Graph-based Depth Video Denoising and Event Detection for Sleep Monitoring

Cheng Yang [#1], Yu Mao [*2], Gene Cheung [*3], Vladimir Stankovic [#4], Kevin Chan [%5]

[#] *Department of Electronic and Electrical Engineering, University of Strathclyde Glasgow, UK*

[1,4] {cheng.yang, vladimir.stankovic}@strath.ac.uk

[*] *The Graduate University for Advanced Studies, National Institute of Informatics Tokyo, Japan*

[2,3] {mao, cheung}@nii.ac.jp

[%] *School of Medicine, University of Western Sydney, Camden and Campbelltown Hospitals Sydney, Australia*

[5] drkevinchan@bigpond.com

***Abstract*—Quality of sleep greatly affects a person's physiological well-being. Traditional sleep monitoring systems are expensive in cost and intrusive enough that they disturb the natural sleep of clinical patients. In our previous work, we proposed a non-intrusive sleep monitoring system to first record depth video in real-time, then offline analyze recorded depth data to track a patient's chest and abdomen movements over time. Detection of abnormal breathing is then interpreted as episodes of apnoea or hypopnoea. Leveraging on recent advances in graph signal processing (GSP), in this paper we propose two new additions to further improve our sleep monitoring system. First, temporal denoising is performed using a block motion vector smoothness prior expressed in the graph-signal domain, so that unwanted temporal flickering can be removed. Second, a graph-based event classification scheme is proposed, so that detection of apnoea / hypopnoea can be performed accurately and robustly. Experimental results show first that graph-based temporal denoising scheme outperforms an implementation of temporal median filter in terms of flicker removal. Second, we show that our graph-based event classification scheme is noticeably more robust to errors in training data than two conventional implementations of support vector machine (SVM).**

I. Introduction

It is well documented [1] that a sleep-deprived person carries a number of health-related risks, including increase in body weight, increased risk of diabetes and heart diseases, increased risk of psychiatric illness such as depression, etc. Further, it is not simply the *quantity* of sleep that affects a person's physiological well-being, but also the *quality* of sleep. In particular, sleep-disordered breathing is a common problem [2], and repeated episodes of *apnoea* (temporary suspension of external breathing) and *hypopnoea* (overly shallow breathing or low respiratory rate) can significantly disturb a person's sleep. It is thus important to monitor and detect episodes of apnoea and hypopnoea during a patient's sleep for quick diagnosis and appropriate treatment. We address the problem of sleep monitoring and apnoea / hypopnoea detection in this paper.

Existing sleep monitoring systems can be categorized into two classes: i) vibration-sensing wristbands like Fitbit[1] and Jawbone UP[2]; and ii) full multi-sensing monitoring devices like Philips Alice PDx[3]. While the first class of systems are minimally intrusive, they mostly record sleep time, *i.e.*, the *quantity* rather than the *quality* of sleep. On the other hand, the second class are accurate in measuring various vital signs such as oxygen intake, airflow, etc, but are expensive and intrusive with multiple body straps and tubes. (PDx requires up to 10 minutes to set up.) The numerous sensors attached to various body parts disturb the natural sleep of a patient during monitoring (who had enough difficulties sleeping).

In our previous work [3], we proposed a non-intrusive sleep monitoring system based on depth video coding and analysis. Not relying on the lighting condition of a dark sleeping room, we used a MS Kinect camera projecting infrared light patterns to capture depth images of the sleep patient. We then analyzed the recorded depth data to track the patient's chest and abdomen movements over time; detection of abnormal breathing was then interpreted as episodes of apnoea or hypopnoea. Unlike vibration-sensing wristbands, our system can quantify the quality of one's sleep by detecting episodes of apnoea / hypopnoea during the night. Yet unlike full monitoring devices, our system is entirely non-contact, and thus is completely non-intrusive to the patient's sleep.

Leveraging on recent advances in graph signal processing (GSP) [4], in this paper we propose two new additions to further improve our sleep monitoring system. First, we perform temporal denoising using a block motion vector smoothness

2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), Sep. 22-24, 2014, Jakarta, Indonesia.
978-1-4799-5896-2/14/$31.00

[1] http://www.fitbit.com/
[2] https//jawbone.com/up/
[3] http://alicepdx.respironics.eu

prior expressed in the graph-signal domain, so that unwanted flickering can be removed. Second, we propose a graph-based event detection scheme to detect apnoea / hypopnoea accurately and robustly. Experimental results show first that our graph-based denoising scheme outperforms an implementation of temporal median filter in terms of flicker removal without over-smoothing. Second, we show that our graph-based event classification scheme is noticeably more robust to errors in training data than two conventional implementations of support vector machine (SVM) [5].

