Living-Skin Classification via Remote-PPG

Wenjin Wang, Sander Stuijk, and Gerard de Haan

***Abstract*—Detecting living-skin tissue in a video on the basis of induced color changes due to blood pulsation is emerging for automatic region of interest localization in remote photoplethysmography (rPPG). However, the state-of-the-art method performing unsupervised living-skin detection in a video is rather time-consuming, which is mainly due to the high complexity of its unsupervised on-line learning for pulse/noise separation. In this paper, we address this issue by proposing a fast living-skin classification method. Our basic idea is to transform the time-variant rPPG-signals into signal shape descriptors called "Multi-resolution Iterative Spectrum" (MIS), where pulse and noise have different patterns enabling accurate binary classification. The proposed technique is a proof-of-concept that has only been validated in lab conditions but not in real clinical conditions. The benchmark, including synthetic and realistic (non-clinical) experiments, shows that it achieves a high detection accuracy better than the state-of-the-art method, and a high detection speed at hundreds of frames per second in Matlab, enabling real-time living-skin detection.**

***Index Terms*—Biomedical monitoring, remote sensing, photoplethysmography, supervised learning, face detection.**

I. Introduction

REMOTE photoplethysmography (rPPG) enables contactless monitoring of blood volume pulse using a regular RGB camera. Recent advances in rPPG [1]–[5] trigger interest of using this technique to detect living-skin pixels in a video [6]–[10]. Given a video without any prior information related to the subjects, living-skin detection is defined as the procedure of locating the image regions containing physiological signals (e.g., pulsatile blood) of a living human being[1]. As compared to the conventional methods of using skin-appearance features to detect skin-regions (e.g., face) [11] in an image, this approach prevents the false detection of human-similar objects and guarantees that detected regions contain pulse-signals. This is particularly attractive for fully automatic video-based heart-rate monitoring, as demonstrated in [9], [10]. It is also desirable for other medical applications that require living skin-tissue localization, such as (color-based) respiration monitoring [12], SpO2 measurement [12], and PPG-imaging [13]. In this paper, we consider living-skin detection as an independent task/function for general video health monitoring applications.

W. Wang and S. Stuijk are with the Electronic Systems Group, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, 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).

[1]"Living-skin detection" is different from the general image-based skin detection as more often studied in computer vision, which finds the skin-similar objects in a static image but cannot indicate their liveness, e.g., a doll or a photograph of a person may be detected as well.

Earlier methods in living-skin detection [6]–[10] use a common scheme consisting of three steps: (i) segmenting the video into spatio-temporal regions to extract locally independent rPPG-signals; (ii) exploiting intrinsic properties of the pulse-signal to differentiate pulse and noise from extracted rPPG-signals; and (iii) labeling the regions containing pulse as skin. In this scheme, the core function is step (ii) that *separates pulse and noise*, which is also the key component to distinguish different methods in literature. In 2013, Gibert *et al.* [6] used a pre-defined threshold to select the regions with high spectrum energy within the pulse-rate band as skin, which is further used in [8]. Meanwhile, Lempe *et al.* [7] employed a relative pulse amplitude mapping approach to find the Region of Interest (RoI), which is closely related to PPG-imaging. However, it relies on the facial landmark detection, and thus cannot work fully automatic as it is restricted to human face-like objects. The methods in [6]–[8] have limited accuracy since their pulse/noise classification is only based on a single value (e.g., spectrum amplitude), as shown by the comparison in [10]. In 2014, van Luijtelaar *et al.* [9] constructed a joint multi-dimensional feature space using different properties of pulse and skin, and applied a clustering method to find skin. This approach enriches the feature representation of skin as compared to [6]–[8]. In 2015, Wang *et al.* [10] proposed a similarity-based living-skin detection method "Voxel-Pulse-Spectral" (VPS), to detect the regions sharing pulse similarities (e.g., frequency and phase) as one human being, which shows superior performance in dealing with practical challenges.

Essentially, the core function of the state of the art VPS [10] is its unsupervised learning step, which exploits the intrinsic properties (e.g., frequency and phase) of the extracted pulse-signals to build a similarity matrix and clusters the regions sharing similar pulsatile properties as skin using sparse PCA [14] or sparse subspace clustering [15]. Although attractive, the high computational complexity of similarity learning makes real-time processing demanding. For example, given n rPPG-signals, the complexity of similarity matrix factorization is $O(n^3)$, without counting the time spent on constructing the similarity matrix and sparsity optimization [10]. This complexity issue motivated us to explore supervised learning for the pulse and noise differentiation as a mean to move the bulk of the computational load to an off-line learning stage.

We show the feasibility of this approach by transforming the rPPG-signals into signal shape descriptors that enable the fast supervised classification between pulse and noise. The proposed method consists of two steps: (i) transforming the rPPG-signals extracted from a (segmented) video into signal shape descriptors called "Multi-resolution Iterative Spectrum" (MIS); (ii) utilizing the dictionary learning on the transformed descriptors for fast training and prediction. The presented technique is a proof-of-concept that has only been validated

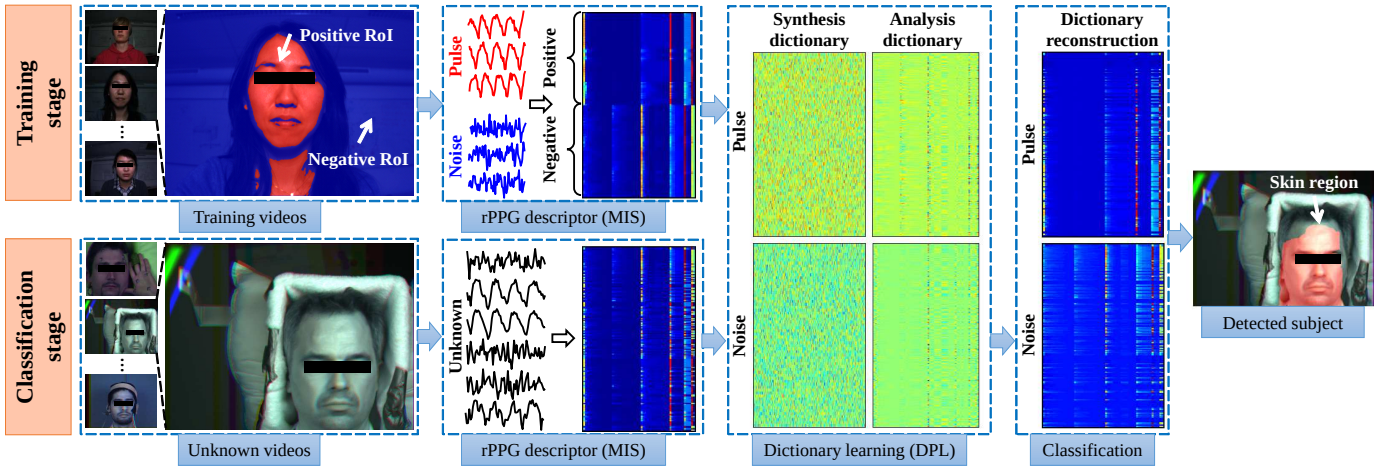


Fig. 1. The flowchart of the living-skin classification method. In the training stage, the positive RoI (skin) and negative RoI (non-skin) are defined in a training video for extracting and labeling the rPPG-signals (pulse or noise). Based on the proposed MIS transformation algorithm, the labeled rPPG-signals are transformed into descriptors for dictionary learning (DPL). In the classification stage, an unknown video is segmented into spatio-temporally coherent regions (e.g., super-voxels) for rPPG-signal extraction and transformation. The transformed descriptors are fed into the learned dictionaries for classification. Finally, we use the classification results to label the video segments as skin (in red) or non-skin. In this example, the subjects (e.g., female recorded by an RGB camera) used for classifier training are different from the subjects (e.g., male recorded by an infrared camera) used for classifier testing.

in lab conditions but not in real clinical conditions. It demonstrates state-of-the-art performance in our benchmark with significantly improved processing speed (e.g., hundreds of frames per second in Matlab). The key contribution of this work is a new non-parametric transformation method (MIS) that converts rPPG-signals into signal shape descriptors for fast supervised classification.

