View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Strathclyde Institutional Repository

University of Strathclyde Glasgow

Strathprints Institutional Repository

Yang, Cheng and Cheung, Gene and Stankovic, Vladimir (2015) Estimating heart rate via depth video motion tracking. In: ICME-2015, 2015-06-29 - 2015-07-03, Italy. ,

This version is available at http://strathprints.strath.ac.uk/55061/

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Unless otherwise explicitly stated on the manuscript, Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Please check the manuscript for details of any other licences that may have been applied. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (http://strathprints.strath.ac.uk/) and the content of this paper for research or private study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to Strathprints administrator: strathprints@strath.ac.uk

ESTIMATING HEART RATE VIA DEPTH VIDEO MOTION TRACKING

Cheng Yang[†], *Gene Cheung*[‡], *Vladimir Stankovic*[†]

[†]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK
[‡]The Graduate University for Advanced Studies, National Institute of Informatics, Tokyo, Japan

ABSTRACT

Depth sensors like Microsoft Kinect can acquire partial geometric information in a 3D scene via captured depth images, with potential application to non-contact health monitoring. However, captured depth videos typically suffer from low bit-depth representation and acquisition noise corruption, and hence using them to deduce health metrics that require tracking subtle 3D structural details is difficult. In this paper, we propose to capture depth video using Kinect 2.0 to estimate the heart rate of a human subject; as blood is pumped to circulate through the head, tiny oscillatory head motion can be detected for periodicity analysis. Specifically, we first perform a joint bit-depth enhancement / denoising procedure to improve the quality of the captured depth images, using a graph-signal smoothness prior for regularization. We then track an automatically detected nose region throughout the depth video to deduce 3D motion vectors. The deduced 3D vectors are then analyzed via principal component analysis to estimate heart rate. Experimental results show improved tracking accuracy using our proposed joint bit-depth enhancement / denoising procedure, and estimated heart rates are close to ground truth.

Index Terms— health monitoring, image enhancement, graph signal processing

1. INTRODUCTION

As the general population ages in the developed countries, cheap and non-invasive health monitoring has become more in demand. Among available health monitoring systems are image-based systems with the distinct advantage of being completely non-contact and thus non-intrusive. Of particular interest are systems based on new depth sensors like Micosoft Kinect that can acquire fairly accurate 3D geometric data from captured depth images, and can be fully functional even in dark rooms—useful for applications such as sleep monitoring. Previous depth-image-based systems [1, 2] have demonstrated that certain human vital signs like respiratory rate can be accurately estimated, so that medically urgent events like sleep apnoea (temporary suspension of breathing) can be detected. However, due to limitations of the depth sensing technologies, captured depth videos typically suffer from low bit-depth representation (*e.g.*, Kinect 2.0 has bit-depth of 13 bits for each captured depth pixel) and sensory noise corruption. This means that it is difficult to design depth-image-based systems to estimate health metrics that require tracking subtle 3D structural details in the scene.

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. It has been previously shown [3] that as blood is pumped from the heart to the head for circulation, the head will oscillate slightly due to Newtonian mechanics, and tracking this oscillatory movement can lead to a heart rate estimate. Unlike previously used high-resolution color video [3], the key challenge using depth video is to overcome the low bit-depth representation and sensory noise inherent in the observed data. Towards this end, we first propose a joint bit-depth enhancement / denoising procedure to improve the quality of the captured depth images, using a graph-signal smoothness prior for regularization [4]. We then track an automatically detected nose region throughout the depth video to deduce 3D motion vectors of the subject. Finally, the deduced 3D motion vectors are analyzed via principal component analysis (PCA) to estimate heart rate. Experimental results show improved tracking accuracy using our joint bit-depth enhancement / denoising procedure, and our estimated heart rates are close to ground truth.

The outline of the paper is as follows. We first discuss related work on Section 2. We then overview our heart rate detection system in Section 3. We present our depth video pre-processing algorithms in Section 4, and the heart rate estimation algorithm in Section 5. We present experimental results and conclude remarks in Section 6 and 7, respectively.

