Automatic RoI Detection for Camera-based Pulse-rate Measurement

Ron van Luijtelaar[1], Wenjin Wang[2], Sander Stuijk[2] and Gerard de Haan[2]

[1]Profit Consulting, Netherlands
[2]Electronic System Group, Electrical Engineering Department, Eindhoven University of Technology, Netherlands

Abstract. Remote photoplethysmography (rPPG) enables contactless measurement of pulse-rate by detecting pulse-induced colour changes on human skin using a regular camera. Most of existing rPPG methods exploit the subject face as the Region of Interest (RoI) for pulse-rate measurement by automatic face detection. However, face detection is a sub-optimal solution since (1) not all the subregions in a face contain the skin pixels where pulse-signal can be extracted, (2) it fails to locate the RoI in cases when the frontal face is invisible (e.g., side-view faces). In this paper, we present a novel automatic RoI detection method for camera-based pulse-rate measurement, which consists of three main steps: subregion tracking, feature extraction, and clustering of skin regions. To evaluate the robustness of the proposed method, 36 video recordings are made of 6 subjects with different skin-types performing 6 types of head motion. Experimental results show that for the video sequences containing subjects with brighter skin-types and modest body motions, the accuracy of the pulse-rates measured by our method (94%) is comparable to that obtained by a face detector (92%), while the average SNR is significantly improved from 5.8 dB to 8.6 dB.

1 Introduction

Home-based healthcare monitoring is growing in popularity as modern technologies allow the medical devices to be comfortably and easily accessible. In order to assess the most basic body functions, vital physiological signals such as pulse-rate, respiration rate or blood oxygen saturation have to be measured. An extensively employed noninvasive method for pulse-rate monitoring is the Photoplethysmography (PPG) introduced in 1937 [1]. It uses contact probes, consists of a dedicated light source and photo-detector, to detect vital fluctuation of optical absorption in skin tissues. However, the contact probes that need to be attached to human body is highly inconvenient and uncomfortable to use, i.e., it cannot be used in scenario like fitness or surveillance; it is strictly prohibited in case of sensitive subjects with damaged skin or neonates. In contrast, non-contact based vital signs monitoring is preferred for its unobtrusiveness and noninvasiveness. A promising alternative is the recently introduced camera-based pulse-rate monitoring - remote photoplethysmography (rPPG) [2, 3].

In the human cardiovascular system, blood pulse propagating throughout the body changes the blood volume in vessels. rPPG measures the pulse-signal by detecting the optical absorption of haemoglobin variation across the light spectrum using a regular camera. Although the recent developments in rPPG technology have demonstrated promising performance in measuring pulse-signals under ambient light conditions [4–7], a basic problem is not addressed in all these existing approaches: how to optimally initialise the Region of Interest (RoI) for pulse-rate monitoring. In previous rPPG work, the most commonly used approach for selecting the RoI is to (1) assume that the face contains essential skin pixels for pulse-signal extraction, (2) perform automatic face detection using a Viola-Jones face detector [6, 4]. However, this approach is restricted to the face-similar RoI and likely to fail when the face is invisible or occluded. Other skin segmentation techniques that only based on the assumptions of human skin appearance (e.g., skin colour or texture) [8] cannot distinguish the skin regions from background scenes with skin-similar patterns.

In 2011, Schmitz [5] proposed an automatic skin detection method using the pulse-signal as a feature, since characteristics of pulse-signals can prevent the false detection of skin-similar objects in the scene. But this method is typically designed for stationary subjects, which cannot tackle with subjects' body motions. Therefore, in this paper, we aim to improve the motion robustness of automatic RoI detection for rPPG monitoring. The proposed method consists of three steps: (1) Subregion tracking: the frames in a pre-defined time interval are segmented into subregions for local pulse-signal extraction; (2) Feature extraction: a set of spatiotemporal features (e.g., pulse-signal frequency and skin chromaticity) are constructed to discriminate the skin/non-skin regions ; and (3) Skin-region clustering: the subregions are classified into the skin/non-skin clusters using the feature vectors, and the skin cluster is found as the RoI for pulse-rate monitoring.

The organisation of this paper is as follows: In section 2, we provide an overview of the proposed method and illustrate the steps in detail. In section 3, the experiment and evaluation metric are set up to verify the proposed method. We discuss the experimental results and show the findings in section 4 and finally in section 5, we draw the conclusions of this work.

2 Method

2.1 Subregion generation

Given a video sequence containing a subject, our goal is to find the RoI that is optimal for pulse-rate measurement. In fact, the pulse-signal can only be extracted from *skin pixels* in the video, so the non-skin pixels (e.g., background) should be discarded for deriving a clean and accurate pulse-signal. Therefore, we propose to initialise a RoI that only contains the skin pixels. Since no prior knowledge/assumption is posed on the location of skin pixels, we first segment the whole video into subregions, where the pulse-signals can be locally measured without inference. Afterwards, the subregions that present clean pulse-signals are automatically selected to initialise the RoI for pulse-rate monitoring.

