Skin Segmentation using Active Contours and Gaussian Mixture Models for Heart Rate Detection in Videos

Alexander Woyczyk Vincent Fleischhauer
Sebastian Zaunseder
University of Applied Sciences and Arts Dortmund
Dortmund, Germany
alexander.woyczyk@fh-dortmund.de

Abstract

Current research focuses on non-contact means to capture physiological signals like the heart rate. One promising approach uses videos (imaging PPG, iPPG). The common procedure to derive the heart rate by iPPG comprises three steps: segmentation of a region of interest, usage of colour information from that region to yield a pulse signal and analysis of that signal to estimate the heart rate. This contribution proposes a novel approach to yield a region of interest using a Gaussian mixture model based level set formulation. The proposed method aims to segment a homogeneous region on an individual basis. To that end, we model the probability distributions for the pixel skin and non-skin class by two separate Gaussian mixture models. The proportion of the posterior probabilities are then included in the formulation of the level set function. The procedure yields a region of interest, which is used to derive a pulse signal from its average intensity or additional processing steps. We tested the method on own data and data of the 1st Challenge on Remote Physiological Signal Sensing. It is shown that the proposed method can improve the results for heart rate estimation on moving subjects. The potential of our approach is underlined by the promising result in the challenge.

1. Introduction

The heart rate is one of the most important vital parameters and an essential element of modern medicine. imaging photoplethysmography (iPPG) uses videos to yield the heart rate without any contact. The technique captures the varying light absorption due to subtle blood volume changes in skin tissue. iPPG thus allows to obtain a pulse signal from which the heart rate can be derived [24, 28]. The path from video to heart rate usually consists of a three step procedure: (1) segmentation of a region of interest (ROI), i.e. an area defined for further processing, (2) signal extraction, i.e. colour information from that region used to yield a pulse signal and (3) heart rate estimation, i.e. the pulse signal analyzed to estimate the heart rate.

Many studies focus on the signal extraction. An early solution was introduced by Verkruysse *et al.*. The method uses the green channel and averages its intensity of all pixels within the ROI. The green channel offers a beneficial trade-off between absorption by haemoglobin and penetration depth and thus typically shows the highest signal-to-noise-ratio considering a single colour channel [24, 29]. Other more complex solutions combine colour channels to yield a better signal-to-noise ratio. Generally one can distinguish between data-driven channel combinations and channel combinations based on a priori knowledge [28]. Data-driven methods make use of source separation techniques like independent component analysis [17]. Though widely implemented, the results using such methods are equivocal and their practical application is further complicated by additional constraints like permutation indeterminacy [26]. Methods based on a priori knowledge rely on the observation, that the effect of blood pulsation and artifacts are differently pronounced in channels of different colour or projection spaces. Researchers have proposed multiple algorithms to exploit such differences by combining colour channels. CHROM [6] and POS [25] are amongst the most popular methods. Both methods combine the red, green and blue colour channel to yield a signal, which reflects the blood pulsation very well while suppressing artifacts and specular reflections.

Heart rates be estimated by detecting single heart beats in the pulse signal. However, due to the typically low signal quality they are identified most often in the frequency domain, using fast Fourier transform (FFT) to transform the filtered signal and examine its frequency components, as in [1, 23].

To segment the ROI, most works rely on either static

bounding boxes, e.g. Verkruysse and Lewandowska [24, 13] or use tracking of feature points to compensate for subject motion [20, 8]. Respectively statistical skin detection [18], neuronal-network based skin detection [12] or static colour thresholds [3] help to reduce the number of pixels not contributing to the signal. Another approach defines the ROI using facial landmarks to increase the number of skin pixels, e.g. [11]. Trumpp *et al.* used Bayesian skin detection in combination with a level set approach to segment skin from background [23]. This approach not only considers skin regions but aims for homogeneous areas because they are assumed to contribute to an improved signal-to-noise-ratio.

Similar to Trumpp, this contribution proposes a level set approach as well, but uses different constraints and substantially differs by relieving the dependency on pre-trained skin classification and expressing fore- and background models as multivariate distributions, allowing a more precise segmentation of skin in cases of non-uniform fore- and background. It uses Gaussian mixture model (GMM) to train an individual model for the pixels' classes (skin and non-skin) for each video and embeds this model into a level set function. Since the algorithm is initialized in the face region at the start of the video, the derived ROI contains skin areas of the face but is not limited to them. Due to its independence from facial landmarks or other shape constraints, it is further adaptable to movement, rotation or partial occlusion of the subject.