2. RELATED WORK

In [5], [6], [7], the human subject is recorded using a conventional RGB camera, and the heart rate is extracted from the recorded video using the subtle colour changes in the facial skin due to blood circulation. In contrast to our approach, all these approaches require high-resolution $coloured$ video of the skin. In [3], similarly to our work, the detection of subtle head oscillations in videos during the cyclical movement of blood from the heart to the head is used to measure the pulse rate. In contrast to our work, [3] uses coloured video to extract feature points, which are tracked throughout the video to deduce motion. The motion of the feature points are then analysed using PCA to estimate heart rate. Though also motion-based, again we differ from [3] in that only depth video is used for analysis, which is not affected by external lighting conditions.

In [8], a thermal infrared sensor (TIRS) is used to capture subtle temperature changes in the sub-nasal skin surface for heart rate detection. However, a good TIRS (over \$1000) is far more expensive than a Kinect sensor. In [9], 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 [1] and [2], an MS Kinect 1.0 depth sensor is used for detecting episodes of sleep disorder, namely 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 signal processing, is then used in [1] and [2], respectively, to detect episodes of apnea / hypopnoea. Oscillatory head movements due to heart beat are much smaller than respiratory chest movements and much harder to detect in depth videos, however, and hence the challenge in this paper.

3. SYSTEM OVERVIEW

We first overview our depth-video-based heart rate detection system in Section 3.1. We then derive a simple depth image noise model from collected observed data in Section 3.2. We discuss the graph-signal smoothness prior we employ for joint bit-depth enhancement / denoising in Section 3.3. Finally, we describe our selection of target region for head tracking in depth video in Section 3.4.

3.1. Heart Rate Estimation System

In terms of hardware, our system is composed of a Kinect 2.0 camera connected to a standalone laptop. For simplicity, we assume that the camera is placed in front of the human subject at a distance of roughly 75 to 80cm. Depth video is captured at 30 frames per second (fps) at 512×424 spatial resolution. Each captured depth image is corrupted by sensory noise, and thus denoising is one important pre-processing task. Derivation of an appropriate noise model for Kinect 2.0 is discussed in details in Section 3.2.

Each captured pixel is represented by 13 bits, which translates to a depth granularity of no smaller than 1mm (the granularity varies according to the physical distance between the captured subject and the capturing camera). Because the head movement due to heart beat is very slight (roughly 5mm according to [3]), this granularity is coarse for our tracking algorithm. Thus another key challenge is to enhance bit-depth in the captured depth video prior to analysis for improved heart rate detection.

Algorithmically, our method can be divided into three parts. First, we define a *target region* $\mathbf{x}_1$ within the human subject's face in frame 1—one that is amenable to robust head tracking in the captured depth video. Second, we jointly enhance the bit-depth and denoise each of the depth frames using our proposed pre-processing algorithm. Finally, we track the target region throughout the depth video, so that the deduced 3D vectors can be analyzed via PCA to estimate heart rate. The joint bit-depth enhancement / denoising optimization is discussed in Section 4, while the heart rate estimation procedure is discussed in Section 5.

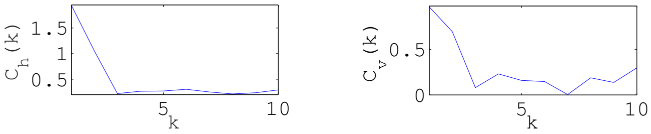
3.2. Derivation of Noise Model

We first derive a suitable noise model for Kinect 2.0 captured pixels in a depth video frame, which we will use later for our to-be-described denoising algorithm. For model derivation, we placed statically a flat board on a table and recorded a depth video of T frames. Let $x_{i,j}^t$ 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 $\frac{1}{T}\sum_{t=1}^{T} x_{i,j}^t$, i.e., the average pixel intensity value at the same location over all T frames. Given image size of $M \times N$ pixels, we can estimate the *horizontal auto-correlation* $C_h(k)$ as:

$$C_h(k) = \frac{\sigma^{-2}}{TM(N-k)} \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N-k} (x_{i,j}^t - \mu_{i,j})(x_{i,j+k}^t - \mu_{i,j+k}) \tag{1}$$

