Fully-automatic camera-based pulse-oximetry during sleep

Tom Vogels[1] Mark van Gastel[1] Wenjin Wang[2] Gerard de Haan[2]
[1]Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Philips Research, Eindhoven, The Netherlands
tommel_vogels@hotmail.com,m.j.h.v.gastel@tue.nl,{wenjin.wang,g.de.haan}@philips.com

Abstract

Current routines for the monitoring of sleep require many sensors attached to the patient during a nocturnal observational study, limiting mobility and causing stress and discomfort. Cameras have shown promise in the remote monitoring of pulse rate, respiration and oxygen saturation, which potentially allows a reduction in the number of sensors. Applying these techniques in a sleep setting is challenging, as it is unknown upfront which portion of the skin will be visible, there is no unique skin-color outside the visible range, and the pulsatility is low in infrared. We present a fully-automatic living tissue detection method to enable continuous monitoring of pulse rate and oxygen saturation during sleep. The system is validated on a dataset where various typical sleep scenarios have been simulated. Results show the proposed method to outperform the current state-of-the-art, especially for the estimation of oxygen saturation.

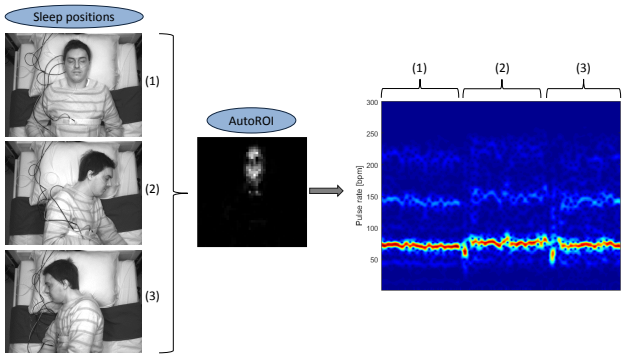

Figure 1: Overview of our approach: from the input frames physiological features are calculated to create a weight map for automatic living pixels detection. The video-frames weighted with this map are used for the extraction of the cardiac pulse signal and estimation of oxygen saturation levels.

1. Introduction

A third of US adults report that they often sleep less than the recommended daily sleep time [15]. Insufficient sleep on a regular basis is linked with many chronic diseases and conditions such as diabetes, heart disease, obesity, and depression. There are over 100 different types of sleep disorders like *e.g.* difficulty falling asleep, staying asleep, or excessive day-time sleepiness. It is therefore critical to receive the correct diagnosis and work with a qualified physician to develop a treatment plan. Unfortunately, most sleep disorders, estimated 95% [13], go undiagnosed and untreated, simply because people do not realize they have a problem or are unaware their problems can be reduced.

Polysomnography (PSG) is the current gold-standard for the diagnosis of sleep disorders. This multi-parametric diagnostic tool requires a large collection of sensors attached to the skin of the patient to monitor sleep. After a nocturnal observational study a trained clinician annotates and scores the PSG data for events related to sleep disorders. Although informative to the clinician, PSG has several drawbacks. The sensors attached to the subject not only are cumbersome, but also often wired limiting the mobility of the patient. This causes stress, discomfort and adds to the sleep problem affecting the diagnostic value.

Various physiological parameters monitored during a standard PSG could potentially also be monitored remotely with a camera. Especially the vital signs extracted from the optical detection of blood-induced skin color variations, remote photoplethysmography (rPPG), have recently been shown feasible in near-infrared (NIR) on healthy patients in a lab setting [17, 19, 18, 20]. In one of the most common sleep disorders, apnea, the patient frequently stops breathing during so-called apnea-events, which tends to be visible in a lowering of the heart rate. Involuntary reflexes cause the person to startle awake at the end of such apnea-event causing a quick rise of the heart rate and blood-pressure. Besides these variations in heart (pulse) rate, the deprivation of oxygen also causes a desaturation of the arterial oxygenation to

levels often lower than 80% [16]. It is therefore important to monitor both parameters accurately and continuously for the monitoring of sleep disorders in order to diagnose this important sub-type.

In this paper we present a framework which: 1) can detect living tissue from NIR recordings fully-automatically during sleep, 2) can monitor the sleep-relevant parameters pulse rate and oxygen saturation (SpO_2) simultaneously, and 3) combines living tissue detection with conventional tracking allowing a fast recovery after substantial patient movement, *e.g.* after a change in sleep position. Our method relies on physiological features only for the detection of living tissue because of the absence of discriminative color or facial appearance features in NIR in combination with the sleep setting, where it is unknown upfront which portion of the skin will be visible.

The remainder of the paper is organized as follows: in Section II the related work in skin detection for physiological measurements is discussed. Details about the proposed method are presented in Section III and the experimental results in Section IV. Finally we draw our conclusions and give our recommendations in Section V.

