Unsupervised Subject Detection via Remote-PPG

Wenjin Wang, Sander Stuijk, and Gerard de Haan

***Abstract*—Subject detection is a crucial task for camera-based remote healthcare monitoring. Most existing methods in subject detection rely on supervised learning of physical appearance features. However, their performances are highly restricted to the pre-trained appearance model while still suffering from false detection of human-similar objects. In this paper, we propose a novel unsupervised method to detect alive subject in a video using physiological features. Our basic idea originates from the observation that only living skin tissue of a human presents pulse-signals, which can be exploited as the feature to distinguish human skin from non-human surfaces in videos. The proposed VPS method, named Voxel-Pulse-Spectral, consists of three steps: it (1) creates hierarchical voxels across the video for temporally parallel pulse extraction; (2) builds a similarity matrix for hierarchical pulse-signals based on their intrinsic properties; and (3) utilizes incremental sparse matrix decomposition with hierarchical fusion to robustly identify and combine the voxels that correspond to single/multiple subjects. Numerous experiments demonstrate the superior performance of VPS over a state-of-the-art method. On average, VPS improves 82.2% on the precision of skin-region detection; 595.5% on the Pearson correlation and 542.2% on Bland-Altman agreement of instant pulse-rate. ANOVA shows that in all-round evaluations, the improvements of VPS are significant. The proposed method is the first method that uses pulse to robustly detect alive subjects in realistic scenarios, which can be favorably applied for healthcare monitoring.**

***Index Terms*—Biomedical monitoring, remote sensing, photoplethysmography, face detection, object segmentation.**

I. Introduction

The task of detecting subjects in a video has been extensively studied in the past decades in the context of computer vision. In the emerging field of camera-based healthcare monitoring, there is a growing interest in applying subject detection to locate image region of living skin-tissue of a patient for clinical diagnosis, i.e., remote heart-rate measurement. Most existing works in subject detection exploit appearance features of human skin to discriminate between subject and background in a supervised training mechanism. However, a common problem faced by these methods is that their trained features are not unique to human beings; any feature that is similar to human skin can be misclassified. Moreover, supervised methods are usually restricted to prior-known samples and tend to fail when unpredictable samples occur, i.e., the Viola-Jones face detector trained with frontal faces cannot locate faces viewed from the side [1], while a skin classifier trained with bright skin fails with dark skin [2].

W. Wang and S. Stuijk are with the Electronic Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Einhoven, The Netherlands, e-mail: (W.Wang@tue.nl, S.Stuijk@tue.nl).

G. de Haan is with the Philips Innovation Group, Philips Research, Eindhoven, The Netherlands, e-mail: (G.de.Haan@philips.com).

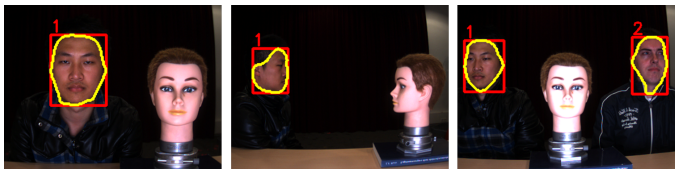

Fig. 1. An example of the detected alive subjects by the proposed VPS method, where (1) the real human face can been distinguished from an artificial face (found in yellow non-rigid contour); (2) multiple subjects can be differentiated from each other (identified in red bounding-box with an ID number).

Inspired by the recent progress in remote photoplethysmography (rPPG) [3]–[6], we observe that as compared to physical appearance features, the invisible physiological features (e.g., pulse) can better differentiate human skin from non-human surfaces. In the natural environment, only the skin tissue of an alive subject exhibits pulsatility, so any object showing no pulse-signal can be safely classified into the non-skin category. It prevents the false detection of objects with an appearance similar to human skin, as shown for example in Figure 1. Moreover, as compared to some high-level feature descriptors like Haar [1] and HOG [7], the patterns of a pulse-signal are more unitary and intuitively recognizable, i.e., all pulse-signals present a significant spectrum peak in certain frequency-bands. Essentially, it allows the unsupervised detection of human skin without training.

In this paper, we propose a novel method to detect alive subjects (e.g., living skin-tissue) in a video using the pulse as a feature. Given a video without any prior information related to the subjects (e.g., location, size, and number), our strategy is to first densely segment the whole video into hierarchical voxels in the spatio-temporal domain. Each voxel is considered as an independent pulse-sensor for temporally parallel pulse extraction without interference. Afterwards, a similarity matrix is built to describe the pairwise relationship of hierarchical voxels based on intrinsic properties (e.g., frequency and phase) of extracted pulse-signals. Since the voxels pointing at the same subject are mutually correlated in the similarity matrix, we develop an incremental sparse matrix decomposition algorithm to factorize and select the voxels corresponding to the skin-tissues of different subjects. Finally, hierarchical voxels are robustly fused into a single objectness map of human skin-tissues.

The key contributions of our work are two-fold: (1) we propose a similarity-based method that exploits the hierarchical voxel-based segmentation and intrinsic properties of human pulse for unsupervised alive subject detection; (2) we develop a spectral analysis algorithm to robustly decompose and update the similarity matrix in the temporal domain, which enables automatic subject number definition. The Voxel-Pulse-Spectral (VPS) method proposed in this work is the first complete solution that uses the pulse for unsupervised alive subject

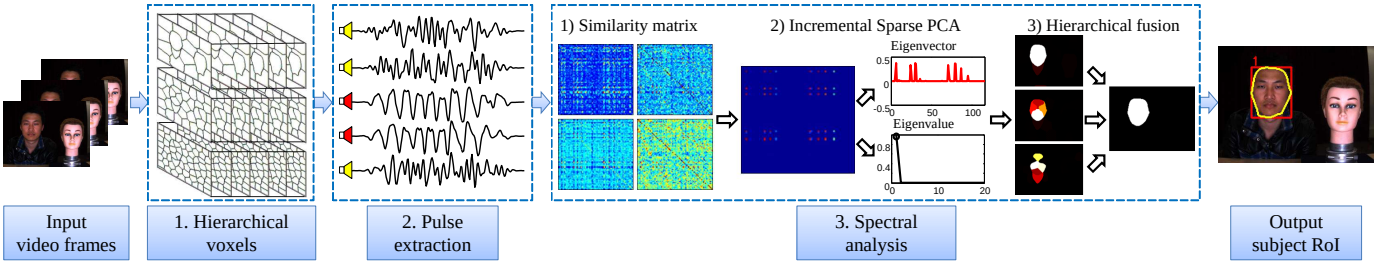

Fig. 2. The flowchart of the proposed VPS method: (1) it takes an input video and constructs the hierarchical voxels across the video frames; (2) each voxel simulates an independent pulse-sensor in a parallel pulse extraction process; and (3) all voxels in the hierarchy are pairwisely connected in a similarity matrix based on the measured pulse, while the sparse similar entries denoting the voxel connections are incrementally factorized and fused into a human objectness map. Finally, it outputs the detected non-rigid subject RoI.

detection in videos considering realistic challenges. It has been thoroughly evaluated with numerous challenging videos and demonstrates robustness to practical challenges, i.e., body-motion, skin-tone, etc. The state-of-the-art performance of VPS indicates that it can be applied in a large-scale camera-based healthcare monitoring system that requires automatic alive subject detection or living skin-tissue detection, i.e., the remote monitoring of heart-rate, SPO2, respiration, etc.