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

$$C_v(k) = \frac{\sigma^{-2}}{T(M-k)N} \sum_{t=1}^{T} \sum_{i=1}^{M-k} \sum_{j=1}^{N} (x_{i,j}^t - \mu_{i,j})(x_{i+k,j}^t - \mu_{i+k,j}) \tag{2}$$



(a) $C_h(k)$ (b) $C_v(k)$

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

Fig. 1 shows the auto-correlation plots tested on a sequence of $T = 15000$ frames computed on a flat 30×30 $(M \times N)$ square surface at a distance 77.1cm from the camera. We observe that the auto-correlation in both cases decrease 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) noise model, which was heuristically found to model well the measured noise, the likelihood $Pr(\mathbf{y}|\mathbf{x})$ of observing a depth pixel patch $\mathbf{y}$ given the original patch is $\mathbf{x}$ is:

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

where $\mathbf{P}$ is the precision matrix (inverse of the covariance matrix). To model neighboring pixel correlation using GMRF, we set the entries in $\mathbf{P}$ as follows [10]:

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

$\mathbf{P}$ will be used in our denoising algorithm in a later section. 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 [11] for depth image noise modelling for time-of-flight cameras.

3.3. Graph-signal Smoothness Prior

As in other inverse imaging problems, a signal prior for the desired signal is needed for regularization. As done in [4, 2], in this paper we also employ a *graph-signal smoothness prior*; *i.e.*, a depth block $\mathbf{x}$ is piecewise smooth if $\mathbf{x}^T\mathbf{L}\mathbf{x}$ is small, where $\mathbf{L}$ is the *graph Laplacian* for 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:

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

where I_i is the pixel intensity of pixel i and σ_I^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 *combinatorial graph Laplacian* $\mathbf{L}$ is then defined as the difference between the degree matrix $\mathbf{D}$ and the adjacency matrix $\mathbf{W}$:

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

It can be shown 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.

Given $\mathbf{L}$ 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:

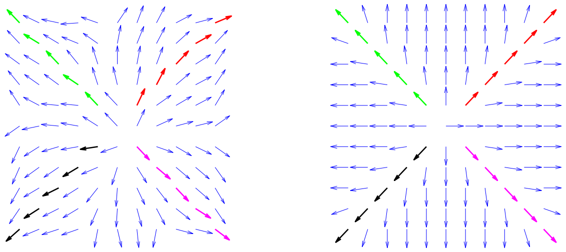
$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_k \lambda_k \alpha_k^2 \tag{8}$$

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.

3.4. Target Region Selection

We discuss next how we select the *target region* $\mathbf{x}_1$. The region needs to be sensitive to head movements due to blood circulation and easily tractable from frame to frame. For simplicity, we assume that the target region is of fixed size $H \times H$ pixels, where H is an odd number. When the subject is facing the camera, the *nasal tip* is typically the closest point and contains sharp edges that can be tracked. Thus, we select the target region to be the nasal tip surface area.

Specifically, we treat the nasal tip as a 3D object with its corresponding cross section that is parallel with the image plane as its base. The shape of this object resembles that of a $\mathbf{C_{4v}}$-symmetry [12] square pyramid. Thus to identify the nasal tip surface area, we find the best-matched block to the $\mathbf{C_{4v}}$-symmetry square pyramid. A strong feature of a $\mathbf{C_{4v}}$-symmetry square pyramid is that the gradient direction of each apex-connected edge of the $\mathbf{C_{4v}}$-symmetry square pyramid is constant. We thus formulate the following gradient direction-based target region selection process. For each candidate block $\mathbf{x}_1^c$ within the face region denoted by $\mathbf{X}_1$, we first obtain the $H \times H$ gradient direction map, $\nabla^{\mathbf{x}_1^c}$, where the element in the ith row and jth column, $\nabla^{\mathbf{x}_1^c}_{i,j}$, is calculated counterclockwise from the direction of increasing column coordinates and $-\pi \leq \nabla^{\mathbf{x}_1^c}_{i,j} \leq \pi$, $1 \leq i,j \leq H$. Fig. 2(a) shows the gradient direction map of the sample nasal tip area. It can be seen from the figure that the main diagonal and anti-diagonal entries of $\nabla^{\mathbf{x}_1^c}$, shown counterclockwise in red, green, black, and magenta, without considering the central one $\nabla^{\mathbf{x}_1^c}_{(H+1)/2,(H+1)/2}$, are close to those of a $\mathbf{C_{4v}}$-symmetry square pyramid with $H \times H$ square base shown in Fig. 2(b). Thus, we select the target region $\mathbf{x}_1$ as a region within $\mathbf{X}_1$ whose main diagonal and anti-diagonal entries of the gradient direction map without the central one are closest to those of $\mathbf{C_{4v}}$-symmetry square pyramid with $H \times H$ square base.