In the next section, we will first provide the foundations of the algorithm, namely details on GMM and level sets. Based on that, we will present the implementation of the algorithm. Afterwards, testing data sets and evaluation methods are presented. At last, our results are shown and discussed.

2. Methods

2.1. Gaussian Mixture Models

GMM is a common way to model a unknown data distribution. A GMM models the data by multivariate normal distributions, which are called components of the mixture model. Thereby a mixture of K Gaussians is described by a set of normal density functions

$$\sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2) \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and standard deviation σ. Expanding this formulation to n-dimensional orders, σ_k^2 is replaced with the n×n covariance matrix Σ_k. Each component is assigned a weight π_k, which add up to one. The weights themselves are the prior probability for each component. The Gaussian distribution represents the likelihood function for a given component. Using Bayes' theorem the posterior probability of an observation x belonging to component k is given by

$$p(z_k = 1|x) = \pi_k \cdot \frac{\mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x|\mu_j, \Sigma_j)} \tag{2}$$

Here $z_k = 1$ denotes that x belongs to component k. Since the probability distribution is not known, it has to be estimated from the observed data. Given N observations $x_{1,\dots,N}$, this estimation is performed by the expectation-maximization algorithm (EM-Algorithm).
For now let y_{ik} be shorthand for $p(z_k = 1|x_i)$ from equation 2. Starting with an initial estimation for the mixture components given by the K-means algorithm, the expectation step consists of evaluating equation 2 for all observations and components [21].
The maximization step then updates each components configuration according to the new estimates. The priori probability is replaced by the average of new posteriors for this component according to

$$\pi_k{}' = \frac{1}{N} \sum_{i=1}^{N} y_{ik} \tag{3}$$

Means and variances are updated accordingly by calculating

$$\mu_k{}' = \frac{\sum_{i=1}^{N} y_{ik} x_i}{\sum_{i=1}^{N} y_{ik}} \tag{4}$$

and

$$\Sigma_k{}' = \frac{\sum_{i=1}^{N} y_{ik} (x_i - \mu_i{}')(x_i - \mu_i{}')^T}{\sum_{i=1}^{N} y_{ik}} \tag{5}$$

The process of alternating expectation and maximization steps is repeated until the log-likelihood of the model, given by equation 6, converges to a local maximum.

$$\ln p(X) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \tag{6}$$

2.2. Active Contours with Level Set

Active contours describe the method of evolving a curve in order to segment an object from its surroundings within an image. From a starting initialization, the curve moves inwards or outwards, guided by a set of restraints, until it reaches some kind of object boundary [22].

Level set functions use an implicit representation of the object boundary, *i.e.* they define a continuous function over the image space, so that every pixel can be labelled as back- or foreground. The level set function is propagated towards the object boundary over a series of iteration steps. The level set function $\Phi(\mathbf{x}, t_i)$ is therefore defined for every

pixel $\mathbf{x}$ at iteration step t_i. The inside or foreground is then defined by all pixels where $\Phi(\mathbf{x}) > k$ for a given threshold k [22].

One fundamental approach to a level set based foreground-background segmentation was formulated as energy minimization term by Chan and Vese [5]. For a contour C the energy F is described as

$$\begin{aligned} F(C, c_1, c_2) = \mu \cdot length(C) + \nu \cdot area(insideC) \\ + \lambda_1 \cdot \int_{insideC} |u_0 - c_1| dxdy \\ + \lambda_2 \cdot \int_{outsideC} |u_0 - c_2| dxdy \end{aligned} \tag{7}$$

where u_0 is the intensity of the current pixel, c_1 and c_2 are the average intensities for for- and background and μ, ν, λ_1 and λ_2 are weighting parameters. A larger μ penalises the curvature of the contour, positive ν favours shrinking of the contour over growing.