The remainder of this paper is organized as follows. In Section II, we analyze the considered problem and describe the proposed method in detail. In Section III and IV, the proposed method is experimentally evaluated, compared and discussed. Finally in Section V, we draw our conclusions.

II. Method

A. Problem definition

Our goal is to perform fast supervised living-skin detection using rPPG. However, rPPG-signals, combining pulse and noise signals, are time-variant samples that have different properties in frequency, amplitude and phase. If the rPPG-signal can be converted into a different representation, where (i) the within-class variance of pulse or noise is minimized, (ii) the between-class variance of pulse and noise is maximized, it would be possible to learn the transformed patterns for binary classification. Therefore, the most critical problem concerning our task is to find a *transformation* that achieves this goal. Our exploration results in an intuitively simple but effective transformation method called "Multi-resolution Iterative Spectrum" (MIS). Based on the MIS-descriptor (i.e., the transformed rPPG signal), we introduce a complete strategy for supervised living-skin detection, which involves dictionary learning for off-line training and on-line classification (see Fig. 1).

B. Multi-resolution Iterative Spectrum (MIS) Descriptor

Since this work focuses on addressing the core problem in living-skin detection (i.e., pulse and noise separation), we assume that we have a method in video segmentation [10], [16] that can segment the video sequences, including the training and testing data, into spatio-temporally coherent regions. From these local segments, we can extract multiple rPPG-signals in parallel using existing rPPG algorithms [1]–[5]. Among them, (i) Blind Source Separation (BSS) based approaches (PCA [2] and ICA [1]) use different criteria to de-mix temporal RGB traces into uncorrelated or independent signal sources to retrieve the pulse; (ii) data-driven approach (2SR [5]) measures the temporal hue change from the spatial subspace rotation of skin-pixels as the pulse; and (iii) model-based approaches (CHROM [3], PBV [4] and POS [17]) exploit characteristic color absorptions due to blood pulsation (i.e., the G-channel has the largest pulsatile variation, followed by the B- and R-channels [4]) as the prior knowledge to design a projection function, from which the pulse-signal is extracted (i.e., color property driven signal decomposition).

Here we particularly choose CHROM for local pulse extraction. The reasons are threefold: (i) model-based approaches show superior robustness than BSS-based approaches [3], [4]; (ii) 2SR requires a clean skin-mask and sufficient number of skin-pixels for estimating the pulse, which is not preferable for local skin regions/patches that contain less skin-pixels and large spatial quantization noise; and (iii) in model-based approaches, PBV is particularly designed for challenging fitness use-cases and requires sufficient amount of distortions (to stabilize the RGB-covariance matrix inversion [4]) for pulse extraction, which performs sub-optimal in (nearly) static subjects with little distortions. An additional reason for choosing CHROM lies in the fact that we, in this way, keep the pulse extraction step in line with the most direct competitor [10] for benchmarking. This allows for a direct and clean comparison on the function of pulse and noise separation as proposed in this paper compared to [10]. Note that we directly use the raw rPPG-signals output by CHROM, disabling the band-pass filtering function to preserve the noise pattern for two-class discrimination. At this point, our main focus is in transforming the rPPG-signal into a different representation that allows

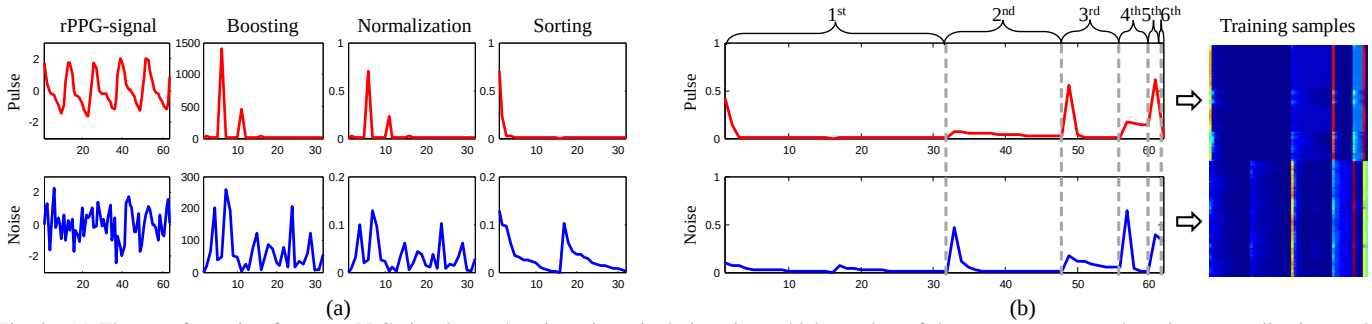


Fig. 2. (a) The transformation from an rPPG-signal to a descriptor in a single iteration, which consists of three steps: spectrum boosting, normalization, and sorting. (b) The transformation is iteratively performed multiple times, where the input for the next round is the transformed signal from the last round.

supervised learning and classification.

Considering pulse and noise as two classes, the transformed representation (i.e., signal shape descriptor) requires three properties in rPPG-signals to be eliminated:

- **Frequency** The descriptor should be independent of the pulse-rate;
- **Amplitude** The descriptor should not depend on the amplitude of the rPPG-signal;
- **Phase** The descriptor should be invariant to phase changes in the rPPG-signal.

The typical pulse-signal (with certain periodicity in a short time-interval) has a peaked pattern in the frequency spectrum as compared to that of the background noise-signal. We tend to emphasize their difference/contrast in the spectral flatness for discriminating the pulse and noise. Given these requirements, we transform the rPPG-signals using the following steps:

1) Spectrum boosting: Based on the assumption that the pulse is a periodic signal[2], all existing methods [6]–[10] transform the rPPG-signal from the time-domain to the frequency-domain for analysis. The transformed pulse presents a significant peak[3] in the pulse frequency-band ($[40, 240]$ beats per minute (bpm))[4], whereas the transformed (background) noise consists of irregular spectrum components that do not show such a pattern. This has been used in earlier work [6]–[10], and we also perform a Fourier Transform:

$$\overrightarrow{F}^L = \mathcal{F}(\overrightarrow{P}^L), \tag{1}$$

where $\overrightarrow{P}^L$ denotes the $1 \times L$ rPPG-signal in the time domain; $\overrightarrow{F}^L$ denotes the $1 \times L$ transformed frequency spectrum; $L = 2^n$ denotes the signal length where $n \geq 2$ and $n \in \mathbb{Z}$; $\mathcal{F}(\cdot)$ denotes the Fast Fourier Transform (FFT) operator. Considering a 20 frames per second (fps) video camera, we define $L = 64$, which is a short time-interval that includes at least 2-3 cardiac cycles, allowing the analysis of the pulse frequency. Although a long time-interval (e.g., $L = 256$) would allow more accurate statistical analysis and has a better frequency resolution, it increases the detection latency and may include other physiological signals (e.g., respiration) or low frequency distortions (e.g., motion drift) which harms the pulse extraction. Note that the operation in (1) is a short-time FFT, where the number of points used in the transformation is the number of time-samples (e.g., 64 points in case of a signal estimated from 64 frames) without decimation or interpolation.

The real and imaginary parts of $\overrightarrow{F}^L$ contain phase information, which can be eliminated by using the amplitude or power spectrum. Here we particularly choose the power spectrum as it boosts the pulsatile energy, which is in line with prior art [9], [10]. Since $\overrightarrow{F}^L$ is a mirrored spectrum with half redundancy, we first halve[5] it and then derive the power spectrum:

$$S^{L/2} = \overrightarrow{F}^{1 \to L/2} \odot \mathbf{conj}(\overrightarrow{F}^{1 \to L/2}), \tag{2}$$

where $S^{L/2}$ denotes the $1 \times L/2$ power spectrum; $\mathbf{conj}(\cdot)$ denotes the conjugation; $\odot$ denotes the element-wise product. In $S^{L/2}$, the phase information disappears, while the frequency peak of the pulse is boosted as compared to that of the noise (see Fig. 2 (a)). Note that the same effect of (2) can be obtained by calculating the Power Spectral Density of $\overrightarrow{P}^L$.