2. Related work

Most existing works in subject detection exploit appearance features of human skin to discriminate between subject and background in a supervised training mechanism. Sometimes these features are not unique, like color, or trained classifiers may fail with samples not presented in the training data, like non-frontal faces for the Viola-Jones face-detector [22]. Still they are useful in niche-application areas. Main problem for our application is that most solutions are limited to the visible spectrum [22, 10, 2]. Although [2] seems most promising it would require a large annotated dataset for our field of use, with still a high chance of failure because of missing color features in NIR.

A unique discriminative feature of skin pixels is the presence of cardiac-synchronous color variations induced by blood volume variations, which we refer to as 'living-skin', which was first exploited by Jeanne *et al.* [9]. Elaborating on this idea, most methods in living-skin detection [9, 5, 11, 21, 12, 1, 24, 25] use a common scheme consisting of three steps: (1) segmenting the video into spatio-temporal regions to extract locally independent rPPG-signals; (2) exploiting intrinsic properties of the pulse signal to differentiate pulse and noise from extracted rPPG signals; and (3) labeling the regions containing pulse as skin. In this scheme, the core function is step (2) that separates pulse and noise, which is also the key component to distinguish different methods in literature. Gibert *et al.* [5] 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 [12]. Meanwhile, Lempe *et al.* [11] 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 [5, 11, 12] 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 [24]. Van Luijtelaar *et al.* [21] constructed a joint multi-dimensional feature space using different properties of pulse and skin, and applied a clustering method to find skin. A similarity-based living-skin detection method "Voxel-Pulse-Spectral" (VPS) has been proposed by Wang *et al.* [24] 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. Later they proposed a supervised method [25] for living tissue detection. These methods however again includes color features as Wang *et al.* worked in the visible spectrum.

Although good results have been reported in visible light for pulse extraction, the performance in NIR is not up to par and furthermore the adequacy of the proposed methods for SpO_2 estimation has not been demonstrated yet. In NIR the much lower contrast between skin and background hampers the creation of relevant spatio-temporal regions to extract locally independent rPPG-signals (step (1)). Furthermore, the much lower pulse strength in NIR renders it more difficult to differentiate between pulse (skin) and noise (background), step (3).

3. Method description

For our sleep monitoring application the three most important criteria for ROI detection are: 1) the invariance to sleep position, 2) the invariance to the environment (*e.g.* the color of the pillow and sheets), and 3) the exclusion of non-skin pixels within the ROI. This third criterion is very important for the SpO_2 estimation as will be explained later on in this section. Movements for a longer period of time are unlikely to occur for our scenario. We therefore make our design decisions for the processing pipeline to primarily fulfil the three main criteria. The proposed processing pipeline is visualized in Fig. 2, divided into five groups. In this section we will describe the various processing processing steps sequentially.

3.1. Pre-processing

Gaussian smoothing

After acquisition, smoothing with a 2-D Gaussian kernel is performed on the input frames with 968×728 pixel resolution. This reduces the effect of sharp edges in the images, *e.g.* the boundaries between skin/non-skin. The effect of the Gaussian smoothing is shown in Fig. 2. The standard deviation of the Gaussian kernel, σ, is empirically determined

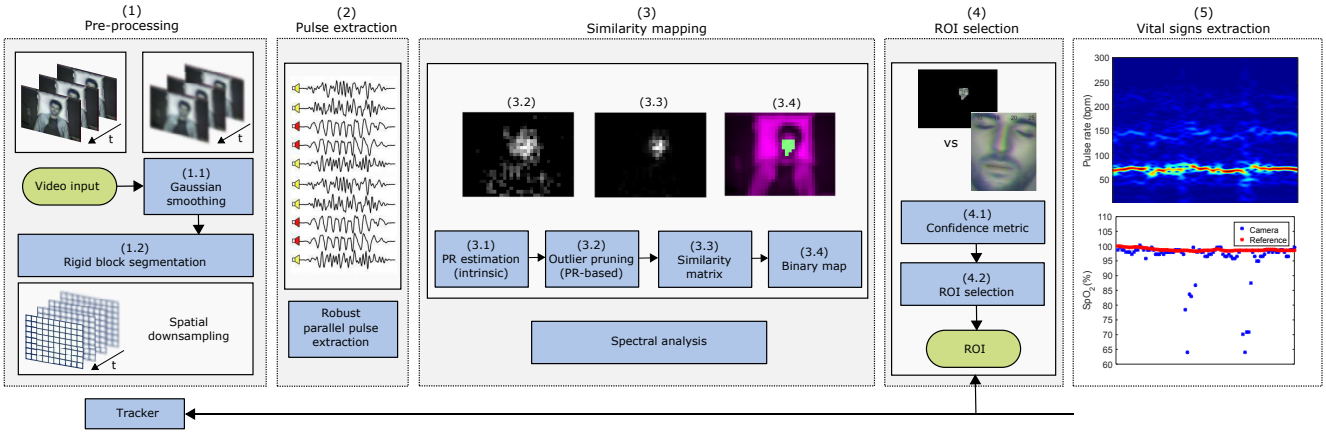

Figure 2: rPPG-based living tissue detection processing pipeline for pulse extraction and SpO_2 estimation in NIR.

to be $\sigma = 5$ for the evaluation of the dataset, based on its effect on the pulse extraction and consecutively the resulting ROI-mask.

