Automatic region-based heart rate measurement using remote photoplethysmography

Benjamin Kossack[1] Eric Wisotzky[1,2] Anna Hilsmann[1] Peter Eisert[1,2]
[1]Fraunhofer Heinrich Hertz Institute [2]Humboldt-Universität zu Berlin
Berlin, Germany
benjamin.kossack@hhi.fraunhofer.de

Abstract

This paper presents a model-based approach to measure the vital signs from RGB video files focusing on the heart rate. We use the plane-orthogonal-to-skin (POS) remote photoplethysmography (rPPG) transformation performed individually at five well-defined regions of interest (ROI) in the face. We extract the heart rate information by a correlation of the different rPPG signals in these ROIs and a magnitude-based reliability calculation. This increases the robustness of the heart rate extraction from videos. With this method, we achieve a mean of all calculated mean-absolute-errors of $8.324\,\mathrm{BPM}$ *in the V4V-Challenge data (averaged over all videos of the training and validation set).*

1. Introduction

This paper is part of the submission to the V4V – Vision for Vitals-Challenge [13]. This challenge and the associated workshop are in conjunction with ICCV 2021 (Montreal, Canada).

Telehealth or telemedicine as a remote clinical service for diagnosis and medical monitoring is a fast-growing public health section. Especially, the COVID-19 pandemic moved it into focus. For all emerging issues in a telehealth session, the face of the patient is an essential source of information about well-being as this is most visible for the health professional. Therefore, it would be beneficial to determine the patient's condition directly from the face.

The patient's condition can be classified by measuring important medical signs. The four primary vital signs are body temperature, blood pressure, heart rate (HR), and respiratory rate (RR). These signs can be measured using a patient monitoring system. For example, an optical measuring technique called photoplethysmography (PPG) is commonly used to measure the human pulse rate [20]. The principle of PPG is based on human blood circulation and the fact that blood absorbs more light than surrounding tissue. Thus, variations in blood volume affect light transmission or reflectance accordingly [14]. A PPG sensor placed directly on the skin optically detects these changes in blood volume [14].

However, telehealth does not allow measuring the vital signs using patient contact. Hence, remote photoplethysmography (rPPG) allowing contactless measurements of the pulse rate with a regular camera has been developed [20] as the blood flow through the circulatory system of a human leads to a continuous change in skin color. Using rPPG techniques, this color variation is detectable in a video signal, and the heart rate (HR) can be determined. According to medical definitions, the rate extracted from an rPPG signal is the pulse rate [1]. However, the term heart rate will be used in this paper, since healthy people usually have an identical pulse and heart rates, HR is often stated as the same as pulse rate [8, 12, 15], and the V4V-Challenge uses that term.

The extraction of vital signs from video recordings of the face is an emerging topic that has increased in recent years. The majority of rPPG-related literature extract vital signs globally [20]. The HR can be extracted by analyzing subtle color changes in the skin area [11, 4, 16, 15]. Poh et al. [11] presented an automated and motion tolerant heart rate measurement technique from video images based on blind source separation using independent component analysis (ICA). To robustly extract an rPPG signal regardless of the subject's skin tone and illumination color (non-white illumination), a chrominance-based calculation of the rPPG signal has been developed [4]. Another knowledge-based color channel combination method is proposed by Wang et al. [16], entitled as Plane-Orthogonal-to-Skin (POS), projecting a three-channel (R-G-B) image onto a plane orthogonal to the [1, 1, 1] direction to create a two-channel image. These two channels are then fused to the desired rPPG signal.

As global model-based methods can be affected, e.g., by noise, compression artifacts, or masking, several of the lat-

est rPPG related publications focus on using convolutional neural networks (CNN) to extract the heart rate from video [3, 19, 18]. In [17], these three deep learning-based methods (Deepphys [3], rPPGNet [19], and Physnet [18]) are tested using the publicly available UBFC-rPPG dataset [2]. All three networks outperform model-based approaches as ICA [11], CHROM [4], and POS [16]. However, [17] shows that under different lighting conditions, these model-based approaches (ICA, CHROM and POS) show better results than the CNNs. Thus, the authors concluded that traditional methods are more robust to light variations.