The outline of the paper is as follows. We discuss related works in Section II and overview our sleep monitoring system in Section III. We discuss the two novelties in our system, graph-based temporal denoising and sleep event detection, in Section IV and V, respectively. Finally, we present experimental results and conclusion in Section VI and VII.

II. Related Work

To the best of our knowledge, there are only two other works [6], [7] that also used captured depth video for sleep monitoring. [6] claimed that a *Time-of-Flight* (ToF) camera was used to detect chest and abdomen movements for apnoea detection, but there is no mention of what ToF camera was used and how chest and abdomen movements were deduced from collected depth measurements. There is also no performance analysis of the proposal against ground truth data. This renders a direct comparison with [6] impossible.

[7] described a sleep monitoring system using a Kinect camera, where chest movements over time are tracked by observing the nearest depth measurement of the patient to a virtual camera directly above the patient. Our system [3] is different from [7] in that we propose a more accurate dual-ellipse model, so that individual chest and abdominal movements can be tracked, as recommended in standard sleep medicine [1], even if the patient is sleeping sideway.

New GSP tools [4] such as *graph Fourier transform* (GFT) have been shown recently to be useful in applications such as depth map coding [8] and spatial denoising [9]. In this paper, we show how GFT can also be used for temporal depth video denoising, which is more complex than the more straight-forward spatial denoising case and involves the joint optimization of motion vectors and noise-corrupted pixels in the target frame, as described in Section IV. GSP tools have been also used for data classification [10]. Though similar in spirit to [10], our graph-based classification methodology, described in Section V, is more intuitive and less complex—each optimization instance is formulated as an unconstrained quadratic programming problem, solvable in closed form.

III. System Overview

We first overview our proposed sleep monitoring system in [3], which we have set up at Bondi Junction Private Sleep Laboratory in Sydney, Australia, to capture depth videos of patients with suspected sleep problems. The system is composed of a first-generation MS Kinect depth capturing camera and a Lenovo X220 laptop. As shown in Fig. 1, the camera is set up at a higher elevation above and away from the head of the patient lying down. Kinect captures depth image of resolution 640×480 at 30fps at 11-bit pixel precision. In [3], we proposed an efficient H.264 implementation of Kinect-captured video, where different 8 bits per pixel are extracted from 11 available bits of different temporal frames for encoding. At decoder, the uncoded 3 bits are recovered from neighboring frames using a block motion search procedure. We will assume full 11 bits per pixel are recovered at the decoder for further processing. See [3] for details.

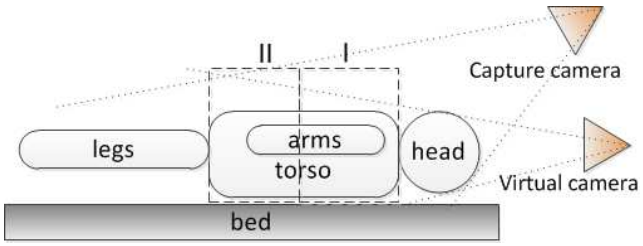


Fig. 1. Side view of sleep patient. Torso is divided into two cross sections, each modeled by an ellipse.

We stress that, beyond automatic apnoea / hypopnoea detection, recorded depth videos are useful as a visual inspection tool of detected sleep events for doctors and an educational tool for patients. The recorded video can also be used to detect other sleep-related events beyond apnoea, such as irregular leg movements, frequent turning / tossing, etc [1]. Thus, annoying flickering removal in sleep video is important; we discuss our graph-based approach in Section IV.

A. Dual Ellipse Model

We next discuss how apnoea or hypopnoea is detected given a recorded depth video. There are three basic steps. First, we back-project depth pixels from the captured camera view to the 3D space and reproject them to a virtual viewpoint image, where the virtual camera is located parallel to the sleeping patient above his/her head, as shown in Fig. 1. This back-projection / reprojection procedure requires intrinsic and extrinsic camera parameters, which can be computed using standard camera calibration procedures[4] [11].

Next, given the computed coordinates (u, v, d)—*i.e.,* pixel location (u, v) with depth value d—in the head-on view, we classify observations into two cross-sections that correspond to the patient's chest and abdomen using the depth values d. It is recommended in standard sleep medicine [1] to track chest and abdominal movements for detection of apnoea; in *central apnoea*, there is a lack of respiratory effort and hence a corresponding lack of chest and abdominal movements, while in *obstructive apnoea* there can be very slight movements in chest and abdomen but in opposite phase. Though we do not distinguish between central and obstructive apnoea (central apnoea takes place only 0.4% of the time), we nonetheless follow the medical recommendation and track chest and abdominal movements separately.