Temporal rigid blocks

For the segmentation and consequently initialization and tracking of candidate regions, many state-of-the-art methods [24, 25] in visible light rely on the available color information, which is absent in NIR, leading to non-discriminative spatio-temporal clusters. Therefore, a rigid block scheme as originally used by Gibert *et al.* [5] is used to overcome this inability as this approach uses neither temporal nor spatial information available in the frames for the definition of the subregions.

Each frame of 968×728 pixels is downsampled into blocks (subregions) of fixed size using a box-shaped kernel. Each subregion represents the value of 29×29 pixels, yielding a total of 850 subregions. This downsampling factor is chosen as a trade-off between accuracy of the overall subject-detection and runtime considerations. Additionally, the subjects in the intended use-case are of a relatively fixed size and distance such that the chosen granularity is considered sufficient. By employing a spatial averaging technique the camera quantization error and sensor noise are reduced. The values of the downsampled image are concatenated for each frame and color channel, resulting in a data cube with dimension $34 \times 25 \times 3xN$, where N is the number of frames. For the pulse extraction from the data in the cube we use a temporal sliding window of 10 seconds (150 frames) with 8 seconds overlap. Due to the rigidness of the grid, it becomes evident what influence sharp edges may have on the computation of the mean. Especially edges that appear both in and out of a subregion (*e.g.* due to movements) over the course of the temporal window can significantly affect the temporal signal. For our sleep monitoring application movements are however limited, and the main challenge is to accurately detect the living tissue with the unpredictable appearance of the skin. Furthermore, we will present a hybrid approach later on which allows tracking of the detected living tissue to improve motion robustness.

3.2. Pulse extraction

For robust pulse extraction in NIR we use the PBV method [4], as the other two state-of-art methods CHROM [3] and POS [23] are based on assumed knowledge of the main distortion directions in visible light, which requires non-obvious adaptation for NIR-wavelengths, *e.g.* there is no standardized skin-color in NIR. A possible drawback of the PBV approach is the dependency of the pre-determined pulse signature $\vec{P}_{bv}$ on the pulse quality. The optimal $\vec{P}_{bv}$ varies according to camera specifications and light conditions. For the intended setup (sleep scenarios) both these influences are controlled and therefore the pulse signature can be accurately determined. This method of pulse extraction is performed on each of the available subregions resulting in 850 signals of which, depending on the subject size in the frame, a portion contain a pulse signal.

3.3. Similarity mapping

After extracting the pulse-signals $S_{i,j}$, where i, j indicates the row and column index of the specific subregion, the next step is to effectively exploit the pulse-signal as feature to distinguish living and non-living tissue. It is previously noted by Wang *et al.* [24] that skin regions belonging to the subject share pulse-similarities with each other in terms of phase and frequency, whereas the ones extracted from non-living tissue are uncorrelated. Spectral analysis of the created signals therefore revolves around finding the similarities between the available pulse signals to identify living-tissue by computing a similarity matrix. The high

computational load of the similarity matrix is reduced by employing two pre-processing steps using the intrinsic signal characteristics. We will first summarize the calculation of the similarity matrix and hereafter describe the proposed pruning steps.

Similarity matrix

A reduced and modified version of a similarity matrix as originally developed by Wang *et al.* [24] is used to identify similarities between the pulse signals of the subregions using four similarity features:

1. *Spectrum peak amplitude:*

$$F = \max_{f \in [40,240]} (\mathscr{F}(\vec{S}^L_{ij}) \circ \mathscr{F}(\vec{S}^L_{i'j'})^*), \tag{1}$$

where f is the frequency in beats-per-minute (BPM), $\circ$ denotes the element-wise product, $*$ is the conjugation and $\mathscr{F}(\cdot)$ represents the Fourier transform.

2. *Spectrum phase:*

$$P = max(\mathscr{F}^{-1}(NCC)), \tag{2}$$

with

$$NCC = \frac{\mathscr{F}(\vec{S}^L_{ij}) \circ \mathscr{F}(\vec{S}^L_{i'j'})^*}{||\mathscr{F}(\vec{S}^L_{ij}) \circ \mathscr{F}(\vec{S}^L_{i'j'})^*||_2}, \tag{3}$$

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

3. *Spectrum entropy:*

$$E = \frac{\sum_{f=40}^{240} NCC(f) log(NCC(f))}{log(240-40)} \tag{4}$$

4. *Inner product:*

$$I = < \frac{\vec{S}^L_{ij}}{||\vec{S}^L_{ij}||_2}, \frac{\vec{S}^L_{i'j'}}{||\vec{S}^L_{i'j'}||_2} >, \tag{5}$$

