Estimating Heart Rate and Rhythm via 3D Motion Tracking in Depth Video

Cheng Yang *Student Member, IEEE*, Gene Cheung *Senior Member, IEEE*, Vladimir Stankovic *Senior Member, IEEE*

***Abstract*—Low-cost depth sensors, such as Microsoft Kinect, have potential for non-intrusive, non-contact health monitoring that is robust to ambient lighting conditions. However, captured depth images typically suffer from low bit-depth and high acquisition noise, and hence processing them to estimate biometrics is difficult. In this paper, we propose to capture depth video of a human subject using Kinect 2.0 to estimate his/her heart rate and rhythm (regularity); as blood is pumped from the heart to circulate through the head, tiny oscillatory head motion due to Newtonian mechanics can be detected for periodicity analysis. Specifically, we first restore a captured depth video via a joint bit-depth enhancement / denoising procedure, using a graph-signal smoothness prior for regularization. Second, we track an automatically detected head region throughout the depth video to deduce 3D motion vectors. The detected vectors are fed back to the depth restoration module in a loop to ensure that the motion information in two modules are consistent, improving performance of both restoration and motion tracking in the process. Third, the computed 3D motion vectors are projected onto its principal component for 1D signal analysis, composed of trend removal, band-pass filtering, and wavelet-based motion denoising. Finally, the heart rate is estimated via Welch power spectrum analysis, and the heart rhythm is computed via peak detection. Experimental results show accurate estimation of the heart rate and rhythm using our proposed algorithm as compared to rate and rhythm estimated by a portable oximeter.**

I. Introduction

As the general population ages, cheap and non-invasive health monitoring has become essential. Among many health monitoring systems available on the market are image-based systems [2], [3], with the distinct advantage of being completely non-contact and thus non-intrusive. Further, unlike passive sensors (*e.g.*, conventional RGB cameras), depth sensors (*e.g.*, Microsoft Kinect) acquire depth images—per-pixel distance between physical objects in the 3D scene and the sensing device—by actively projecting infrared rays into the scene and observing the feedback, and thus are robust to ambient lighting conditions. Previous depth-image-based systems [4]–[6] have demonstrated that certain biometrics like respiratory rate can be accurately estimated for sleep apnoea detection (temporary suspension of breathing). However, due to the limitations of today's depth sensing technologies, captured depth videos typically suffer from low bit-depth (*e.g.*, Kinect 2.0 has bit-depth of 13 bits for each captured depth pixel) and acquisition noise. Thus it is difficult to process acquired depth images to estimate biometrics that require tracking of subtle 3D motion of a human subject.

In this paper, we strive to overcome this difficulty and propose to capture depth video of a human subject using Kinect 2.0 to estimate his/her heart rate and rhythm (regularity of heart beats over time [7]). As blood is pumped from the heart to the head for circulation, the head will oscillate slightly due to Newtonian mechanics (typically 5mm or less), and tracking this tiny oscillatory movement can lead to an estimate of heart rate and rhythm [8]. Unlike previously used high-resolution color video [8], the key challenge using depth video is to overcome the low bit-depth and acquisition noise inherent in the observed data.

Towards this goal, we propose to first restore depth images via a joint bit-depth enhancement / denoising procedure, using a graph-signal smoothness prior for regularization [9], [10]. We then track an automatically detected head region throughout the depth video to deduce 3D motion vectors. The detected vectors are fed back to the depth restoration module in a loop to ensure that the motion information in the two modules are consistent, resulting in a boost in performance for both restoration and motion tracking. Third, the computed 3D motion vectors are projected onto its principal component via principal component analysis (PCA) for 1D signal analysis: trend removal, band-pass filtering and wavelet-based motion denoising. Finally, the heart rate is estimated via Welch power spectrum analysis, and the heart rhythm is computed via peak detection. Experimental results show accurate estimation of the heart rate and rhythm using our proposed algorithm as compared to rate and rhythm estimated by a portable oximeter.

The outline of the paper is as follows. We first discuss related work on Section II. We then overview our heart rate detection system in Section III. We present our depth

This work was supported in part by the JSPS Grant-in-Aid for Challenging Exploratory Research (15K12072). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 734331. This paper was presented in part at the IEEE International Conference on Multimedia and Expo, Torino, Italy, June-July 2015 [1].

C. Yang and V. Stankovic are with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XQ, UK (e-mail: {cheng.yang,vladimir.stankovic}@strath.ac.uk).

G. Cheung is with National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101–8430, Japan (e-mail: cheung@nii.ac.jp).

Digital Object Identifier 10.1109/TMM.2017.2672198

video joint bit-depth enhancement / denoising algorithms in Section IV, the tracking algorithm in Section V, and the heart rate and rhythm estimation algorithms in Section VI. We present experimental results and conclude remarks in Section VII and VIII, respectively.

II. Related Work

A. Heart Rate Estimation Systems

In 2008, Verkruysse *et al.* [11], estimated heart rate from an RGB video, recorded using a single conventional digital camera, by analyzing subtle changes in the skin color caused by blood circulation. Since then, there has been an increased interest in contact-less heart rate monitoring using imaging techniques.

In [12]–[16], a human subject is recorded, sitting still, using a conventional RGB camera, and the heart rate is extracted from the recorded video using the subtle color changes in the facial skin due to blood circulation. The approach of [13] collects the time-series of color information from any pixel in the facial region, performs temporal filtering followed by independent component analysis (ICA), and obtains the heart beat rate estimate from the frequency of the maximum power in the resulting spectrum. In [14], using bandpass filtering and localized spatial pooling, the time waveform of the electrocardiogram (ECG) signal is estimated. In [15], region of interest (ROI) in the face region is adaptively updated, making the approach robust to random head movements during recording. In [16] the pulse rate is estimated using PCA on the averaged R-,G-,B-components from a manually selected ROI in the face region. In [17], a Bayesian approach is proposed that uses mechanical, ballistocardiographic (BCG) signal, skin colour variation and head motion extracted from a video recorded by a webcam, to provide a beat-to-beat heart rate estimation.

In [18], an RGB-camera based system is proposed that is robust to head movement, eye blinking, smiling, and illumination conditions changes, based on face tracking and normalized least mean square adaptive filtering. Based on the framework of [11], [13], [15], [16], a smartphone application and a mobile service robot application is proposed in [19] and [20], respectively. All of the above approaches and another three methods [21]–[23] require high-resolution color video of the facial skin, and thus the systems are all sensitive to the ambient lighting conditions.

Thermal imaging has been used in the past for contact-less heart beat estimation (see [24]–[29] and references therein). However, a good thermal infrared sensor is far more expensive than a Kinect sensor.

In [8], [30], [31], similar to our work, the detection of subtle head oscillations in videos due to blood circulation from the heart to the head is used to measure the pulse rate. In contrast to our work, [8], [30], [31] uses color video to extract feature points, which are tracked throughout the video to deduce motion. We differ from [8], [30], [31] in that we use only depth video for analysis, which is robust to ambient lighting conditions.

B. Depth Video Health Monitoring Systems

In [32], a depth map generated by a Kinect sensor is used to estimate respiratory and heart rates. However, the system is very restrictive and impractical, requiring a subject laying supine with chest unclothed to observe the neck and thorax areas used for motion tracking. In [4], [5], [33], an MS Kinect 1.0 depth sensor is used for detecting episodes of apnoea and hypopnoea, by extracting the respiratory rate from the tracked chest and abdomen movements. The depth video of the patient sleeping is recorded in complete darkness, temporal denoising is performed to mitigate effects of temporal flickering, and support vector machine or graph-based classifiers, is then trained in [4] and [5], respectively, to detect episodes of apnea / hypopnoea. Oscillatory head movements due to heart beats are much smaller than respiratory chest movements and much harder to detect in depth videos, however, and hence the challenge in this paper.

C. Comparison with Our Previous Work

Compared to our previous proposed depth video heart rate estimation system [1], we have the following improvements and additions. First, we improve the performance of our depth video restoration module by sharing motion information with the head region tracking module. Second, we show that our system is robust to viewing angle of the human subject, including front, side and back views, even when the subject is wearing a mask. Third, in addition to heart rate estimation, we quantify the heart rhythm via appropriate peak detection.

III. System Overview

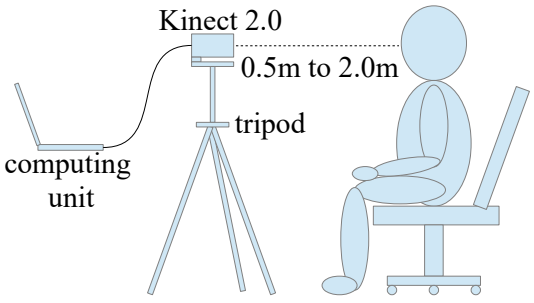


Fig. 1. System setup (a front facing example).

As shown in Fig. 1, our system is composed of a Microsoft Kinect 2.0 camera connected to a standalone laptop. Ideally, the camera is placed 0.5m to 2.0m away from the human subject. Depth video is captured at 30 frames per second (fps) at 512×424 spatial resolution and bit-depth of 13 bits per pixel.

Figs. 2 and 3 show example captured depth images of front, side and back views of a human subject, and a front view of a subject wearing a face mask. Different from our previous work [1], we will show our system is robust to viewing angle of the human subject, and

operational even if the subject wears facial coverings that obscure facial features from the camera.