II. Related work

A. Camera-based pulse extraction

In the human cardiovascular system, blood pulse propagating throughout the body changes the blood volume in skin tissue. Since the optical absorption of hemoglobin in blood varies across the light spectrum, detecting color variations of skin reflection can reveal the pulse-rate [5]. Recent remote photoplethysmography (rPPG) techniques demonstrate encouraging results by detecting pulse-induced color variations on human skin using a regular RGB camera. In 2008, Verkruysse *et al.* found that in an ambient light condition, the PPG-signal has different relative amplitudes in the RGB channels of human skin-pixels [6]. Based on this finding, Blind Source Separation methods (e.g., PCA-based [4] and ICA-based [5]) were proposed to independently factorize the temporal RGB signals for finding the pulse. In 2013, de Haan *et al.* introduced a Chrominance-based rPPG method to define the pulse as a linear combination of RGB channels under a standardized skin-tone assumption [3], which is one of the most accurate rPPG methods in dealing with realistic challenges (e.g., various skin-tones). More recently, Wang *et al.* proposed a complete framework to significantly improve the motion robustness of rPPG [8], which profits from the spatially redundant pixels of a camera sensor. Nevertheless, all these rPPG methods rely on a pre-defined skin area (e.g., face) for pulse extraction.

B. Pulse-based RoI detection

Given the fact that the human pulse can be measured by rPPG in videos, the pulse-signal can thus be used to assist the subject detection, i.e., detecting alive subjects by locating their living skin tissue. In 2013, Gibert *et al.* proposed a face detection method based on the pulse-signal [9]. This method slices the video into fixed rigid-grids for local pulse extraction. It sets a hard threshold to find the grids with high spectrum energy and label them as the face region. It is limited to videos in which the stationary face needs to be placed at a pre-defined distance from the camera. Our VPS method does not suffer from these limitations. Meanwhile, Lempe *et al.* presented a Region of Interest (RoI) selection method on the face to enhance the rPPG monitoring [10]. However, their RoI is constrained to pre-defined facial landmarks, which is not a general solution for subject detection, i.e., it cannot detect other body parts (e.g., hands) that might be visible in a video. In contrast, our VPS method does not make such an assumption and can detect all body parts with pulsatile blood.

III. Voxel-Pulse-Spectral (VPS) method

The overview of the proposed VPS method is shown in Figure 2, which takes an input video and outputs the subject RoI. There are three main steps in the flowchart: hierarchical voxels, pulse extraction and spectral analysis. Each step is discussed in detail in the following subsections.

A. Hierarchical voxels

Given a video without any prior information about the subject, our strategy is to first segment the video into dense local regions where a pulse can be independently measured. To some extent, such strategy has already been exploited in [9] by slicing the video into fixed rigid-grids. However, the subject size is quantized by the grid geometry, which fails when the subject is small or when there is body motion. Therefore, we propose to use a superior video segmentation method for pulse extraction called *hierarchical voxels*.

In our method, the hierarchical voxels consist of spatio-temporally coherent clusters in multiple scales, where pixels sharing appearance and spatial similarities in the temporal domain are grouped together. Starting from one scale, constructing the voxels is defined as the procedure of minimizing the chromatic energy E_c and spatial-distance energy E_s between adjacent pixels in a short interval $T \in \{2n+1, n \in \mathbb{N}^+\}$ [11] as:

$$\arg\min \Big(\sum_{t-\frac{T-1}{2}}^{t+\frac{T-1}{2}} \sum_{p \in P(t)} (1-\lambda) E_c^t(p,k) + \lambda E_s^t(p,k) \Big), \quad (1)$$

where $p \in P(t)$ is the set of pixels in the t-th frame. The representation of p is a 4-dimensional feature vector (x, y, u, v), where (x, y) and (u, v) are respectively the coordinates in the image plane and the chromatic plane (e.g., UV plane of

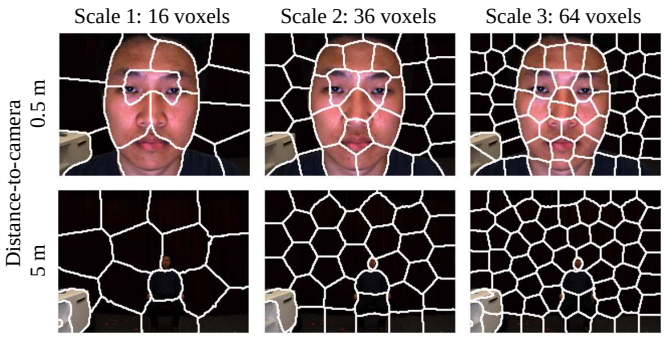

Fig. 3. An example of hierarchical voxels segmentation in videos with a subject at different distances to the camera. It consists of three scales of voxels with $k = 16, 32, 64$.

YUV space, the empirical space for skin segmentation). K-means clustering is performed to assign pixels into k clusters for minimizing the total energy during T. λ is the parameter controlling the balance between two energy terms.

Furthermore, the single scale voxels are extended to multiple scales by initializing different k in Eq. 1 simultaneously, where each scale is an independent clustering. Considering that the voxels in separate scales have different resolutions and energy variations, the λ_i in i-th scale is adaptively self-tuned based on its own energy variations at t as:

$$\lambda_i^t = \log(k)\sqrt{\frac{\sigma\big(\phi(u_i^t)\big)\cdot\sigma\big(\phi(v_i^t)\big)}{\sigma\big(\phi(x_i^t)\big)\cdot\sigma\big(\phi(y_i^t)\big)}}, \tag{2}$$

where $\sigma(\cdot)$ denotes the standard deviation operator; $\phi(\cdot)$ represents the set of cluster means; $\log(k)$ controls the voxel compactness, i.e., voxels with higher resolution (larger k) should be more compact. The real-time tuning of λ in different scales avoids volatile and flickering clustering, which preserves the fine-grained segmentation.

To our purpose, there are four benefits of using hierarchical voxels for video segmentation, as illustrated in Figure 3: it (1) establishes the spatio-temporally coherent "tubes" for pulse measurement; (2) enables the scale-invariant subject detection in a video, where the motion of a subject can be quantized by voxels with different resolutions in multiple scales; (3) maintains high boundary recall of subject shapes; and (4) creates a statistical observation of skin-regions, since the pulse measured from voxels with different resolutions have different quantized qualities.