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

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

$$\Sigma = 1 - \exp\left(- \frac{(F \circ P \circ E \circ I)^2}{2\sigma^2_{F,P,E,I}} \right), \tag{6}$$

where $\sigma_{F,P,E,I}$ represents the entry-wise standard deviation between four matrices. The similarity matrix Σ contains the mutually connected subregions. In order to find the subregions belonging to the subject, a matrix decomposition technique is used to factorize Σ into multiple eigenvectors and eigenvalues by using the singular value decomposition:

Figure 3: (left) Eigenvalues after decomposition of the similarity matrix, (middle) projection of the first eigenvector similarity matrix, and (right) weight map of intrinsic signals

$$\Sigma = U \Lambda V^T. \tag{7}$$

The result is a set of eigenvectors U and eigenvalues Λ; each eigenvector indicating the correlation between subregions whereas the eigenvalues indicate the strength of this correlation. By inspecting the values of the sorted eigenvalues as displayed in Fig. 3, the number of subjects can be identified; for the intended scenario this mostly is one, indicated by the large drop after the first eigenvalue. This results in the first eigenvector U describing the correlation within $S_{i,j}$ between the subregions.

While this similarity approach works well for estimation of living-tissue, its runtime leaves to be desired due to the calculation of the many similarity features as described earlier; the runtime scales exponentially with the number of subregions. Especially for our scenario where only a small subset of the subregions contains skin, there are many redundant calculations. In order to reduce the runtime for the creation of the similarity matrix, a preprocessing step was devised which already prunes many non-relevant subregions prior to the similarity matrix calculation.

Pruning

The similarity matrix describes the correlation *between* signals, while the pruning is only based on the intrinsic properties of the pulse signals. By strictly focusing on the intrinsic signal properties, computations remain within bounds while still being able to prune many non-skin pixels to reduce the number of calculations for the similarity matrix. The pruning consists of two steps: 1) a pulse rate estimate is made based on intrinsic signal properties, and 2) outliers are pruned and signal quality is rated on signal-to-noise ratio (SNR).

1. **Pulse rate estimation (intrinsic)** $S_{i,j}$ is evaluated in order to estimate the pulse rate of the subject based on the combination of two different methods: 1) Principal Component Analysis (PCA) of $S_{i,j}$, and 2) the average pulse rate of the signals with the highest signal quality, where the selection of the principal component is guided by the previous pulse rate estimates.

2. **Outlier pruning** The estimated pulse rate ($\hat{PR}$) serves as the basis for the pruning stage. Subregions should have a frequency peak at or close to $\hat{PR}$ (a small margin of 6 BPM is accepted to adjust for small inaccuracies) in order to qualify as a candidate for the similarity matrix, all other subregions are removed. Additionally, the signals $S_{i,j}$ get weighted based on the SNR of the accepted subregions; resulting in an SNR-based weight-map as shown in Fig. 3. The unpruned, weighted, pulse-signals serve as the input to the similarity matrix as described in the previous paragraph.

Binary mask
The similarity map provides a weight-map which indicates the strength of correlation between subregions. This weight-map is used for estimating the final pulse- and SpO_2-signal of the available living tissue later on, but for now a binary mask is created from this mapping using an automated threshold which maximizes intra-class variability [14] to provide a proper indication of the living and non-living tissue within the frame. A simple threshold and morphological closing method is used for this:

$$simmap_{i,j} = binarize(U_{i,j})$$
$$simmap_{i,j} = simmap_{i,j} \oplus B$$
$$simmap_{i,j} = simmap_{i,j} \ominus B,$$

where $U_{i,j}$ indicates the re-mapped first eigenvector of the similarity matrix and B a $2x2$ square structure element.

3.4. ROI selection

As mentioned before, motion robustness is limited when using rigid blocks, whereas the commonly used detectors/trackers cannot cope with the unpredictable appearance of the skin during sleep. We therefore developed a novel hybrid method combining the rPPG-based subject detection with a tracker in which the tracker's ROI is updated when a valid region has been found. Vice versa, when the rPPG-based subject detection temporarily fails, *e.g.* due to movement, the tracker takes over for extracting the pulse signal until a proper region has been recaptured by the rPPG-based method. Firstly the tracker is briefly described, which is followed by the description of the method used to define a confidence metric for the ROI estimated by the similarity matrix.

Tracker
A relatively simple tracker is employed based on Kernelized Correlation Filters (KCF) [6]. This tracker provides a reasonable trade-off between accuracy and runtime. The tracker is initialized for each reliable ROI detected by the rPPG-based method. Note that tracker is only used for the extraction of the pulse signal during and just after the motion event and not for the estimation of the SpO_2-levels since SpO_2 measurements are highly susceptible to tracking inaccuracies and inclusion of non-skin pixels due to the rectangular bounding box.

Confidence metric
Similar to the pulse rate estimation used during the pruning step of the similarity matrix, a combination of several metrics is used to define a confidence metric for the selection of either the tracker or the rPPG-based estimated region. The quality of the estimated region is determined by three metrics, all related to the binary mask created from the mapping provided by the similarity matrix: 1) the SNR of the signal from the remaining ROIs in the binary mask, 2) the size of the largest cluster in the binary mask, and 3) the sparsity of binary mask.