(a) front view (b) side view (c) back view

Fig. 2. Examples of depth images with three viewing angles.

(a) front view with mask (b) experimental mask

Fig. 3. Example of a depth image of front view with a mask.

Algorithmically, as shown in Fig. 4, our method can be divided into three steps. The first step (orange block) is restoration of captured depth images via a joint bit-depth enhancement / denoising procedure. Denoising is necessary since it is known that depth sensors are susceptible to acquisition noise [34]. Bit-depth enhancement is necessary because the granularity of a captured depth pixel by the Kinect sensor (1.0 ± 1.5mm when capturing at the distance of 0.5m to 2.0m [35])[1] is not sufficiently fine-grained to capture subtle head motion due to heart beat (roughly 5mm [8]) without processing. The joint bit-depth enhancement / denoising optimization is discussed in Section IV.

In the second step (blue block), we specify the head region of the human subject in frame 1, using which we perform robust head region tracking in the following frames. Subsequently, we feed the tracked motion vectors back to our joint bit-depth enhancement / denoising module in a loop, so that the motion information in the two modules are consistent. The requirement to enforce consistency in motion information in the modules results in better depth image restoration quality *and* better head region tracking. The head region tracking algorithm and the motion feedback loop are discussed in Section V.

In the third step (green blocks), we project obtained 3D motion vectors along principal component via PCA, and then perform 1D analysis: i) non-linear trend removal, ii) band-pass filtering, and iii) wavelet-based motion denoising. Non-linear trend removal reduces non-stationary trends of the signal [36], [37]. As done in previous image-based heart-rate estimation systems [13], [18], band-pass filtering removes motion of frequencies outside the band of interest. Wavelet-based motion denoising futher reduces noise in the motion signal via wavelet-domain soft-thresholding [38]. The above post-processing procedures are discussed in Section VI.

Finally, we compute heart rate via Welch PSD estimation, and compute heart rhythm [7] via peak detection. Our estimated heart rate and rhythm are compared against a portable oximeter in Section VII.

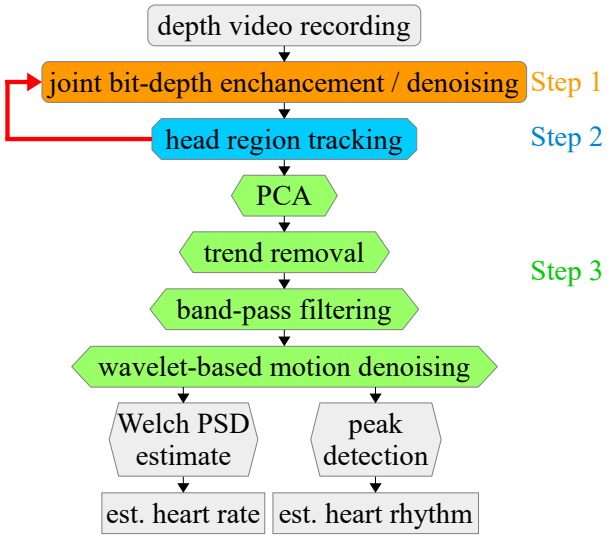


Fig. 4. Three-step system overview. 1) Orange: pre-processing component; 2) Blue: tracking component, and red arrow: the loop with motion prior feedback; 3) Green: post-processing components.

IV. Depth Video Restoration

We first restore depth video via a joint bit-depth enhancement and denoising procedure described in this section.

A. Derivation of Noise Model

We first derive a suitable noise model for Kinect 2.0 captured pixels in a depth video frame. For model derivation, we placed a flat static board on a table and recorded a depth video of T frames. Let $x^t_{i,j}$ be the depth pixel intensity at location (i, j) of frame t. For each location (i, j), we first compute the empirical mean $\mu_{i,j}$ as $\frac{1}{T}\sum_{t=1}^{T} x^t_{i,j}$, *i.e.*, the average pixel intensity value at the same location over all T frames. Given an image size of $E \times F$ pixels, we estimate the *horizontal auto-correlation* $C_h(k)$ as:

$$C_h(k) = \frac{\sigma^{-2}}{TE(F-k)} \sum_{t=1}^{T} \sum_{i=1}^{E} \sum_{j=1}^{F-k} (x^t_{i,j} - \mu_{i,j})(x^t_{i,j+k} - \mu_{i,j+k}), \quad (1)$$

where we assume that the variance σ^2 is the same for all pixel locations. One can estimate the *vertical auto-correlation* $C_v(k)$ similarly:

$$C_v(k) = \frac{\sigma^{-2}}{T(E-k)F} \sum_{t=1}^{T} \sum_{i=1}^{E-k} \sum_{j=1}^{F} (x^t_{i,j} - \mu_{i,j})(x^t_{i+k,j} - \mu_{i+k,j}). \quad (2)$$

[1]The granularity varies according to the physical distance between the subject and the camera.

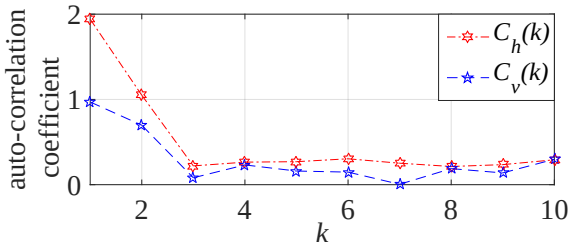


Fig. 5. Empirically computed $C_h(k)$ and $C_v(k)$ $(1 \leq k \leq 10)$ for the horizontal and vertical dimension, respectively.

Fig. 5 shows the auto-correlation plots tested on a sequence of $T = 15000$ frames computed on a flat 30×30 $(E \times F)$ square surface at a distance 0.78m from the camera. We observe that the auto-correlation in both cases decreases rapidly as k increases, which means that the correlation with immediate neighboring pixels is strong but weakens considerably thereafter. We can thus construct a suitable noise model as follows. Assuming a Gaussian Markov Random Field (GMRF) [39] noise model, the likelihood $Pr(\tilde{\mathbf{x}}|\mathbf{x})$ of observing a depth pixel patch $\tilde{\mathbf{x}}$ given the original patch is $\mathbf{x}$ is:

$$Pr(\tilde{\mathbf{x}}|\mathbf{x}) = \exp\left(-\frac{(\tilde{\mathbf{x}}-\mathbf{x})^T\mathbf{P}(\tilde{\mathbf{x}}-\mathbf{x})}{\sigma^2}\right), \tag{3}$$

where $\mathbf{P}$ is the precision matrix (inverse covariance matrix). To model neighboring pixel correlation using GMRF, we set the entries in $\mathbf{P}$ as follows [40]:

$$P_{i,j} = \begin{cases} 1/\sigma^2 & \text{if } i = j, \\ -\frac{C_h(1)}{\sigma^2} & \text{if } i \text{ and } j \text{ are horizontal neighbors,} \\ -\frac{C_v(1)}{\sigma^2} & \text{if } i \text{ and } j \text{ are vertical neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

$\mathbf{P}$ will be used in our denoising algorithm. We note that, to the best of our knowledge, Kinect 2.0 acquisition noise has not been studied formally. However, our results are consistent with those of [41] for depth image noise modelling for time-of-flight cameras.

B. Graph-signal Smoothness Prior

As in other inverse imaging problems, a signal prior for the desired signal is needed for regularization. As done in [9], [10], we employ a *graph-signal smoothness prior*; *i.e.*, a depth block $\mathbf{x}$ is piecewise smooth if the *graph Laplacian regularizer* $\mathbf{x}^T\mathbf{L}\mathbf{x}$ is small, where $\mathbf{L}$ is the *graph Laplacian* for a graph that connects neighboring pixels in block $\mathbf{x}$. Specifically, we first construct a graph $\mathcal{G}$ where the nodes in the graph correspond to pixels in block $\mathbf{x}$. We connect each node to its horizontal and vertical neighbors to yield a 4-connected graph. The edge weight $w_{i,j}$ between two nodes i and j is the exponential of their pixel intensity difference:

$$w_{i,j} = \exp\left(-\frac{|I_i - I_j|^2}{\rho^2}\right), \tag{5}$$

where I_i is the pixel intensity of pixel i and ρ^2 is a scaling parameter.

Having defined edge weights, one can define the *adjacency matrix* $\mathbf{W}$ where the (i,j)-th entry is $W_{i,j} = w_{i,j}$. The *degree matrix* $\mathbf{D}$ is a diagonal matrix where the i-th diagonal entry is $D_{i,i} = \sum_j W_{i,j}$. The graph Laplacian $\mathbf{L}$ is then defined as the difference between $\mathbf{D}$ and $\mathbf{W}$:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \tag{6}$$

It can be shown [42] that the Laplacian regularizer $\mathbf{x}^T\mathbf{L}\mathbf{x}$ is a measure of variation in the signal $\mathbf{x}$, modulated by weights $w_{i,j}$:

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_{i,j} w_{i,j}(x_i - x_j)^2. \tag{7}$$

Thus $\mathbf{x}^T\mathbf{L}\mathbf{x}$ is small if the squared signal variations $(x_i - x_j)^2$ are small *or* the modulating weights $w_{i,j}$ are small.