B. Pulse extraction

Each voxel in the hierarchy is assumed to be an independent pulse-sensor in parallel pulse extraction. Based on our study in the state-of-the-art rPPG methods, we rely on the Chrominance-based method (CHROM) [3] for pulse measurement. Different from CHROM that uses the spatially averaged RGB of *all* pixels to derive the pulse-signal in a local region, we combine the pixel RGB values in a voxel by *weighting* them based on their distance to the voxel boundary, i.e., pixels close to the voxel boundary are less reliable due to occasional jittering artifacts between neighboring voxels and thus should be less weighted. Assuming that the closest Euclidean distance from a pixel k to the voxel boundary is d_k, the average RGB of j-th voxel in i-th scale at t is combined as:

$$(\bar{R}_{ij}^t, \bar{G}_{ij}^t, \bar{B}_{ij}^t) = \frac{\sum_{k=0}^{N}\big(d_k\cdot(R_{ijk}^t, G_{ijk}^t, B_{ijk}^t)\big)}{\sum_{k=0}^{N} d_k}, \tag{3}$$

where N denotes the number of pixels in j-th voxel. In a constant lighting environment, human skin tissue shows the same relative PPG-amplitude, but the chromatic differences in voxels lead to the variations in pulse-amplitudes. So different from [3], we use the temporal derivatives of average RGB in a voxel, i.e., $dC_{ij}^t = C_{ij}^t - C_{ij}^{t-1}, C \in \{\bar{R}, \bar{G}, \bar{B}\}$, to derive its chrominance-signals. In the interval T (defined in Eq. 1), the normalized chrominance derivatives are calculated as:

$$\begin{cases} \overrightarrow{dX}_{ij}^T = 3\frac{\overrightarrow{d\bar{R}}_{ij}^T}{\sum_{t=0}^{T} d\bar{R}_{ij}^t} - 2\frac{\overrightarrow{d\bar{G}}_{ij}^T}{\sum_{t=0}^{T} d\bar{G}_{ij}^t} \\ \overrightarrow{dY}_{ij}^T = 1.5\frac{\overrightarrow{d\bar{R}}_{ij}^T}{\sum_{t=0}^{T} d\bar{R}_{ij}^t} + \frac{\overrightarrow{d\bar{G}}_{ij}^T}{\sum_{t=0}^{T} d\bar{G}_{ij}^t} - 1.5\frac{\overrightarrow{d\bar{B}}_{ij}^T}{\sum_{t=0}^{T} d\bar{B}_{ij}^t} \end{cases}, \tag{4}$$

where $(dR_{ij}^t, dG_{ij}^t, dB_{ij}^t)$ denotes the temporal derivatives of RGB in a voxel between two frames. The chrominance derivatives estimated in each interval are linearly combined into pulse derivatives and further integrated. Afterwards, different pulse intervals are overlap added to a complete pulse-signal $\overrightarrow{S}_{ij}^L$ with length L. This procedure is interpreted as:

$$\overrightarrow{S}_{ij}^L = \sum_{t=0}^{L-T+1} \overrightarrow{S}_{ij}^{t+T} + w\cdot\textbf{csum}(\overrightarrow{dX}_{ij}^T - \frac{\sigma(\overrightarrow{dX}_{ij}^T)}{\sigma(\overrightarrow{dY}_{ij}^T)}\overrightarrow{dY}_{ij}^T), \tag{5}$$

where **csum**$(\cdot)$ denotes the cumulative sum of temporal derivative signals; w is the Hanning window for smoothing the overlap adding [3]. Consequently, the parallel extracted pulse-signals $\overrightarrow{S}_{ij}^L$ (from j-th voxel in i-th scale) are centralized and normalized as:

$$\overrightarrow{S}_{ij}^L = \frac{\overrightarrow{S}_{ij}^L - \mu(\overrightarrow{S}_{ij}^L)}{\sigma(\overrightarrow{S}_{ij}^L)}, \tag{6}$$

where $\mu(\cdot)$ denotes the averaging operation. Note that the pulse-signal is the only feature used in our method. No other appearance features like color or texture are used.

C. Spectral analysis

After extracting the pulse-signals from hierarchical voxels, a more critical question concerning our task is: how to effectively exploit the pulse-signal as a feature to distinguish skin and non-skin in an unsupervised manner? We observe that the pulse-signals extracted from the skin-regions belonging to the same subject share *similarities* in many aspects such as phase and frequency, whereas the ones extracted from non-skin regions (e.g., background) are irregular noises without correlation. Therefore, we propose to use the pairwise similarities of pulse-signals to find alive subjects. This is also applicable to the case of multiple subjects, because the pulse measured from different subjects can be differentiated in phase and frequency as well.

1) Similarity matrix: In this step, we create a similarity matrix $\Sigma = (D, C)$ to interconnect the hierarchical voxels based on the measured pulse. In Σ, the entries D in the diagonal trace contain all voxels in different scales; the remaining entries C denote the pairwise connection between any pair of voxels.

To build such a similarity matrix, the distance metric for measuring the pulse similarity needs to be defined. The most commonly used distance metrics, i.e., L1 and L2 distances, are not applicable to the pulse-feature. However, compared to other appearance features (e.g., Haar and HOG), we notice an essential and unique character in the pulse-feature: it contains *periodicity*.

According to our observation, the pulse-signals from the same subject show the following relations: (1) they have similar frequency and thus their cross-correlation presents a significant spectrum peak; (2) they have no significant phase shift; (3) their frequency correlation is regular and less disordered; and (4) if considering pulse-signals as multi-dimensional vectors, the included angle between two similar vectors is small. Therefore, we propose a new distance metric to build the similarity matrix for pulse-signals by emphasizing above connections, which is composed of four different measurements:

- **Spectrum peak** In the frequency domain, we define a pulse-rate band $f \in [40, 240]$ BPM (Beats Per Minute) for voxels to communicate, which is a broad range for healthy subjects including neonates and sporting subjects. The spectrum peak of two cross-correlated pulse-signals is defined as:

$$F = \arg \max_{f \in [40,240]} \big(\mathcal{F}(\overrightarrow{S}^L_{ij}) \circ \mathcal{F}(\overrightarrow{S}^L_{i'j'})^*\big), \tag{7}$$

where $\circ$ denotes the element-wise product; * is the conjugation; $\mathcal{F}(\cdot)$ represents the Fast Fourier Transform (FFT).

- **Spectrum phase** Two similar pulse-signals are also in the same phase, so their normalized cross-correlation should show a strong response in the time domain as:

$$P = \max\big(\mathcal{F}^{-1}(NCC)\big), \tag{8}$$

with

$$NCC = \frac{\mathcal{F}(\overrightarrow{S}^L_{ij}) \circ \mathcal{F}(\overrightarrow{S}^L_{i'j'})^*}{\big\|\mathcal{F}(\overrightarrow{S}^L_{ij}) \circ \mathcal{F}(\overrightarrow{S}^L_{i'j'})^*\big\|_2}, \tag{9}$$

where $\|\cdot\|_2$ is the L2-norm; $\mathcal{F}^{-1}(\cdot)$ denotes the inverse FFT.