Since pulse-signals need to be measured in the time-domain, the subregions generated in each video frame have to be concatenated in a time interval (e.g., the interval defined for pulse-signal extraction). Considering the influence of subject motions, the subregions formulated in the time domain may shift the location and lose the temporal consistency. To solve this problem, we propose to segment the video based on the feature corners, which are relatively easy to be tracked between consecutive frames. A spatial mesh, consisting of multiple triangles, is constructed to segment the whole image into subregions by connecting the tracked feature corners.

In this step, we employ the Harris corner detection [9] to find the salient feature points for tracking, which is only ran on the first frame of the pre-defined interval. Subsequently, the detected feature corners are temporally tracked across the remaining frames within the interval using the pyramid optical flow developed by Lucas & Kanade [10]. The outliers, i.e., the corners that disappear or exhibit large tracking errors, are rejected in a forward-backward validation procedure for maintaining the tracking coherence [11]. Finally, we rely on the Delaunay triangulation [12] to connect the tracked feature corners for segmenting the whole image into subregions spatially. The reason for using Delaunay triangulation is that it attempts to maximise the minimum angle of all the angles of the triangles during the triangulation, and thus leads to a more compact image partition, i.e., skinny triangles are avoided.

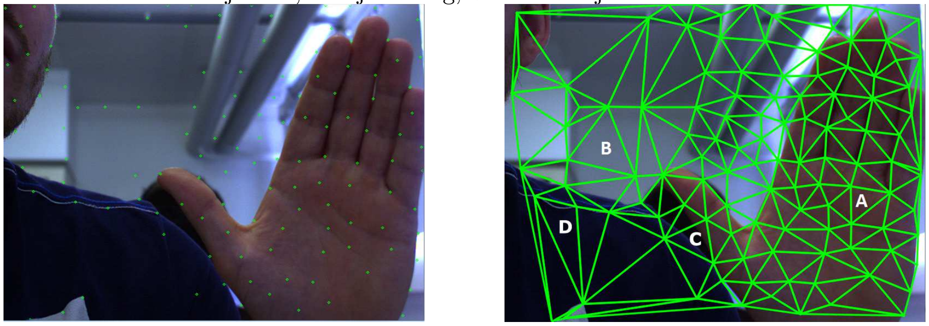
An example of this step is shown in Figure 1: (1) shows the tracked feature corners in a predefined time interval. In textured regions like the hand, more feature corners can be detected than the plain regions (e.g., background wall); and (2) shows that the whole image is segmented into multiple triangular subregions, in which some are skin regions (e.g., subregion A) that contain pulse-signals while others are background scenes (e.g., subregion B, C and D). The following steps aim to distinguish the skin regions from non-skin regions (background) by feature extraction and clustering.

2.2 Feature extraction

Spectral features Following the region tracking, we use the chrominance-based method (CHROM) proposed by de Haan & Jeanne[6] to extract the pulse-signals from subregions in the time interval. CHROM uses a combination of two orthogonal chrominance signals $X = R - G$ and $Y = 0.5R - 0.5G - B$ and is therefore capable of eliminating the motion-induced specular reflection changes in a white light luminance condition. In order to enable the correct functioning with colour light source, skin-tone standardisation is applied by adapting the equation as:

$$\begin{array}{l} X_n = 3R_n - 2G_n \\ Y_n \ = 1.5R_n + G_n - 1.5B_n \end{array}, \tag{1}$$

where R_n, G_n and B_n are colour channels normalised to their mean over the interval, which make the pulse-signal independent of the brightness of the light



(1) (2)

Fig. 1. An example of subregion generation in the video containing a subject hand. (1) the feature corners in the whole image are detected and tracked in the predefined time interval, (2) Delaunay triangulation is performed to connect the feature corners as an oversegmented mesh, where each triangle is a subregion for local independent pulse extraction.

source. The pulse-signal is defined as:

$$S = X_n - \alpha Y_n, \tag{2}$$

with

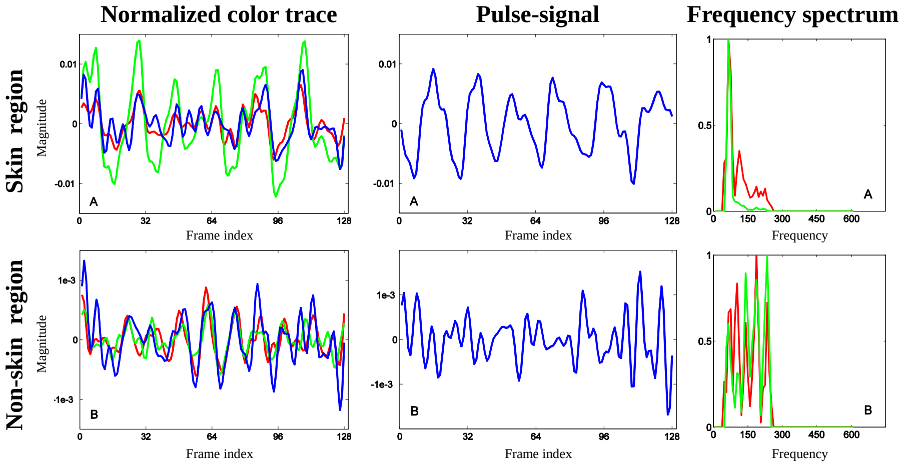
$$\alpha = \frac{\sigma(X_n)}{\sigma(Y_n)}, \tag{3}$$

where $\sigma(\cdot)$ denotes the standard deviation operation. Given the fact that the pulse-rate of a healthy subject is within the frequency-band $[40, 220]$BMP, we band-pass filter the S to eliminate out-of-band frequency noise. Figure 2 shows an example of normalised RGB traces and corresponding pulse-signals extracted from skin and non-skin regions respectively.