Since $\mathbf{L}$ is positive semi-definite, one can perform eigen-decomposition on $\mathbf{L}$ to obtain non-negative eigen-values λ_k and eigen-vectors ϕ_k. We can then express $\mathbf{x}^T\mathbf{L}\mathbf{x}$ alternatively as:

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_k \lambda_k \alpha_k^2, \tag{8}$$

where eigen-value λ_k can be interpreted as the k-th graph frequency, and $\alpha_k = \phi_k^T\mathbf{x}$ is the coefficient for the k-th graph frequency. In this interpretation, a small $\mathbf{x}^T\mathbf{L}\mathbf{x}$ means that the energy of the signal $\mathbf{x}$ is concentrated in the low graph frequencies.

C. Joint Bit-depth Enhancement / Spatial Denoising

We first discuss the procedure to perform joint bit-depth enhancement / spatial denoising for the first frame. Denote the depth values of a target block in the frame, in vector form, by $\mathbf{y}$. It is a quantized (low bit-depth) and noise-corrupted version of the original vector of depth values $\mathbf{x}$:

$$\mathbf{y} = \text{round}\left(\frac{\mathbf{x}+\mathbf{n}}{Q}\right)Q, \tag{9}$$

where Q is the quantization parameter due to coarse depth precision by the Kinect sensor, and $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is the additive noise.

The objective is to recover the original $\mathbf{x}$ given $\mathbf{y}$. Using a *maximum a posteriori* (MAP) formulation, we can derive the objective as follows. Let $\mathbf{z} = \mathbf{x} + \mathbf{n}$ be the noise corrupted signal before quantization. Using the total probability theorem, likelihood $Pr(\mathbf{y}\,|\,\mathbf{x})$ can be written as:

$$Pr(\mathbf{y}\,|\,\mathbf{x}) = \int_{\mathbf{z}} Pr(\mathbf{z}\,|\,\mathbf{x})\, Pr(\mathbf{y}\,|\,\mathbf{z},\mathbf{x})\, d\mathbf{z}. \tag{10}$$

$Pr(\mathbf{y}\,|\,\mathbf{z},\mathbf{x}) = Pr(\mathbf{y}\,|\,\mathbf{z})$ evaluates to 1 if $\mathbf{y} = \text{round}\left(\frac{\mathbf{z}}{Q}\right)Q$ and 0 otherwise. Equivalently, condition $\mathbf{y} - Q/2 \leq \mathbf{z} < \mathbf{y} + Q/2$ must be satisfied for $Pr(\mathbf{y}\,|\,\mathbf{z},\mathbf{x})$ to be non-zero. Thus, likelihood $Pr(\mathbf{y}\,|\,\mathbf{x})$ can be rewritten as:

$$Pr(\mathbf{y}\,|\,\mathbf{x}) = \int_{\mathbf{z}\in\mathcal{R}_\mathbf{y}} \exp\left[-\frac{(\mathbf{z}-\mathbf{x})^T\mathbf{P}(\mathbf{z}-\mathbf{x})}{\sigma^2}\right] d\mathbf{z}, \tag{11}$$

where $\mathbf{P}$ is the precision matrix defined in (4), σ^2 is the noise variance, and $\mathcal{R}_\mathbf{y} = \{\mathbf{z} \,|\, y_i - Q/2 \le z_i < y_i + Q/2\}$.

$Pr(\mathbf{y}\,|\,\mathbf{x})$ in the form (11) is still difficult to use. We thus approximate it as:

$$Pr(\mathbf{y}\,|\,\mathbf{x}) \propto \max_{\mathbf{z}\in\mathcal{R}_\mathbf{y}} \exp\left[-\frac{(\mathbf{z}-\mathbf{x})^T\mathbf{P}(\mathbf{z}-\mathbf{x})}{\sigma^2}\right]. \tag{12}$$

We first note that as $Q \to 0$, the area $\mathcal{R}_\mathbf{y}$ over which the integration of $\mathbf{z}$ in (11) is performed shrinks, and $Pr(\mathbf{y}|\mathbf{x})$ becomes roughly constant $Pr(\mathbf{y} = \mathbf{z}\,|\,\mathbf{x})$ over $\mathcal{R}_\mathbf{y}$. Similarly, the maximization in (12) also approaches $Pr(\mathbf{y} = \mathbf{z}\,|\,\mathbf{x})$.

Suppose now that Q is non-negligibly large. We observe that (11) and (12) have similar shapes. $Pr(\mathbf{y}\,|\,\mathbf{x})$ in (11) integrates $\mathbf{z}$ over $\mathcal{R}_\mathbf{y}$, a Q-neighborhood of $\mathbf{y}$, where the integrating exponential function is large if $\mathbf{z}$ is close to $\mathbf{x}$. Hence $Pr(\mathbf{y}\,|\,\mathbf{x})$ is large if $\mathbf{y}$ is close to $\mathbf{x}$ for given Q, or if Q is large for given $\mathbf{y}$. This is also true for (12).

1) Objective Function: Given likelihood in (12) and the graph-signal smoothness prior, one can now derive the MAP objective by minimizing the negative log of the likelihood and prior:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & (\mathbf{z}-\mathbf{x})^T\mathbf{P}(\mathbf{z}-\mathbf{x}) + \mu\, \mathbf{x}^T\mathbf{L}\,\mathbf{x} \\ \text{s.t.} \quad & y_i - \tfrac{Q}{2} \le z_i < y_i + \tfrac{Q}{2}, \quad \forall i, \end{aligned} \tag{13}$$

where μ is a parameter to trade off the first fidelity term and the second signal smoothness prior term that depends on the signal-to-noise ratio (SNR).

2) Optimization Procedure: With two inter-dependent variables $\mathbf{x}$ and $\mathbf{z}$ and a constraint on $\mathbf{z}$, the optimization (13) is difficult to solve directly. We hence propose to alternately solve for one variable while keeping the other fixed and iterate. In particular, when $\mathbf{z}$ is fixed, the optimal $\mathbf{x}$ can be solved in closed form by taking the derivative of (13) with respect to $\mathbf{x}$ and setting it to zero:

$$\mathbf{x}^* = (\mathbf{P} + \mu\mathbf{L})^{-1}\mathbf{P}\mathbf{z}. \tag{14}$$

On the other hand, when $\mathbf{x}$ is fixed, the optimal $\mathbf{z}$ to minimize the fidelity term (the graph-signal smoothness term does not involve $\mathbf{z}$) while satisfying the constraint is:

$$z_i^* = \begin{cases} y_i + Q/2 - \epsilon & \text{if} \quad x_i \ge y_i + Q/2 \\ y_i - Q/2 & \text{if} \quad x_i < y_i - Q/2 \\ x_i & \text{otherwise} \end{cases} \tag{15}$$

where ϵ is a small positive constant. The two variables are optimized alternately until the solution converges[2]. Note that the edge weights $w_{i,j}$ in the graph Laplacian $\mathbf{L}$ needs to be updated using (5) each time a new signal $\mathbf{x}$ is computed.

D. Joint Bit-depth Enhancement / Temporal Denoising

For restoration of subsequent depth frames, we formulate the following optimization problem. Denote by $\mathbf{y}_t$ an observed target block at time instant t. Denote by $\mathbf{z}_t$ and $\mathbf{x}_t$ the noise-corrupted, pre-quantized version and the restored version of $\mathbf{y}_t$, respectively. A *motion vector* (MV) $\mathbf{v}$ points to a matching block $\mathbf{x}_{t-1}(\mathbf{v})$ in the previous restored frame $t-1$ that is most similar to restored $\mathbf{x}_t$. See Fig. 6 for an illustration. The optimization thus becomes the search for MV $\mathbf{v}$ and denoised patch $\mathbf{x}_t$ that minimize three terms: i) a fidelity term with respect to observation $\mathbf{y}_t$, ii) a graph-signal smoothness term $\mathbf{x}_t^T\mathbf{L}\mathbf{x}_t$, and iii) a motion estimation term $\|\mathbf{x}_{t-1}(\mathbf{v}) - \mathbf{x}_t\|_2^2$ that measures the difference between two matching blocks in the two frames:

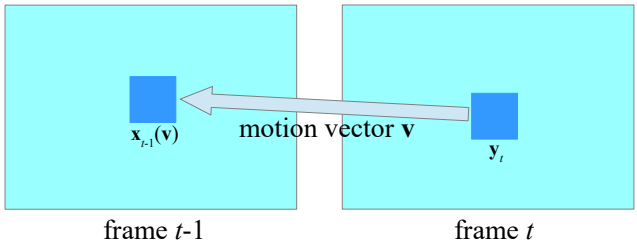


Fig. 6. Illustration of matching between blocks in current frame t and previous frame $t-1$.

$$\begin{aligned} \min_{\mathbf{v},\mathbf{z}_t,\mathbf{x}_t} \quad & (\mathbf{z}_t-\mathbf{x}_t)^T\mathbf{P}(\mathbf{z}_t-\mathbf{x}_t) + \mu\, \mathbf{x}_t^T\,\mathbf{L}\,\mathbf{x}_t + \gamma\, \|\mathbf{x}_{t-1}(\mathbf{v}) - \mathbf{x}_t\|_2^2 \\ \text{s.t.} \quad & \mathbf{y}_t - \frac{Q}{2} \le \mathbf{z}_t < \mathbf{y}_t + \frac{Q}{2}. \end{aligned} \tag{16}$$

1) Optimization Procedure: To solve (16), we use a similar alternating method as follows. We first search for the optimal $\mathbf{v}$ that minimizes the motion estimation term $\|\mathbf{x}_{t-1}(\mathbf{v}) - \mathbf{y}_t\|_2^2$. Then we fix $\mathbf{v}_t$, and alternately solve for $\mathbf{z}_t$ and $\mathbf{x}_t$, where the optimal $\mathbf{x}_t$ given $\mathbf{v}_t$ and $\mathbf{z}_t$ is:

$$\mathbf{x}_t^* = (\mathbf{P} + \mu\mathbf{L} + \gamma\,\mathbf{I})^{-1}(\mathbf{P}\mathbf{z}_t + \gamma\,\mathbf{x}_{t-1}), \tag{17}$$

where $\mathbf{I}$ is the identity matrix. The optimal $\mathbf{z}_t$ given fixed $\mathbf{x}_t$ is solved using (15).