[4]Camera calibration software can be downloaded here: http://www.vision.caltech.edu/bouguetj/calib_doc/htmls/ref.html

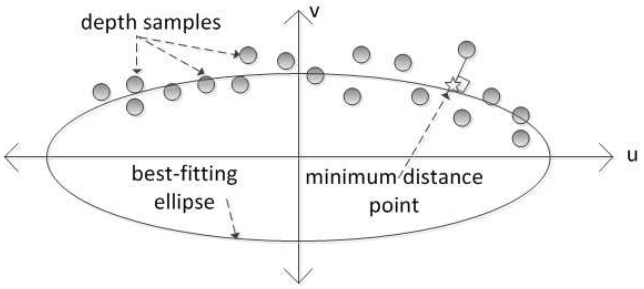


Fig. 2. Best-fitting ellipse from multiple depth observations of the cross section. The closest ellipse point to each observation is perpendicular to the tangent of ellipse at that point.

Finally, we estimate chest and abdomen movements by fitting two ellipses to the available depth observations in the two cross-sections of the patient's body, as ilustrated in Fig. 2. An ellipse in 2D space, with major and minor axes denoted at u and v respectively, can be described as:

$$\left(\frac{u}{a}\right)^2 + \left(\frac{v}{b}\right)^2 = r^2 \tag{1}$$

a and b are called the *major* and *minor radius*, respectively. For simplicity, we assume the center of the ellipse is at origin $(0, 0)$. r is determined based on the patient's waist size.

Assuming a Gaussian noise model for the observed depth data, we formulated a *maximum likelihood* (ML) problem to find two best fitting ellipses with parameters (a, b) per frame. In [3], computed variances of a and b in a window of frames for the two cross-sections were used as training data to design an SVM to detect episodes of hypopnoea. In this paper, we will instead use computed vairances of a and b as input to perform graph-based classification, to be discussed in Section V.

IV. Temporal Depth Video Denoising

We now discuss how we perform temporal denoising using a graph-signal formulation. We first discuss how we can formulate an optimization problem for the motion field in a frame t given previous frame $t-1$ and a motion smoothness prior. Then we discuss how the problem can be modified if frame t is corrupted by noise, and present an efficient algorithm to solve it.

A. Finding Motion Field

For simplicity, we assume first that neither target frame t nor previous frame $t-1$ is corrupted by noise. The goal is to find an accurate motion field for all $K \times K$ pixel blocks in frame t. Let the ith $K \times K$ block in frame t, with upper-left pixel at $\mathbf{p}_i$, be denoted by $b_{\mathbf{p}_i}(t)$. Let the *motion vector* (MV) of the ith block be $\mathbf{v}_i = (x_i, y_i)$; the MV field of all N blocks in the frame is expressed in vector form as $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$.

We first assume a *spatial motion smoothness prior*: a block's MV will be similar to MVs of neighboring blocks if they belong to the same object; *i.e.*, the MV field is *piecewise smooth* (PWS). One way of expressing piecewise smoothness is through a graph [8]. We first construct a 4-connected graph, where each node i represents a block $b_{\mathbf{p}_i}(t)$, and the node is connected to nodes that correspond to neighboring blocks of $b_{\mathbf{p}_i}(t)$. We compute the weight $w_{i,j}$ of an edge connecting two nodes (blocks) i and j as follows:

$$w_{i,j} = \exp\left\{-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|_2^2}{\sigma_v^2}\right\} \tag{2}$$

where σ_v is a chosen parameter. Given the constructed graph, we can define the degree and adjacency matrices, $\mathbf{D}$ and $\mathbf{A}$, correspondingly [4]. The *graph Laplacian* is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. If the MV field is PWS, we say the *graph variation* term, $\|\mathbf{v}^T\mathbf{L}\mathbf{v}\|_2^2$, is small:

$$\mathbf{v}(x)^T\,\mathbf{L}\,\mathbf{v}(x) = \sum_{i,j} w_{i,j}\left(\mathbf{v}_i(x) - \mathbf{v}_j(x)\right)^2 \tag{3}$$

Note that because $\mathbf{v}_i$ is a multi-valued sample (contains x- and y-coordinates of the MV), $\|\mathbf{v}^T\mathbf{L}\mathbf{v}\|_2^2$ means computing $\mathbf{v}^T\mathbf{L}\mathbf{v}$ for the x- and y-coordinates $\mathbf{v}(x)$ and $\mathbf{v}(y)$ of $\mathbf{v}$ separately, then computing the resulting vector magnitude square.