2) Spectrum normalization: Since the descriptor has to be invariant to the spectrum amplitude, we normalize $S^{L/2}$:

$$\bar{S}^{L/2} = \frac{S^{L/2}}{\|S^{L/2}\|_p}, \tag{3}$$

where $\bar{S}^{L/2}$ denotes the $1 \times L/2$ normalized spectrum; $\|\cdot\|_p$ denotes the Lp-norm, which can either be the L1-norm or L2-norm. Note that normalizing the standard deviation is not preferred, since we only want to eliminate the absolute energy but keep the relative energy between spectral components for pulse/noise discrimination. Here we use the L1-norm that can better suppress noise w.r.t. the total energy[6]. The normalized

[2]Although heart-rate variability may exists in a pulse-signal, this assumption is still valid especially when analyzing the pulse in a short time-interval. The reason is that the human heart-rate does not vary randomly in a short time period, but rather varies around a specific frequency. When looking at such a short-time signal in the frequency domain, the pulsatile energy usually spread in a few spectral components around a certain pulse-frequency, but not equally spread in the entire spectrum.

[3]The "significant peak" refers to the spectral components with higher amplitudes (i.e., at least 50% higher qualitatively) as compared to the ones containing noise within the pulse frequency-band. Note that it cannot differentiate the significant and periodic in-band motion-signals with the same amount of relative color changes in RGB-channels as a typical pulse.

[4]The assumed pulse frequency-band covers a broad range of possible heart-rates that can be achieved by a human being, including the neonates and adults performing fitness exercises.

[5]Halving a spectrum means cutting/splitting a spectrum into two parts with half length and only keeping the first/lower part for analysis.

[6]The L1-norm, also known as the Manhattan distance, is commonly used in the regularization to recover sparse signals or coefficients. Since we aim at increasing the local contrast/difference between the patterns of pulse and noise, we use the L1-norm here. If there is a frequency-peak in the FFT-spectra, the L1-normalization will more significantly suppress the remaining components w.r.t. to the peak component as compared to the L2-norm (i.e., the Euclidean distance).

$\bar{S}^{L/2}$ is independent of the spectrum amplitude, whereas the relative energy distribution of its entries is unchanged (see Fig. 2 (a)).

3) Spectrum sorting: The next step is to eliminate the frequency dependency of $\bar{S}^{L/2}$. We notice that although different individuals may have different heart-rates, their pulse frequencies are mostly peaked and concentrated in a certain (lower) band (e.g., $[40, 240]$ bpm).This is a common underlying assumption of existing rPPG algorithms [1], [3]–[5], [18] that use the Signal-to-Noise-Ratio (SNR) as an evaluation metric to measure the quality of the extracted pulse-signal. This assumption has also been used by existing living-skin detection methods [6], [9] to label the skin (with high SNR) and non-skin (with low SNR) regions, assuming that the in-band periodic motions can be eliminated by rPPG[7]. In contrast, the frequency of (non-skin) background signals are usually the white-noise spreading in the complete frequency-band.

To capture this character, we separate $\bar{S}^{L/2}$ into the lower and upper parts to approximate the "in-band" and "out-band" frequencies, where the pulsatile property is implicitly exploited here. To eliminate the frequency dependency, we propose to sort the separated spectra and concatenate them:

$$\hat{S}^{L/2} = [\mathbf{sort}(\bar{S}^{1\rightarrow L/4}), \mathbf{sort}(\bar{S}^{L/4\rightarrow L/2})], \tag{4}$$

where $\hat{S}^{L/2}$ denotes the $1 \times L/2$ spectrum with sorted entries; $\mathbf{sort}(\cdot)$ denotes the sorting operator that sorts elements in a vector in the descending order; $[\cdot,\cdot]$ denotes the concatenation operator. The Equation (4) can eliminate the inter-subject frequency variance, but preserve the frequency difference between pulse and noise in the lower-/upper- bands (see Fig. 2 (a)). An essential difference between our approach and prior art [6]–[8] is that they only use a single value (e.g., normalized spectrum peak) to separate pulse and noise, whereas we exploit all entries in a sorted spectrum for classification, which is in fact a *signal shape descriptor*. This is a unique step of this work.

4) Iteration: Using the previous three steps, we arrive at a transformed signal $\hat{S}^{L/2}$ given the input rPPG-signal $\overrightarrow{P}^{L}$, where pulse and noise have different patterns, i.e., the transformed pulse has a peaked pattern with sudden drops, while the transformed noise has a relatively flat/smooth pattern.

To obtain better classification performance, we can increase the between-class variance of pulse and noise by repeating the same procedure (boosting, normalization and sorting) on the transformed signals. The two iterations combined provide an anti-phase pattern between two classes, which could lead to easier separation. The mathematical meaning of this procedure is: the FFT transformation of a sorted and peaked spectrum results in a flat spectrum, while the FFT transformation of a sorted and flat spectrum results in a peaked spectrum (see the proof in Appendix A).

[7]The applied rPPG algorithm CHROM can deal with the (motion-induced) distortions to some extent. CHROM uses the prior knowledge of specular and intensity distortions to eliminate the color variations that are not due to pulse. This is based on the physiological property of PPG that it has the largest color absorption in the G-channel, followed by the B- and R- channel, i.e., PPG shows a specific color variation direction in the DC-normalized RGB space. When noise-induced color changes (e.g., artificial light flickering or periodic movement in background) do not follow such a color variation direction, they will be suppressed by CHROM. However, if the noise distortion has a similar color variation direction as the pulse, it will be less suppressed in the resulting rPPG-spectrum. Practically speaking, it is however very uncommon to find an object emitting/reflecting periodic light that has the same relative RGB variations as a typical pulse-signal.

Algorithm 1 Multi-resolution Iterative Spectrum (MIS)

Input: the rPPG-signal $\overrightarrow{P}^{L}$ with length L

Initialize: $X = [\cdot]$
for $i = 1, 2, ..., log_2(L) - 1$ **do**
  $\overrightarrow{F}_i^{L} = \mathcal{F}(\overrightarrow{P}^{L})$
  $S_i^{L/2} = \overrightarrow{F}_i^{1\rightarrow L/2} \odot \mathbf{conj}(\overrightarrow{F}_i^{1\rightarrow L/2}) \rightarrow$ boosting
  $\bar{S}_i^{L/2} = \frac{S_i^{L/2}}{\|S_i^{L/2}\|_1} \rightarrow$ normalization
  $\hat{S}_i^{L/2} = [\mathbf{sort}(\bar{S}_i^{1\rightarrow L/4}), \mathbf{sort}(\bar{S}_i^{L/4\rightarrow L/2})] \rightarrow$ sorting
  $X_{i+1} = [X_i, \hat{S}_i^{L/2}] \rightarrow$ concatenation
  $\overrightarrow{P}^{L} = \hat{S}_i^{L/2} \rightarrow$ updating
end for
$X = \frac{X}{\|X\|_2} \rightarrow$ normalization

Output: the rPPG-descriptor X with length $\sum_{i=1}^{log_2(L)-1} L/2^i$

Based on this idea, we iteratively perform the same transformation (step 1) - step 3)) on the transformed signals, i.e., the output of last iteration is the input for the next iteration. In the end, a long descriptor X is created by concatenating the transformed patterns obtained in different iterations:

$$X_{i+1} = [X_i, \hat{S}_i^{L/2^i}], \{i | i \in \mathbb{Z}, 1 \leq i < \log_2(L)\}, \tag{5}$$

where $\hat{S}_i^{L/2^i}$ denotes the $1 \times L/2^i$ transformed signal in the i-th iteration; X_{i+1} denotes the $1 \times \sum_{j=1}^{i} L/2^j$ descriptor concatenated from 1-st to i-th iterations. When the iteration is finished, the complete descriptor X is further normalized by the L2-norm[8] (see Fig. 2 (b)). Essentially, the proposed descriptor is built on the hypothesis that iterative transformation can improve the discriminativity of the descriptor in terms of pulse and noise. This hypothesis is experimentally verified in Appendix B.