(a) a sample nasal tip block (b) a square pyramid

Fig. 2: Quiver plots of gradient direction maps.

Mathematically, we divide $\nabla^{\mathbf{x}_1^c}$ into four quadrants, denoted as $\nabla^{\mathbf{x}_1^c}_{q_i}$, $1 \leq i \leq 4$, and formulate the following optimization problem to select $\mathbf{x}_1$:

$$\begin{aligned}\mathbf{x}_1 = \arg\min_{\mathbf{x}_1^c \in \mathbf{X}_1} \sum_{p=1}^{(H-1)/2} & (\nabla^{\mathbf{x}_1^c}_{q_{1,p}} - \frac{\pi}{4})^2 + (\nabla^{\mathbf{x}_1^c}_{q_{2,p}} - \frac{3\pi}{4})^2 \\ & + (\nabla^{\mathbf{x}_1^c}_{q_{3,p}} + \frac{3\pi}{4})^2 + (\nabla^{\mathbf{x}_1^c}_{q_{4,p}} + \frac{\pi}{4})^2\end{aligned} \tag{9}$$

where $\nabla^{\mathbf{x}_1^c}_{q_{i,p}}$, $1 \leq p \leq (H-1)/2$ denotes the main diagonal or anti-diagonal entries of $\nabla^{\mathbf{x}_1^c}$ that are in the ith quadrant, for even and odd i, respectively; *e.g.*, $\nabla^{\mathbf{x}_1^c}_{q_{1,p}}$ denotes the anti-diagonal entries of $\nabla^{\mathbf{x}_1^c}$ that are in the first quadrant, shown in red in Fig. 2(a), and $\nabla^{\mathbf{x}_1^c}_{q_{2,p}}$, $\nabla^{\mathbf{x}_1^c}_{q_{3,p}}$, and $\nabla^{\mathbf{x}_1^c}_{q_{4,p}}$, are shown in Fig. 2(a), in green, black, and magenta, respectively.

4. DEPTH VIDEO PRE-PROCESSING

4.1. Joint Bit-depth Enhancement / Spatial Denoising

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

$$\mathbf{y} = \text{round}\left(\frac{\mathbf{x}+\mathbf{n}}{Q}\right) Q \tag{10}$$

where Q is the quantization parameter due to coarse depth precision by the Kinect sensor, and $\mathbf{n}$ 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:

$$Pr(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{z}} Pr(\mathbf{z}|\mathbf{x}) Pr(\mathbf{y}|\mathbf{z},\mathbf{x}) d\mathbf{z} \tag{11}$$

$Pr(\mathbf{y}|\mathbf{z},\mathbf{x})$ 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 simplified to

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

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

$Pr(\mathbf{y}|\mathbf{x})$ in the form (12) is still difficult to use. We thus approximate it as:

$$Pr(\mathbf{y}|\mathbf{x}) \propto \max_{\mathbf{y}-\frac{Q}{2} \leq \mathbf{z} < \mathbf{y}+\frac{Q}{2}} \exp\left[-\frac{(\mathbf{z}-\mathbf{x})^T \mathbf{P} (\mathbf{z}-\mathbf{x})}{\sigma^2}\right] \tag{13}$$

One can see that (12) and (13) have similar shapes. $Pr(\mathbf{y}|\mathbf{x})$ in (12) must integrate $\mathbf{z}$ over region $\mathcal{R}_{\mathbf{y}}$ within 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 $\mathbf{x}$ or Q is large. This is also true for $Pr(\mathbf{y}|\mathbf{x})$ in (13).