After discretization and linearization, Chan and Vese arrive at a calculation scheme of the ϕ function for location (i, j) [4]

$$\begin{aligned} &\frac{\phi_{i,j}^{n+1} - \phi_{i,j}^{n}}{\Delta t} = \\ &\delta_h(\phi_{i,j}^n) \cdot \left[\frac{\mu}{h^2} \Delta_-^x \left(\frac{\Delta_+^x \phi_{i,j}^{n+1}}{\sqrt{\frac{(\Delta_+^x \phi_{i,j}^n)^2}{h^2} + \frac{(\phi_{i,j+1}^n - \phi_{i,j-1}^n)^2}{(2h)^2}}} \right) \right. \\ &+ \frac{\mu}{h^2} \Delta_-^y \left(\frac{\Delta_+^y \phi_{i,j}^{n+1}}{\sqrt{\frac{(\phi_{i+1,j}^n - \phi_{i-1,j}^n)^2}{(2h)^2} + \frac{(\Delta_+^y \phi_{i,j}^n)^2}{h^2}}} \right) \\ &\left. - \nu - \lambda_1 (u_{i,j} - c_1(\phi^n))^2 + \lambda_2 (u_{i,j} - c_2(\phi^n))^2 \right] \end{aligned} \tag{8}$$

Here the gradient of the function ϕ is approximated by the finite differences

$$\begin{aligned} \Delta_-^x \phi_{i,y} = \phi_{i,j} - \phi_{i-1,j}, \quad \Delta_+^x \phi_{i,j} = \phi_{i+1,j} - \phi_{i,j}, \\ \Delta_-^y \phi_{i,y} = \phi_{i,j} - \phi_{i,j-1}, \quad \Delta_+^y \phi_{i,j} = \phi_{i,j+1} - \phi_{i,j} \end{aligned} \tag{9}$$

Furthermore Δt is an artificial time step between two consecutive iteration steps, h is the artificial distance between two pixels and δ_h is the following approximation of the dirac function:

$$\delta_h(\phi_{i,j}^n) = \frac{1}{\pi \cdot (1 + (\phi_{i,j}^n)^2)} \tag{10}$$

3. Skin Segmentation with GMM and Level Sets

3.1. Overview

The aim of this contribution is to extend the level-set formulation for skin segmentation and to define a suitable ROI for iPPG, i.e. define a foreground that serves as ROI. One specific limitation of the method described by Chan consists in the handling of background as homogeneous region. Background is not bound to be homogeneous, but can feature different coloured objects, e.g. doors, paintings or furniture. Additionally, the subject can wear different coloured clothing. Since these areas belong to the background class in the special case of skin segmentation, the approach of segmenting two homogeneous regions proposed by the original level set segmentation is not optimal. In order to perform the skin segmentation necessary for further signal extraction we combine the colour based GMM with a level set approach to account for spatial dependencies. The proposed procedure requires an individual initialization, i.e. training a skin and a non-skin model, and applies a modified level set formulation to segment the ROI using the model based probabilities.

3.2. Global Parameters of Skin and Non-Skin Models

To model different coloured objects for skin and non-skin regions, the number of components for the skin class and non-skin class have to be set beforehand. After visual observation of the colour distribution histograms of sampling videos, the number of GMM components was set to four for the non-skin class and to two for the skin class. The observation confirmed the assumption that at least three different colour components are present in the non-skin region: Hair, clothing and background where the background itself may contain other objects as well.

As displayed in figure 1 the face region may consists of several dominant colours. Experiments varying the number of components for the skin class have shown that using only one component tends to dismiss parts of the skin region, where shadows lead to darker areas. On the other hand, three or more components increased the risk of adding facial hair and eyes to the skin area. Accordingly the skin model was set to two components, taking into account skin inhomogeneity and discriminating against hair or other occlusions.

Figure 2 shows the colour cloud of a whole frame and the corresponding skin and non-skin GMMs. Note that one component of the skin class is mainly shadowed by a non-skin component, indicating that the colour is also present in the background, e.g. hair.