The complete transformation using the described four steps is called "Multi-resolution Iterative Spectrum" (MIS), where "Multi-resolution" means that the spectrum is processed in different resolutions after being halved in each iteration, i.e., the "Multi-resolution" is achieved by (2) where the length of the spectrum is reduced by $\frac{1}{2}$ in each iteration. The complete algorithm of the proposed MIS method is shown in Algorithm 1.

C. Dictionary learning

When the rPPG-signals are transformed into descriptors, we use them to train a binary classifier. For this purpose, we adopt a recently introduced classifier, namely Dictionary Learning (DL), for our task. It has demonstrated the state-of-the-art

[8]Given the generated complete descriptor, the purpose of this step is to normalize the global (energy) difference/variance between individual descriptors for training, which is different from the previous step using L1-normalization to increase the local contrast between pulse and noise. Therefore, we use the L2-normalization here to generate globally consistent/smooth representations with the same local pulse/nose contrast remained.

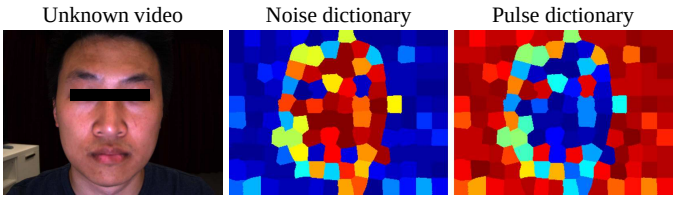


Fig. 3. An example of the reconstruction error maps obtained by noise and pulse dictionaries on MIS-descriptors measured from an unknown video (segmented by the super-voxels [10] in a single scale). The red color denotes the regions with large reconstruction errors, i.e., the skin-regions containing pulse-signals can hardly be reconstructed by the noise dictionary.

performance in face/object/action/digit recognition over a large number of classes [19], [20] (i.e., over hundreds of classes for face recognition and object recognition), and significantly outperforms commonly used classifiers, e.g., Support Vector Machine (SVM) and Sparse Representations Classifier (SRC), and does not require a large-scale training dataset as required for Deep Learning. Particularly, we apply an accurate but efficient DL method proposed by [20], namely "projective Dictionary Pair Learning" (DPL). The classification procedure consists of two stages that are described below:

1) Offline training: Given a set of training samples $X_i^n \in \mathbb{X}$ drawn from binary classes comprised of pulse and noise descriptors, where X_i^n denote n samples in the i-th class, DPL aims to learn an effective dictionary from $\mathbb{X}$ to enhance the discriminative capability of classification using known class labels. The training model of DPL is:

$$\arg\min_{P,D} \sum_{i=1}^{2} \|X_i - D_i P_i X_i\|_F^2 + \lambda \|P_i X_{j(j\neq i)}\|_F^2, \quad (6)$$

where D_i and P_i are synthetic and analytic dictionaries learned from X_i that allow efficient classification/reconstruction. In (6), $\|X_i - D_i P_i X_i\|_F^2$ ensures the representative capability by minimizing the reconstruction errors of sub-dictionaries on samples from the same class, while $\lambda \|P_i X_{j(j\neq i)}\|_F^2$ promotes the discriminative capability by preventing the representation of sub-dictionaries on samples from different class.

The essential difference between the supervised learning in (6) and the unsupervised similarity learning in [10] is that (6) exploits the labels of training samples to enhance the discriminativity of learning, where noise is explicitly treated as a separate class.

2) Online classification: Using the trained dictionary pairs D and P, we can classify the unknown samples $\mathbb{Y}$ that consist of transformed descriptors measured from a new video. Feeding $\mathbb{Y}$ to the trained DPL model, we assign each sample $y_j \in \mathbb{Y}$ a label l_j based on the reconstruction errors of pulse and noise dictionaries:

$$l_j = \begin{cases} 1 & \text{, if } \|y_j - D_1 P_1 y_j\|_2^2 \geq \|y_j - D_2 P_2 y_j\|_2^2 \\ 0 & \text{, else} \end{cases}, \quad (7)$$

where $\|y_j - D_i P_i y_j\|_2^2$ represents the error of using the i-th dictionary to reconstruct y_j. For example, D_1 and P_1 represent the noise dictionary. $\|y_j - D_1 P_1 y_j\|_2^2 \geq \|y_j - D_2 P_2 y_j\|_2^2$ means that the sample y_j has larger reconstruction errors on the noise dictionary than that on the skin dictionary. Thus this y_j belongs to the skin dictionary, i.e., the label value $l_j = 1$ denotes the skin class. Fig. 3 shows an example of the reconstruction error maps using a trained dictionary on an unknown test video. The l_j obtained by (7) is used to label the video segments as skin or non-skin.

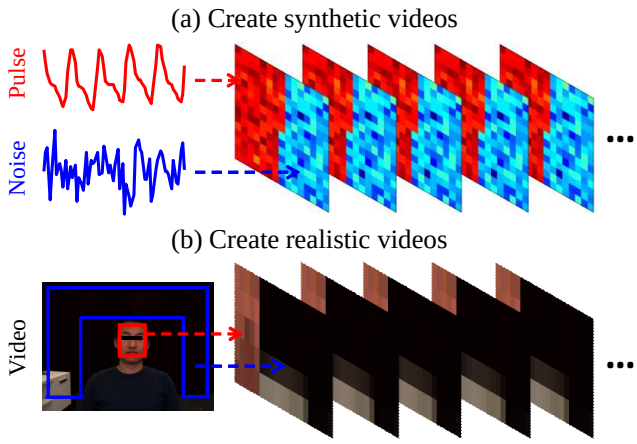


Fig. 4. Illustration of creating synthetic and realistic videos: (a) in synthetic videos, the red and blue blocks denote skin regions (with pulse) and non-skin regions (with noise) respectively, where the color saturation represents the Signal to Noise Ratio (SNR) of the contained signal. Note that the pulse is simulated by the PPG-signal sampled from a subject with a length 1000 frames, which is also the temporal length of the phantom sequence; and (b) in realistic videos, the skin/non-skin blocks are sampled from skin/background regions in real video recordings. Note that each video sequence is recorded in 20 frames per second and contains 600 frames.

III. Experiments

A. Benchmark dataset

Since this work focuses on the core problem of *pulse/noise separation*, the front-end steps as present in the framework of [10] are not addressed in this paper. These include the steps of video segmentation, pulse extraction, and region fusion. For benchmarking, we (i) create *synthetic videos* to investigate the strength/weakness of the proposed method; (ii) use *realistic videos* (e.g., in RGB and near infrared (NIR)) to verify its practical functionality. This study has been approved by the Internal Committee Biomedical Experiments of Philips Research, and the informed consent has been obtained from each subject.

Synthetic videos To simulate and control the factors influencing the living-skin detection, we create synthetic videos (i.e., video phantoms), which are chessboard-like image sequences comprised of pulse and noise (see Fig. 4 (a)). Since the purpose of this experiment is to verify the function of pulse/noise classification but not the video segmentation or rPPG-signal extraction, we assume that the pulse-/noise- signals have been obtained after the steps of video segmentation and pulse extraction, thus the skin motion and physiological properties of rPPG are not considered. The video phantoms only contain a single signal-channel (including pulse and noise) instead of RGB-channels, so the effects of skin-color and lighting-spectra are not simulated (i.e., it is also not realistic to manually simulate all the physiological and optical behaviors of skin-reflections without fully understanding rPPG at this stage). Note that the simulated pulse-signals are contact-PPG signals sampled from a middle-age healthy subject, while

TABLE I
DESCRIPTION OF FOUR CHALLENGES DEFINED FOR REALISTIC VIDEO RECORDINGS.