Obtaining the pulse-signal for each subregion, we can extract the spectral features (e.g., frequency and phase) to summarise the unique characteristics of skin regions. A commonly used term to describe a pulse-signal is the "periodicity", so spectral analysis is performed on S using the Discrete Fourier Transfer (DFT). To retrieve an acceptable spectral resolution, the length of the interval for deriving a pulse-signal is set to $N = 128$ (6.4 sec duration in 20 FPS camera) while its frequency spectrum is interpolated by 512 FFT bins. In practice, subject motions and camera noise may degrade the pulse-signals' quality. To reduce the influence of such noise, the pulse-signals are auto-correlated before extracting the periodicity features as:

$$A[\tau] = \sum_{n=0}^{N-1} S[n]S[n-\tau], \quad -\frac{1}{2}N < \tau < \frac{1}{2}N. \tag{4}$$

Figure 2 shows the spectrum magnitude of the pulse-frequency and its auto-correlation. Apparently, the spectrum derived from skin region shows a clear frequency-peak within the pulse-rate frequency-band, whereas the spectrum derived from non-skin region does not show a pattern (random noise). In addition,



(1) (2) (3)

Fig. 2. The signals in the first row are derived from a skin region while the signals in the second row are derived from a non-skin region. (1) the normalised colour channels R_n (red), G_n (green) and B_n (blue); (2) the extracted pulse-signals S using CHROM; and (3) the normalised frequency spectrums (red) and corresponding auto-correlation A (green) of the pulse-signal.

it shows that auto-correlation can significantly suppress the frequency noise in the spectrum.

The frequency F corresponding to the highest peak in the spectrum is considered to be the pulse frequency, which is calculated as:

$$F = \hat{k} \cdot \frac{f_s}{N}, \tag{5}$$

with

$$\hat{k} = \arg\max_{k} \left\{ |\hat{M}[k]| \mid k \in [0, (N-1)/2] \right\} \tag{6}$$

where f_s represents the sampling frequency; N is the parameter controlling the resolution of frequency spectrum; $\hat{M}$ denotes the spectrum magnitude. Furthermore, the pulse-signals extracted from different subjects mostly undergo a different phase, so another commonly used term “spectrum phase” is employed as a discriminate feature, which is calculated as:

$$\theta = \arctan \left(\frac{\mathrm{Im}(S[\hat{k}])}{\mathrm{Re}(S[\hat{k}])} \right) \tag{7}$$

where $\theta \in [-\pi, \pi]$. At this point, the extracted pulse-signal is summarised by two spectral features: frequency (F) and phase (θ), which is sufficient to distinguish between a pulse-signal and a noise signal.

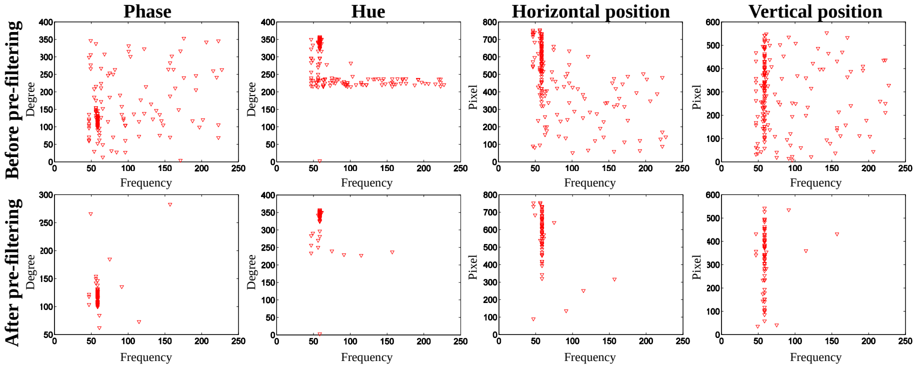
Spatial features Inspired by the colour-based skin segmentation, we exploit the chromaticity property of skin for skin/non-skin regions classification. In spatial domain (each single frame), image pixels are transformed from RGB colorspace to Hue-Saturation-Value (HSV) colorspace, where the intensity is separated from the intrinsic information of object chromaticity. After that, we only adopt H (hue) component as the spatial feature. Note that in this study, the skin chromaticity feature is exploited in an unsupervised way, which is different from earlier works that used prior knowledge or pre-assumed threshold for skin segmentation. There are two observations behind this step: (1) the skin pixels belonging to the same subject are homogeneous in hue, (2) eliminating S and V components can enhance the robustness towards illumination changes and shadows.