- **Spectrum entropy** We use the term "entropy" to measure the regularity of correlation between two pulse-signals as:

$$E = \frac{\sum_{f=40}^{240} NCC(f) \log\big(NCC(f)\big)}{\log(240 - 40)}, \tag{10}$$

where the interpretation of E is consistent with the other measurements, i.e., larger E denotes better correlation.

- **Inner product** In the time domain, we use the inner product to measure the cosine angle between two pulse-signals as:

$$I = < \frac{\overrightarrow{S}^L_{ij}}{\|\overrightarrow{S}^L_{ij}\|_2}, \frac{\overrightarrow{S}^L_{i'j'}}{\|\overrightarrow{S}^L_{i'j'}\|_2} >, \tag{11}$$

where $<,>$ denotes the inner product operation.

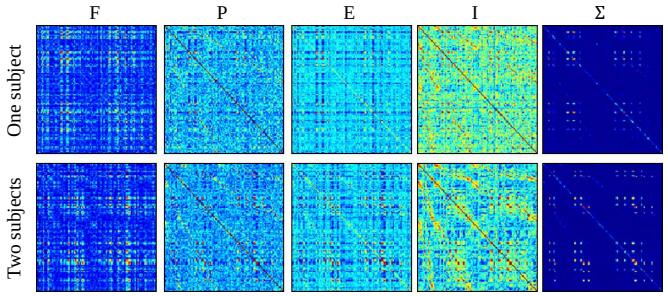

Fig. 4. An example of four measurements and their fused similarity matrix Σ. The entries with higher energy represent the index of similar voxels in the hierarchy.

Finally, these four measurements are normalized to the range $[0, 1]$ and fused together with a Gaussian kernel as:

$$\Sigma = 1 - \exp\big(-\frac{(F \circ P \circ E \circ I)^2}{2\sigma^2_{I,F,P,E}}\big), \tag{12}$$

where $\sigma_{I,F,P,E}$ represents the entry-wise standard deviation between four matrices. Note that the four measurements are not completely independent from each other, the redundancy between measurements is beneficial for reducing the uncertainty in similarity estimation. Figure 4 shows an example of four measurements and their fused similarity matrix. In our distance metric, two well-aligned pulse-signals show boosted frequency energy during the cross-correlation, which can effectively suppress the noise entries (e.g., voxels without pulse). In contrast, previous distance metrics are all objective measurements that cannot enhance the connection between similar entries in the comparison.

In the end, all voxels in the hierarchy are mutually connected in the similarity matrix. The task of detecting an alive subject in voxels can be reformulated as finding a subspace partition of the similarity matrix such that the entries in the same subspace have identical similarity direction.

2) Incremental sparse matrix decomposition: In fact, the similarity matrix Σ can be interpreted as a linear combination of $\lambda_1 x_1 x_1^\top + \lambda_2 x_2 x_2^\top + ...\lambda_n x_n x_n^\top$, where $x_i \in X$ is a set of orthogonal vectors in the multi-dimensional space. In order to find the voxels belonging to the same subject, we use the matrix decomposition technique to factorize Σ into X, where different subjects are separated into different eigenvectors. Since Σ is a sparse matrix with many zero entries (e.g., the voxels pointing at background share no similarity), we apply the Sparse PCA [12] to decompose Σ into X by seeking a trade-off between expressive power and data interpretability. The Sparse PCA finds the first sparse eigenvector with the maximum variance in Σ by optimizing the following non-convex objective function:

$$\arg\max_X (X^\top \Sigma X) \text{ subj.to } \|X\|_2 = 1, \|X\|_1 \leq n, \tag{13}$$

where $\|\cdot\|_1$ is the L1-norm; $n > 0$ controls the cardinality of X. However, computing sparse eigenvectors with maximum variance is a combinatorial problem and numerically hard to solve, so we drop the non-convex rank constraint in Eq. 13

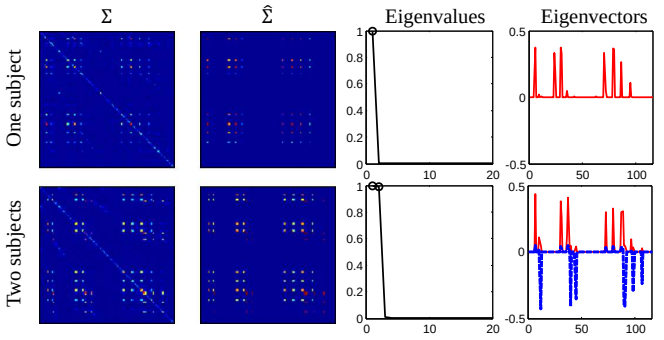

Fig. 5. An example of similarity matrix decomposition using incremental sparse PCA, where similar voxels are factorized into the same direction in the selected eigenvectors.

following the lifting procedure for semidefinite relaxation with l_1 penalization [12] as:

$$\arg\max_{\hat{\Sigma}} \mathbf{Tr}(\Sigma\hat{\Sigma}) - \rho\|\hat{\Sigma}\|_1 \text{ subj.to } \mathbf{Tr}(\hat{\Sigma}) = 1, \hat{\Sigma} \succeq 0, \quad (14)$$

where $\mathbf{Tr}(\cdot)$ denotes the matrix trace operation; $\rho > 0$ controls the sparsity; $\hat{\Sigma} = XX^\top$ is a symmetric matrix approximated by the first leading eigenvector. At this point, we adopt a recent algorithm named Hybrid Conditional Gradient Smoothing (HCGS) [13] to solve Eq. 14. The merit of HCGS is the fast convergence in convex relaxation using conditional gradient approaches.

However in practice, Σ may consists of *multiple* sparse eigenbasis in case of multiple subjects, whereas Eq. 14 only promotes the sparsity in the first leading eigenvector. To address this issue, we estimate the succeeding sparse eigenvectors x_i by sequentially deflating Σ with preceding sparse eigenvectors $x_1, x_2, ..., x_{i-1}$ using Hotelling's deflation as:

$$\Sigma_i = \Sigma_{i-1} - (x_i^\top \Sigma_{i-1} x_i) x_i x_i^\top, \ i \in [1, m], \quad (15)$$

with

$$m = \arg\max_i \left(\frac{x_{i-1}^\top \Sigma_{i-1} x_{i-1}}{1 + x_i^\top \Sigma_i x_i}\right), \quad (16)$$

where $x_i \in X$ can be derived by the power iteration [13]; m is the automatically found number of most expressive eigenvectors, which also implies the number of subjects in a video, i.e., m is usually found at the largest eigenvalue gap. Figure 5 shows an example of the factorized and selected eigenbasis from a similarity matrix: the noisy entries in the original Σ are eliminated in $\hat{\Sigma}$; the eigenvalues clearly show the number of most expressive eigenvectors in $\hat{\Sigma}$.