Challenge	Description
Seg. scale	A fixed skin area (with skin-type III) is segmented into different scales, i.e., 5×5, 10×10 and 15×15 grids.
Skin-tone	Subjects with different skin-types (according to the Fitzpatrick scale) are included, i.e., skin-type I-II, III and IV-V.
Body-motion	Different subjects (with skin-type I-III) performing irregular motions (e.g., talking) are included.
Infrared	Subjects with different skin-types are registered by 3 separate NIR cameras centered at 675 nm, 800 nm and 905 nm.

the simulated noise-signals are created from Gaussian white-noise. Based on [10], we manipulate the parameters that could have a large impact on the detection performance (i.e., the spatial position of skin in an image does not vary the results much and thus is not simulated). These include (see Fig. 5 (a)):

(i) Noise level The noise distortion is a main challenge for rPPG and thus also for living-skin detection. To investigate this, we increase the noise level of the simulated pulse in a phantom as:

$$\overrightarrow{P}_i^L = \overrightarrow{P}_i^L + \beta \cdot \mathcal{N}(0,1), \tag{8}$$

where $\overrightarrow{P}_i^L$ denotes the pulse in the i-th block of a phantom (with normalized mean and standard deviation); $\mathcal{N}(0,1)$ denotes the Gaussian white noise with zero mean and unit variance; β controls the noise level, which is set to five levels $[0.1, 0.5, 0.75, 1.0, 1.5]$.

(ii) Phantom resolution To investigate the detection speed, we changed the phantom resolution to $[5 \times 5, 10 \times 10, 15 \times 15, 20 \times 20, 25 \times 25]$ grids. This simulates different number of segmentations in a video, where the lower/higher resolutions represent the coarse/fine segmentations. Note that the quantization noise is not increased when increasing the resolution.

(iii) Skin percentage We changed the percentage of skin in a phantom to $[10\%, 25\%, 50\%, 75\%, 90\%]$ to investigate the methods' robustness to different skin percentages.

In each of the above synthetic challenges, only the investigated parameter is changed while the rest remains constant. This is to investigate the independent influence of the parameter to the detection performance (i.e., control variate method). If several parameters are changed together, it is hard to draw solid conclusion on which parameter contributes to the performance change. The default parameter settings are: noise level is 0.75; phantom size is 15×15 grids; skin percentage is 50%.

Realistic videos The feasibility of using rPPG to detect skin has been thoroughly investigated in [10], even in challenging scenarios with different body-motions, occlusions, positions, etc. For the real videos used in this paper, we focus on analyzing the core functions (i.e., pulse/noise separation) for living-skin detection. So we eliminate the challenges that are not addressed in this paper, such as video segmentation. To this end, the recorded videos are manually segmented into locally independent spatio-temporal grids and re-organized into chessboard-like phantom sequences (see Fig. 4). The videos are recorded by a professional video recording system provided by Philips Research, with a regular RGB[9]/infrared[10] camera recording in an uncompressed bitmap format and constant frame-rate[11]. Considering a recording setup where the subject sits in front of a camera, we assume that the skin (25%) occupies a relatively smaller portion of the video as compared to the background (75%). To facilitate the percentage calculation, we define 100 grids for each phantom frame, where 25/75 grids are related to the skin/background regions manually cropped from a video. Based on the experimental results of [10], we recognize that the skin pulsatility and signal noise have a large impact on the detection performance. Thus four challenges are included in our recordings (see Fig. 5 (b)):

(i) Segmentation scale Different segmentation scales result in different segmentation resolutions and also in different quantization noise in locally extracted rPPG-signals. We investigated this challenge by changing the segmentation scale to $[5 \times 5, 10 \times 10, 15 \times 15]$ grids in a fixed skin area.

(ii) Skin-tone The lower pulsatility of dark skin-tone makes the detection more challenging [10], as the higher melanin contents in dark skin absorbs a portion of diffuse reflections carrying pulsatile information but not reduces the specular reflections. To investigate this challenge, subjects with different skin-tones are recorded and categorized into three skin-types according to the Fitzpatrick scale, i.e., participants are from West Europe (skin-type I-II), East Asia (skin-type III) and Sub-Sahara Africa (skin-type IV-V).

(iii) Body-motion Body motion is considered as one of the most significant challenges for rPPG [3], since the motion-induced color changes can easily disrupt the subtle pulse-induced color variations in skin reflections. We investigated this challenge by recording subjects with body motions, where the motion-type is talking (i.e., irregular motion). Note that the motion tracking addressed by video segmentation is outside the scope of this work. Recorded videos are translated to phantoms as explained earlier.

(iv) Infrared Pulse-rate monitoring in infrared is emerging for clinical environment as it can work in full darkness, such as sleep monitoring and neonatal monitoring [21]. However, the skin pulsatility is much lower in infrared wavelengths [22] and thus more challenging for living-skin detection. In this challenge, we use three monochrome infrared cameras centered at wavelength 675 nm, 800 nm and 905 nm to record subjects with different skin-types.

Table I summarizes the protocol of these four challenges. Note that the feasibility of using rPPG to detect other body-parts, such as the palm and arm, has been demonstrated in [10], we shall not repeat these experiments in this paper but focus on

[9]Global shutter RGB CCD camera USB UI-2230SE-C from IDS, with 768×576 pixels, 8 bit depth, and 20 fps.

[10]Manta of Allied Vision Technologies GmbH, with 968×728 pixels, 8 bit depth, and 15 fps.

[11]If the video frame-rate is not constant but time-varying, one can interpolate/resample the RGB measurements according to the time-stamp of each video frame provided by the system.

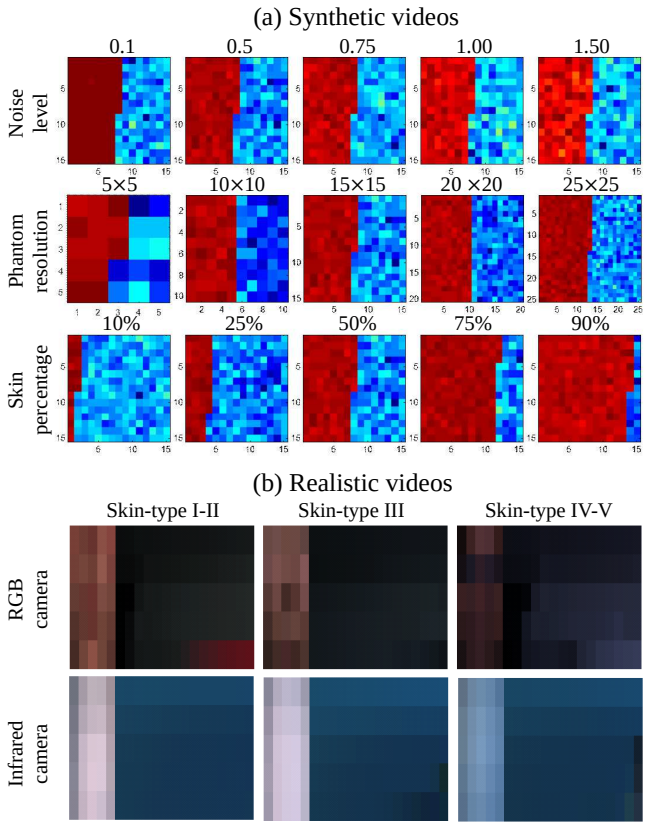


Fig. 5. Example of created (a) synthetic video phantoms and (b) realistic video phantoms.

benchmarking the core functions in living-skin classification. We mention that the claim of the other body-parts detection is restricted to [10].

B. Evaluation metrics

Since the living-skin detection, studied in this paper, is considered as an independent task for general video health monitoring applications, our evaluation is therefore geared to comparing the performance of pulse/noise separation, instead of the pulse extraction accuracy as [9], [10]. All tested methods are qualitatively evaluated by the following two metrics:

Detection accuracy We adopt the same metric as used in [10] to measure the detection accuracy, which is the Area Under Curve (AUC) of the precision curve representing the percentage of successfully detected regions in a video. The successfully detected region is defined as the overlap between the found RoI and ground-truth (e.g., a binary image labeled as skin and non-skin), where the precision is calculated as the ratio between the overlap/intersection area and union area of the found RoI and ground-truth, i.e., intersection-over-union. The precision curve is generated by changing the threshold determining a successful detection between $[0.1, 1]$, where 1 means the fully matched detection with intersection-over-union 1.