V. Head Region Tracking

We now discuss how we select and track the ROI given restored depth frames, and then how using motion information obtained from the tracking module as a motion prior, we can further refine the depth video restoration process described in Section IV.

A. ROI Selection

Similar to others in the object tracking literature [8], [14], we first specify an ROI—a subject's head region to deduce the motion caused by heart beat—for tracking in subsequent frames. Since only part of the upper body is present in the depth frames, conventional RGB-plus-depth based human body part detection methods like [43] and depth-only based methods like [44], [45] that require most of the body to be visible do not work well. We thus adopt a template-matching based method in [46], [47] to detect a human head. First, Canny edge detection [48] is applied on the restored depth frame, followed by scale-invariant Chamfer matching [49] with

[2]One can easily prove that the alternating algorithm converges: at each step the objective in (13) is decreased, and the objective—a sum of two quadratic terms—is lower-bounded by 0.

a pyramid of binary head-templates, which returns the potential locations of the human subject. Next, a circular region is extracted around each detected location and fit with a hemisphere model [46] to locate the probable head position. Finally, for simplicity, we designate a square area $\boldsymbol{p}$ of size $K \times K$ pixels centered inside the hemisphere as our ROI for tracking. See Fig. 7 for an illustration.

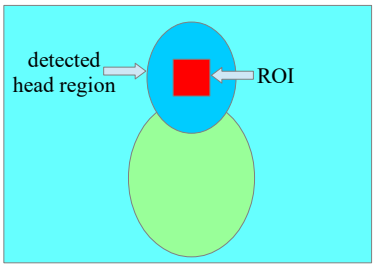


Fig. 7. ROI centered inside a hemisphere model [46] that detects the head of the human subject.

B. Head Region Tracking

Once the ROI is selected, we adopt *kernelized correlation filter* (KCF) [50], [51], one of the state-of-the-art computationally efficient and accurate trackers [52], [53], for tracking ROI in subsequent depth video frames.

We first briefly review KCF [51]. KCF tracks a target based on extracted features from the selected $K \times K$ ROI $\boldsymbol{p}_t$ in frame t. It contains the following three steps: model construction, target detection, and model update. Given a tracking region (a region that is larger than ROI to provide negative samples [51]) $\boldsymbol{q}_t$ of size $L \times L$ in frame t, where $\boldsymbol{q}_t$ has the same centre coordinate as $\boldsymbol{p}_t$, and a feature matrix $\hat{\boldsymbol{q}}_t$ extracted from $\boldsymbol{q}_t$, KCF first constructs a target model based on kernel correlation [51] of $\hat{\boldsymbol{q}}_t$ with itself and kernel regression [54] that minimizes the squared error over $\hat{\boldsymbol{q}}_t$ and a pre-defined regression target with Gaussian distribution. Then, an $L \times L$ response map $\boldsymbol{\varpi}_t$ is computed based on the constructed target model and kernel correlation of $\hat{\boldsymbol{q}}_t$ with a corresponding feature matrix $\hat{\boldsymbol{q}}_{t+1}$ extracted from frame $t+1$. The coordinates within the tracking region $\boldsymbol{q}_{t+1}$ in frame $t+1$ are ranked based on the response map, and the one for which the response map reaches maximum $\max\{\boldsymbol{\varpi}_t\}$ is identified as the centre of the target. Finally, the model and feature matrices are updated based on linear interpolation from the previous frame to the following frames.

In Section VII, we discuss our choice of the kernel and features used for KCF tracking. The above KCF tracking process returns the 2D image coordinates of the target centre and the maximums $\max\{\boldsymbol{\varpi}_t\}$'s of the response map in the T frames.

C. Enforcing Consistency in Motion Information

We next improve the depth video quality and tracking accuracy by enforcing consistency in motion information in the depth video restoration and tracking modules with a *motion prior feedback loop*. We do this by first reformulating our depth video restoration problem, *i.e.*, joint bit-depth enhancement / denoising objective in Section IV, with the deduced 2D motion vectors for each frame $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_T]$ from the 2D image coordinates of the target centre in the T frames in Section V-B, where we treat $\mathbf{U}$ as the motion prior. In particular, we add a motion prior term in (16):

$$\begin{aligned} \min_{\mathbf{v},\mathbf{z}_t,\mathbf{x}_t} \quad & (\mathbf{z}_t - \mathbf{x}_t)^T \mathbf{P} (\mathbf{z}_t - \mathbf{x}_t) + \mu\, \mathbf{x}_t^T \mathbf{L}\, \mathbf{x}_t \\ & + \gamma\, \|\mathbf{x}_{t-1}(\mathbf{v}) - \mathbf{x}_t\|_2^2 + \beta_t\, \|\mathbf{v} - \mathbf{u}_t\|_2^2 \\ \text{s.t.} \quad & \mathbf{y}_t - \tfrac{Q}{2} \le \mathbf{z}_t < \mathbf{y}_t + \tfrac{Q}{2}, \end{aligned} \tag{18}$$

where $\beta_t = \kappa \cdot \max\{\boldsymbol{\varpi}_t\}$. We can interpret $\max\{\boldsymbol{\varpi}_t\}$ as the tracker's confidence in the estimated motion vector $\mathbf{u}_t$. κ is a scaling constant.

To solve (18), we first search for the optimal $\mathbf{v}^*$ that minimizes the sum of the motion estimation and motion prior terms:

$$\min_{\mathbf{v}} \quad \gamma\, \|\mathbf{x}_{t-1}(\mathbf{v}) - \mathbf{x}_t\|_2^2 + \beta_t\, \|\mathbf{v} - \mathbf{u}_t\|_2^2. \tag{19}$$

Instead of an exhaustive search, for fast implementation we assume simply that the optimal solution $\mathbf{v}^*$ to (19) is a convex combination of $\mathbf{u}_t$ and previously computed $\mathbf{v}'$ from solving (16) without the motion prior before entering the feedback loop, *i.e.*, $\mathbf{v}^* = \zeta \mathbf{u}_t + (1-\zeta)\mathbf{v}'$, where $0 \le \zeta \le 1$. We argue that the above assumption is reasonable because $\mathbf{v} = \mathbf{u}_t$ would minimize the motion prior term in (19), while $\mathbf{v} = \mathbf{v}'$ would minimize the motion estimation term, so a convex combination of these two solutions would in general lead to a better one. The optimal value ζ^* can be easily found via binary search.

Then, we fix $\mathbf{v}^*$, and alternately solve for $\mathbf{z}_t$ and $\mathbf{x}_t$, where the optimal $\mathbf{x}_t$ given $\mathbf{v}^*$ and $\mathbf{z}_t$ is (17), and the optimal $\mathbf{z}_t$ given fixed $\mathbf{x}_t$ is solved using (15). Next, we re-apply KCF tracking described in Section V-B using the re-restored depth video computed using the above procedure. The above motion prior feedback loop gives us the updated 2D image coordinates of the target centre in the T frames, which could be iteratively used again for joint bit-depth enhancement / denoising. Empirically, we found that there is no substantial improvement after two iterations. Finally, we perform mapping of the target centres with their depth intensities from the image space to the 3D space (real-world space), using Kinect for Windows 2.0 software development kit [55], and deduce the corresponding 3D motion $\boldsymbol{\Delta} = [\Delta_1, ..., \Delta_T]$. We next discuss how to use $\boldsymbol{\Delta}$ for estimating heart rate and rhythm as discussed in Section VI.

VI. Estimating Heart Rate and Rhythm

Given the deduced 3D motion $\boldsymbol{\Delta}$, in this section, we first discuss how we project $\boldsymbol{\Delta}$ along the principal component via PCA, then perform 1D analysis, including trend removal, band-pass filtering, and wavelet-based motion denoising. We then estimate heart rate using Welch PSD estimation, and estimate heart rhythm with peak detection.

A. PCA and 1D Signal Analysis

Note that when the subject is facing the camera, it is reasonable to discard the horizontal component in $\mathbf{\Delta}$ before PCA, as done in [1] and [8], since the horizontal component contains most of body-balancing movement unrelated to heart rate [8]. However, in this paper, since we are addressing the more challenging problem of estimating heart rate from heterogeneous viewing angles of the human subject with respect to the camera, we keep all three components of $\mathbf{\Delta}$. This means that we have to pro-actively remove the body-balancing component from the PCA-projected 1D signal in a separate step before actual heart rate estimation.