As a matter of fact, some intrinsic (e.g., pulse-rate variation) and extrinsic (e.g., luminance changes) factors may occasionally change the similarity matrix in subsequent frames, which leads to an instability of the sparse eigenvectors estimated from each single frame. To solve this problem, we employ the incremental subspace updating [14] to smoothly adapt the $x_i \in X$ to real-time changes in the time domain. Basically, it considers the time-varying similarity matrix $\hat{\Sigma}_{new}$ as a new observation, and use multiple observations $[\hat{\Sigma}_{old}, \hat{\Sigma}_{new}]$ from different frames to enrich the subspace model as:

$$[U, D, V] = \mathbf{SVD}([\hat{\Sigma}_{old}, \hat{\Sigma}_{new}]), \quad (17)$$

where $\mathbf{SVD}(\cdot)$ denotes the Singular Value Decomposition; U and D are incrementally updated eigenvectors and eigenvalues. Eventually, we arrive at an algorithm to incrementally estimate multiple sparse eigenvectors from a time-varying similarity matrix, as shown in Algorithm 1.

3) Hierarchical fusion: By projecting the estimated sparse eigenvectors onto the hierarchical voxels, we obtain the voxel-based human objectness map in multiple scales, where each scale has a different quantized description to the subject, as illustrated in Figure 6: the eigenvector not only decides the subject direction (sign) in subspaces, but also decides the pulsatility (amplitude) of corresponding skin-regions, i.e., forehead and cheek show relatively high pulsatility in projection, which aligns with the findings in [6].

The final step is to fuse multiple objectness maps into a single output. Due to the fact that hierarchical measurement creates a statistical observation for skin-regions, our basic idea is exploiting this redundancy to derive a single output for which all scales have the highest agreement. In this sense, the hierarchical fusion is cast as the energy minimization between multiscale objectness maps as:

$$\arg\min_{\hat{o}} (\gamma E_1 + (1-\gamma) E_2), \quad (18)$$

with

$$\begin{cases} E_1 = \sum_i \sum_j \|o_{ij}, \hat{o}\|_2 \\ E_2 = \sum_i (\sum_j \underset{o_{ij} \subseteq o_{i-1,j}}{\|o_{ij}, o_{i-1,j}\|_2} + \sum_j \underset{o_{i+1,j} \subseteq o_{ij}}{\|o_{ij}, o_{i+1,j}\|_2}) \end{cases}, \quad (19)$$

Algorithm 1 Incremental Sparse PCA

Input: similarity matrix $\Sigma \in \mathbb{R}^{n \times n}$, eigenvectors U, eigenvalues D

$\rho = 0.25$ (sparsity), $N = 100$ (iteration times)
for $k = 1, 2, ..., N$ **do**
  $\beta_k = \frac{1}{n}\mathbf{Tr}(\Sigma\Sigma_k) + \frac{\rho}{n}\|\Sigma_k\|_1$
  $Z_k = \Sigma - \frac{\sqrt{k}\rho}{2\sqrt{2}}\Sigma_k + \frac{\sqrt{k}\rho}{2\sqrt{2}}\text{sign}(\Sigma_k) \odot (|\Sigma_k| - \frac{2\sqrt{2}}{n\sqrt{k}})$
  $X = \{x_1, x_2, ..., x_m\} \leftarrow$ multiple eigenvectors of Z_k, where m is determined by Eq. 15 and Eq. 16
  $\hat{\Sigma} = XX^\top$
  $\Sigma_{k+1} = (1 - \frac{2}{k+1})\Sigma_k + \frac{2}{k+1}\hat{\Sigma}$
  if $\frac{|\beta_k - \beta_{k-1}|}{\beta_k} < 10^{-3}$ **then**
    break
  end if
end for
if $U, D == 0$ **then**
  $[U, D, V] = \mathbf{SVD}(\hat{\Sigma})$
else
  $[U, \hat{\Sigma}^{'}]R = \mathbf{QR}([UD, \hat{\Sigma}]) \leftarrow$ solved by QR decomposition
  $[U^{'}, D^{'}, V^{'}] = \mathbf{SVD}(R)$
  $U_m^{'} \in U^{'}$, $D_m^{'} \in D^{'} \leftarrow$ select top m eigenvectors and eigenvalues, where m is determined by Eq. 16
  $U = \text{sign}(U) \odot |[U^{'}, \Sigma_{new}^{'}]U_m^{'}|$, $D = D_m^{'} \leftarrow$ update subspace model
end if

Output: updated U and D

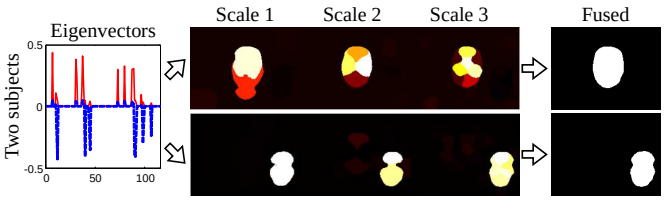

Fig. 6. The selected sparse eigenvectors are projected onto the hierarchical voxels and further fused into a single human objectness map, where different subjects are isolated.

where o_{ij} corresponds to the objectness value of j-th voxel in i-th scale that determined by the eigenvector elements; $\hat{o}$ denotes the fused objectness map; γ controls the balance between two energy terms. In Eq. 19, E_1 minimizes the energy between inputs and output, while E_2 minimizes the energy between spatially overlapped voxels in different scales, i.e., an implicit tree structure. Figure 6 shows an example of the fused result in a video with two alive subjects.

IV. Experiments

A. Benchmark dataset

A benchmark dataset consisting of 40 video sequences has been built to evaluate the proposed VPS method. The videos are recorded by a regular RGB camera [1] in an uncompressed data format, at a frame rate of 20 Hz, 768×576 pixels, 8 bit depth. During recordings, the subjects, staying in front of the camera with skin visible, are illuminated by the constant office light. In each video, 400 frames are used for evaluation and thus **16000** frames are measured in total.

In order to assess the method's robustness to realistic challenges, our recordings simulate 10 different challenging scenarios as described below (the bold number in bracket denotes the number of frames simulated for this challenge):

- **Skin-tone (2400)** Six subjects with different skin-tones are recorded, i.e., participants are from West Europe, East Asia, Sub-Sahara Africa and India.
- **Scale (2000)** Subject has five distances to the camera during recordings, i.e., 0.5 m, 1 m, 1.5 m, 3 m and 5 m.
- **Motion (2000)** Subject performs five basic types of head motion during recordings, i.e., translation, rotation, scaling, non-rigid talking and mixture of all motions.
- **Position (1600)** Subject has four different 2D positions (horizontal and vertical) in videos, i.e., top, bottom, right and left.
- **Posture (1200)** Subject has three different postures in videos, i.e., sitting, lying and standing.
- **Occlusion (1200)** Subject has different parts of face occluded by non-human objects (e.g., book and cloth).
- **Fake (2000)** Non-human objects that could be falsely recognized as human are recorded together with subject, i.e., artificial face, luminance source with flickering frequency.
- **Background (800)** Subject is recorded in both the colorful-clutter background and skin-similar background.
- **Body-part (1600)** Different body parts are recorded, i.e., both the frontal and back sides of the palm and arm.
- **Multi-subjects (1200)** Different number of subjects are recorded in the same video.