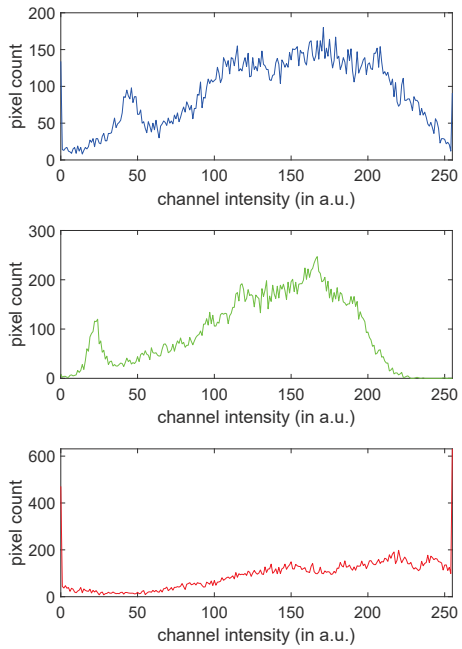
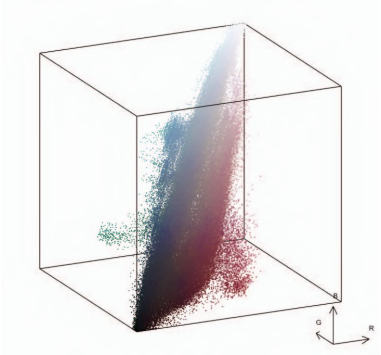
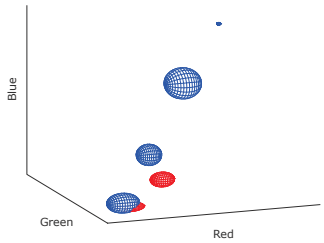


Figure 1: Histogram of pixels in the face region of a single subject, projected on blue (top), green (middle) and red (bottom) colour channel.



(a) Colour cloud



(b) Representation of skin (red) and non-skin (blue) GMMs. Each blob is centred at its mean with radii of its diagonal elements of Σ

Figure 2: Colour cloud and calculated GMMs using EM algorithm for a single subject.

3.3. Initialization

The individual GMMs for skin and non-skin are both generated from the first frame of the sample video. This initialization step has to be executed once for each video file. At first a face detection algorithm determines the position of the subjects face. The presented algorithm identifies the objects position with a cascade classifier using Haar features. Here the pre-trained Haar cascade, supplied by the OpenCV framework [21], is implemented. Pixels within the resulting bounding box are marked for training the skin GMM, while pixels on the outside of the box contribute to the estimation of the non-skin GMM.

3.4. Segmentation

After calculating the models for skin and non-skin, a level set function is used to derive a implicit segmentation. Therefore the data term $(u_{i,j} - c(\phi^n))^2$ in equation 8 of the Chan & Vese model, which penalizes the euclidean distance between pixel value and average fore- or background intensity in RGB space, is replaced by a proportion of posterior probabilities originating form the GMM models. Especially given a RGB triple u_0, the penalizing terms for the skin and non-skin class are given by

$$dist_{skin}(u_0) = \frac{p_{skin}(u_0)}{p_{skin}(u_0) + p_{nonSkin}(u_0)} \tag{11}$$

and

$$dist_{nonSkin}(u_0) = \frac{p_{nonSkin}(u_0)}{p_{skin}(u_0) + p_{nonSkin}(u_0)} \tag{12}$$

Posterior probabilities p_{skin} and $p_{nonSkin}$ are the cumulative posteriors of the respective model, e.g. p_{skin} is calculated by

$$p_{skin}(u_0) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(u_0 | \mu_k, \Sigma_k) \tag{13}$$

In order to limit the effect on locations where both skin and non-skin models yield a low posterior probability but a high proportion, additional weighting factors are introduced by

$$w_{skin}(u_0) = -\frac{1}{log(p_{skin}(u_0))} \tag{14}$$

and

$$w_{nonSkin}(u_0) = -\frac{1}{log(p_{nonSkin}(u_0))} \tag{15}$$

Substituting the data term in equation 8, the following

iterative solution to calculate ϕ is employed

$$\begin{aligned}\phi_{i,j}^{n+1} = \Big[\phi_{i,j}^{n} + \delta_h \cdot (\mu \cdot \Big(\phi_{i+1,j}^{n} \cdot divR \\ + \phi_{i-1,j}^{n} \cdot divL + \phi_{i,j+1}^{n} \cdot divU + \phi_{i,j-1}^{n} \cdot divD\Big) \\ - \nu - \lambda_1 \cdot w_{skin}(u_{i,j}) \cdot dist_{skin}(u_{i,j}) \\ + \lambda_2 \cdot w_{nonSkin}(u_{i,j}) \cdot dist_{nonSkin}(u_{i,j}))\Big] \\ \cdot \frac{1}{1 + \delta_h \cdot \mu \cdot (divR + divL + divU + divD)}\end{aligned} \tag{16}$$