4.1.1. Objective Function

Given likelihood in (13) and the graph-signal smoothness prior, one can now derive the MAP objective by minimizing the negative log of the likelihood and prior:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & (\mathbf{z}-\mathbf{x})^T \mathbf{P} (\mathbf{z}-\mathbf{x}) + \mu \, \mathbf{x}^T \mathbf{L} \, \mathbf{x} \\ \text{s.t.} \quad & y_i - \tfrac{Q}{2} \leq z_i < y_i + \tfrac{Q}{2}, \ \ \forall i \end{aligned} \tag{14}$$

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).

4.1.2. Optimization Procedure

With two inter-dependent variables $\mathbf{x}$ and $\mathbf{z}$ and a constraint on $\mathbf{z}$, the optimization (14) 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 in (14) with respect to $\mathbf{x}$ and setting it to zero:

$$\mathbf{x}^* = (\mathbf{P} + \mu \mathbf{L})^{-1} \mathbf{P} \mathbf{z} \tag{15}$$

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:

$$z_i^* = \begin{cases} y_i + Q/2 - \epsilon & \text{if } \ x_i \geq y_i + Q/2 \\ y_i - Q/2 & \text{if } \ x_i < y_i - Q/2 \\ x_i & \text{o.w.} \end{cases} \tag{16}$$

where ϵ is a small positive constant. The two variables are optimized alternately until the solution converges. 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.

4.2. Joint Tracking / Temporal Denoising

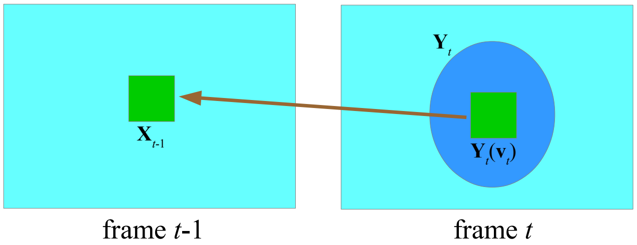


Fig. 3: Illustration of tracking target region $\mathbf{Y}_t(\mathbf{v})$ in frame t, given tracked region $\mathbf{X}_{t-1}$ in previous frame $t-1$.

To track a target region $\mathcal{T}$ over a sequence of frames, we perform the following procedure. We first perform joint bit-depth enhancement / spatial denoising on new frame t as described in the previous section. We then formulate the following optimization for joint tracking / temporal denoising. Let $\mathbf{Y}_t$ be the observed frame at time instant t. A *motion vector* (MV) $\mathbf{v}_t$ points to a sub-region $\mathbf{Y}_t(\mathbf{v}_t)$ inside $\mathbf{Y}_t$ that corresponds to the target region $\mathcal{T}$ in frame t. Let $\bar{\mathbf{x}}_{t-1}$ be the denoised target region $\mathcal{T}$ in previous frame $t-1$. The optimization thus becomes the search for MV $\mathbf{v}_t$ and denoised patch $\mathbf{x}_t$ that minimize three terms: i) a fidelity term with respect to observation $\mathbf{Y}_t(\mathbf{v}_t)$, ii) a graph-signal smoothness term $\mathbf{x}_t^T \mathbf{L} \mathbf{x}_t$, and iii) a motion estimation term $\|\bar{\mathbf{x}}_{t-1} - \mathbf{x}_t\|_2^2$ that measures how well the designated target regions match in the two frames:

$$\begin{aligned} \min_{\mathbf{v}_t,\mathbf{z}_t,\mathbf{x}_t} \quad & (\mathbf{z}_t-\mathbf{x}_t)^T \mathbf{P} (\mathbf{z}_t-\mathbf{x}_t) + \mu \, \mathbf{x}_t^T \mathbf{L} \, \mathbf{x}_t \\ & + \gamma \, \|\bar{\mathbf{x}}_{t-1} - \mathbf{x}_t\|_2^2 \\ \text{s.t.} \quad & \mathbf{Y}_t(\mathbf{v}_t) - \tfrac{Q}{2} \leq \mathbf{z}_t < \mathbf{Y}_t(\mathbf{v}_t) + \tfrac{Q}{2} \end{aligned} \tag{17}$$