Moreover, the skin regions corresponding to a particular body part are closely located from each other between adjacent frames. The position of the geometric centre $\overline{P}$ of each subregion is calculated as:

$$\overline{P} = \frac{1}{N} \sum_{i=0}^{N-1} P_i, \tag{8}$$

where N is the number of tracked feature corners belonging to a specific subregion; P_i is the spatial position of ith feature corner of this subregion.

To this end, a multi-dimensional feature vector, $\overrightarrow{v} = [F, \theta, H, \overline{P}]$ is constructed as the representation of a subregion, which takes the advantage of pulse and chromaticity properties of skin. Figure 3 shows the scatter plots of subregions' distribution in the feature space, where each point represents a subregion. As can be seen, there is an area has higher density than others, which implies that a group of feature vectors show similar patterns in the feature space.



(1) (2) (3) (4)

Fig. 3. An example of subregions' distribution in multi-dimensional feature space, where each point represents a subregion. The scatter plots in the first row represent all the oversegmented subregions without pre-filtering; the scatter plots in the second row represent the remaining subregions after the pre-filtering step, where the subregions that explicitly belong to background are filtered out.

2.3 Skin-region clustering

Pre-filtering To reduce the obvious non-skin subregions before the skin-region clustering, a pre-filtering step is performed to identify the explicit background regions using certain characteristics of the pulse-signal. Here we introduce several essential conditions for classifying a subregion into non-skin category in a prior stage.

Due to the optical absorptions of blood pulse, the amplitudes of normalised colour traces from skin regions are significant larger than that from non-skin regions, as can been seen in Figure 2. So the colour variations induced by pulse and noise can be differentiated by the maximum amplitude of normalised colour traces as:

$$\delta_{C_n,max} = \max\{C_n[i] \mid i \in [0, N-1]\} \tag{9}$$

where N is the interval length; $C_n \in \{R_n, G_n, B_n\}$. In our method, the regions whose maximum amplitudes are not in the range $[0.005, 0.15]$ are identified as background.

Essentially, pulse-signals extracted from skin regions exhibit different amplitudes but remain similar phase due to the different blood absorption rates of the light spectrum. Assuming x and y to be two traces (e.g., R_n and G_n) belonging to different colour traces in a subregion, the normalised Sum of Absolute Differences (SAD) is calculated to assess the amplitude differences between the colour traces as:

$$\delta_{SAD} = \frac{\sum_i \mid x[i] - y[i] \mid}{\sum_i \mid x[i] \mid}, \tag{10}$$

where the subregions with δ_{SAD} outside the range $[0.05, 0.5]$ are identified as background. Afterwards, the Normalised Correlation Coefficient (NCC) between x and y is calculated to find the phase similarities between the colour traces as:

$$\delta_{NCC} = \frac{\sum_i x[i] \cdot y[i]}{\sqrt{\sum_i x[i]^2 \cdot \sum_i y[i]^2}}, \tag{11}$$

where the subregions with δ_{NCC} outside the range $[0.5, 0.99]$ are identified as background. Given that fact that the subregion with real pulse-signal presents a strong frequency-peak in the frequency domain, the ratio between the two largest peaks in the frequency spectrum is calculated to determine the presence of dominant frequency of the pulse as:

$$PR = \frac{\hat{M}_{max2}}{\hat{M}_{max1}}, \tag{12}$$

where $\hat{M}$ represents the amplitude-normalised frequency spectrum. For $\hat{M} \in [0, 1]$, the upper-bound suggests that the amplitudes of the two largest peaks are identical, while the lower-bound indicates that the spectrum is dominated by one frequency. Therefore, the regions where $PR > 0.6$ are identified as background. Moreover, the amplitude of frequency-peak is calculated as:

$$\delta_{\hat{M},max} = \max\{\hat{M}[i] \mid i \in [0, N-1]\}, \tag{13}$$

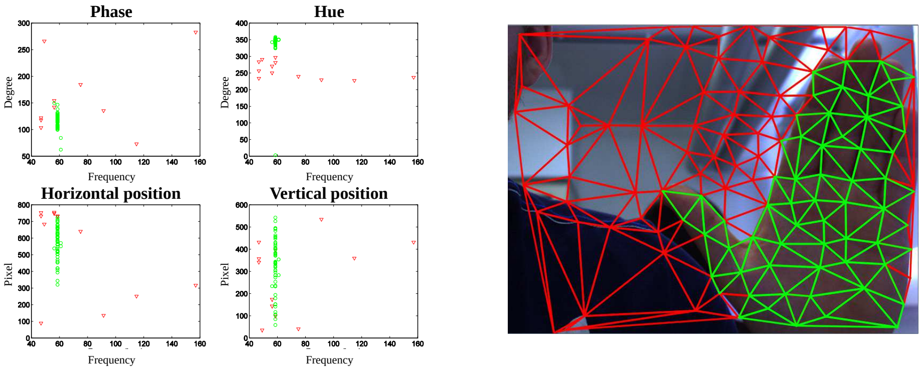
where the regions with low frequency energy ($\delta_{\hat{M},max} < 0.005$) are identified as background. Note that parameters discussed in this section are specified *empirically.* Figure 3 shows an example of subregions' distribution in feature space after pre-filtering step. The pre-filtering step can effectively prune the subregions that are obviously not belonging to the skin category.