1. Should the detected region be inaccurate, chances are that ROIs with lower SNR are selected. An empirically determined selected threshold-SNR value is selected to indicate the quality of the signals in the binary mask. The signals are weighted according to the binary mask $simmap_{i,j}$ and similarity mask $U_{i,j}$. The SNR of the mean of its non-zero entries is calculated and compared to the threshold-value (default: 4dB).

2. Living tissue is usually spatially clustered. The largest cluster should be of reasonable size relative to the total number of subregions ($N_{subregions}$) and should still be accepted when a part of the face is occluded due to body rotations and/or sheets (default: $0.0075 \cdot N_{subregions}$).

3. The sparsity of the binary mask is related to the previous metric in the sense that for one subject a good rPPG-mask usually only contains one significant cluster and not multiple smaller, more wide-spread clusters. The total number of ROIs present in the binary mask is compared to the largest available cluster and used to decline the rPPG-based ROI when these two quantities deviate too much (default: 0.75).

If all three criteria are satisfied the rPPG-based ROI is used for pulse extraction and SpO_2 estimation, otherwise the tracker's ROI is used.

3.5. Vital signs extraction

Pulse extraction
The presence of periodic pulse signal of the skin pixels is already extensively used to determine the presence of living-tissue in frame. A final step is made to extract the subject's pulse signal by combining the pulse signals from the subregions of the binary mask. This is achieved by using the

previously calculated SNR-map to weight the pulse signals within the binary mask. When the tracker is used, the pulse signal is simply the signal extracted from the ROI of the tracker. As mentioned in the subregion creation, a pulse signal segment is obtained every 2 seconds based on a moving window of 10 seconds. This pulse signal segment is overlap-added to its previous iterations to obtain the final pulse signal.

3.6. SpO_2 estimation

Compared to pulse-extraction, the requirements for the ROI are significantly more strict for SpO_2-estimation due to its high susceptibility towards data pollution. Any inclusion of non-skin affects the *relative* (DC-normalized) amplitudes and hence the measurement. Therefore, the estimation method differs from pulse-extraction on the following aspects:

- Only the binary mask $simmap_{i.j}$ and weight map $U_{i,j}$, created by the rPPG-based pipeline using the pulsatile information serve as bases for the SpO_2 estimation.
- The Gaussian smoothing is skipped as an initial step where background inherently gets mixed in with the PPG signal. Therefore the signals extracted from the downsampled, unsmoothed, frames are fed to the APBV method [18] for SpO_2 estimation.

The reported SpO_2 estimates in the results section are "raw", meaning that these are not post-processed, *e.g.* filtered or any type of outlier-rejection.

4. Experiments

4.1. Experimental setup

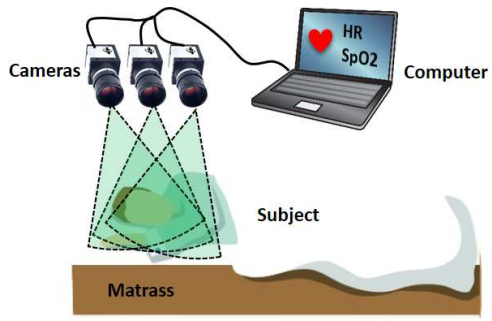

Figure 4: Experimental setup.

The experimental setup consists of three identical monochrome cameras, type Manta of Allied Vision Technologies GmbH, which capture the frames synchronously at 15fps with a resolution of 968×728 pixels and with 8-bit depth. Each camera is equipped with a 14mm lens and an optical filter to capture a specific part of the light spectrum. For our benchmark dataset filters with center-wavelengths of 760, 800 and 890nm are used. The reasoning for this wavelength selection is twofold: 1) the clinical desire to measure the vital signs in darkness (i.e. sleep scenarios) and 2) the wavelengths are sufficiently spaced to provide enough amplitude and SpO_2 *contrast* between the different channels while remaining within the spectral sensitivity of the camera sensor. The three cameras are placed perpendicular to the subject and are located above the pillow. The distance between the subject and cameras is 1.7m and is selected such that the cameras cover the typical width of a bed, as displayed in Fig. 4. The frames are registered by an affine transformation. As reference, a pulse-oximeter (Philips M1191B) is attached to the right index finger and connected to a Philips MP50 patient monitor, which data is stored synchronously with the video data. Two light units with incandescent light bulbs placed at both sides of the subject's head at a distance of 1.6m are used to provide diffuse, homogeneous illumination. Broadband light sources are used as they cover the selected wavelengths and sources limited to NIR which satisfy our requirements on homogeneity were not available at the time of the recordings.

4.2. Dataset

To simulate realistic sleep scenarios we asked five healthy volunteers to sleep in different supine positions. The three main sleep positions are side, back and stomach, where side is by far the most common [7]. Since sleeping on the stomach leads to occlusion of all skin pixels, it was not included in the protocol. Besides the different sleep positions we asked the subjects to perform out-of-frame movements, from supine position on the back to upright position, to simulate a bed-exit event and verify the capability of the methods to re-capture the ROI. Institutional Review Board approval and informed consent were obtained prior to measurements.