[1]The global shutter RGB CCD camera with type USB UI-2230SE-C of IDS.

B. Evaluation metrics

Since the output of VPS is a binary human objectness map, the Ground-Truth (GT) for each video is also a binary video sequence containing *non-rigid* human RoI, which is created by machine-assisted manual annotation (e.g., assisted by object tracker and skin classifier). The performance of VPS is measured from the following two aspects.

Overlap region The overlap region between the RoI obtained by VPS and GT is defined as a ratio:

$$r = \frac{\text{Area}(\text{ROI}_{\text{GT}} \cap \text{ROI}_{\text{VPS}})}{\text{Area}(\text{ROI}_{\text{GT}} \cup \text{ROI}_{\text{VPS}})} \in [0, 1] \tag{20}$$

The precision of overlap region in a video is represented by the success rate, i.e., the percentage of frames where r the exceeds a threshold $t \in [0, 1]$. Consequently, we use the Area Under Curve (AUC) of precision curve and the precision at $t = 0.5$ to show the results.

Instant pulse-rate When the RoI obtained by VPS and GT are well-aligned, the extracted pulse-signals should be very similar. So we apply the Pearson correlation (e.g., r-value) and Bland-Altman (e.g., 1.96σ)[2] metrics to show the agreement between VPS and GT on instant pulse-rate. The instant pulse-rate, defined as the inverse of the peak-to-peak interval of the pulse-signal, is derived by a peak detector in the time-domain and smoothed by a 3-point mean filter. It can capture the instantaneous change of the pulse and reflect the real-time differences between two signals.

C. Compared methods

We compare the proposed VPS method to the most recent FDR method [9], named the "Face Detection based on RPPG". Both methods are implemented by us in C++ using the OpenCV 2.4 library [15] and ran on a laptop with an Intel Core i7 processor (2.70 GHz) and 8 GB RAM. The parameters in FDR are remained identical to [9], while the parameters in VPS are empirically defined as: (1) 3 scales segmentation with $k = 16$, 36 and 64 in Eq. 1; (2) T in Eq. 1 is 3; (3) L in Eq. 5 is 128; and (4) γ in Eq. 18 is 0.5. For fair comparison, all the parameters are fixed without tuning when processing different videos.

V. Results and Discussion

A. Comparison per challenge

The experimental results obtained by VPS and FDR in 10 challenging categories are shown in Figure 7 and summarized in Figure 8. In the "scale", "posture" and "position" categories where influences are from the subject spatial location, VPS demonstrates the superior performance over FDR by achieving the highest score in each evaluation, i.e., when t is set to 0.5 in comparing overlap regions, VPS obtains 98.8%, 99.1% and 100.0% precision against FDR's 26.7%, 6.0% and 7.5%. It shows that FDR fails with videos where the subject has variant spatial positions or postures, whereas the hierarchical segmentation in VPS allows the subject to have different

[2]Coefficient of repeatability (1.96σ) is inverted to remain the interpretation consistent with other metrics.

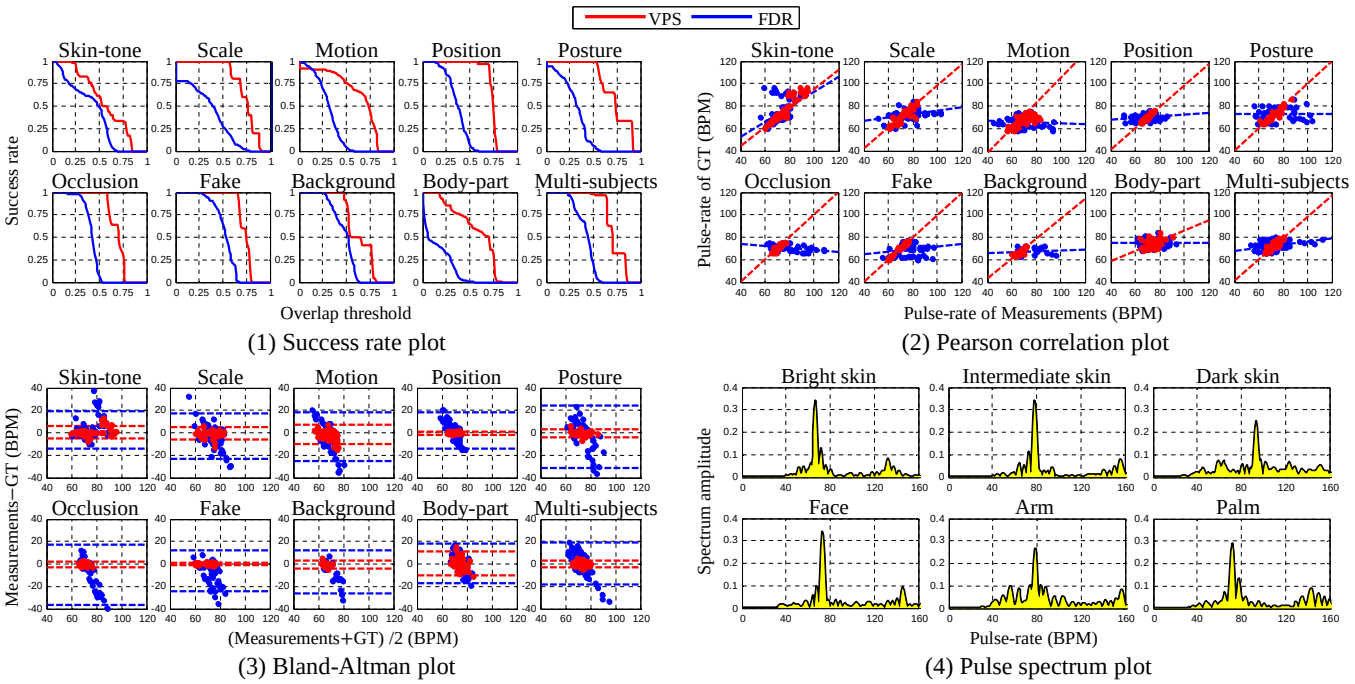

Fig. 7. Results obtained by VPS and FDR in 10 challenge categories, which are measured by (1) success rate of overlap region; (2) Pearson correlation of instant pulse-rate, where bold lines denote the regression coefficients; and (3) Bland-Altman agreement of instant pulse-rate, where bold lines denote the coefficient of repeatability. (4) shows the spectrum amplitude of pulse extracted from both the subjects with different skin-tones and different body-parts in a single subject.

distances to the camera, spatial positions or postures when being recorded. In the "scale" category, we notice that the longer distance between subject and camera actually increases the uncertainty for detection, which is mainly due to the fact that less skin reflections can be received by the remote camera.