We can now define an optimal MV field $\mathbf{v}$ as one that finds good block matches in previous frame $t-1$ *and* is smooth:

$$\min_{\mathbf{v}} \sum_i \|b_{\mathbf{p}_i+\mathbf{v}_i}(t-1) - b_{\mathbf{p}_i}(t)\|_2^2 \;+\; \lambda\,\|\mathbf{v}^T\mathbf{L}\mathbf{v}\|_2^2 \tag{4}$$

where λ is a chosen weighting parameter that trades off the motion estimation term (first term) and the MV smoothness term (second term).

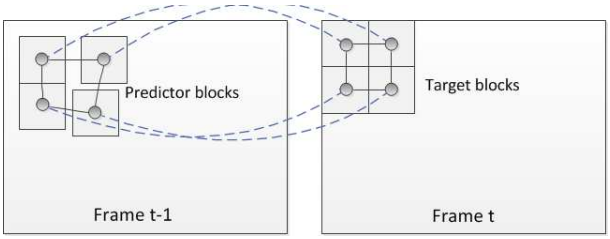


Fig. 3. Example graph construction given four blocks in target frame t and four corresponding predictor blocks in previous frame $t-1$.

B. Temporal Denoising

We now remove the assumption that target frame t is noise-less, meaning we have to find MV field $\mathbf{v}$ *and* denoise blocks $b_{\mathbf{p}_i}(t)$ at the same time. Beyond spatial motion smoothness prior, we now assume further a *temporal motion smoothness prior*; *i.e.*, if ith block at position $\mathbf{p}_i$ of frame t has MV $\mathbf{v}_i$, then the predictor block at position $\mathbf{p}_i + \mathbf{v}_i$ of frame $t-1$ will have a MV $\mathbf{u}_{\mathbf{p}_i+\mathbf{v}_i}$ that is similar to $\mathbf{v}_i$. We can again express this notion of smoothness via a graph. In particular, in additional to the graph constructed for MV $\mathbf{v}_i$ in frame t, we create additional nodes to represent predictor blocks in frame $t-1$. We draw an edge between node representing block $b_{\mathbf{p}_i}(t)$ in frame t and node representing corresponding predictor block $b_{\mathbf{p}_i+\mathbf{v}_i}(t-1)$ with weight computed using (2).

In addition, we draw an edge between two predictor blocks at locations $\mathbf{p}$ and $\mathbf{q}$ in frame $t-1$, if $\|\mathbf{p}-\mathbf{q}\|_2^2 \le \delta$, with weight:

$$w_{i,j} = \exp\left\{-\frac{\|\mathbf{u}_\mathbf{p} - \mathbf{v}_\mathbf{q}\|_2^2}{\sigma_v^2}\right\}\exp\left\{-\frac{\|\mathbf{p}-\mathbf{q}\|_2^2}{\sigma_g^2}\right\} \tag{5}$$

where σ_g is a chosen parameter. This weight assignment is similar to one done in *bilateral filtering* [12]. See Fig. 3 for an example graph constructed from four blocks in the target frame t and four corresponding predictor blocks in the previous frame $t-1$.

Without loss of generality, we define the combined motion vector ζ to be a concatenation of MV $\mathbf{u}$ of predictor blocks of frame $t-1$ and MV $\mathbf{v}$ of target blocks of frame t, *i.e.* $\zeta^T = [\mathbf{u}^T\ \mathbf{v}^T]$. We can also define degree and adjacency matrices $\mathbf{D}$ and $\mathbf{A}$ as done previously for the larger graph. The resulting Laplacian $\mathbf{L}$ is again $\mathbf{L} = \mathbf{D} - \mathbf{A}$.

With these definitions, we can define the new objective to find MV $\mathbf{v}$ and denoised blocks $b_{\mathbf{p}_i}(t)$ as a composition of three terms: i) motion estimation term, ii) MV smoothness term, and iii) fidelity term with respect to observed noisy blocks $d^o_{\mathbf{p}_i}(t)$, *i.e.*,

$$\min_{\mathbf{v}, b(t)} \left\{ \begin{array}{l} \sum_i \|b_{\mathbf{p}_i+\mathbf{v}_i}(t-1) - b_{\mathbf{p}_i}(t)\|_2^2 \ + \ \lambda \, \|\zeta^T \mathbf{L} \zeta\|_2^2 \\ \quad + \ \mu \sum_i \|b_{\mathbf{p}_i}(t) - b^o_{\mathbf{p}_i}(t)\|_2^2 \end{array} \right\} \tag{6}$$

where μ is weighting parameter for the fidelity term. We discuss how we solve (6) next.