4.3. Benchmark methods

We compare the performance of our proposed "Hybrid" and "rPPG-only" methods with four benchmark methods: 1) the "Voxel-Pulse-Spectral" (VPS) method of Wang *et al.* [24], 2) the PPG-based method of Gibert *et al.* [5], 3) the Viola-Jones face detector [22], and 4) a tracker [6] where the rectangular ROI is manually initialized at the first frame.

4.4. Evaluation metrics

The performance of the different methods is evaluated for each evaluation window of 150 samples with the following metrics for the cardiac pulse signal and oxygen saturation:

Pulse

- Signal-to-noise ratio (SNR) $\Rightarrow$ $SNR = 10\log_{10}\left(\frac{\sum_{f=40}^{240}(U(f)S_f(f))^2}{\sum_{f=40}^{240}(1-U(f)S_f(f))^2}\right)$, where S_f is the frequency spectrum of the pulse signal, S, f the frequency in BPM and U is a binary template mask with ones around the fundamental frequency and harmonics, and zeros elsewhere.
- Mean absolute error (MAE) $\Rightarrow$ $MAE = \frac{\sum_{i=1}^{N}|PR^{cam}(i)-PR^{ref}(i)|}{N}$, where PR^{cam}, PR^{ref} indicate the estimated pulse rate extracted from the camera and reference PPG signal, respectively.
- Root-mean-square error (RMSE) $\Rightarrow$ $RMSE = \sqrt{\frac{(PR^{cam}(i)-PR^{ref}(i))^2}{N}}$
- Coverage (C): the percentage where a ROI could be detected.

SpO$_2$

- MAE: see earlier description for pulse, where for SpO$_2$ the pulse rates are replaced with oxygenation levels.
- MedAE: the median of the absolute error to reduce the effect of outliers, *e.g.* during the change in sleep position.
- PERC: the clinically acceptable accuracy criterion is specified in the International Standard for pulse-oximeter manufacture ISO 80601-2-61-2011 [8], which requires an accuracy of $\leq 4\%$ within the saturation range $70-100\%$. PERC expresses the percentage where the measurement satisfies this criterion.

5. Results and Discussion

An overview of the performance of our method is displayed in Fig. 5, whereas a comparison with the benchmark methods is displayed in Fig. 6. We will now discuss the pulse, SpO$_2$ and runtime results separately.

Pulse The results of the pulse extraction are summarized in Table 1. For the sleep positions scenario it can be observed that although our proposed methods provide the best results, also most of the benchmark methods are reasonably capable of extracting the pulse signal. The Viola-Jones method fails during sleeping on the side leading to a coverage of only 38%, which was expected because the detector is mostly trained with images of frontal faces. From Fig. 6 the contribution of the tracker can be recognized during the changes in sleep position; whereas the "rPPG-only" method is temporarily unable to extract the pulse signal, the "hybrid" method is capable to extract the signal during the turning event. For the performance evaluation of the "out-of-frame" scenario we discarded the moments where the subject was absent. An example of the detected ROIs with our proposed method is visualized in Fig. 7.

	Sleep positions				Out-of-frame		
Method	SNR (dB)	MAE (BPM)	RMSE (BPM)	C (%)	SNR (dB)	MAE (BPM)	RMSE (BPM)
Hybrid	**5.74**	**1.85**	**3.28**	100	6.13	**1.85**	**2.75**
rPPG-only	5.58	2.81	5.23	100	**6.47**	2.22	3.24
VPS [24]	4.24	3.33	5.75	100	4.14	2.02	3.14
Gibert [5]	4.40	3.32	5.51	100	3.52	2.86	5.31
Viola-Jones [22]	2.14	7.87	11.4	38.4	1.89	6.55	15.9
Tracker [6]	3.18	3.90	7.37	100	1.52	19.4	31.9

Table 1: Results pulse extraction.

	Sleep positions			Out-of-frame		
Method	MAE (pp)	MedAE (pp)	PERC (%)	MAE (pp)	MedAE (pp)	PERC (%)
rPPG-only	**4.11**	**1.67**	**79.3**	**2.55**	**1.60**	**87.2**
VPS [24]	6.73	4.82	39.3	5.09	4.07	44.8
Gibert [5]	5.29	3.62	56	5.75	4.42	47.1
Viola-Jones [22]	10.1	9.43	9.78	7.71	6.66	28.7
Tracker [6]	5.06	3.37	60.9	5.28	3.24	58.6

Table 2: Results SpO$_2$ estimation.