In the "occlusion", "fake" and "background" categories where influences are from the natural environment, both methods show favorable performance in the AUC of overlap region precision. Since pulse-signal is the only feature used in both methods, non-human objects showing no pulsatility (e.g., artificial face, occluded book and skin-similar background) can be safely excluded. Although the color lamp in the "fake" category also exhibits flickering frequencies, its color changes do not align with the pulsatile direction in RGB space and thus can be eliminated. In the instant pulse-rate evaluation, VPS significantly outperforms FDR by showing strong correlation and high agreement with ground-truth, i.e., in Pearson correlation, VPS gains 0.89, 0.99 and 0.85 correctness against FDR's -0.43, 0.19 and 0.14. It shows that FDR almost has no correlation with ground-truth signals. This can be explained by the RoI's temporal consistency: the variant shapes and locations of RoI estimated by FDR in subsequent frames leads to abrupt changes in the instant pulse-rate. This problem has been solved in VPS by using the incremental subspace updating.

The temporal consistency maintained by VPS shows extraordinary benefits in the "motion" category, i.e., in Pearson correlation, VPS achieves 0.75 correctness against FDR's -0.10. First, the voxel-based segmentation in VPS strengthens the temporal coherence of space-time "tubes" for pulse extraction. Second, the hierarchical segmentation can quantize the subject motions with different extent, i.e., vigorous body motion can be captured by voxels with lower resolutions. Third, the incremental subspace updating in similarity factorization rejects inconsistent motion-induced outliers. In comparison, FDR completely fails with the subject with even slight body motions. However, a limiting factor for voxel-based motion tracking could be the "subject size". For example, if a subject is too far away from the camera (e.g., 10 meters) and moving in high speed, the segmentations in all scales cannot quantize such motion and thus not voxel can track this subject. In this case, a higher resolution camera is preferred.

Besides, it has to be noted that VPS works properly when the pulse-rate of a subject is changed during the monitor-

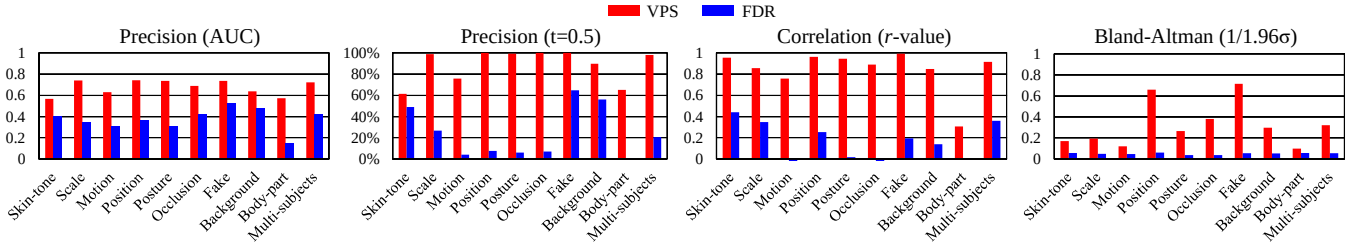

Fig. 8. The categorical comparison between VPS and FDR using four evaluation criteria, where larger values denote the better performance.

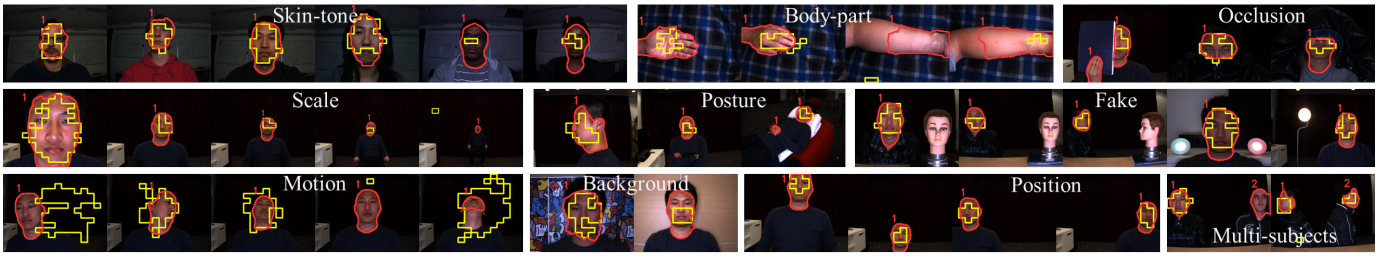

Fig. 9. The snapshot of detected alive subject RoI in benchmark videos using the proposed VPS method (red non-rigid contours) and FDR method (yellow rigid grids). The red ID number denotes the subjects' identity found by VPS.

ing (e.g., while doing gymnastic exercises). This is because humans only have one cardiovascular system. Therefore the pulse-rate changes simultaneously in all skin-regions of one subject. Since VPS is essentially based on pulse similarities, skin-regions showing similar pulses at a certain moment are mutually connected in the similarity matrix, i.e., it does not matter how pulses are changed, once they are changed together.

"Skin-tone" and "body-part" are the two most challenging categories where both methods show declined performance compared to other categories. This is due to the fundamental physiological limitations in color-based rPPG: (1) dark skin has higher melanin contents than bright skin. The higher spectral absorption of melanin contents in dark skin limits the amount of light entering the deeper skin layer with pulsatile blood vessels; (2) the skin tissue in different body parts show different portions of pulsatile blood volume, which is mainly due to the difference in skin composition, i.e., the skin-regions with more fat are more difficult to be detected. As a proof, we show the pulse frequency spectrums measured from the subjects with different skin-tones and also the different body-parts of a single subject in Figure 7 (4): it is apparent that the spectrum of dark skin and arm are more noisy. In Figure 9, we also notice that the palm is easier to be detected than the arm, and only the skin-regions close to the wrist in an arm can be found with pulse. Although both methods suffer from performance degradation in these two categories, VPS is still clearly much better. Especially for subjects with dark skin, the similarity-based measurement in VPS effectively boosts the voxels containing weak pulse-signals while suppressing the noisy ones.

In the "multi-subjects" category, VPS again obtains superior results in comparison, i.e., in Pearson correlation, it improves FDR's 0.36 correctness to 0.91. Although FDR can roughly find multiple subjects via thresholding the grids with higher spectrum energy, it only finds the regions with living beings instead of identifying them. In contrast, our VPS method, factorizing the different similarities of pulse-features into independent directions, can identify/separate different subjects with even similar pulse-rates, i.e., subtle phase shift in two pulse-signals can be revealed by inner product metric in VPS. Figure 9 shows an example of identified subjects in videos with multiple living beings. Obviously, our VPS method cannot distinguish two individuals if their pulse would have exactly the frequency and phase.

TABLE I
ANOVA FOR OVERALL BENCHMARK DATASET. BOLD ENTRIES INDICATE THE EVALUATION METRIC WITH ρ-VALUE SMALLER THAN 0.05.

Evaluation metric	MS-within	MS-between	ρ-value
Precision (AUC)	0.0078	0.4663	$\mathbf{4.2139e^{-7}}$
Precision (t=0.5)	0.0406	2.0804	$\mathbf{1.1536e^{-6}}$
Correlation	0.0545	2.6032	$\mathbf{1.8420e^{-6}}$
Bland-Altman	0.0227	0.3706	$\mathbf{0.0008}$

B. Overall comparison

To understand the significance of difference between VPS and FDR, we apply the balanced oneway Analysis of Variance (ANOVA) to analyze the results obtained by each evaluation metric. In ANOVA, the p-value is used for interpretation and a common threshold 0.05 is specified to determine whether the difference is significant, i.e., if p-value < 0.05, the difference is significant.