C. Optimization Algorithm

(6) is difficult to solve as it involves a large set of optimization variables. Our strategy is to alternately solve one set of variables at a time while keeping the other set fixed, until convergence. Suppose first we initialize MV $\mathbf{v}$ using conventional ME, then fix $\mathbf{v}$ and solve for optimal blocks $b_{\mathbf{p}_i}(t)$. The MV smoothness term is not affected by the selection of $b_{\mathbf{p}_i}(t)$, and so (6) reduces to:

$$\min_{b(t)} \sum_i \|b_{\mathbf{p}_i+\mathbf{v}_i}(t-1) - b_{\mathbf{p}_i}(t)\|_2^2 + \mu \sum_i \|b_{\mathbf{p}_i}(t) - b^o_{\mathbf{p}_i}(t)\|_2^2 \tag{7}$$

Let $b_{\mathbf{p}_i}(t)$ be a convex combination of $b_{\mathbf{p}_i-\mathbf{v}_i}(t-1)$ and $b^o_{\mathbf{p}_i}(t)$, *i.e.*,

$$b_{\mathbf{p}_i}(t) = \epsilon \, b_{\mathbf{p}_i-\mathbf{v}_i}(t-1) + (1-\epsilon) \, b^o_{\mathbf{p}_i}(t) \tag{8}$$

By substituting (8) into (7), taking the derivative with respect to ϵ and setting the equation to zero, one can see that the optimal ϵ^* is:

$$\epsilon^* = \frac{1}{1+\mu} \tag{9}$$

This agrees with intuition; if $\mu = 0$, then $\epsilon^* = 1$ and $b_{\mathbf{p}_i}(t)$ is set to predictor block $b_{\mathbf{p}_i-\mathbf{v}_i}(t-1)$, and if $\mu = 1$, then $\epsilon^* = 1/2$, and $b_{\mathbf{p}_i}(t)$ is the average of predictor block $b_{\mathbf{p}_i-\mathbf{v}_i}(t-1)$ and observed noisy block $b^o_{\mathbf{p}_i}(t)$.

Now we fix blocks $b_{\mathbf{p}_i}(t)$ and solve for the optimal MV $\mathbf{v}$. The fidelity term is not affected by MV $\mathbf{v}$, so (6) reduces to:

$$\min_{\mathbf{v}} \sum_i \|b_{\mathbf{p}_i+\mathbf{v}_i}(t-1) - b_{\mathbf{p}_i}(t)\|_2^2 \ + \ \lambda \, \|\zeta^T \mathbf{L} \zeta\|_2^2 \tag{10}$$

(10) is still difficult to solve, as each change in MV $\mathbf{v}_i$ induces a change in corresponding predictor block $b_{\mathbf{p}_i+\mathbf{v}_i}(t-1)$, resulting in a different predictor MV $\mathbf{u}_{\mathbf{p}_i+\mathbf{v}_i}$ and a modified Laplacian $\mathbf{L}$. Our strategy then is to first find the optimal MV $\mathbf{v}^*$ that minimizes the smoothness term, then try to install $\mathbf{v}^*_i$ one-by-one into (10) to see if the objective is reduced.

Given ζ is a concatenation of predictor MV $\mathbf{u}$ and target MV $\mathbf{v}$, we can rewrite the smoothness term as:

$$\begin{aligned} & \underbrace{\begin{bmatrix} \mathbf{u} & \mathbf{v} \end{bmatrix}^T}_{\zeta^T} \underbrace{\begin{bmatrix} \mathbf{L_{uu}} & \mathbf{L_{uv}} \\ \mathbf{L_{vu}} & \mathbf{L_{vv}} \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}}_{\zeta} \\ = \ & \mathbf{u}^T \mathbf{L_{uu}} \mathbf{u} + \mathbf{u}^T \mathbf{L_{uv}} \mathbf{v} + \mathbf{v}^T \mathbf{L_{vu}} \mathbf{u} + \mathbf{v}^T \mathbf{L_{vv}} \mathbf{v} \end{aligned} \tag{11}$$

The first term is a constant and not influenced by $\mathbf{v}$. Thus to find $\mathbf{v}^*$ that minimizes the smoothness term, we write:

$$\min_{\mathbf{v}} \mathbf{v}^T \mathbf{L_{vv}} \mathbf{v} + 2 \mathbf{u}^T \mathbf{L_{uv}} \mathbf{v} \tag{12}$$

This is an unconstrained quadratic programming problem, with closed form solution [13]:

$$\mathbf{v}^* = \mathbf{L}^{\#}_{\mathbf{vv}} \left(-\mathbf{u}^T \mathbf{L_{uv}} \right)^T \tag{13}$$

where $\mathbf{L}^{\#}_{\mathbf{vv}}$ is the pseudo-inverse of $\mathbf{L_{vv}}$.