The inverse gradient in equation 16 of the ϕ function is defined by following functions

$$\begin{aligned} divR &= \frac{1}{\sqrt{(\Delta_{+}^{x}\phi_{i,j}^{n})^2 + (\frac{\phi_{i,j+1}^{n} - \phi_{i,j-1}^{n}}{2})^2}} \\ divL &= \frac{1}{\sqrt{(\Delta_{-}^{x}\phi_{i,j}^{n})^2 + (\frac{\phi_{i,j+1}^{n} - \phi_{i,j-1}^{n}}{2})^2}} \\ divU &= \frac{1}{\sqrt{(\Delta_{+}^{y}\phi_{i,j}^{n})^2 + (\frac{\phi_{i+1,j}^{n} - \phi_{i-1,j}^{n}}{2})^2}} \\ divD &= \frac{1}{\sqrt{(\Delta_{-}^{y}\phi_{i,j}^{n})^2 + (\frac{\phi_{i+1,j}^{n} - \phi_{i-1,j}^{n}}{2})^2}} \end{aligned} \tag{17}$$

The calculation is repeated until the solution converges to a local maximum, i.e. the difference between the last function and the current one is lower than a previously set threshold, or a maximum number of iterations is reached.

4. Data and Evaluation Strategy

4.1. Video Data

To evaluate the proposed algorithm a set of validation videos was recorded. Seven subjects were recorded for a duration of 30 seconds. During this time the subjects were instructed to move their head. This data set will further be denoted as **movement data set**. The videos were captured by a RGB camera (IDS UI-3370CP-C-HQ) positioned at a distance of 30 to 50 cm to the subjects head. The resulting field of view covered the head and parts of the upper body of the subject, as in figure 3. Recordings took place in an office building under uncontrolled daylight and ceiling lighting. Daylight intensity and subjects' position relative to ceiling differed, resulting in uncontrolled illumination. The recording parameters were set to a video size of 320×420 pixels, recorded at 100 fps. The recorded 12 bit videos were transformed to colour depth of 8 bit per channel for image segmentation and signal extraction. For validation purposes, a reference photophlethysmogram (PPG) was recorded during the video recording using a finger clip (ADInstruments MLT1020FC).

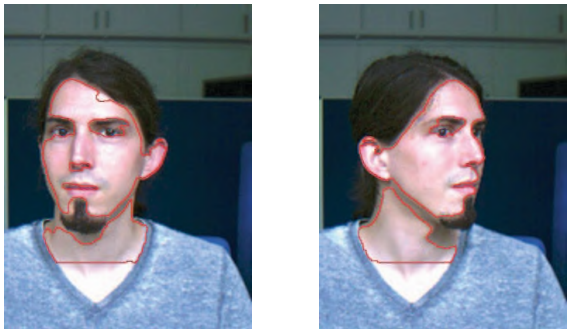


Figure 3: Frames and segmented ROI before (left) and after (right) movement.

Additionally, the proposed algorithm was tested on the video data supplied for the first Challenge on Remote Physiological Signal Sensing (RePSS) [14]. The data consists of videos from the VIPL-HR V2 database [16] and Oulu BioFace (OBF) Database [14, 27]. The supplied data set contains 500 subjects with five 10 second video sequences each. The provided samples are recorded in various locations, under different illumination conditions, with varying resolutions and video qualities. This data will further be denoted as **RePSS data set**.

4.2. Signal Processing

The aforementioned procedure yields a ROI that can be used as base for the subsequent processing steps, namely signal extraction and heart rate estimation.

For signal extraction, as discussed in the introduction, multiple strategies have been described. In this contribution we choose the green channel as it doesn't require further calculation steps. We add CHROM [6] and POS [25] as more complex solutions that are widely acknowledged and have been shown to be powerful.

For heart rate estimation, we consider sequences of 10 seconds duration. The heart rate is estimated for each segment independently by the following procedure: First, signal segments are band pass filtered in order to remove trends and high frequency components originating from image noise or artefacts. The cut-off frequencies are set to 0.5 Hz for the high pass filter and 4 Hz for the low pass filter. Secondly, we transform the filtered signal to frequency domain using the FFT. Lastly, the frequency component having the highest amplitude within the range of 50 to 180 bpm is presumed to originate from the heart rate. This procedure is equally applied to all methods of signal extraction, i.e. the green channel, CHROM or POS. Figure 5 demonstrates the procedure.