PCA [56], [57] performs eigen-decomposition to determine three eigenvectors of $\mathbf{\Delta}$ and arranges them in descending order of the magnitude of the corresponding three eigenvalues. We project $\mathbf{\Delta}$ onto each eigenvector separately, resulting in three 1D projected trajectories $\mathbf{S}_i$. Next, we follow [8] and choose the most periodic among $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$, by finding for each $\mathbf{S}_i$ the frequency with maximum power and calculating the ratio of the sum of the spectral power at this frequency and its k following harmonics over the total spectral power, and choosing $\mathbf{S}_i$ that gives the highest ratio.

Denote by $\mathbf{S}$ the most periodic 1D projected trajectory. $\mathbf{S}$ contains various irrelevant movements due to respiratory motion, blinking, and body-balancing motion as described earlier. We adopt the following three modules to reduce the effect of these irrelevant movements. First, similar to [18], [37], we remove the non-linear trend (due to low-frequency movements) in $\mathbf{S}$ by fitting a 9th-order polynomial to $\mathbf{S}$ and subtracting the polynomial. Second, we pass the trend-removed signal to a 5th-order Butterworth [58] band-pass filter to isolate a normal-heart-rate frequency band of interest, which we choose to be 0.7Hz to 4Hz [18].

Third, we perform wavelet decomposition [59], [60] with an s-sec sliding window and e-sec overlapping on the band-pass filtered signal. Specifically, we choose Daubechies orthogonal wavelets that are optimal in the sense that they have a minimum support size for a given number w of vanishing moments [38]. We then heuristically choose Daubechies $w = 4$ wavelets which balances the regularity of the singularities and support size [38]. We decompose the signal into 4 levels given the 30 Hz depth image frame rate: Level-4 [7.5,15]Hz, Level-3 [3.75,7.5]Hz, Level-2 [1.875,3.75]Hz, Level-1 [0.9375,1.875]Hz, and detail [0,0.9375]Hz. Next, we perform wavelet denoising with minimax soft thresholding and multiplicative threshold rescaling [61] followed by wavelet reconstruction.

B. Heart Rate Estimation

Let $\boldsymbol{r}$ be the wavelet reconstructed signal at Level-1, which contains most of the signal of interest. We perform Welch PSD estimation [62] on $\boldsymbol{r}$, with a s_l-sample segment length, s_o-sample overlapping, and s_d DFT points. The frequency f_h with the maximum PSD is designated as the heart-beat frequency, and we estimate the heart rate as $60 \times \mathrm{f}_h$ beats/minute.

C. Heart Rhythm Estimation

Recall that heart rhythm is another essential measure to assess one's health condition [7]. We estimate heart rhythm using peak detection on $\boldsymbol{r}$. This first requires appropriately setting the following peak detection parameters: minimum distance between two neighbouring peaks, denoted as $\mathtt{d}$, minimum height of each peak, denoted as $\mathtt{h}$, and minimum numerical drop on both sides of each peak, *i.e.*, minimum prominence, denoted as $\mathtt{p}$. We clarify these parameter settings in Section VII. Given 30 depth image frame rate, the heart rhythm is then estimated as $\psi/30$ seconds, where ψ denotes the standard deviation of the distances between each two neighbouring peaks within an s-sec sliding window.

VII. Experimentation

A. Experimental Setup

We recorded depth videos of length 44.1~124.6-sec, of 13 healthy volunteers (between 21 and 37 years of age), who were sitting still, 0.64~1.3m away from the camera. Our experimental setup is shown in Fig. 8, which also shows a finger pulse oximeter (ANAPULSE ANP100, Ana Wiz Ltd, UK), used only for collecting ground truth data. We collected three depth videos for each sitting volunteer—front, side and back views captured by the camera separately in three different trials (see Fig. 2). Three of these volunteers performed one trial with front view while wearing a mask (see Fig. 3). Thus, we performed 42 trials in total. Mean heart rates, measured by the oximeter, varied from 66 to 87bpm, as shown in Fig. 9. The proposed heart rate and rhythm estimation algorithm is implemented in Matlab R2015b on a laptop running Windows 10, with Core i7 4600U 2.1GHz CPU and 8GB RAM. The mean computational time is 0.783s per frame.

First, we show in Fig. 10 a sample of restored images using our proposed joint bit-depth enhancement / denoising scheme and motion prior feedback loop. Compared to Fig. 10(a), the restored image block in Fig. 10(b) obtained using our proposed joint bit-depth enhancement / denoising scheme presents much less noise while preserving sharp edges. After enforcing the consistency in motion information in the restored depth video, the final output image block in Fig. 10(c) shows further improvement in terms of noise level and edge sharpness.

Next, in Figs. 11 and 12, we show an example of head detection result for ROI selection before head motion tracking (see Section V-A). Specifically, we find that it is sufficient to use an ellipse as the binary head template for Chamfer matching (see Fig. 11(a)). We show in Fig. 12 an example of Chamfer matching result in front, side and back views, where the rectangles in green denote

the matched head candidate blocks. We also find that it is more robust to use a hemi-ellipsoid model (see Fig. 11(b)) instead of a hemisphere model [46] to locate the probable head position based on the Chamfer matching result. The final detected head blocks are highlighted by rectangles in red in Fig. 12.

Next, we justify the kernel and features adopted for tracking. Note that, to the best of our knowledge, the majority of state-of-the-art features used for tracking [63]–[67] have only been adopted in texture images. We thus use four candidate features that have already been used in depth images [68]–[74], namely, Haar-like [68], histogram of oriented gradients (HOG) [69], [73], [74], histogram of oriented normal vectors (HONV) [70], and local depth pattern (LDP) [72], together with three low-level features, namely, raw pixels, gray level (raw pixels after subtracting the mean), and histogram of the raw pixels, for accuracy testing in terms of the heart rate estimation. We test the above seven candidate features using KCF tracker, with three different kernels, namely, Gaussian, polynomial, and linear kernels, on 3 randomly selected experimental video sequences. Fig. 13 shows the root mean square error (RMSE) of the estimated heart rate by using all combinations of the candidate features and kernels for our tracking process. It can be seen from the figure that the combination of gray level feature with Gaussian kernel performs best. Thus, this combination is used in our tracking component for all video sequences in the sequel.

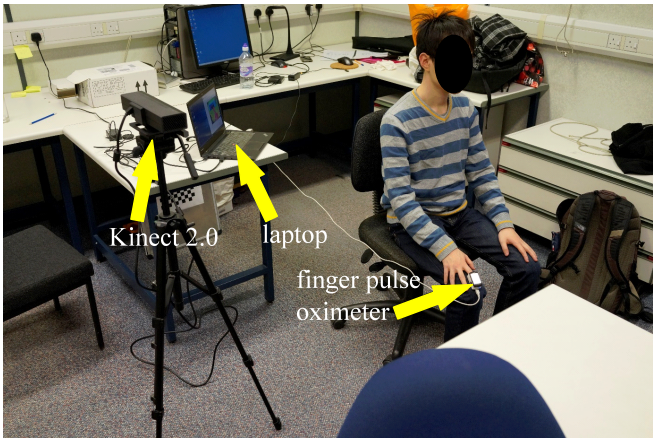


Fig. 8. Experimental setup (a side facing example).

In the following two subsections, we present estimation results of the heart rate and rhythm using our proposed scheme, and five competing schemes that are simplified versions of our proposed scheme: scheme 1) without joint bit-depth enhancement / denoising, the motion prior feedback loop and wavelet-based motion denoising, 2) without the motion prior feedback loop and wavelet-based motion denoising, 3) without wavelet-based motion denoising, 4) without joint bit-depth enhancement / denoising and the motion prior feedback loop, and 5) without the motion prior feedback loop. These schemes are denoted as Schemes A-E, respectively. Additionally, we compare our proposed method to motion-based schemes from [8], [30] and [31]. To the best of our knowledge, there are no other contemporary works that estimate heart rate using depth image sequences only. Although [8], [30], [31] have only been tested for colour image sequences, we test these schemes for depth image sequences, since all these remote heart rate estimation system (and the proposed) rely on a feature-based tracking component to acquire the motion trajectories. The main differences among these three motion-based competing schemes are as follows: [8], [30] and [31] adopt a Lucas-Kanade tracking scheme, where [8] and [30] focus on tracking multiple feature points while keeping only the vertical component of the head motion and [31] single feature point while keeping both the horizontal and vertical head motion components. [30] adopts a moving average filter for motion smoothing, and applies discrete cosine transform for heart rate estimation, unlike [8] and [30]

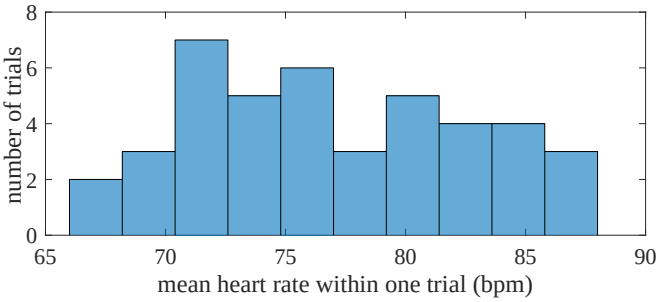
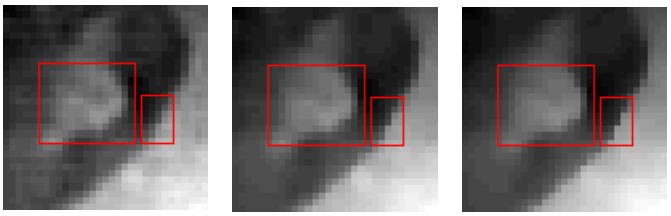
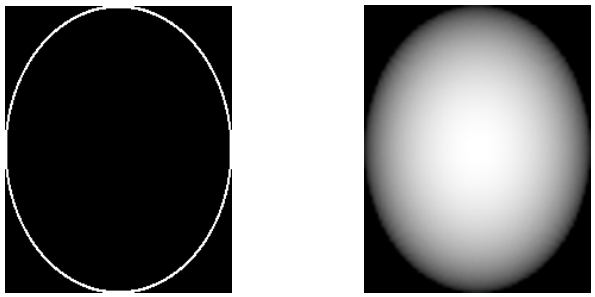


Fig. 9. Mean heart rates (from the oximeter) of all trials.