Table I shows the ANOVA results of each evaluation metric in the complete benchmark dataset, from which we conclude that VPS is significantly different from FDR in all round evaluations, i.e., p-values are all much smaller than 0.05. This implies that the improvement from FDR to VPS is substantial. Additionally, the overall comparison in Figure 10 shows that VPS significantly outperforms FDR in all evaluations and achieves the state-of-the-art performance, i.e., on average, the percentage of improvements obtained by VPS are respectively the 82.2%, 265.7%, 595.5% and 542.2% on precision (AUC), precision (t=0.5), correlation and Bland-Altman.

In addition, we would like to stress the practical functions of VPS in real use-cases. As can be seen, the key idea of VPS is the use of physiological signals (e.g., pulse) in detecting human beings, which is conceptually different from conventional subject detection methods that rely on physical appearance. The proposed method can directly lead to further advances in rPPG: it can substitute the commonly used Viola-Jones face detector or skin classifier in existing rPPG methods

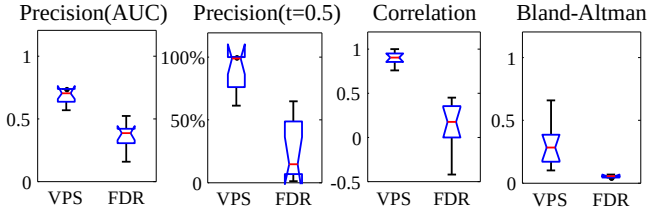

Fig. 10. The overall comparison between VPS and FDR in all evaluation metrics using ANOVA, which shows the comparison of median (red bar), standard deviation (blue box), minimum and maximum (black bar) values.

[3], [8]. The benefits are two-fold: (1) the accurate skin-region localization (e.g., non-rigid RoI) could improve the quality of the extracted rPPG-signal, since only the living skin-pixels showing pulsatility are detected, tracked and finally measured; and (2) the continuous pulsatile-region detection avoids drift of an object tracker during long-term tracking.

VI. Conclusion

In this study, we have presented a novel method for alive subject detection using rPPG. In essence, our method creates hierarchical voxels for parallel pulse extraction, builds a sparse similarity matrix based on pulse characters, and incrementally factorizes it for finding the living skin-tissues of alive subjects. Experiments show the superior performance of our method in comparison with a state-of-the-art method. On average, our method improves 82.2% on the precision of skin-region detection; and 595.5% and 542.2% on the Pearson correlation and Bland-Altman of instant pulse-rate. Besides, ANOVA shows that in all-round evaluations, the improvements obtained by our method are significant, i.e., all p-values < 0.05. It has been proved that only using pulse-features can robustly detect real human being in videos without supervised training. The superior robustness of the proposed method demonstrates it to be the first approach that successfully uses the pulse-signal to detect real human being in realistic scenarios. The proposed method is very valuable for camera-based healthcare monitoring systems that require automatic alive subject or living skin-tissue detection.

Acknowledgment

The authors would like to thank Dr. Ihor Kirenko at Philips Research for his support, and also the volunteers from Eindhoven University of Technology for their efforts in creating the benchmark dataset.

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition (CVPR), 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, Dec. 2001, pp. I–511–I–518.

[2] N. A. Ibraheem, R. Z. Khan, and M. M. Hasan, "Comparative study of skin color based segmentation techniques," *International Journal of Applied Information Systems*, vol. 5, no. 10, pp. 24–34, Aug. 2013.

[3] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *Biomedical Engineering, IEEE Trans. on*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[4] M. Lewandowska, J. Ruminski, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, Sept. 2011, pp. 405–410.

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

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

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893.

[8] W. Wang, S. Stuijk, and G. de Haan, "Exploiting spatial redundancy of image sensor for motion robust rppg," *Biomedical Engineering, IEEE Trans. on*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

[9] G. Gibert, D. D'Alessandro, and F. Lance, "Face detection method based on photoplethysmography," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, Aug. 2013, pp. 449–453.

[10] G. Lempe, S. Zaunseder, T. Wirthgen, S. Zipser, and H. Malberg, "ROI selection for remote photoplethysmography," in *Bildverarbeitung fur die Medizin*, ser. Informatik aktuell. Springer Berlin Heidelberg, 2013, pp. 99–103.

[11] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann, "Temporally consistent superpixels," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec. 2013, pp. 385–392.

[12] Y. Zhang, A. dAspremont, and L. Ghaoui, "Sparse PCA: Convex relaxations, algorithms and applications," vol. 166, pp. 915–940, 2012.

[13] A. Argyriou, M. Signoretto, and J. A. Suykens, "Hybrid conditional gradient-smoothing algorithms with applications to sparse and low rank regularization," *Regularization, Optimization, Kernels, and Support Vector Machines*, p. 53, 2014.

[14] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[15] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.

Wenjin Wang received his BSc in Biomedical Engineering (in top class) from Northeastern University, China in 2011 and his MSc in Artificial Intelligence (with full scholarship) from University of Amsterdam, Netherlands in 2013. Currently, he is a PhD candidate at Eindhoven University of Technology, Netherlands and cooperates with the Vital Signs Camera project at Philips Research Eindhoven.

Wenjin works on problems in computer vision, i.e., remote photoplethysmography (rPPG).

Sander Stuijk received his M.Sc. (with honors) in 2002 and his Ph.D. in 2007 from the Eindhoven University of Technology. He is currently an assistant professor in the Department of Electrical Engineering at Eindhoven University of Technology. He is also a visiting researcher at Philips Research Eindhoven working on bio-signal processing algorithms and their embedded implementations. His research focuses on modelling methods and mapping techniques for the design and synthesis of predictable systems with a particular interest into bio-signals.

Gerard de Haan received BSc, MSc, and PhD degrees from Delft University of Technology in 1977, 1979 and 1992, respectively. He joined Philips Research in 1979 to lead research projects in the area of video processing/analysis. From 1988 till 2007, he has additionally taught post-academic courses for the Philips Centre for Technical Training at various locations in Europe, Asia and the US. In 2000, he was appointed "Fellow" in the Video Processing & Analysis group of Philips Research Eindhoven, and "Full-Professor" at Eindhoven University of Technology. He has a particular interest in algorithms for motion estimation, video format conversion, image sequence analysis and computer vision. His work in these areas has resulted in 3 books, 2 book chapters, 170 scientific papers and more than 130 patent applications, and various commercially available ICs. He received 5 Best Paper Awards, the Gilles Holst Award, the IEEE Chester Sall Award, bronze, silver and gold patent medals, while his work on motion received the EISA European Video Innovation Award, and the Wall Street Journal Business Innovation Award. Gerard de Haan serves in the program committees of various international conferences on image/video processing and analysis, and has been a Guest-Editor for special issues of Elsevier, IEEE, and Springer.