Because $\mathbf{v}^*$ only minimizes the second term in objective (10), we perform the following greedy procedure using $\mathbf{v}^*$ to reduce the overall objective function value: we iteratively try to install a "beneficial" component of $\mathbf{v}^*$ into the current vector $\mathbf{v}$—one that decreases the objective (10). We stop when no more beneficial component in $\mathbf{v}^*$ exists.

Pixels in frame t, $b(t)$, and MV $\mathbf{v}$ are alternately optimized using the two procedures described above, until the solution converges. Experimentation shows this only requires a few iterations in practice.

V. Sleep Event Detection

We now discuss how given the computed variances for the ellipse major and minor radius a and b, we can formulate a graph-based classification problem to identify if a new data sample is normal or abnormal breathing (for simplicity, we classify hypoponoea and apnoea into the same abnormal breathing category). First, we split the depth video sequence into 10-second windows. This size was chosen as a good trade-off between complexity and performance, since the respiratory rate of both normal breathing and overly-shallow-breathing (hypopnoea) are approximately 3 breaths per 10 seconds. For each 10-second window, we compute the variances of all ellipse parameters, *i.e.* $(\hat{a}_1, \hat{b}_1)$ and $(\hat{a}_2, \hat{b}_2)$, for the two ellipses corresponding to the chest and abdomen cross-sections of the patient as discussed in Section III-A. As training data, we assume the availability of a length-M sample vector $\mathbf{y}^o$ with ground truth classification—obtained using more intrusive sleep monitoring equipments, for example. Specifically, for each four-dimensional sample i, $y^o_i = [\hat{a}^i_1, \hat{b}^i_1, \hat{a}^i_2, \hat{b}^i_2]$, there is a classification $C(y^o_i)$ of y^o_i to $\{1, -1\}$, indicating the event of normal or abnormal breathing, respectively.

Let z be a new four-dimensional sample with no classfication yet. Further, let $\xi = [(\mathbf{y}^o)^T \ z]^T$ be a concatenation of classified sample vector $\mathbf{y}^o$ and unclassified sample z. We can treat ξ as a graph-signal: we draw an edge between any two

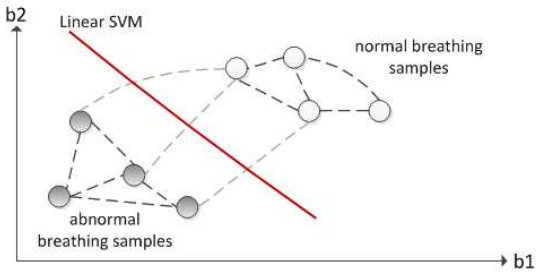


Fig. 4. Example graph construction given 8 classified training samples (4 normal, 4 abnormal). Linear SVM classifier is also shown.

samples (nodes) y_i^o and y_j^o (resulting in a complete graph), where the edge weight $w_{i,j}$ is:

$$w_{i,j} = \exp\left\{-\sum_{k=1}^{4} \frac{d_k \|x_k^i - x_k^j\|_2^2}{\sigma_c^2}\right\} \tag{14}$$

where x_k^i represents the four parameters $\hat{a}_1^i$, $\hat{b}_1^i$, $\hat{a}_2^i$ and $\hat{b}_2^i$ for $k = 1, \ldots, 4$ respectively, and d_k is a parameter weight. See Fig. 4 for an example graph constructed from eight two-dimensional samples. Edge weights connecting normal and an abnormal breathing samples will typically be small.

Given the defined edge weights, we can compute the adjacent and degree matrices, $\mathbf{A}$ and $\mathbf{D}$, as defined earlier, and subsequently the graph Laplacian $\mathbf{L}$ as well. Our objective is to find classification of unclassified sample z such that the graph variation term is minimized (*i.e.*, the graph-signal is smooth):

$$\min_{z} \ \xi^T \mathbf{L} \xi \tag{15}$$

Note that unlike ζ in (11), ξ is a vector of single-value samples (each with value either 1 or -1), and so taking the l_2-norm of the graph variation term is not necessary. The intuition of (15) is that classified samples in $\mathbf{y}^o$ with ellipse parameters close to unclassified sample z will have large edge weights $w_{i,j}$, and so a smooth graph-signal prior will ensure z to have similar classification as these similar samples in $\mathbf{y}^o$.