Detection speed Since the major motivation of this work is to improve the detection speed of the unsupervised method in [10], it is important to know how fast the proposed method is, i.e., whether the supervised detection can run in real-time. In our experiment, the detection speed is measured in frames per second (fps). This is an important indicator for practical usage that has not been considered in prior art [6]–[10].

C. Compared methods

We compare the proposed method (e.g., MIS-descriptor transformation and dictionary learning) to the core steps of the other two existing methods FDR [6] and VPS [10] (the state-of-the-art), which are respectively the frequency-peak based method and similarity-learning based method. All three methods are implemented in Matlab and ran on a laptop with an Intel Core i7 processor (2.70 GHz) and 8 GB RAM.

The parameters in FDR and VPS remained identical to the default settings in the original papers. In the proposed method, the length of the pulse-signal used for deriving the MIS-descriptor is 64 frames, measuring the video contents within 3.2 s given a 20 fps video camera. The number of frames used for deriving the descriptor needs to contain at least 2-3 cardiac cycles for determining the pulse-features in the FFT-domain. Although using more frames (or longer window length) can increase the FFT-resolution, it also increases the detection latency and may include more low frequency components (e.g., respiration and motion drift) which harms the classification. Therefore, 64 frames is a compromise considering the quality of MIS-descriptor and latency of detection. The DPL classifier does not require a huge dataset to train but still needs some variance in the training data. Hence, we use 6 subjects (e.g., 6 videos) with different skin-tones as a default setting to train the DPL, including the inter-subject variance. For each subject, 100 pulse-signals are extracted from skin-regions using CHROM, including the time variance of an individual subject. We have 600 samples for the positive class (pulse). Correspondingly, we create a negative class (noise) with the same amount of samples (e.g., 600) using the CHROM-signals extracted from the background-regions in 6 videos. Therefore, the DPL classifier has been trained on a dataset containing a total of 1200 samples (600 samples per class).

More specifically, the skin/non-skin regions, in each training video frame, are labeled by the machine-assisted manual annotation (e.g., assisted by an object tracker and OV-SVM skin classifier used by [18]). Note that there is only one annotator who checks/repairs the segmentation results to guarantee the correct annotation (see an example in Fig. 1), i.e., the inter-annotator comparison is not included. For the testing dataset, the total number of samples created from videos is 885000 (signals), with 345000 samples for the positive class (pulse) and 540000 samples for the negative class (noise). The training dataset and testing dataset are independent of each other as they are created from different subjects at different times to avoid over-fitting, i.e., the 6 subjects used for creating the training data do not appear in the testing data. Also, the subjects are randomly assigned to each of the training and testing groups. The training parameters (e.g., dictionary size) in DPL remained identical to the default settings in [20]. Note that the proposed MIS transformation algorithm is non-parametric itself, i.e., it does not require parameter setting/tuning.

TABLE II
DETECTION ACCURACY (AUC) ON SYNTHETIC VIDEOS

Challenge	Parameter	FDR	VPS	MIS
Noise level	0.10	100.0%	100.0%	100.0%
	0.50	99.7%	100.0%	100.0%
	0.75	88.6%	94.9%	95.3%
	1.00	57.0%	69.1%	86.7%
	1.50	12.2%	63.2%	66.8%
Phantom resolution	5×5	87.2%	96.4%	95.4%
	10×10	90.1%	95.2%	95.3%
	15×15	89.5%	95.2%	95.4%
	20×20	88.6%	94.5%	95.5%
	25×25	88.8%	93.9%	95.1%
Skin percentage	10%	86.1%	94.7%	91.6%
	25%	89.2%	97.6%	93.5%
	50%	89.5%	95.2%	92.8%
	75%	89.1%	90.0%	93.0%
	90%	88.7%	88.3%	93.5%
Overall	Average	83.0%	91.2%	92.7%

TABLE III
DETECTION SPEED (FPS) ON SYNTHETIC VIDEOS

Challenge	Parameter	FDR	VPS	MIS
Noise level	0.10	4441.0	8.6	635.3
	0.50	3796.4	7.9	617.1
	0.75	4430.8	4.3	577.7
	1.00	4442.2	3.1	572.4
	1.50	4397.3	2.7	573.7
Phantom resolution	5×5	15123.6	70.8	909.5
	10×10	7079.7	21.7	733.2
	15×15	3903.0	4.5	584.2
	20×20	1692.5	1.1	387.8
	25×25	1077.0	0.4	266.1
Skin percentage	10%	4585.7	4.0	622.9
	25%	4602.4	4.5	600.5
	50%	4497.8	4.3	581.0
	75%	4628.5	3.6	578.3
	90%	4634.0	3.2	566.2
Overall	Average	4889.8	9.6	587.1

IV. RESULTS AND DISCUSSION

In this section, we first benchmark the proposed living-skin classification method with the two existing methods using synthetic and realistic videos. Next, we put it into a (super-voxel based) video segmentation framework to visualize its practical functionality. To simplify the illustration, the proposed method is referred to as MIS, including the steps of MIS-transformation and DPL classification.

A. Comparison on synthetic videos

Table II and III show the results of detection accuracy and detection speed obtained by three compared methods on synthetic videos. When the noise level is low $(0.1 - 0.5)$, all three methods achieve almost 100% detection accuracy. When the noise level increases $(0.75 - 1.5)$, FDR and VPS have more quality drops than MIS. FDR that only employs a single value (e.g., the frequency peak) for detection obtains the worst performance, which is in line with the findings in [10]. VPS suffers performance degradations when the built similarity matrix is significantly distorted by noise, i.e., the noise-induced phase shift between two pulse-signals may break their correlation in a similarity matrix. MIS transforms the rPPG-signal into a binary class representation, which is less sensitive to the changes in frequency and phase, i.e., pulse-signals with a frequency 70 bpm or 75 bpm will all be converted into descriptors belonging to the pulse class. Moreover, MIS classifies each rPPG-sensor independently and thus the noise-distorted sensors do not affect each other. Besides, the measurement noise affect the quality of the unsupervised learning in VPS, but do not influence the trained dictionary in MIS, which is an advantage of supervised learning. Table III shows that FDR and MIS, running at thousands/hundreds of frames per second, are much faster than VPS. The fast processing speed of (i) FDR is due to its low complexity, (ii) MIS is due to its efficient descriptor transformation and DPL classification. When increasing the noise level, the running speed of FDR and MIS remain relatively stable, whereas VPS significantly slows down. This is due to the online convex optimization in VPS [10], where the noisy entries in the similarity matrix result in a longer overall run-time to converge.

The phantom resolution has limited impact on their detection accuracy, where VPS and MIS have similar high accuracies that are consistently better than FDR. However, the phantom resolution has a large impact on the detection speed, especially for VPS where the speed dramatically decreases from 70.8 fps to 0.4 fps. Although VPS can run at a high-speed when the phantom resolution is 5×5 grids, 25 segments in a video is too coarse for detection and thus not preferred. In comparison, MIS can process over 600 video segments in real-time, without sacrificing the detection accuracy.

The detection accuracies of all three methods are not very sensitive to the challenge of skin percentage, although VPS has a slight quality drop (around 5%) when the skin percentage arrives $75\% - 90\%$. Since VPS exploits the similarity between spatially redundant rPPG-signals to cluster the regions sharing similar pulsatile properties (e.g., frequency and phase) as skin, its detection accuracy depends on the quality of the majority of pulse-signals and is thus a bit more variant to the skin percentage. In contrast, FDR and MIS treat each local region as an independent classification task, i.e., the quality of each region does not influence each other. Nevertheless, the detection speeds of all three methods remain stable in this challenge, as the total phantom resolution is fixed.