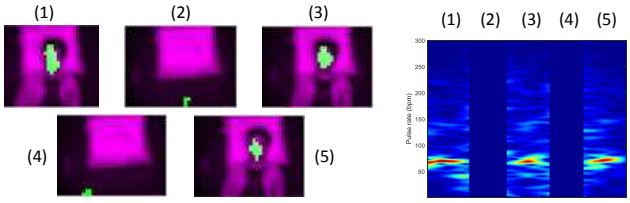

Figure 7: The detected ROI during out-of-frame movements.

SpO$_2$ The results of the SpO$_2$ estimations are summarized in Table 2. It can be observed that our method outperforms all benchmark methods for both scenarios, especially after a change in position. This likely results from the much cleaner ROIs, i.e. ROIs which only contain skin pixels. For SpO$_2$ any inclusion of non-skin pixels within the ROI has a direct effect on the measurement because of the DC-normalization. This effect is best demonstrated by the results of the "tracker" method, visualized in Fig. 6; during the first sleeping pose where the tracker is accurately initialized the correspondence between the SpO$_2$ estimates and the reference is very good, after a change in position the method renders inaccurate due to inclusion of non-skin pixels. When comparing the performance of the rPPG-based methods VPS [24] and Gibert [5], it becomes evident that the absence of relevant color features in NIR leads to non-discriminative spatio-temporal regions.

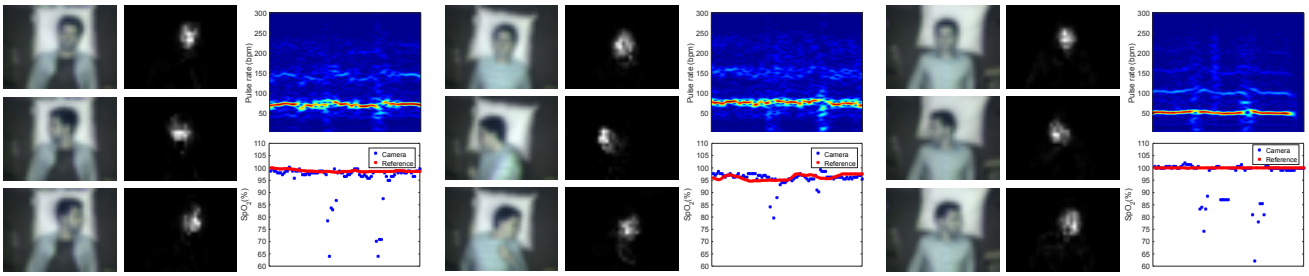

Figure 5: Overview of the results obtained with our proposed "hybrid" method on the recordings with three different sleep positions. The first column is the downsampled input frame, the second column the pixel weight map used for the extraction of the pulse signal and the estimation of the SpO_2 value, as visualized in the third column.

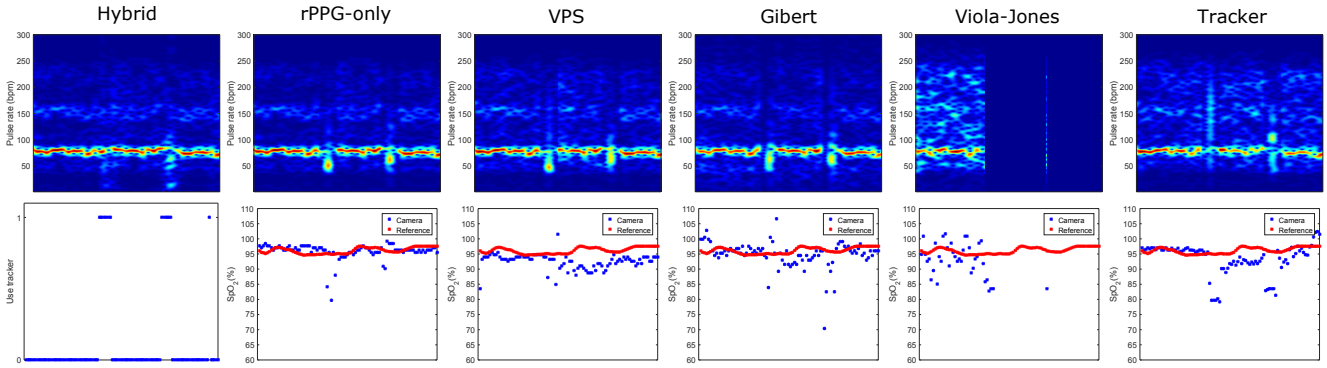

Figure 6: Overview of the results of all evaluated methods on the recordings with three different sleep positions. The top row are spectrograms of the pulse signals and the bottom row the SpO_2 estimates. The first figure on the bottom row indicates when the tracker is used for the proposed hybrid method.

Evaluation	Task	Similarity matrix (pruned)	Similarity matrix (full)	VPS (K=20) [24]	Gibert [5]	Viola-Jones [22]	Tracker [6]
Per frame	Pre-processing	0.07s	0.07s	0.07s	-	0.49s	-
Per frame	Detector/Tracker	0.11s	0.11s	-	0.44s	-	0.11s
Per interval	(Parallel) pulse extraction	0.65s	0.65s	0.65s	0.01s	0.02s	0.01s
Per interval	PR estimation	0.08s	-	-	-	-	-
Per interval	Pruning	0.04s	-	-	-	-	-
Per interval	Similarity matrix	2.0s	40s	-	-	0.02s	-
	Total	**8.17s**	**46.17s**	**2.75s**	**13.5s**	**14.74s**	**3.31s**

Table 3: Runtime overview of the evaluated methods.