DBSCAN Clustering After eliminating the explicit background regions in the pre-filtering step, the remaining regions will be clustered into different groups based on extracted feature vectors $\overrightarrow{v}$. In the scatter plots of the feature space of Figure 3, we can clearly recognise an area where the points' density is considerably higher than other areas. The points from the dense areas are considered to be skin (inliers with similar patterns), whereas the points located outside the dense areas are regarded as background noise (outliers). Based on such observation, we employ a clustering method called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13], the clustering method that typically relies on the density notion of data, to separate the skin/non-skin clusters in the feature space.

DBSCAN separates the data into three different types (e.g., cluster core point, border point and noise point) using a density threshold with two parameters: ε specifies a range and $minPts$ denotes a minimal number of sample in a cluster. In our method, data for clustering are the feature vectors in multi-dimensional space. The range of each element in $\overrightarrow{v}$ is controlled by a scaling factor. To determine whether the data samples are located between each other, a boolean factor is calculated as:

$$B_{i,j} = \begin{cases} \text{true} & \text{—}\overrightarrow{v}_i - \overrightarrow{v}_j\text{—} \leq \boldsymbol{\epsilon} \\ \text{false} & \text{elsewhere} \end{cases} \tag{14}$$

where $\overrightarrow{v}_i$ and $\overrightarrow{v}_j$ are two feature vectors; the range $\boldsymbol{\epsilon} = [3, 10, 10, 100]$ is empirically determined. Note that the DBSCAN algorithm exploits the inner distribution/structure of data (e.g., density) for clustering, which does not require



(1) (2)

Fig. 4. An example of DBSCAN clustering on feature vectors. (1) scatter plots of the feature vectors along different directions (e.g., phase, hue, x and y) after clustering. The green points represent the feature vectors classified to a skin cluster; (2) the corresponding clustering result in the image.

a predefined cluster number in advance. Figure 4(1) shows an example of clustering result using the specified parameters, where the feature vectors further to the dense area are classified as background. Figure 4(2) shows the corresponding result in a image, in which most of the subregions representing the skin are clustered correctly.

Cluster growing We observe that in some cases, a small amount of skin regions may be falsely identified as the background in the pre-filtering step. Although these regions lack the essential conditions for pre-filtering, they still show a pulse-signal which is sufficient to be clustered into skin group using DBSCAN. In order to recover these misclassified skin regions, a cluster growing procedure is performed. In this process, the average pulse-signal of a cluster is compared to the pulse-signals from its surrounding subregions using Eq.(11). When the pulse-signal of a surrounding subregion is highly correlated with the averaged pulse-signal from the skin cluster ($\gamma_{NCC} > 0.5$), the subregion is added to the cluster. Since the amplitudes of pulse-signals extracted from subregions inside the skin cluster are mostly different, the auto-correlated pulse-signals are normalised as:

$$Sn_i = \frac{S_i}{\sigma(S_i)} \tag{15}$$

where $\sigma(\cdot)$ corresponds to the standard deviation operator. To minimise the impact of noise when combining multiple pulse-signals into a single averaged pulse-signal, we perform the *alpha-trimmed mean* to reject the outliers with extreme values, i.e., α is set to 0.5 in our method.

3 Experiments

3.1 Experimental setup

The experimental setup consists of a standard 768x572 pixels, 8bit, global shutter CCD camera (type USB UI-2230SE-C of IDS GmbH) operating at 20 frames per second and focused at the subject's face using a flexible C-mount lens (Tamron 21VM412ASIR). The duration of each recording is set to approximately 90 seconds. The recorded videos are stored uncompressed. We recruit 6 healthy subjects with an equal number of male and female in 3 different skin-types according to the Fitzpatrick scale [14]: *Skin-type II*, *Skin-type III* and *Skin-type V*, as shown in the snapshot of Figure 5. To mimic the real use-case scenarios for recording, each subject is asked to perform 6 types of head motion: *stationary*, *translation*, *scaling*, *non-rigid* (talking), *rotation* and *mixed motion* (a mixture of all those movements). All recordings are made in a controlled environment using a single illumination source located in front of the subject. In parallel to the video, we synchronously record the raw pulse-oximeter data (PPG signal) from a transmissive pulse-oximeter finger clip of Contec Medical Systems (model CMS50E) using the USB protocol on the device. Both the rPPG and PPG signals are treated equally in the post-processing like band-pass filtering.

To calculate the pulse-rates, we detect the bin index of the frequency-peak using a 512 point FFT on the Hanning windowed pulse-signals.