Like ζ in (11), because ξ is a concatenation of known and unknown signal samples $\mathbf{y}^o$ and z, we can similarly derive the optimal classification of z as:

$$z^* = \mathbf{L}_{zz}^{-1} \left(-(\mathbf{y}^o)^T \mathbf{L}_{\mathbf{y}^o z}\right)^T \tag{16}$$

where $\mathbf{L}_{zz}$ and $\mathbf{L}_{\mathbf{y}^o z}$ are the bottom-right and top-right quadrant of the Laplacian matrix respectively, as written in (11). In this case, $\mathbf{L}_{zz}$ is simply a scalar since z is a single sample. Because the classifier is restricted to be in the set $\{1, -1\}$, in practice we set z to be 1 if $z^* > 0$, and -1 otherwise. The magnitude $|z^*|$ can be interpreted as the confidence in the estimated classification.

A. Alternative Graph Formulation for Noisy Samples

If the training data samples $\mathbf{y}^o$ are noisy, we can add an additional fidelity term and optimize the entire sample vector ξ instead:

$$\min_{\xi} \|\mathbf{y} - \mathbf{y}^o\|_2^2 \ + \ \gamma \, \xi^T \mathbf{L} \xi \tag{17}$$

where γ is a parameter to trade off data fidelity term and the graph-signal smoothness prior; γ should set large if the data noise level is large. (17) is still an unconstrained quadratic programming problem, thus the optimal solution can be solved in closed form.

VI. Experimentation

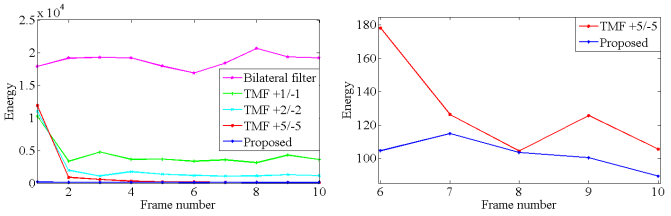
We now demonstrate the effectiveness of our proposed temporal denoising scheme for flicker removal and event detection scheme for apnoea / hypopnoea detection.

A. Experimental Setup

We captured depth videos of six patients diagnosed with *obstructive sleep apnoea* (OSA) [1], at Bondi Junction Private Sleep Laboratory in Sydney, Australia during October and November 2013. Besides our depth video capturing, each patient was connected to a professional-grade sleep monitoring system (expensive and intrusive) that measured various vital signs. This provided ground truth data for validation of our event detection results.

B. Experimental Results

First, we evaluate the performance of our proposed graph-based temporal denoising scheme in terms of flickering reduction. For comparison, we used a scheme that performs bilateral filtering [12], a non-iterative, local method on each frame, where each pixel is replaced by a weighted average of intensity from neighboring pixels. In particular, each weight is determined by a combination of both spatial-domain and intensity-domain Gaussian distributions. We also implemented an algorithm that performs motion estimation and temporal median denoising (TMF) separately similar to existing works such as [14].

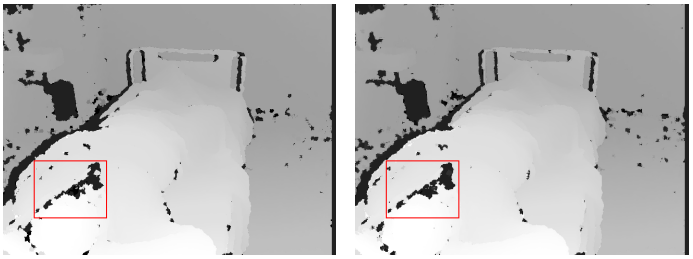


(a) energy vs. frame number (b) close-up of (a)

Fig. 5. Energy of the difference between two consecutive frames, where *+i/-i* denotes the number of future and previous depth images used for TMF.

Figure 5 shows the energy of the difference between two consecutive frames for our proposed scheme and the comparison scheme for the first 10 frames of an acquired sleep video sequence. We observe that our proposed scheme can more effectively reduce frame-difference energy, and thus flickering effects, over the comparison schemes, even if fewer number of frames were used in the processing window.

Figure 6 shows a sample depth frame before and after our proposed graph-based denoising. We observe that while our scheme reduces the flickering effect, it does not over-smooth and preserves sharp edges well.