Runtime We evaluated the runtime of our proposed method(s) and the benchmark methods using a notebook with an Intel Core 2.70GHz i5-6400 CPU, 8GB RAM and an NVIDIA GeForce GTX 970 GPU. The results are displayed in Table 3.

6. Conclusions

In this paper, we presented a framework for fully-automatic remote pulse rate and SpO_2 estimation during sleep. The limited color contrast in NIR and the unpredictable appearance of the available skin-portion during sleep negatively impacts the performance of current state-of-the-art methods. We presented a method which combines the benefits of rPPG-based features to find static "living pixels" with those of a tracker that can bridge relatively short intervals where the subject moves. The framework has been successfully validated on a dataset were realistic sleeping conditions have been simulated by healthy subjects. Especially for the critical estimation of SpO_2 our proposed method outperforms the benchmark methods because of the much cleaner ROI. The next step is to validate the system in a clinical setting on patients with sleep disorders and expected associated SpO_2 variations.

References

[1] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 2017. 2

[2] S. Chaichulee, M. Villarroel, J. Jorge, C. Arteta, G. Green, K. McCormick, A. Zisserman, and L. Tarassenko. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 266–272. IEEE, 2017. 2

[3] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 3

[4] G. De Haan and A. Van Leest. Improved motion robustness of remote-ppg by using the blood volume pulse signature. *Physiological measurement*, 35(9):1913, 2014. 3

[5] 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*, pages 449–453. IEEE, 2013. 2, 3, 6, 7, 8

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015. 5, 6, 7, 8

[7] C. Idzikowski. Sleep position gives personality clue. http://news.bbc.co.uk/2/hi/health/3112170.stm, 2003. 6

[8] Particular requirements for basic safety and essential performance of pulse oximeter equipment. Standard, International Organization for Standardization, Geneva, CH, Apr. 2011. 7

[9] V. Jeanne, F. J. De Bruijn, R. Vlutters, G. Cennini, and D. Chestakov. Processing images of at least one living being, Sept. 24 2013. US Patent 8,542,877. 2

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002. 2

[11] G. Lempe, S. Zaunseder, T. Wirthgen, S. Zipser, and H. Malberg. Roi selection for remote photoplethysmography. In *Bildverarbeitung für die Medizin 2013*, pages 99–103. Springer, 2013. 2

[12] H. Liu, T. Chen, Q. Zhang, and L. Wang. A new approach for face detection based on photoplethysmographic imaging. In *International Conference on Health Information Science*, pages 79–91. Springer, 2015. 2

[13] N. C. on Sleep Disorders Research. Wake up america: a national sleep alert, 1993. 1

[14] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 5

[15] S. Paruthi and J. Josephson. Centers for disease control (U.S.), and centers for disease control and prevention (U.S.). *Morbidity and Mortality Weekly Report (MMWR)*. 1

[16] W. H. Tsai, W. W. Flemons, W. A. Whitelaw, and J. E. Remmers. A comparison of apnea–hypopnea indices derived from different definitions of hypopnea. *American journal of respiratory and critical care medicine*, 159(1):43–48, 1999. 2

[17] M. van Gastel, S. Stuijk, and G. de Haan. Motion robust remote-ppg in infrared. *IEEE Transactions on Biomedical Engineering*, 62(5):1425–1433, 2015. 1

[18] M. Van Gastel, S. Stuijk, and G. De Haan. New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring. *Scientific reports*, 6:38609, 2016. 1, 6

[19] M. van Gastel, S. Stuijk, and G. de Haan. Robust respiration detection from remote photoplethysmography. *Biomedical optics express*, 7(12):4941–4957, 2016. 1

[20] M. Van Gastel, S. Stuijk, and G. De Haan. Camera-based pulse-oximetry-validated risks and opportunities from theoretical analysis. *Biomedical Optics Express*, 9(1):102–119, 2018. 1

[21] R. Van Luijtelaar, W. Wang, S. Stuijk, and G. de Haan. Automatic roi detection for camera-based pulse-rate measurement. In *Asian Conference on Computer Vision*, pages 360–374. Springer, 2014. 2

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001. 2, 6, 7, 8

[23] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017. 3

[24] W. Wang, S. Stuijk, and G. De Haan. Unsupervised subject detection via remote ppg. *IEEE Transactions on Biomedical Engineering*, 62(11):2629–2637, 2015. 2, 3, 4, 6, 7, 8

[25] W. Wang, S. Stuijk, and G. de Haan. Living-skin classification via remote-ppg. *IEEE Transactions on Biomedical Engineering*, 64(12):2781–2792, 2017. 2, 3