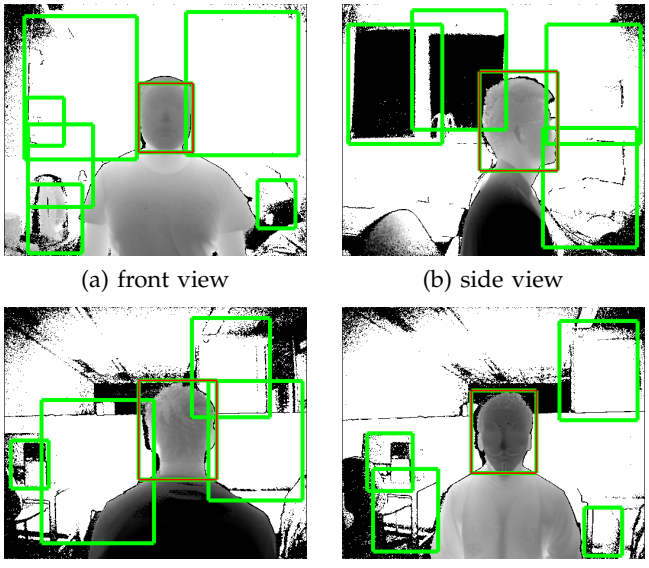
(a)raw depth (b)JBDED (c)JBDED+MPFL

Fig. 10. A sample result of joint bit-depth enhancement / denoising (JBDED) and motion prior feedback loop (MPFL).



(a) template for CM [47] (b) template for head fitting

Fig. 11. Templates used for head detection. CM = chamfer matching.



(a) front view (b) side view

(a) back view (b) front view with mask

Fig. 12. An example of head detection result. Rectangles in green: head candidate blocks using chamfer matching [47]; Rectangles in red: final head fitting result.

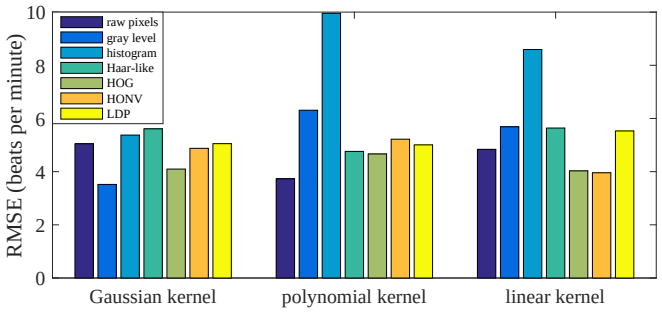


Fig. 13. RMSE (bpm) of the estimated heart rate by using all combinations of the candidate features and kernels for our tracking process, based on randomly selected 3 out of all 42 trials, where the combination of gray level feature and Gaussian kernel is chosen for ROI tracking.

that use fast Fourier transform.

B. Estimating Heart Rate

Table I lists the parameter settings for estimation of the heart rate, obtained empirically, including joint bit-depth enhancement / denoising $(\sigma, \sigma_I, Q, \mu, \gamma)$, tracking (K, L), the motion prior feedback loop (κ), PCA (k), wavelet-based motion denoising (s, e), and Welch PSD estimate (s_l, s_o, s_d).

We first investigate the performance of our proposed scheme that consists of wavelet-based motion deonising and a competing scheme without such component (Scheme C). We do this by comparing our proposed scheme with Scheme C in terms of the RMSE of estimated heart rate, as shown in Fig. 14. Each marker denotes the RMSE of the estimated heart rate during one trial. It can be seen that there is a significant RMSE reduction by adding wavelet-based motion denoising, indicating that the amount of the noise in the motion

TABLE I
PARAMETER SETTINGS FOR ESTIMATING HEART RATE.

sign	parameter	setting
σ	std. of the noise at an arbitrary pixel	1
ρ	scaling factor of the edge weight in a graph	6
Q	quantization factor of the sensor	1
μ	regularization factor for graph smoothness	6
γ	regularization factor for motion estimation	1
K	width and height of ROI / denoising block	32
L	width and height of tracking region	64
κ	scaling factor for motion prior	0.5
k	number of harmonics used for choosing $\mathbf{S}_i$	3
s	sliding window (second) for wavelet denoising	10
e	overlapping (second) for wavelet denoising	9.967
s_l	segment length (sample) for Welch PSD estimate	90
s_o	overlapping (sample) for Welch PSD estimate	45
s_d	DFT points for Welch PSD estimate	300

signal, *e.g.*, irrelevant movement such as respiratory and body-balancing movements, is effectively reduced with wavelet-based motion denoising.

Next, given the fact that wavelet-based motion denoising helps improve the result, we investigate the performance of our proposed scheme with respect to two competing schemes, Scheme D, which does not use joint bit-depth enhancement and denoising and Scheme E, which does not use motion prior feedback loop. We do this by comparing our proposed scheme with Schemes D and E in terms of the estimated heart rate over time from a front view trial, as shown in Fig. 15. It can be seen that 1) our proposed joint bit-enhancement / denoising module effectively reduces the noise from the Kinect 2.0 sensor, and 2) our proposed loop with motion prior feedback further improves the performance of joint bit-enhancement / denoising.

In Table II, we present numerical results for our proposed scheme and all its five simplified versions, mean and variance of the computed RMSE for each view from all 13 subjects, *i.e.*, front, side and back views, and front view with a mask. It can be seen that 1) our proposed scheme is robust with respect to viewing angle, which indicates that the system can also handle the cases when the subject is facing the camera with arbitrary angles, and 2) all the following three components, namely, joint bit-depth enhancement / denoising (JBDED), motion prior feedback loop (MPFL), and wavelet-based motion denoising (WBMD), are essential for minimizing the noise from the sensor and motion in order to accurately estimate the heart rate.

Note that, 3 out of all 13 subjects that participated in the trials with a mask, whereas all 13 subjects participated in the other three types of trials (i.e., front, side, and back view without a mask). That is, the number of subjects participated in each type of trials is generally small. Although this is a probable cause for a larger variance with a smaller mean for the result of the trials with a mask using our proposed method as shown in Table II, the small difference is not of statistical significance.

In Table III, we show the mean percentage error of the heart rate estimation using the proposed method, its five

TABLE II
MEAN AND VARIANCE OF THE RMSE (BPM) OF THE ESTIMATED HEART RATE USING SCHEMES A-E AND PROPOSED SCHEME.

scheme	A		B		C		D		E		**Proposed**	
joint bit-depth enhancement / denoising			✓		✓				✓		✓	
motion prior feedback loop					✓						✓	
wavelet-based motion denosing							✓		✓		✓	
metric	mean	VAR	mean	VAR	mean	VAR	mean	VAR	mean	VAR	mean	VAR
front	17.09	25.86	15.27	25.76	13.81	18.75	10.02	8.36	7.26	5.06	**5.82**	**3.72**
side	16.87	14.79	16.57	35.42	18.19	47.27	9.91	13.95	**6.74**	4.28	6.83	**4.11**
back	16.49	26.37	17.01	30.70	16.60	40.29	9.12	12.86	8.44	8.33	**6.36**	**3.61**
mask	11.47	35.30	13.02	62.98	11.19	50.61	7.19	12.80	7.55	18.96	**5.15**	**3.81**
avg	15.48	25.58	15.47	38.71	14.95	39.23	9.06	12.00	7.50	9.16	**6.04**	**3.81**

TABLE III
MEAN PERCENTAGE ERROR (%) OF THE ESTIMATED HEART RATE USING ALL SCHEMES.

scheme	A	B	C	D	E	**Proposed**	[8]	[30]	[31]
front	18.79	16.63	15.26	10.54	7.72	**6.02**	23.59	24.94	24.10
side	19.01	19.65	22.17	11.13	**7.58**	7.66	27.26	28.29	20.28
back	18.14	18.43	18.23	10.31	8.97	**6.96**	30.34	31.85	26.86
mask	14.25	16.70	13.92	9.25	8.45	**5.77**	15.48	15.41	16.17
avg	17.55	17.85	17.40	10.31	8.18	**6.60**	24.17	25.12	21.85

simplified versions, and the three competing schemes [8], [30] and [31]. We can conclude that [8], [30] and [31] perform significantly worse than our proposed method, it is essential to retain the head motion in all three dimensions (horizontal, vertical, and axial directions) for motion analysis, and all of the proposed JBDED, MPFL and WBMD system components are essential for minimization of the sensor and motion noise for heart rate estimation.