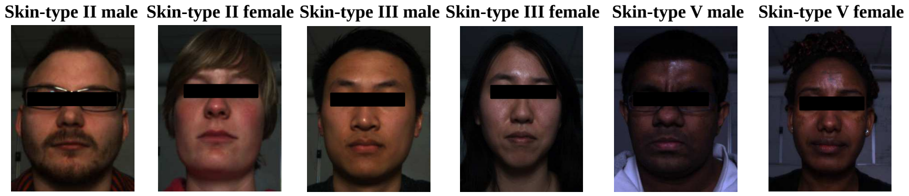


Fig. 5. A snapshot of six subjects with three different skin-types in our recorded video sequences.

3.2 Evaluation metric

In line with [6], the long-term pulse-signal across the complete video sequence is generated by overlap-adding the signal intervals with a hanning window. The quality of the extracted pulse-signal is measured in Signal-to-Noise Ratio (SNR), which is defined as the ratio between the pulse in-band spectrum energy (1st and 2nd harmonics of the spectrum) and the remaining spectrum energy as:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{f=40}^{240}(U_t(f)\hat{M}(f))^2}{\sum_{f=40}^{240}((1-U_t(f))\hat{M}(f))^2}\right), \tag{16}$$

where $\hat{M}$ is the spectrum amplitude of the pulse-signal S; f is the frequency range in Beats Per Minute (BPM); $U_t(f)$ a binary template function made by selecting the first and second harmonics according to the reference PPG-signal. In order to allow for pulse-rate variability ,$U_t(f)$ is defined in such a way that 22 bins in the 512 bin spectrum centred around the PPG pulse-rate (9 bins around the first harmonic and 13 bins around the second harmonic) are passed as in-band pulse-signal.

In addition, the pulse-rate accuracy, the percentage of alignment of the pulse-rates given by rPPG and reference PPG sensor, is measured for comparing the rPPG methods' performance. Note that the pulse-rate is calculated by detecting the frequency of the spectrum peak in a pre-defined time interval.

3.3 Compared method

To benchmark the proposed method, the pulse-signal obtained by our method is compared to that obtained by the rPPG method using automatic face detection [4]. In the compared method, a Viola-Jones face detector [15] is used to locate the face with a bounding box. The middle 60% width and 100% height of the bounding box is used as the RoI. The averaged RGB values of pixels inside the RoI are calculated for deriving the pulse-signal using CHROM. In cases that the face (e.g., side-view faces) cannot be detected, the pulse-signal is interpolated with zeroes to ensure the continuity.

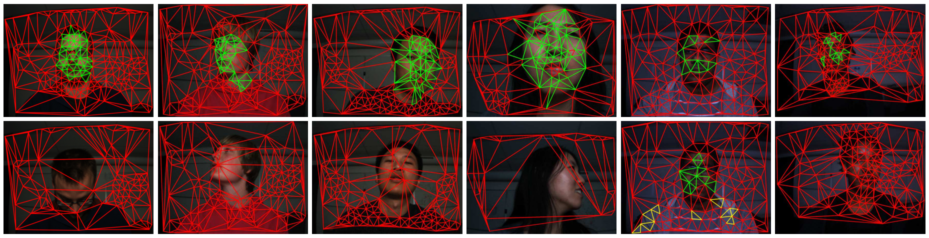


Fig. 6. Examples of detected RoI in recorded video sequences using our method. The first row indicates the cases where skin-regions are correctly found. The second row indicates the cases where our method fails to handle.

4 Results and Discussion

In Figure 6, it shows the examples of the found RoI for the video sequences using the implementation of the proposed method. As can be seen in the first row of Figure 6, most of the skin regions are correctly found for the Skin-type II and Skin-type III subjects, whereas the number of skin regions can be found for Skin-type V subjects is significantly reduced. The reason is related to the compositions of skin tissues: the melanin content is much higher in dark skin (e.g., Skin-type V), which absorbs part of the diffuse light reflections that carry the pulse-signal. Thus it leads to a reduced fraction of the diffusely reflected light from the skin [6]. Particularly for the Skin-type V male, our method fails to find the RoI for the complete video sequences using the fixed parameters. Thus the minimum distance between tracked feature corners is increased to 40 pixels while the minimum number of regions in a cluster is decreased to 3 for this specific subject. However, the assumption that the content (e.g., parameters) of a given region is fixed may not be valid anymore. At the same time, decreasing the cluster size could lead to more clusters, and thus increase the risk of false positive detection. As can be seen in the second row of Figure 6, our method fails to find the skin regions when the subjects' motions are vigorous. The performance degradation is caused by the failure of feature corners tracking. The drifting of feature corners during the large pixel displacement results in the shape change of triangles, which interrupts the temporal consistency for pulse-signal measurement.