4.2.1. Optimization Procedure

To solve (17), we use a similar alternating method as follows. We first search for the optimal $\mathbf{v}_t$ that minimizes the motion estimation term $\|\bar{\mathbf{x}}_{t-1} - \mathbf{Y}_t(\mathbf{v}_t)\|_2^2$. We then 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:

$$\mathbf{x}_t^* = (\mathbf{P} + \mu\mathbf{L} + \gamma\mathbf{I})^{-1}(\mathbf{P}\mathbf{z}_t + \gamma\bar{\mathbf{x}}_{t-1}) \tag{18}$$

where $\mathbf{I}$ is the identity matrix. The optimal $\mathbf{z}$ given fixed $\mathbf{x}$ is solved using (16).

5. HEART RATE ESTIMATION

In this section, we first describe the analysis of the tracked movement vectors via PCA, and then explain the procedure of heart rate estimation based on the PCA decomposition result.

5.1. Principal Component Analysis

Given $\bar{\mathbf{x}}_t$, the tracked and denoised target region $\mathcal{T}$ in frame t, we designate the centre coordinate of $\bar{\mathbf{x}}_t$ as $horizontal$ position h_t and $vertical$ position v_t of $\bar{\mathbf{x}}_t$, and the depth intensity at centre coordinate as $axial$ position a_t of $\bar{\mathbf{x}}_t$. Since h_t contains most of equilibrium movement [3] that can affect heart rate estimation, we remove h_t, and use a 2D vector (v_t, a_t) to denote $vertical$ and $axial$ positions of $\bar{\mathbf{x}}_t$. We find that the granularity of $vertical$ component is approximately 1.6992mm per pixel coordinate, and $axial$ component is approximately 1.0147mm per depth intensity, at capturing distance 77.1cm. Therefore we unify the measurement of v_t and a_t into mm unit, denoted as $\Delta_t = (v_t^{\mathrm{mm}}, a_t^{\mathrm{mm}})^T$, before we apply PCA.

Let $\mathbf{\Delta}$ be the $2 \times T$ tracked *movement matrix*, $\mathbf{\Delta} = [\Delta_1, ..., \Delta_T]$. We calculate the 2×1 mean matrix $\overline{\mathbf{\Delta}}$ and 2×2 covariance matrix $\mathbf{O}$ as:

$$\overline{\mathbf{\Delta}} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{\Delta}_i \tag{19}$$

$$\mathbf{O} = \frac{1}{T}\sum_{i=1}^{T}(\mathbf{\Delta}_i - \overline{\mathbf{\Delta}})(\mathbf{\Delta}_i - \overline{\mathbf{\Delta}})^T = \frac{1}{T}\mathbf{H}\mathbf{H}^T \tag{20}$$

where $\mathbf{H} = [\Delta_1 - \overline{\mathbf{\Delta}}, \ldots, \Delta_T - \overline{\mathbf{\Delta}}]$. PCA [13, 14] determines the eigenvectors of the movement by solving the following algebraic eigenvalue problem:

$$\mathbf{O}\mathbf{E} = \mathbf{E}\mathbf{\Lambda} \tag{21}$$

where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2]$ denotes the eigenvectors of the 2D movement and $\mathbf{e}_1$ and $\mathbf{e}_2$ are in descending order according to the amplitude of their corresponding eigenvalues, and $\mathbf{\Lambda}$ denotes a diagonal matrix of the corresponding eigenvalues λ_1, λ_2.

Next, similarly to [3], we project $\mathbf{\Delta}$ onto $\mathbf{e}_i$, $i = 1,\ 2$, to get time plots of the projected movement $\mathbf{S}_i^{\mathrm{PCA}} = [\|\mathrm{proj}_{\mathbf{e}_i}\Delta_1\|_2, \ldots, \|\mathrm{proj}_{\mathbf{e}_i}\Delta_T\|_2]$, then assess the periodicity of each $\mathbf{S}_i^{\mathrm{PCA}}$ as the percentage of the spectral power on the frequency with maximum power and its first k harmonics over the total spectral power, and finally choose the most periodic $\mathbf{S}_i^{\mathrm{PCA}}$.