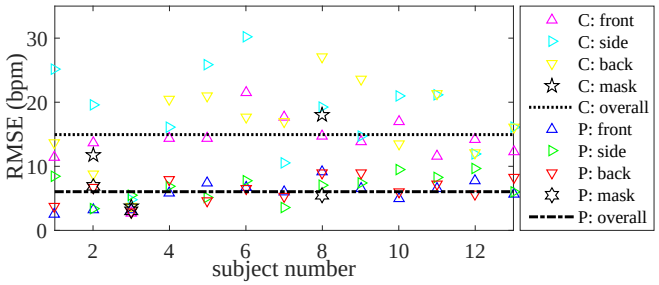


Fig. 14. RMSE of the estimated heart rate using Scheme C (no wavelet denoising) and proposed scheme, *i.e.* for all 13 subjects. The overall average for Scheme C is drawn with a dotted line, and proposed scheme with a dot-dash line.

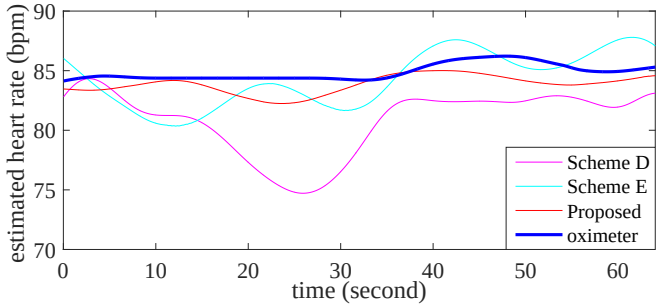


Fig. 15. Estimated heart rate over time from a front view trial by using proposed scheme and Schemes D and E.

C. Estimating Heart Rhythm

For heart rhythm estimation, we first appropriately extract the peaks in each 10-sec sliding window by peak detection. Recall that the depth image frame rate is 30 fps and the maximum frequency of the normal heart rate is 4 Hz (240 bpm) [18], *i.e.*, the minimum distance between each two neighbouring peaks is $\lceil 30/4 \rceil = 8$ samples. Table IV lists the parameter settings of peak detection for estimating heart rhythm. The heart rhythm is subsequently estimated by computing the standard deviation ψ of the distance between each two neighbouring peaks within a 10-sec sliding window, then averaging within a whole trial, and finally divided by 30 (fps).

TABLE IV
PARAMETER SETTINGS OF PEAK DETECTION FOR ESTIMATING HEART RHYTHM.

sign	parameter	setting
`d`	minimum distance (sample) between two peaks	8
`h`	minimum height (amplitude) of each peak	0
`p`	minimum prominence of each peak	0.05

We first investigate the performance of the proposed scheme with respect to two competing schemes, which are the same as the proposed scheme except: one competing scheme that does not use joint bit-depth enhancement nor motion prior feedback loop (Scheme D), and the other only drops motion prior feedback loop (Scheme E). Fig. 16 shows the results in terms of the estimated heart rhythm over time from a back view trial. Similarly to the results presented in Fig. 15, both joint bit-depth enhancement / denoising and motion prior feedback loop are essential before heart rhythm estimation.

Next, in Table V, we present the estimated heart rhythm for each view from all 13 subjects, using the proposed method, its five simplified versions, and the three competing schemes [8], [30] and [31]. It can be seen from the table that the proposed scheme outperforms all benchmark schemes for all cases except the trial with the mask, when Scheme D is slightly better.

TABLE V
ESTIMATED HEART RHYTHM (SECOND) USING ALL SCHEMES.

scheme	A	B	C	D	E	**Proposed**	[8]	[30]	[31]	oximeter
front	0.2713	0.2787	0.2887	0.1808	0.1751	**0.1722**	0.2069	0.2312	0.2107	0.0596
side	0.2866	0.3010	0.3016	0.1795	0.1846	**0.1793**	0.3021	0.3041	0.2259	0.0494
back	0.2723	0.2861	0.2826	0.1798	0.1748	**0.1723**	0.3433	0.3651	0.2209	0.0582
mask	0.2911	0.3143	0.3072	**0.1808**	0.1877	0.1864	0.2475	0.2801	0.3182	0.0477
avg	0.2803	0.2950	0.2950	0.1802	0.1805	**0.1776**	0.2750	0.2951	0.2439	0.0537

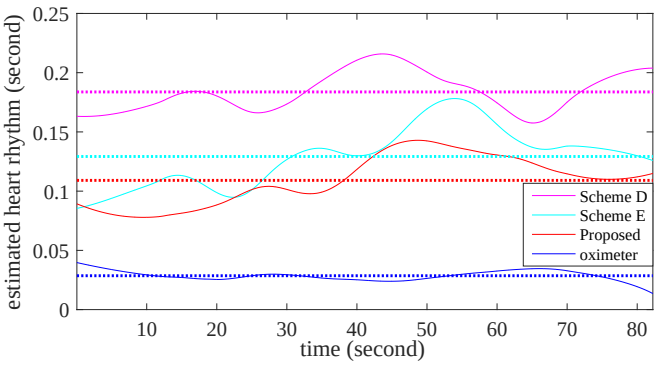


Fig. 16. Estimated heart rhythm over time from a back view trial by using proposed scheme and Schemes D and E, *i.e.*, three schemes with wavelet-based motion denoising. Mean of the estimated heart rhythm for each scheme is drawn with a dash line.

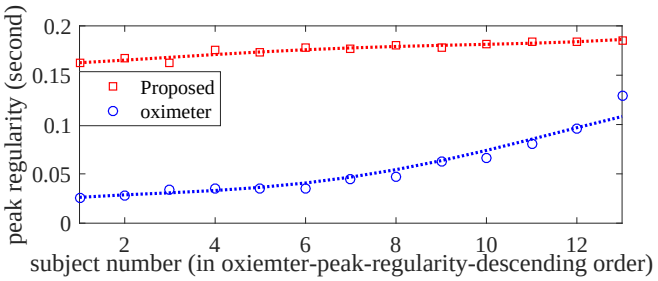


Fig. 17. Estimated heart rhythm using proposed scheme in ascending order of the heart rhythm from the oximeter. Polynomial regressions are drawn with dash lines.

Finally, we investigate the ability of our proposed scheme in distinguishing between the subjects with various heart rhythms and detecting abnormal heart rhythm. We do this by computing the mean of the estimated heart rhythm for each subject, and list the results in the ascending order of the heart rhythm means obtained with the oximeter, as shown in Fig. 17. In particular, we adopt polynomial regression by a 4-th order polynomial curve fitting. One can see that the non-linear trend of the fitted polynomial for our proposed scheme follows that for the oximeter. We also compute the Pearson correlation [75] by $\vartheta_{mn} = \sum_{i=1}^{13}(m_i - \bar{m})(n_i - \bar{n}) / \sqrt{\sum_{i=1}^{13}(m_i - \bar{m})^2 \sum_{i=1}^{13}(n_i - \bar{n})^2} = 0.7864$, where m and n denote the average of the estimated heart rhythm for each subject from our scheme and the oximeter, respectively, in the ascending order of the heart rhythm from the oximeter; $\bar{m} = \frac{1}{13}\sum_{i=1}^{13} m_i, \bar{n} = \frac{1}{13}\sum_{i=1}^{13} n_i$. The above facts show that, although the error of the proposed scheme is often large, the scheme has potential to correctly distinguish the subjects with various heart rhythms and detect abnormal heart rhythm.

VIII. CONCLUSION

In this paper, we propose a non-intrusive three-step heart rate and rhythm estimation system via 3D motion tracking in depth video. First, we restore the low-bit-depth, noise-corrupted depth images via a joint bit-depth enhancement / denoising procedure, using a graph-signal smoothness prior for regularization. Second, we track an automatically detected head region throughout the depth video to deduce 3D motion vectors. The detected vectors are fed back to the depth restoration module in a loop to ensure that the motion information in the two modules are consistent, resulting in a boost in performance for both restoration and motion tracking. Third, we project the computed 3D motion vectors onto its principal component via PCA for 1D signal analysis: trend removal, band-pass filtering and wavelet-based motion denoising. Finally, we estimate the heart rate via Welch power spectrum analysis, and estimate the heart rhythm via peak detection. Experimental results show robustness to different views, and accurate estimation of the heart rate and rhythm using our proposed algorithm compared to the values estimated by a portable finger pulse oximeter. Unlike conventional texture (RGB and grayscale) based methods that require the subject to face the camera for reasonable measurements, our proposed scheme estimates the heart rate and rhythm accurately with various views, such as front, side and back views, and even when the subject is wearing a mask.

We note that, since our system relies on tracking subtle head movements, it is sensitive to any types of movements, *e.g.*, slouching or tightening the abdominal and back muscles when feeling uncomfortable with sitting still, smiling, etc. Also, our system would not perform well when the subject is leaning his/her head on a chair or sleep in bed, where the subtle motion is too small to detect. However, this could be addressed by combinations of the features extracted from both depth and infrared images, where both motion and texture based features could be used for more robust tracking and estimation of the heart rate and rhythm.