B. Comparison on realistic videos

Table IV and V show the results of the detection accuracy and detection speed obtained by three compared methods on realistic videos. It is clear that all three methods perform better in lower scales (i.e., above 95% detection accuracy at 5×5 (25) grids), as the quantization noise is lower. MIS obtains a relatively higher detection accuracy when the skin-region is densely segmented into 15×15 (225) grids. The increased segmentation scales decrease their detection speed,

TABLE IV
DETECTION ACCURACY (AUC) ON REALISTIC VIDEOS

Challenge	Parameter	FDR	VPS	MIS
Segmentation scale	5×5	95.9%	96.6%	96.0%
	10×10	76.6%	72.6%	72.4%
	15×15	54.1%	44.0%	56.1%
Skin-tone	Skin-type I-II	54.8%	76.1%	75.2%
	Skin-type III	94.2%	93.9%	93.3%
	Skin-type IV-V	12.9%	33.8%	44.2%
Body-motion	Subject-1	54.0%	67.5%	87.1%
	Subject-2	66.5%	67.4%	88.2%
	Subject-3	48.1%	64.3%	80.2%
Infrared	Skin-type I-II	32.7%	68.5%	77.0%
	Skin-type III	26.0%	33.8%	58.8%
	Skin-type IV-V	9.5%	22.8%	39.5%
Overall	Average	52.1%	61.8%	72.3%

TABLE V
DETECTION SPEED (FPS) ON REALISTIC VIDEOS

Challenge	Parameter	FDR	VPS	MIS
Segmentation scale	5×5	6565.2	22.1	595.9
	10×10	4370.7	7.7	572.4
	15×15	3531.4	2.1	423.5
Skin-tone	Skin-type I-II	5770.3	19.7	856.0
	Skin-type III	5686.3	27.0	793.2
	Skin-type IV-V	5352.0	12.8	853.9
Body-motion	Subject-1	6497.0	15.7	851.7
	Subject-2	6684.8	14.7	835.4
	Subject-3	6614.1	16.8	847.3
Infrared	Skin-type I-II	6756.8	16.0	854.5
	Skin-type III	6918.5	13.1	757.2
	Skin-type IV-V	6692.4	19.2	829.4
Overall	Average	5953.3	15.6	755.9

as the total number of segments increases. When increasing the segmentation scale, the detection speed of VPS significantly drops from 22.1 fps to 2.1 fps, whereas FDR and MIS still run at thousands/hundreds frames per second.

All three methods can better deal with the bright skin than the dark skin. They all suffer clear quality drops in skin-type IV-V, where MIS shows improved robustness than FDR and VPS, i.e., the detection accuracy of FDR is only 12.9%. VPS shows a lower detection speed in skin-type IV-V. This is due to the unsupervised learning step in VPS, which requires more time for noisy rPPG-signals/entries to converge in the online convex optimization [10].

Table IV shows that MIS obtains the best detection accuracy in the challenge of body-motion, followed by VPS and FDR. Although FDR has the highest detection speed, it is vulnerable to motion distortions. This is in line with our observation in the synthetic experiment when changing the noise level in the simulated rPPG-signals.

All three methods show lower detection accuracies in infrared than in RGB, especially in skin-type IV-V (dark skin). The difficulty is due to the reduced pulsatile amplitude w.r.t. that of the (motion) noise distortions. However, MIS still shows generally improved robustness, i.e., the frequency-peak/energy based thresholding method in FDR suggests the limitation in infrared with low skin pulsatilities.

Based on the synthetic and realistic experiments, our conclusions to the three compared methods are: (i) FDR has the highest detection speed at thousands of frames per second, but rather low detection accuracy especially in challenging use-cases with motion distortions or low skin pulsatilities (e.g., dark skin or infrared); (ii) VPS improves the detection accuracy[12] of FDR, but sacrifices the detection speed (in the unsupervised learning), with only an averaged 10-15 fps in both experiments; (iii) MIS obtains both the high detection accuracy (i.e., better than VPS) and high detection speed (i.e., hundreds of frames per second) across different challenges.

In the end, we put MIS into an existing video segmentation framework using super-voxels [10] to visualize its practical functionality. Fig. 6 shows the qualitative results (e.g., detected skin-regions) obtained by this new combination on real video recordings. Note that in cases of multiple living-subjects in the same scene, the living-skin detection method proposed in this paper by itself cannot be used to discriminate different subjects. "Living-subject identification and classification" are left as future work, such as combining the living-skin detection with face recognition. This paper focuses on addressing the core problem of classifying the skin and non-skin regions in a video.

Fig. 6. The snapshot of detected skin-regions (red regions) in recorded videos by using the combination of the super-voxel segmentation and proposed method. The different realistic challenges, recorded by RGB and infrared cameras, confirm its practical functionality.

V. CONCLUSION

In this study, we present a novel method that performs supervised learning on rPPG-signals for living-skin detection. Our core idea is to first convert the time-variant rPPG-signals into signal shape descriptors, in which pulse and noise have different patterns, and then treat it as a binary classification problem. The complete method consists of two steps: (i) transforming the rPPG-signals extracted from a (segmented) video into descriptors using the proposed Multi-resolution Iterative Spectrum (MIS); (ii) utilizing the dictionary learning on the transformed descriptors for fast training and classification. The

[12]We notice that the improvement (in detection accuracy) from FDR to VPS is not as large as reported in [10]. The reason is that [10] compares the complete framework of two methods including the video segmentation, i.e., [6] uses the fixed grid-based segmentation while [10] uses the super-voxel segmentation. In our experiments, the improvement only/particularly reflects the difference between their core functions in pulse and noise separation.

presented technique is a proof-of-concept that has only been validated in lab conditions but not in real clinical conditions. The benchmark on synthetic and realistic videos show the superior performance of the proposed method as compared to the existing methods using unsupervised detection. It especially improves the detection speed of the state-of-the-art method (to hundreds of frames per second in Matlab), which is desirable for real-time monitoring systems that require automatic living skin-tissue detection.

APPENDIX A
PROOF OF ITERATIVE FFT TRANSFORMATION

Theorem A.1. *In FFT transformation, a sorted and peaked spectrum leads to a flat spectrum, while a sorted and flat spectrum leads to a peaked spectrum.*

Proof. For simplicity, we assume the power spectrum as a vector with the minimum length required for FFT transformation, i.e., $L = 2$. The vector is denoted as $v = [v_1, v_2], x_i \in \mathbb{R}_0^+$. Since the FFT transformation of v is:

$$V_i = \sum_{j=1}^{L} v_j \cdot (e^{\frac{-2\pi i}{L}})^{(j-1)(i-1)}, \tag{9}$$

V_i can be derived as:

$$\begin{cases} V_1 = v_1 \cdot e^{-\pi i \times 0 \times 0} + v_2 \cdot e^{-\pi i \times 1 \times 0} = v_1 + v_2 \\ V_2 = v_1 \cdot e^{-\pi i \times 0 \times 1} + v_2 \cdot e^{-\pi i \times 1 \times 1} = v_1 - v_2 \end{cases}, \tag{10}$$

and thus $V = [v_1 + v_2, v_1 - v_2]$. Since v_i in a power spectrum is non-negative, V will be (1) a peaked spectrum if v is flat, and (2) a flat spectrum if v is peaked. For example, $v = [1, 1]$ (flat) leads to $V = [2, 0]$ (peaked), while $v = [1, 0]$ (peaked) leads to $V = [1, 1]$ (flat). □

APPENDIX B
HYPOTHESIS VERIFICATION FOR MULTI-RESOLUTION ITERATIVE SPECTRUM (MIS)

The proposed MIS method is built on the hypothesis that iterative transformation can improve the discriminativity of the descriptor. To verify this hypothesis, we apply the criteria of Linear Discriminant Analysis (LDA) [23] to measure the "between-class variance" and "within-class variance" of the MIS-descriptors transformed from rPPG-signals. LDA aims to find a projection ω that maximizes the following objective:

$$\arg\max_{\omega} \frac{\omega^\top \Sigma_{\text{between}} \omega}{\omega^\top \Sigma_{\text{within}} \omega} = \arg\max_{\omega} \frac{(\omega(\mu_p - \mu_n))^2}{\omega^\top (\Sigma_p + \Sigma_n)\omega}, \tag{11}$$

where Σ_{between} and Σ_{within} denote the between-/within-class variance of pulse and noise; μ_p, μ_n and Σ_p, Σ_n denote the mean and covariance of pulse and noise classes, respectively. (11) can be solved by the eigenvector decomposition [23] as:

$$\Sigma_{\text{within}}^{-1} \cdot \Sigma_{\text{between}} \cdot \omega = \lambda \cdot \omega, \tag{12}$$

where λ denotes the eigenvalues. Here λ is used to indicate the discriminativity between two classes. Since Σ_{between} is rank-1 for two classes, only the principal eigenvalue λ is needed.

