MGT achieve competitive average respiration rate and heart rate estimation values against the other competing schemes. Furthermore, our proposed schemes MPT and MGT estimate the vital signs with less fluctuation compared to their four variants, as can be seen from Fig. 10. The Bland–Altman plots [78] and correlation plots in Figs. 11 and 12 further show the superior performance of our proposed MGT scheme, with the largest $R^2$-value, the smallest 95% confidence interval for the Bland–Altman plot, among all experimented schemes. As expected, the SC scheme (single channel) performs worse than the MC scheme (multichannel) due to the fact that more observations provide more information about the vital signs. Again, our proposed MGT performs better than MPT due to potentially better graph structure representation. Furthermore, MGT performs better than MG with an additional trace term in both respiration and heart rates estimation, and MPT performs better than MP in respiration rate estimation, which shows the superior generality of MPT and MGT than MP and MG, respectively.

### D. Discussion on Model Complexity

We compare the model complexities among the above six state of the art, our proposed schemes, and the four variants of our schemes compared in the above ablation study. As shown in Table V, SC has the lowest time complexity but also has the lowest $R^2$-values for both respiration rate and heart rate estimation. VMD [35] and EEMD [36] have much higher time complexities than SC. PD [53] and MC have the same time complexity at $O(M)$, where MC has slightly larger $R^2$ values, which is due to the advantage of spatial-averaging within multiple subregions in MC over the entire ROI in PD [53]. CPC [39] consists of a dedicated face detector that requires $E$ (can be very large, i.e., $M ≪ E$) parameters, and thus can be very time consuming in practice. DTW [52] consists of repeated pairwise signal segment comparison, which has suboptimal $R^2$ values with a relatively low time complexity. MSC [54] consists of an MSC algorithm, which has a very high time complexity $O(M^3N^3)$ and significantly worse $R^2$ value in heart rate estimation compared to CPC [39]. MP and MG are the most computationally expensive schemes given $M$ blocks (channels) and length-$N$ time series, which is due to the fact that the optimization with a trace constraint generally requires a binary search step [79].
Although our proposed schemes MPT and MGT require eigendecomposition during the optimization of (13) in general, by conditioning $\beta$ in (13), the eigendecomposition can be bypassed given Lemma 1, i.e., the $Q^3$ ($Q$ denotes the segment length of the signal for denoising) term is not required in Table V, and thus MPT and MGT have relatively low time complexity compared to the above two schemes, with improved performance on the estimation of both respiration rate and heart rate.

V. Conclusion

In this article, we propose a vital signs estimation system during sleep via a thermal camera that bridges the gap between practicability and accuracy among existing vital signs estimation systems. The system first captures thermal image sequences of a sleeping subject and then processes the facial regions within image sequences for vital signs signal extraction. Specifically, we propose a graph-based, spatial–temporal signal denoising scheme that leverages on the inherent graph structure among subregions of the facial
area. Experimental results confirmed that our graph-based denoising scheme effectively reduces the noise level introduced by cameras and subjects, and our proposed system outperforms state-of-the-art nonintrusive vital signs monitoring systems. Our system can be deployed efficiently at the edge devices in a smart home setting thanks to relatively low time complexity. The extracted vital signs can then be used for sleep abnormality detection and disease screening. Future works will focus on simultaneous monitoring the respiration pattern and heart beat pattern when the facial region of the subject is not available, and also attempt to use event cameras [80] for more accurate vital signs tracking.

REFERENCES


Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009.

He was a student intern with the Institute for InfoComm Research, Singapore, from 2006 to 2007. From 2007 to 2008, he was a visiting student with the School of Computer Engineering, Nanyang Technological University, Singapore. From 2008 to 2009, he was a visiting student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Postdoctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen–Nuremberg, Erlangen, Germany. Since 2012, he has been with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, where he is currently a Professor. His current research interests include multimedia signal processing and perceptual signal processing.

Prof. Zhai has received the Humboldt Fellowship in 2011, the Award of National Excellent Ph.D. Dissertation from the Ministry of Education of China in 2012, the Best Student Paper Award of Picture Coding Symposium in 2015, the Best Student Paper Award of IEEE International Conference of Multimedia and Expo (ICME) 2016, the Best Paper Award of IEEE TRANSACTIONS ON MULTIMEDIA in 2018, the Grand Prize of Saliency360! Challenge of ICME in 2018, the Best Paper Award of IEEE Mobile Multimedia Computing Workshop in 2019, the Best Paper Award of IEEE CVPR Dyna-Vis Workshop in 2020, the Best Paper Runner-Up of IEEE TRANSACTIONS ON MULTIMEDIA in 2021, and the First Place of UGC VQA Contest FR Track in IEEE ICME 2021. He also received the “Eastern Scholar” and “Dawn” Program Professorship of Shanghai, the NSFC Excellent Young Researcher Award, and the National Top Young Researcher Award in China. He serves as an Editor-in-Chief for Displays (Elsevier) and is on the editorial board of Science China: Information Science (Springer). He is a member of IEEE CAS VSPC TC, MSA TC, and SPS IVMSP TC.

Xiao-Ping Zhang (Fellow, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1992 and 1996, respectively, and the M.B.A. degree (Hons.) in finance, economics, and entrepreneurship from the University of Chicago Booth School of Business, Chicago, IL, USA.

Since Fall 2000, he has been with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of the Communication and Signal Processing Applications Laboratory. He has served as the program director of graduate studies. He is cross-appointed with the Finance Department, Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent consultant for biotech companies and investment firms. His research interests include sensor networks and IoT, machine learning, statistical signal processing, image and multimedia content analysis, and applications in big data, finance, and marketing.

Dr. Zhang received the 2020 Sarwan Sahota Ryerson Distinguished Scholar Award—the Ryerson University highest honor for scholarly, research, and creative achievements. He is the General Chair for the IEEE International Workshop on Multimedia Signal Processing in 2015. He is the Publicity Chair for the International Conference on Multimedia and Expo 2006 and the Program Chair for International Conference on Intelligent Computing in 2005 and 2010. He is the General Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing in 2021, GlobalSIP Symposium on Signal and Information Processing for Finance and Business in 2017, and GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business in 2019. He served as a Guest Editor for Multimedia Tools and Applications and the International Journal of Semantic Computing. He was a Tutorial Speaker of the 2011 ACM International Conference on Multimedia, the 2013 IEEE International Symposium on Circuits and Systems, the 2013 IEEE International Conference on Image Processing, the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, the 2017 International Joint Conference on Neural Networks, and the 2019 IEEE International Symposium on Circuits and Systems. He is the Editor-in-Chief for the IEEE Journal of Selected Topics in Signal Processing. He is a Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON SIGNAL PROCESSING. He served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE Signal Processing Letters. He was an Elected Member of the ICME Steering Committee. He is selected as the IEEE Distinguished Lecturer by the IEEE Signal Processing Society for the term 2020–2021, and by the IEEE Circuits and Systems Society for the term 2021–2022. He is a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada. He is a Registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society.