4.3. Reference Methods

In order to allow a meaningful assessment of the proposed approach to define a suitable ROI, we compare it against state of the art methods. The compared methods differ in the definition of the ROI and its tracking, respectively, as follows: **VJ_static**: The ROI is obtained by using the Viola & Jones face detection implemented in OpenCV on the first frame of each video [21]. The resulting bounding box is kept static over the whole video.
VJ_dynamic: We use the same initial ROI as VJ_static, but instead of keeping it at a static position it is moved according to an average displacement vector. The average displacement is obtained by averaging over all displacement vectors calculated with the Kanade-Lucas-Tomasi tracker on a set of feature points within the initial ROI [9, 2]. The displacement is calculated between every frame of the sequence.
VJ_skin: This method equals VJ_dynamic but in addition a skin classifier is applied to the dynamic ROI. The method additionally tries to exclude non-skin pixel from entering the ROI. Skin is classified by a Bayesian classifier, as described by Jones *et al.* [10].

Subsequent steps after ROI segmentation, i.e. signal extraction and heart rate detection, are identical for all ROI handlings and include using the green channel, CHROM and POS.

5. Results and Discussion

5.1. Evaluation Metrics

In order to compare different approaches and methods, the accuracy and mean absolute error (MAE) are evaluated.

The MAE is calculated as the average of absolute differences between the reference and estimated heart rate over all video sequences of a data set.

To quantify the accuracy of an algorithm, the estimated heart rate is compared to the reference. If the estimated heart rate lies within a predefined range around the reference heart rate, it is assumed as true positive. This range is set to $\mathrm{HR}_{ref} \pm 5$ bpm as in Rasche *et al.* [19]. The accuracy is defined as the number of true positives divided by the overall number of estimation.

5.2. Results

The results on the movement data set, as given in table 1, show that our method outperforms standard procedures for ROI segmentation. Using the green channel on a static ROI performs slightly better than a moving ROI. Moving of the ROI probably introduces additional artefacts, which pollute the signal. In turn, the proposed algorithm produces a more smooth ROI, which leads to less disturbances in the extracted signal as shown in figure 5a and 5b.

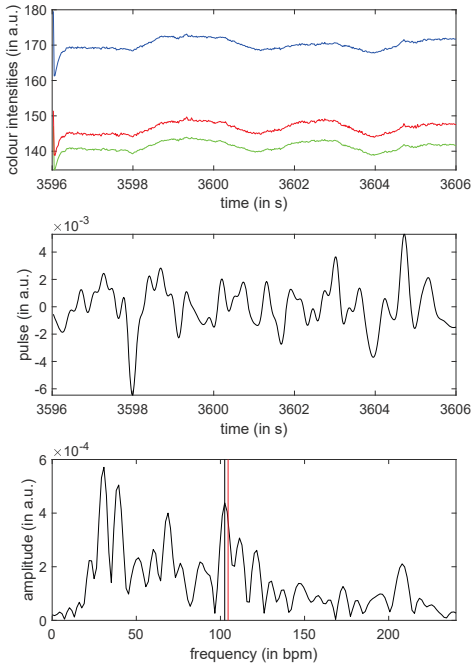
Figure 4: Segmented ROI in pixelated frames from the RePSS challenge data [15].

Heart rate estimations on the RePSS data set are not as accurate as on the movement data set. This was expected, since the videos feature more difficult lighting conditions, stronger image noise, a lower frame rate and were compressed using a lossy video codec. Another source of error was the implemented face detection, which failed on some of the sample videos. Since a precise location of the face is important to train the GMMs, future work will therefore include a more robust face detection. Though low heart rate detection rates limit the significance of the results on the RePSS data set, an improvement using the proposed algorithm can be observed.