REFERENCES

[1] C. Yang, G. Cheung, and V. Stankovic, "Estimating heart rate via depth video motion tracking," in *IEEE International Conference on Multimedia & Expo*, Turin, Italy, Jun. 2015.

[2] F. Erden, S. Velipasalar, A. Z. Alkar, and A. E. Cetin, "Sensors in assisted living: A survey of signal and image processing methods," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 36–44, Mar. 2016.

[3] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos, "A survey on ambient intelligence in healthcare," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2470–2494, Dec. 2013.

[4] C. Yang, G. Cheung, K. Chan, and V. Stankovic, "Sleep monitoring via depth video recording & analysis," in *IEEE International Workshop on Hot Topics in 3D*, Chengdu, China, Jul. 2014.

[5] C. Yang, Y. Mao, G. Cheung, V. Stankovic, and K. Chan, "Graph-based depth video denoising and event detection for sleep monitoring," in *IEEE International Workshop on Multimedia Signal Processing*, Jakarta, Indonesia, Sept. 2014.

[6] C. Yang, G. Cheung, V. Stankovic, K. Chan, and N. Ono, "Sleep apnea detection via depth video audio feature learning," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2016.

[7] B. Zaret, L. Cohen, M. Moser, and Yale University. School of Medicine, *Yale University School of Medicine Heart Book*, 1st ed. NY: William Morrow and Company, Mar. 1992.

[8] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, Jun. 2013.

[9] J. Pang, G. Cheung, W. Hu, and O. C. Au, "Redefining self-similarity in natural images for denoising using graph signal gradient," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Siem Reap, Cambodia, Dec. 2014.

[10] J. Pang, G. Cheung, A. Ortega, and O. C. Au, "Optimal graph Laplacian regularization for natural image denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, Apr. 2015.

[11] W. Verkruysse, L. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21 434–21 445, Dec. 2008.

[12] K. Y. Lin, D. Y. Chen, and W. J. Tsai, "Face-based heart rate signal decomposition and evaluation using multiple linear regression," *IEEE Sensors Journal*, vol. 16, no. 5, pp. 1351–1360, Mar. 2016.

[13] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, Jul. 2010.

[14] H. Y. Wu et al., "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, Jul. 2012.

[15] H. E. Tasli, A. Gudi, and M. Uyl, "Remote PPG based vital sign measurement using adaptive facial regions," in *IEEE International Conference on Image Processing*, Paris, France, Oct. 2014.

[16] M. Lewandowska et al., "Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity," in *Federated Conference on Computer Science and Information Systems*, Szczecin, Poland, Sept. 2011.

[17] C. H. Antink et al., "Beat-to-beat heart rate estimation fusing multimodal video and sensor data," *Biomedical Optics Express*, vol. 6, no. 8, pp. 2895–2907, Aug. 2015.

[18] X. Li et al., "Remote heart rate measurement from face videos under realistic situations," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, Jun. 2014.

[19] S. Kwon et al., "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Diego, CA, Aug. 2012.

[20] R. Stricker et al., "Non-contact video-based pulse rate measurement on a mobile service robot," in *IEEE Int. Symp. Robot and Human Interactive Commun.*, Edinburgh, UK, Aug. 2014.

[21] G. Tabak and A. C. Singer, "Non-contact heart rate detection via periodic signal detection methods," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2015.

[22] S. M. Imaduddin, Y. Athar, A. A. Khan, M. M. Khan, and F. M. Kashif, "A computationally efficient heart rate measurement system using video cameras," in *Emerging Technologies (ICET), 2015 International Conference on*, Peshawar Pakistan, Dec. 2015.

[23] R.-Y. Huang and L.-R. Dung, "Measurement of heart rate variability using off-the-shelf smart phones," *Biomedical Engineering Online*, vol. 15, no. 11, pp. 1–16, Jan. 2016.

[24] M. Yang et al., "Vital sign estimation from passive thermal video," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, Jun. 2008.

[25] L. Boccanfuso et al., "Collecting heart rate using a high precision, non-contact, single-point infrared temperature sensor," in *International Conference on Social Robotics*, Chengdu, China, Oct. 2012.

[26] D. Cardone et al., "Thermal infrared imaging-based computational psychophysiology for psychometrics," *Computational and mathematical methods in medicine*, vol. 2015, Jan. 2015.

[27] S. Y. Chekmenev, A. A. Farag, and E. A. Essock, "Thermal imaging of the superficial temporal artery: An arterial pulse recovery model," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, Jun. 2007.

[28] T. R. Gault and A. A. Farag, "A fully automatic method to extract the heart rate from thermal video," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Portland, OR, Jun. 2013.

[29] S. Y. Chekmenev, H. Rara, and A. A. Farag, "Non-contact, wavelet-based measurement of vital signs using thermal imaging," *Journal of Graphics, Vision and Image Processing*, vol. 6, pp. 25–30, 2006.

[30] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Improved pulse detection from head motions using DCT," in *International Conference on Computer Vision Theory and Applications*, vol. 3, Jan. 2014, pp. 118–124.

[31] L. Shan and M. Yu, "Video-based heart rate measurement using head motion tracking and ICA," in *International Congress on Image and Signal Processing*, vol. 01, Dec. 2013, pp. 160–164.

[32] N. Bernacchia et al., "Non contact measurement of heart and respiration rates based on Kinect™," in *IEEE Int. Symp. Medical Measurements and Applications*, Lisbon, Portugal, Jun. 2014.

[33] M.-C. Yu et al., "Multiparameter sleep monitoring using a depth camera," in *Biomedical Engineering Systems and Technologies*, J. Gabriel et al., Ed. Springer, 2013, vol. 357, pp. 311–325.

[34] E. Lachat et al., "First Experiences with Kinect v2 Sensor for Close Range 3d Modelling," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 93–100, Feb. 2015.

[35] D. Pagliari and L. Pinto, "Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors," *Sensors*, vol. 15, no. 11, pp. 27 569–27 589, Oct. 2015.

[36] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng, "On the trend, detrending, and variability of nonlinear and nonstationary time series," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 38, pp. 14 889–14 894, Sept. 2007.

[37] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[38] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Cambridge, MA: Academic Press, 2008.

[39] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, FL: CRC Press, 2005.

[40] W. Sun, G. Cheung, P. Chou, D. Florencio, C. Zhang, and O. Au, "Rate-constrained 3D surface estimation from noise-corrupted multiview depth videos," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3138–3151, Jul. 2014.

[41] Y. S. Kim et al., "Parametric model-based noise reduction for ToF depth sensors," in *Three-Dimensional Image Processing (3DIP) and Applications II*, Burlingame, CA, Jan. 2012.

[42] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," in *IEEE Signal Processing Magazine*, vol. 30, no.3, May 2013, pp. 83–98.

[43] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Apr. 2012.

[44] J. Shotton and T. Sharp et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, Jan. 2013.

[45] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas, "Real-time head and hand tracking based on 2.5D data," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 575–585, Jun. 2012.

[46] L. Xia et al., "Human detection using depth information by kinect," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Colorado Springs, CO, Jun. 2011.

[47] M. Y. Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa, "Fast directional chamfer matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2010.

[48] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[49] H. G. Barrow et al., "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *International Joint Conference on Artificial Intelligence*, Cambridge, MA, Aug. 1977.

[50] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European Conference on Computer Vision*, Florence, Italy, Oct. 2012.

[51] ——, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[52] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, Sept. 2015.

[53] M. Kristan and J. M. et al., "The visual object tracking VOT2015 challenge results," in *IEEE International Conference on Computer Vision Workshops*, Santiago, Chile, Dec. 2015.

[54] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least-squares classification," *Nato Science Series Sub Series III Computer and Systems Sciences*, vol. 190, pp. 131–154, 2003.

[55] "Kinect for Windows software development kit," accessed: Apr. 2016. [Online]. Available: https://dev.windows.com/en-us/kinect/develop

[56] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[57] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, January 1991.

[58] S. Butterworth, "On the theory of filter amplifiers," *Experimental Wireless and the Wireless Engineer*, vol. 7, pp. 536–541, Oct. 1930.

[59] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.

[60] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[61] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 613–627, May 1995.

[62] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-15, no. 2, pp. 70–73, Jun. 1967.

[63] Y. Yuan et al., "Visual object tracking by structure complexity coefficients," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1125–1136, Aug. 2015.

[64] B. Ma et al., "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1818–1828, Oct. 2015.

[65] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, Aug. 1981.

[66] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[67] H. Bay et al., "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, Jun. 2008.

[68] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, Dec. 2001.

[69] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, Jun. 2005.

[70] S. Tang et al., "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Asian Conference on Computer Vision*, Daejeon, Korea, Nov. 2012.

[71] M. Pietikäinen et al., *Computer Vision Using Local Binary Patterns*. Reading, Massachusetts: Springer-Verlag London, 2011, vol. 40.

[72] S. Awwad, F. Hussein, and M. Piccardi, "Local depth patterns for tracking in depth videos," in *ACM International Conference on Multimedia*, New York, NY, Oct. 2015.

[73] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[74] S. Song and J. Xiao, "Tracking revisited using RGBD camera: Unified benchmark and baselines," in *IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 2013.

[75] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York, NY, USA: Cambridge University Press, 2007.