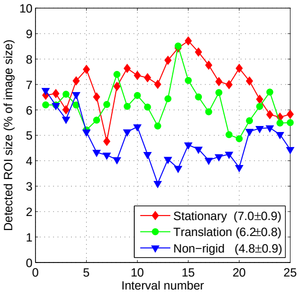
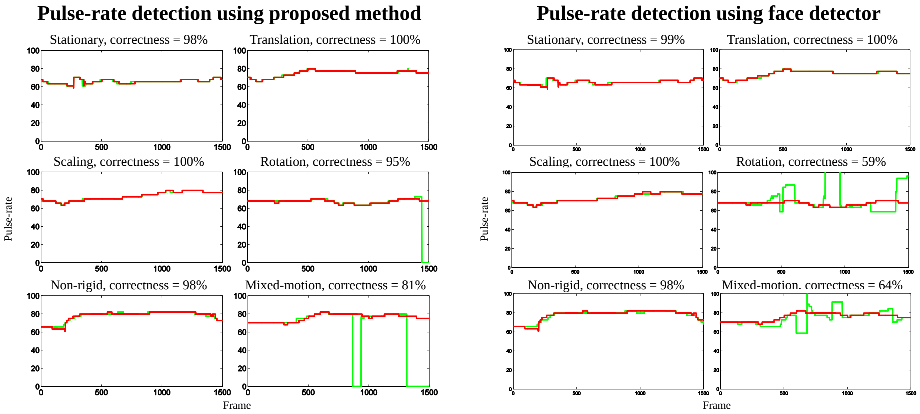


Fig. 7. The consistency of the found RoI is measured in videos of Skin-type II male with stationary, translation and non-rigid motion, where the subject skin captured by camera is assumed to be constant.



(1) (2)

Fig. 8. Example of pulse-rates estimated from Skin-type II male with 6 types of motion. (1) pulse-rates obtained by the proposed method, (2) pulse-rates obtained by the method using Viola-Jones face detector. The x-axis depicts the frame number where 1800 frames correspond to 80 seconds in the video (20 FPS video recording).

To investigate the performance consistency of the proposed method, we analyse the number of skin pixels in the found RoI for the video sequence of Skin-type II male with 3 types of motion: stationary, translation and non-rigid, where the amount of skin pixels captured by camera is assumed to be constant. In Figure 7, the consistency of the found RoI is denoted as a percentage between the RoI size and image size (e.g., 768 × 572 pixels). As expected, the proposed method gains the best consistency in videos with stationary subject. In contrast, the results obtained in videos with translation and non-rigid motions are less consistent, which is due to the fact that some skin pixels or regions appear/disappear during the head movements. In general, the proposed method remains fairly consistent performance (with a maximum standard deviation of $\sigma = 1.0$) in dealing with subject motions like translation and non-rigid motion.

In Figure 8, the pulse-rate obtained by the proposed method is compared to that obtained by the rPPG method with Viola-Jones face detector. The comparison is performed on the videos of Skin-type II male with 6 types of motion. As can be seen that for videos without subject rotation, the pulse-rates obtained by our method remain 98% of the time consistency with the reference pulse-rates (difference smaller than 3 BPM). For the videos including subject rotation, the accuracy of the pulse-rate obtained by our method decreases to 81% as the result of the failure in feature corners tracking. The decrease in pulse-rate accuracy is also significant for the face detector (59%) as the Viola-Jones detector trained with frontal face samples is unable to find the side-view faces. Nevertheless, in situations where the actual pulse-rate is missing, our method outputs a momentary pulse-rate of 0 BPM, whereas the face detector still produces a significant frequency-peak in the spectrum that does not correspond to the pulse-rate. The reason is that the stronger frequency components are introduced into the pulse-signal due to the repeated and abrupt detection failure of the face detector. Figure 9 provides an additional view of the frequency spectrum of the pulse-

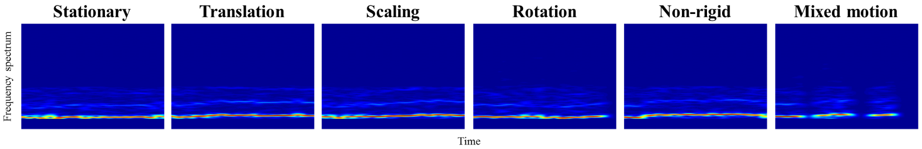
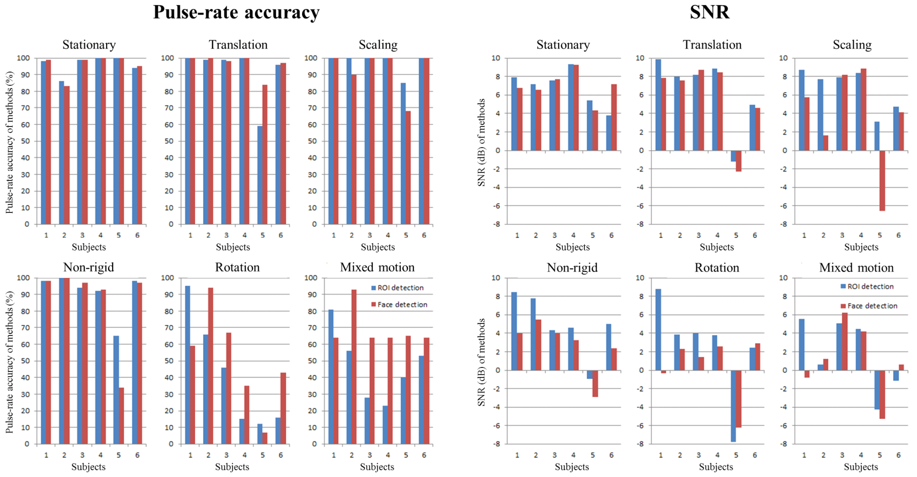


Fig. 9. The frequency spectrum of the pulse-signals obtained from Skin-type II male with 6 types of motion.

signals extracted from 6 types of motion of Skin-type II male, which shows clear pulse-rate frequencies except for the mixed motion.