5.2. Heart Rate Estimation

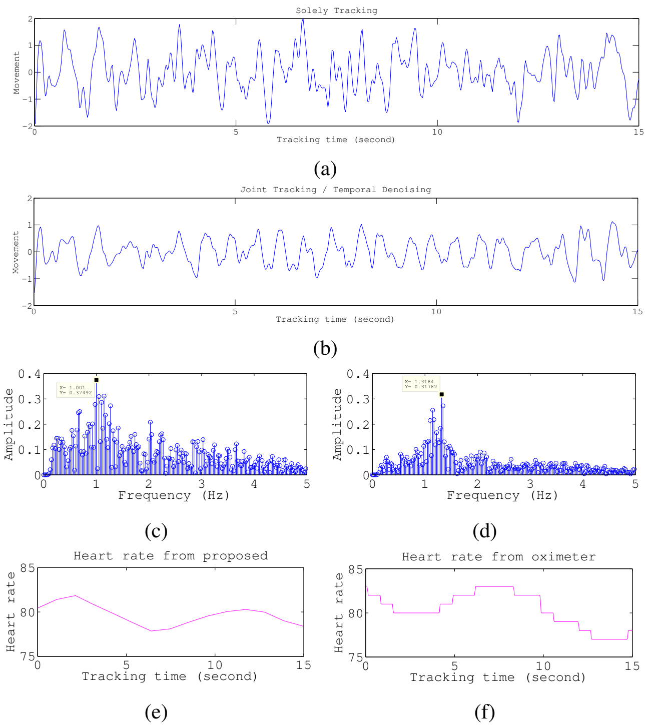
We first pass the most periodic $\mathbf{S}_i^{\mathrm{PCA}}$ through a second-order Butterworth lowpass filter with 0.25 normalized cutoff frequency, and then remove the linear trend of the filtered signal. We denote the resulting signal as $\widehat{\mathbf{S}}_i^{\mathrm{PCA}}$. Next, we estimate the heart rate (HR) by first applying Fast Fourier Transform (FFT) on $\widehat{\mathbf{S}}_i^{\mathrm{PCA}}$, to get single-sided amplitude spectrum of $\widehat{\mathbf{S}}_i^{\mathrm{PCA}}$. The sampling frequency is 30Hz, to be consistent with the 30fps depth video frame accusation, and the window size is 15secs. Then we find the frequency f^* with the largest peak in the single-sided amplitude spectrum of $\widehat{\mathbf{S}}_i^{\mathrm{PCA}}$, and estimate HR as $\mathrm{HR} = 60 \times f^*$ beats/minute.

6. EXPERIMENTAL RESULTS

In the experiments, we test 7 subjects. Data are collected simultaneously using a Kinect 2.0 depth camera and a finger pulse oximeter (ANAPULSE ANP100, Ana Wiz Ltd, UK) which we assume provides ground truth HR reading. The system is implemented in Matlab R2014a on a laptop running Windows 8.1, with Core i7 2820QM 2.3GHz CPU and 16GB RAM. The mean computational time is 0.847s per frame. In this section, we first present tracking results without and with the proposed depth video pre-processing, and then compare the estimated heart rate based on the tracking results with depth video pre-processing to the ground truth.

Fig. 4(a) shows a sample result of the middle 15-second session of the 30-second tracking without the proposed depth video pre-processing (only tracking as described in Section 5). The corresponding tracking result with depth video pre-processing (joint tracking / temporal denoising), shown in Fig. 4(b), indicates cleaner subtle head movements than only tracking. Fig. 4(c) and (d) show the single-sided amplitude spectra of Fig. 4(a) and (b), respectively, which further indicates clearer subtle head movements by applying joint tracking / temporal denoising than solely tracking. Fig. 4(e) shows estimated 15-second HR by applying FFT with 15-second window size on the corresponding whole 30-second sample result of Fig. 4(b) at 30Hz sampling frequency, denoted as $\mathrm{HR_P}$.