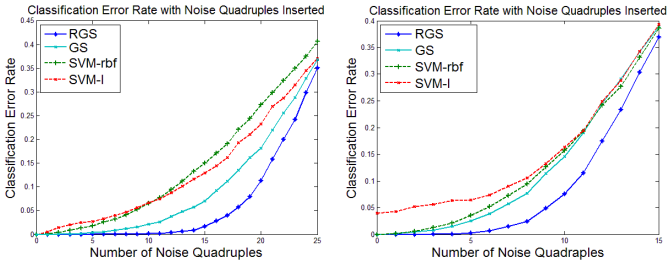
(a) before denoising (b) after denoising

Fig. 6. Two consecutive frames before and after denoising.

For sleep event detection, we compared our two proposed graph-based classification schemes, *graph smoothness* (GS) in (15) and *robust graph smoothness* (RGS) in (17) with two conventional implementations of SVM: linear kernel (SVM-l) and *radial basis function* kernel (SVM-rbf) available in Matlab, and used in our prior work [3]. The training dataset had 50 or 30 quadruples of ellipse parameters with ground-truth classification to normal / abnormal breathing. Additional 50 quadruples were used as test dataset.

We first tested different schemes using the original training dataset. As shown in Fig. 7, all four schemes achieved perfect classification for 50 training quadruples, while for 30 training quadruples, SVM-l had a 4% error rate.

Next we examined the robustness of the four methods. We added noise to the training dataset using the following procedure. Using a uniform distribution, we randomly mis-classified a subset of training data, and then ran the classification schemes on the noise-corrupted training set. For a given number of corrupted quadruples in the training set, we repeated the mis-classification procedure 2500 times and then computed the average. We observe that GS and RGS are more robust under noise in the training set than two SVM methods, and RGS is more noise-robust than GS, as expected. We conjecture that the reason why graph-based event detection is more robust than SVM in training data error is because a full graph considers correlation between every pair of data points, while an SVM only maximizes the distance between the classifier and the boundary data points of the two events.



(a) 50 training samples (b) 30 training samples

Fig. 7. Sleep Event Classification Results

VII. Conclusion

Leveraging on recent advances in graph signal processing, we propose two additions to improve our previously proposed sleep monitoring system based on depth video coding and analysis. First, towards the goal of temporal denoising of depth images, we express spatial and temporal smoothness of block motion vectors in graph-signal domain, and propose an efficient algorithm to denoise blocks and find motion vectors at the same time. Second, to classify new data sample into normal or abnormal breathing based on (possibly noisy) training data, we construct a graph with appropriate weights reflecting data sample similarities and minimize a graph variation term. Experimental results show that both graph-based temporal denoising and classification can outperform existing techniques in existing denoising and classification literature respectively.

References

[1] A. Malhotra and D. P. White, "Obstructive sleep apnoea," in *The Lancet*, vol. 360, no.9328, July 2002, pp. 237–245.

[2] P. P. et al., "Prospective study of the association between sleep-disordered breathing and hypertension," in *The New England Journal of Medicine*, vol. 342, no.19, May 2000.

[3] C. Yang, G. Cheung, K. Chan, and V. Stankovic, "Sleep monitoring via depth video recording & analysis," in *(accepted to) 5th IEEE International Workshop on Hot Topics in 3D (Hot3D)*, Chengdu, China, July 2013.

[4] 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.

[5] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] D. Falie, M. Ichim, and L. David, "Respiratory motion visualization and the sleep apnea diagnosis with the time of flight (tof) camera," in *1st WSEAS International Conference on Visualization, Imaging and Simulation*, Bucharest, Romania, November 2008.

[7] M.-C. Y. et al., "Breath and position monitoring during sleeping with a depth camera," in *International Conference on Heath Informatics*, Vilamoura, Portugal, February 2012.

[8] W. Hu, G. Cheung, X. Li, and O. Au, "Depth map compression using multi-resolution graph-based transform for depth-image-based rendering," in *IEEE International Conference on Image Processing*, Orlando, FL, September 2012.

[9] 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*, Pula, Italy, October 2013.

[10] A. Sandryhaila and J. Moura, "Classification via regularization on graphs," in *Symposium on Graph Signal Processing in IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Austin, TX, December 2013.

[11] J. Keikkila and O. Silven, "A four-step camera calibration procedure with implicit image correction," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 2007.

[12] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision*, Bombay, India, 1998.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, 2004.

[14] S. Matyunin, D. Vatolin, Y. Berdnikov, and M. Smirnov, "Temporal filtering for depth maps generated by kinect depth camera," in *3D TV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, Turkey, May 2011.