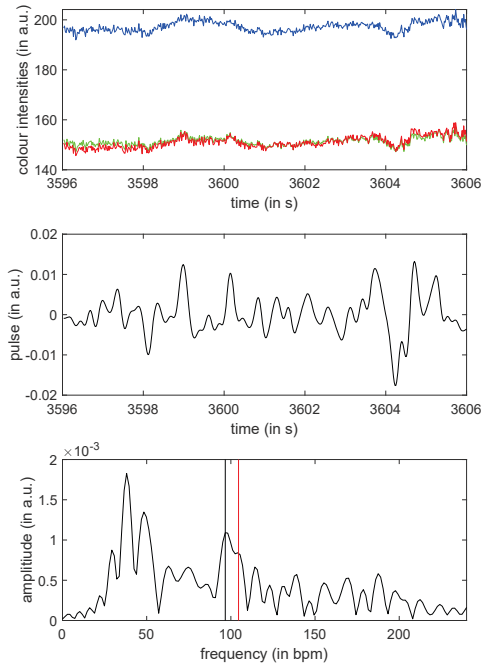
5.3. Discussion

The presented approach was shown to improve the heart rate estimation compared to state of the art methods. The obtained results still do not suffice the needs, particularly in the RePSS data set. This finding can be explained, in parts, by the challenging nature of the used data sets. Accuracies of 90 percent and above, as reported when using other data, will be difficult to obtain. However, the presented approach leaves space for improvements. As stated before, the used face detection limits the performance of the method and will be replaced in the future. In addition, we are going to pursue to include additional constraints in order to improve the level set formulation.

The proposed algorithm to segment a ROI also participated in the First Challenge on Remote Physiological Signal Sensing [15]. It featured 1000 10-second video sequences for which the heart rate had to be estimated. Rankings were based on the MAE metric. The heart rates submitted to the challenge by us were obtained by using the presented method to segment a ROI and further processed by the CHROM pulse extraction. Though all videos were pixelated around the mouth and eye area and challenging lighting conditions, as shown in figure 4, the method yielded a promising result underlining the potential of the proposed approach.



(a) Extracted signal using the proposed algorithm. Top: raw signal, middle: filtered signal, bottom: frequency spectrum with estimated HR (black) and reference HR (red).



(b) Extracted signal using the tracked ROI with skin detection. Top: raw signal, middle: filtered signal, bottom: frequency spectrum with estimated HR (black) and reference HR (red).

Figure 5: Signal processing steps for subject 6.

	Green channel		CHROM		POS	
ROI ID	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE
VJ_static	0.2381	24.87	0.6667	12.77	0.7143	11.06
VJ_dynamic	0.1905	24.61	0.5714	15.28	0.6667	12.49
VJ_skin	0.3333	18.82	0.6190	10.18	0.6190	12.05
LS_GMM	**0.4762**	**18.19**	**0.8571**	**5.91**	**0.8571**	**4.13**

Table 1: Accuracy and MAE for different combinations of ROIs and pulse extraction methods on the movement data set.

6. Conclusion

We have shown that the selection of skin region has a high impact on the performance of heart rate estimation algorithms. Furthermore our experiments suggest, that segmenting a continuous and homogeneous region is more valuable to pulse extraction than pixelwise skin classification. Our calculated ROI can further be used to increase the performance of other pulse extraction methods as CHROM or POS significantly, as it constructs a ROI which inhabits less noise and motion artefacts than static skin detection or tracking of feature points in order to transform the ROI. Every method profits from better video quality and stable illumination. Our experiments have confirmed, that combinations of the colour channels can lead to better heart rate detection. Still the raw signals from the green channel can be used for other applications, e.g. analysis of the pulse waves morphology as conducted by Fleischhauer *et al.* [7]. Our work demonstrated that heart rate estimation is possible even under severe motion of the subject, but for other applications using the pulse wave, movement still poses a difficult challenge, which we will continue to address.

7. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 401786308.

	Green channel		CHROM		POS	
ROI ID	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE
VJ_static	0.1709	23.53	0.3315	16.89	0.3331	16.32
VJ_dynamic	0.1693	23.36	0.3283	16.39	**0.3455**	**15.95**
VJ_skin	0.1385	26.28	0.3411	16.77	0.3279	17.21
LS_GMM	**0.1845**	**22.92**	**0.3579**	**13.61**	0.3283	16.62

Table 2: Accuracy and MAE for different combinations of ROIs and pulse extraction methods on the RePSS data set.

References

[1] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, jun 2019.

[2] Jean-Yves Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker. Technical report, Intel Corporation, 2001.

[3] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. *Biomedical Signal Processing and Control*, 8(6):568–574, 2013.