In Figure 10, the pulse-rate accuracy and SNR obtained by both methods (the proposed method and the method with Viola-Jones face detector) are given for 36 video sequences (6 subjects with 6 types of motion). Figure 10(1) shows that in the videos without subject rotation (e.g., stationary, translation, scaling, and non-rigid motion), our method achieves an average pulse-rate accuracy of 94%, which is comparable to that obtained by the rPPG method with face detector (92%). By comparing the SNR for those video sequences, our method significantly improves the method using face detector from 5.8dB to 8.6dB, which is clearly better. The improvement in SNR is probably due to the reason that our method adopts only the skin pixels that present strong pulse-frequency feature for combination, whereas the method with face detector use all the pixels within RoI (including non-skin pixels like nostril and eyebrows) for averaging. This is further demonstrated in the comparison of videos with stationary and non-rigid motion: both methods achieve high pulse-rate accuracy and SNR in videos with stationary subjects, where no significant motion noise is introduced. However, in the videos with non-rigid motion (talking), the performance of the proposed method is apparently better, because it excludes the noisy regions around the



(1) (2)

Fig. 10. The pulse-rate accuracy and SNR obtained by the compared methods (the proposed method and the method using Viola-Jones face detector) on 36 video sequences (6 subjects performing 6 types of motion). (1) pulse-rate accuracy, (2) SNR quality.

subjects' moving mouth/lips where the pulse-signals are distorted by motion. However, both methods suffer from dramatic performance degradation in videos with vigorous head motions such like rotation and mixed motion. As explained before, the major reason causes the failure of the proposed method is due to the problem of feature corners tracking.

5 Conclusions

In this paper, an automatic RoI detection method for camera-based pulse-rate measurements is presented. The proposed method consists of three main steps, namely subregion tracking, feature extraction and skin region clustering. In the first step, subregions are established using triangulation on the tracked feature corners during a pre-defined interval in the video. In the second step, pulse-signals are extracted from the subregions and the spatiotemporal features (e.g., pulse-signal and skin chromaticity) are formulated as discriminative representation. In the third step, subregions containing similar features are classified into the same clusters using a density-based clustering method.

The proposed method is experimentally verified on 36 video sequences consisting of 6 subjects with 3 skin-types and 6 motion-categories. For the videos of subjects with brighter skin and modest motions (without rotation), the proposed method obtains the comparable pulse-rate accuracy but better SNR as compared to the rPPG method using the face detector, which demonstrates the effectiveness of our strategy of only exploiting the skin pixels containing pulse-signals for pulse-rate measurements. However for videos of subjects with darker skin-types or vigorous motions, the pulse-rate accuracy and SNR obtained by both methods decline. In general, compared to the rPPG method using face detection, the novel automatic RoI detection method developed in this study, which takes the advantage of pulse-signals, demonstrates a favourable performance on subjects with brighter skin-tone, yet it is robust to modest body motions.

References

1. Hertzman, A.B.: Photoelectric plethysmography of the fingers and toes in man. Experimental Biology and Medicine **37** (1937) 529–534
2. Takano, C., Ohta, Y.: Heart rate measurement based on a time-lapse image. Medical engineering & physics **29** (2007) 853–7
3. Huelsbusch, M., Blazek, V.: Contactless mapping of rhythmical phenomena in tissue perfusion using ppgi. Volume 4683. (2002) 110–117
4. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18** (2010) 10762–74
5. Schmitz, G.: Video camera based photoplethysmography using ambient light. M.S. thesis dissertation, Electrical Engineering. Technische Universiteit Eindhoven, Netherlands (2011)
6. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. Biomedical Engineering, IEEE Transactions on **60** (2013) 2878–2886

7. Lewandowska, M., Ruminski, J., Kocejko, T., Nowak, J.: Measuring pulse rate with a webcam - a non-contact method for evaluating cardiac activity. In: Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on. (2011) 405–410
8. Xu, Z., Zhu, M.: Color-based skin detection: survey and evaluation. In: Multi-Media Modelling Conference Proceedings, 2006 12th International. (2006) 10 pp.–
9. Harris, C., Stephens, M.: A combined corner and edge detector. In: In Proc. of Fourth Alvey Vision Conference. (1988) 147–151
10. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI'81, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1981) 674–679
11. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: Pattern Recognition (ICPR), 2010 20th International Conference on. (2010) 2756–2759
12. Delaunay, B.: Sur la sphère vide. a la mémoire de georges voronoï. Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na **6** (1934) 793–800
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. (1996) 226–231
14. Fitzpatrick, T.: The validity and practicality of sun-reactive skin types i through vi. Archives of Dermatology **124** (1988) 869–871
15. Viola, P., Jones, M.: 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. (2001) I–511–I–518 vol.1