Next, to compare the result with the ground truth HR reading from the finger pulse oximeter, denoted as $\mathrm{HR_G}$, we first unify the sampling frequencies of $\mathrm{HR_P}$ and $\mathrm{HR_G}$ to 30Hz (Fig. 4(f) shows corresponding 15-second frequency-unified $\mathrm{HR_G}$), then compute mean percentage error (MPE) of $\mathrm{HR_P}$ wrt $\mathrm{HR_G}$. In this example, the mean $\mathrm{HR_P}$ is 79.52 beats/minute, the mean $\mathrm{HR_G}$ (ground truth) is 80.61 beats/minute, and the MPE of $\mathrm{HR_P}$ wrt $\mathrm{HR_G}$ is 3.02%. Looking back to the frequencies at the largest peaks in Fig. 4(c) and (d), the estimated HR of Fig. 4(a) (solely tracking), $60 \times 1.001 = 60.06$ beats/minute, is far different from the ground truth, while the estimated HR of Fig. 4(b) (joint tracking / temporal denoising), $60 \times 1.3184 = 79.10$ beats/minute, is close to ground truth.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4: Illustration of the proposed depth video motion tracking and the corresponding heart rate estimation. (a) 15-second solely tracking. (b) 15-second joint tracking / temporal denoising. (c) Single-sided amplitude spectrum of (a) using FFT. (d) Single-sided amplitude spectrum of (b) using FFT. (e) 15-second estimated heart rate of the corresponding whole 30-second sample result of (b). (f) Corresponding 15-second frequency-unified ground truth heart rate reading from the finger pulse oximeter.

Table 1 shows the comparison result of mean $\mathrm{HR_P}$ ($\overline{\mathrm{HR_P}}$) and mean $\mathrm{HR_G}$ ($\overline{\mathrm{HR_G}}$) of 1-minute sample for each subject. Overall, the MPEs are always within 10%, showing that our proposed system can effectively estimate heart rate based on subtle motion tracking in depth videos.

7. CONCLUSION

In this paper, we propose a heart rate estimation system based on motion tracking in Kinect 2.0 depth videos. It can operate in complete darkness, thus is useful in applications such as sleep monitoring. We pre-process captured depth videos via joint bit-depth enhancement / denoising, and detect and track the nasal tip area for head motion via joint tracking / temporal denoising. The tracked motion vectors are then analyzed using PCA. Finally, we estimate heart rate via FFT. Experimental results demonstrate that our depth video pre-processing can effectively enhance tracking accuracy, and our estimated heart rates are close to ground truth measurements. Though we performed experiments using a single depth camera placed in front of a test subjects to track the nasal tip area, in practice, multiple appropriately located depth cameras would cover the majority of the subject's pose to improve system robustness.

Subject	$\overline{\mathrm{HR_P}}$	$\overline{\mathrm{HR_G}}$	MPE
1	68.56	71.92	6.33%
2	84.38	91.42	9.25%
3	89.65	88.77	2.51%
4	77.92	79.63	5.97%
5	78.26	74.15	7.12%
6	82.37	85.26	4.83%
7	78.24	76.85	5.72%

Table 1: Comparison result of $\overline{\mathrm{HR_P}}$ and $\overline{\mathrm{HR_G}}$ of 1-minute sample for each subject with MPEs.

8. REFERENCES

[1] 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, July 2014.

[2] 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, September 2014.

[3] 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, June 2013.

[4] W. Hu, X. Li, G. Cheung, and O. Au, "Depth map denoising using graph-based transform and group sparsity," in *IEEE International Workshop on Multimedia Signal Processing*, October 2013.

[5] 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.

[6] 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, October 2014.

[7] H. Y. Wu et al., "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH*, vol. 31, no. 4, July 2012.

[8] 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, October 2012.

[9] N. Bernacchia et al., "Non contact measurement of heart and respiration rates based on Kinect$^{\mathrm{TM}}$," in *IEEE International Symposium on Medical Measurements and Applications*, Lisbon, Portugal, June 2014.

[10] 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, July 2014.

[11] 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, January 2012.

[12] P. J. Flynn, "3-D object recognition with symmetric models: symmetry extraction and encoding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, pp. 814–818, August 1994.

[13] 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, January 1990.

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