[4] Tony Chan and Luminita Vese. An Active Contour Model without Edges. In *Lecture Notes in Computer Science*, volume 1682, pages 141–151. 1999.

[5] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.

[6] Gerard De Haan and Vincent Jeanne. Robust pulse-rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[7] Vincent Fleischhauer, Alexander Woyczyk, Stefan Rasche, and Sebastian Zaunseder. Impact of Sympathetic Activation in Imaging Photoplethysmography. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, volume c, pages 1697–1705. IEEE, oct 2019.

[8] Luca Iozzia, Luca Cerina, and Luca Mainardi. Relationships between heart-rate variability and pulse-rate variability obtained from video-PPG signal using ZCA. *Physiological Measurement*, 37(11):1934–1944, 2016.

[9] Jianbo Shi and Tomasi. Good features to track. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94*, pages 593–600. IEEE Comput. Soc. Press, 1994.

[10] Michael J Jones and James M Rehg. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46(January):81–96, 2016.

[11] Mayank Kumar, Ashok Veeraraghavan, and Ashutosh Sabharval. DistancePPG: Robust non-contact vital signs monitoring using a camera. 6(5):200–215, 2015.

[12] Kual-Zheng Lee, Pang-Chan Hung, and Luo-Wei Tsai. Contact-Free Heart Rate Measurement Using a Camera. In *2012 Ninth Conference on Computer and Robot Vision*, pages 147–152, 2012.

[13] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jdrzej Nowak. Measuring pulse rate with a webcam - A non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems, FedCSIS 2011*, pages 405–410, 2011.

[14] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. The OBF Database: A Large Face Video Database for Remote Physiological Signal Measurement and Atrial Fibrillation Detection. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 242–249. IEEE, may 2018.

[15] Xiaobai Li, Hu Han, Hao Lu, Xuesong Niu, Zitong Yu, Antitza Dantcheva, Guoying Zhao, and Shiguang Shan. The 1st Challenge on Remote Physiological Signal Sensing (RePSS). *arXiv*, 2020.

[16] Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. *IEEE Transactions on Image Processing*, 29:2409–2423, 2019.

[17] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762, 2010.

[18] Michal Rapczynski, Philipp Werner, and Ayoub Al-Hamadi. Continuous low latency heart rate estimation from painful faces in real time. *Proceedings - International Conference on Pattern Recognition*, (C):1165–1170, 2017.

[19] S. Rasche, A. Trumpp, T. Waldow, F. Gaetjen, K. Plötze, D. Wedekind, M. Schmidt, H. Malberg, K. Matschke, and S. Zaunseder. Camera-based photoplethysmography in critical care patients. *Clinical Hemorheology and Microcirculation*, 64(1):77–90, nov 2016.

[20] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiological Measurement*, 35(5):807–831, 2014.

[21] OpenCV team. OpenCV 4.1.1 Documentation, 2019.

[22] Klaus D. Toennies. *Guide to Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition. Springer London, London, 2017.

[23] Alexander Trumpp, Stefan Rasche, Daniel Wedekind, Martin Schmidt, Hagen Malberg, Thomas Waldow, Frederik Gaetjen, Katrin Pl, Klaus Matschke, and Sebastian Zaunseder. Skin Detection and Tracking for Camera-Based Photo-

plethysmography Using a Bayesian Classifier and Level Set Segmentation. pages 43–48, 2017.

[24] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008.

[25] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, jul 2017.

[26] Daniel Wedekind, Alexander Trumpp, Frederik Gaetjen, Stefan Rasche, Klaus Matschke, Hagen Malberg, and Sebastian Zaunseder. Assessment of blind source separation techniques for video-based cardiac pulse extraction. *Journal of Biomedical Optics*, 22(3):035002, 2017.

[27] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 1, pages 151–160. IEEE, oct 2019.

[28] Sebastian Zaunseder, Alexander Trumpp, Daniel Wedekind, and Hagen Malberg. Cardiovascular assessment by imaging photoplethysmography-a review. *Biomedizinische Technik*, 63(5):529–535, 2018.

[29] George Zonios, Julie Bykowski, and Nikiforos Kollias. Skin Melanin, Hemoglobin, and Light Scattering Properties can be Quantitatively Assessed In Vivo Using Diffuse Reflectance Spectroscopy. *Journal of Investigative Dermatology*, 117(6):1452–1457, dec 2001.