For analysis, we use four RGB videos recorded on different subjects to create a small dataset, where each video has 100 pulse/noise signals extracted from skin/non-skin regions, i.e., 400 training samples per class and thus 800 training samples in total. The length of each rPPG-signal is 64, so the overall length of each fully transformed descriptor is 62 (with 5 iterations according to (5)). To understand whether the iterative transform can increase the discriminativity of the descriptor, we perform two comparisons: (i) trimming the descriptors at the end by taking only the entries 1 to n, denoted as $X_{1\to n\in[1,62]}$, (ii) trimming the descriptors at the front by taking only the entries m to 62, denoted as $X_{m\in[1,62]\to 62}$. For each trimmed descriptor, we measure its eigenvalue using (12) and plot in Fig. 7.

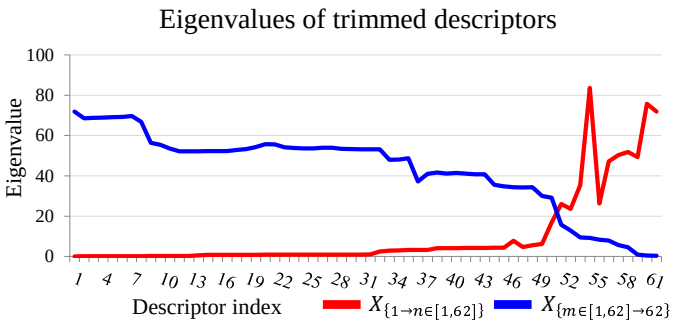


Fig. 7. The two eigenvalue curves (based on criteria of LDA) of trimmed descriptors, where the descriptors are trimmed in two different ways, i.e., $X_{1\to n\in[1,62]}$ and $X_{m\in[1,62]\to 62}$. This is to show the discriminativity of different iterations and partial segments of the descriptor.

Fig. 7 shows that iterative transformation can improve the descriptor's discriminativity: (i) in $X_{1\to n\in[1,62]}$, the eigenvalue obtained in the first iteration (1→32) ranges in $[0.1, 0.9]$, while in the second iteration (33→48) ranges in $[3, 7]$. A sudden jump between 32→34 can be clearly recognized (e.g., from 0.9 to 3), which is the location that the first peak in the noise descriptor appears. In further iterations, the eigenvalues are significantly promoted to ranges $[40, 50]$ and $[70, 80]$ after 4-th and 5-th iterations. However, we also notice that the discriminativity is not consistently improved with the increased entries, i.e., the eigenvalue drops at certain entries after a peak; (ii) in $X_{m\in[1,62]\to 62}$, the eigenvalue drops when frontal entries are removed. However, the discriminativity decreases (blue curve) are much smoother and slower than the discriminativity increases (red curve) in $X_{1\to n\in[1,62]}$. This is because that majority of entries in the tail of the descriptor remained in the trimmed $X_{m\in[1,62]\to 62}$. It implies that the entries generated in succeeding iterations are more discriminative than the ones generated in preceding iterations. However, only using the last few iterations (57→62) cannot make the descriptor more discriminative, where the eigenvalue drops to $[0.3, 8]$.

Based on this investigation, we conclude that our hypothesis is valid. The reasons of improvement could be: (i) the use of noise pattern increases the discriminativity between two classes; (ii) in the iteration, the boosting step makes the energy around the peak more concentrated, the normalization step increase the contrast between peak and remaining entries, and the sorting step further enhance the regularity/consistency of the pattern.

ACKNOWLEDGMENT

The authors would like to thank Dr. Albertus C. den Brinker at Philips Research for reviewing our paper and Mr. Ger Kersten at Philips Research for creating the video recording system, and also the volunteers from Eindhoven University of Technology for their efforts in creating the benchmark dataset.

REFERENCES

[1] M.-Z. Poh *et al.*, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 7–11, Jan. 2011.

[2] M. Lewandowska *et al.*, "Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity," in *Proc. Federated Conf. Comput. Sci. Inform. Syst. (FedCSIS)*, Szczecin, Poland, Sept. 2011, pp. 405–410.

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

[4] G. de Haan and A. van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiol. Meas.*, vol. 35, no. 9, pp. 1913–1922, Oct. 2014.

[5] W. Wang *et al.*, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 9, pp. 1974–1984, Sept. 2016.

[6] G. Gibert *et al.*, "Face detection method based on photoplethysmography," in *Proc. IEEE Conf. Adv. Video Signal Surveillance (AVSS)*, Krakow, Poland, Aug. 2013, pp. 449–453.

[7] G. Lempe *et al.*, "ROI selection for remote photoplethysmography," in *Bildverarbeitung fur die Medizin*, Berlin, Heidelberg, Feb. 2013, pp. 99–103.

[8] H. Liu *et al.*, "A new approach for face detection based on photoplethysmographic imaging," in *Proc. Health Inf. Sci. (HIS)*, Melbourne, Australia, May 2015, pp. 79–91.

[9] R. van Luijtelaar *et al.*, "Automatic roi detection for camera-based pulse-rate measurement," in *Proc. Asian Conf. Comput. Vis. Workshops (ACCV-W)*, Singapore, Nov. 2014, pp. 360–374.

[10] W. Wang *et al.*, "Unsupervised subject detection via remote PPG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 11, pp. 2629–2637, Nov. 2015.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Hawaii, USA, Dec. 2001, pp. I–511–I–518.

[12] L. Tarassenko *et al.*, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiol. Meas.*, vol. 35, no. 5, p. 807, May 2014.

[13] A. V. Moço *et al.*, "Motion robust PPG-imaging through color channel mapping," *Biomed. Opt. Exp.*, vol. 7, no. 5, pp. 1737–1754, May 2016.

[14] Y. Zhang *et al.*, "Sparse PCA: Convex relaxations, algorithms and applications," *Handbook on Semidefinite, Conic and Polynomial Optimization*, vol. 166, pp. 915–940, Sept. 2012.

[15] E. Elhamifar and R. Vidall, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[16] M. Reso *et al.*, "Temporally consistent superpixels," in *Proc. IEEE Conf. Comput. Vis. (ICCV)*, Sydney, Australia, Dec. 2013, pp. 385–392.

[17] W. Wang *et al.*, "Algorithmic principles of remote-PPG," *IEEE Trans. Biomed. Eng.*, vol. PP, no. 99, pp. 1–1, 2016. DOI: 10.1109/TBME.2016.2609282.

[18] ——, "Exploiting spatial redundancy of image sensor for motion robust rPPG," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

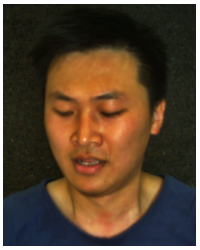
[19] M. Yang *et al.*, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 543–550.

[20] S. Gu *et al.*, "Projective dictionary pair learning for pattern classification," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, Canada, Dec. 2014, pp. 793–801.

[21] S. Fernando *et al.*, "Feasibility of contactless pulse rate monitoring of neonates using google glass," in *Proc. EAI Conf. Wireless Mobile Commun. Healthcare (Mobihealth)*, London, UK, Oct. 2015, pp. 198–201.

[22] L. F. C. Martinez *et al.*, "Optimal wavelength selection for noncontact reflection photoplethysmography," in *Proc. SPIE*, Puebla, Mexico, Aug. 2011, pp. 360–374.

[23] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sept. 1936.



Wenjin Wang received the B.Sc. degree in Biomedical Engineering (in top class) from Northeastern University, Shenyang, China, in 2011 and the M.Sc. degree in Artificial Intelligence (with full scholarship) from University of Amsterdam, The Netherlands, in 2013. Currently, he is a Ph.D. candidate at Eindhoven University of Technology, The Netherlands, and cooperates with the Vital Signs Camera project at Philips Research Eindhoven.

Wenjin Wang 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, 3 book chapters, 180 scientific papers and more than 180 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.