Therefore, we present a model-based approach with a local extraction of these characteristics, allowing a more robust and differentiated recognition of the vital signs.

Recently, rPPG signals have been analyzed locally to visualize the blood flow through a subject's face [7], e.g., for security-related applications to detect presentation attacks with facial masks [6]. In this work, for HR calculation, different regions of interest (ROIs) are chosen (see Figure 1), followed by a reliability determination. The reliability for each ROI is calculated using the correlation coefficient and the maximum magnitude of the heart frequency component. Based on this, the most reliable ROI defines the heart rate.

Figure 1. This figure shows an example of the region-based face segmentation. The face is segmented into five ROIs (full face, forehead, right cheek, left cheek, and nose).

This paper is organized as follows. The proposed method and the reliability determination are explained in Section 2 and 3, respectively. Next, in Section 4, the V4V-Challenge dataset is described. Then, in Section 5, the results are presented, while Section 6 concludes the paper.

2. The Proposed Method

We present a framework that allows analyzing vital signs in RGB video recordings of human faces. As introduced in [7], a person's face can be locally analyzed to extract the blood flow through the face. In order to increase the signal strength, compared to this pixel-based local approach, the face is divided into sub-segments in our method. The major steps of the proposed method are illustrated in Figure 2.

The input to our framework is an RGB video. For each image of this video, we determine the face position. We implemented the face detection and landmark extraction based on an established face analysis algorithm (dlib) [5]. For each detected face, 68 landmarks are extracted. Based on these landmarks, the face is segmented into four symmetric regions that cover the major parts of the face, whereby the eyes are excluded since they mostly do not show visible skin. Figure 1 shows the selected ROIs forehead, right cheek, left cheek, and nose. The fifth ROI is the combination of all sub-regions in Figure 1 (full face region). The mouth and chin areas are not analyzed because their signal quality is relatively low [9]. We apply the symmetrical segments to counteract possible occlusions. For example, if the forehead is covered by hair, the nose is obscured by prominent glasses, or the head is turned sideways so that only one cheek is facing the camera. With the selected segments, the camera sensor always captures one region in its entirety.

The next step extracts the rPPG signal. We choose the POS transformation [16] output as rPPG signal since in [17] it is shown that POS has the best results under different lighting conditions. Further, in [19], POS has the lowest error rate among traditional methods and is only beaten by one convolution network, the rPPGNet (comparison on OBF dataset [10]).

The POS transformation is performed as follows: For every image, each ROI is averaged for each individual R, G, and B channel, and these values are concatenated with the averaged values of that ROI of the previous images. Thus, for each color channel, a one-dimensional time signal $[x_r, x_g, x_b]$ is obtained per region. This signal is then divided by its mean value, resulting in $[\overline{x_r}, \overline{x_g}, \overline{x_b}]$. The output of the POS transformation is the vital signal (e.g., rPPG signal)

$$v = X_s + \frac{\sigma(X_s)}{\sigma(Y_s)} Y_s \tag{1}$$

with

$$X_s = \overline{x_g} - \overline{x_b}, \tag{2}$$

and

$$Y_s = -2\overline{x_r} + \overline{x_g} + \overline{x_b}, \tag{3}$$

where σ is the standard deviation.

The described POS projection [16] is applied to each of the temporal signals (always for the R, G, and B color channels associated with an ROI), which yields to the rPPG signals. The rPPG signal of every ROI is analyzed to extract the HR. A sliding window of 10 s is used to determine the HR, according to [7, 12]. After every calculation, this window is moved forward by 3 s, and the HR calculation is repeated. Then, the estimated HR is assigned to each frame within the three seconds time shift. When first estimating the heart rate for a video file, the determined HR is assigned to each frame in the 10 s window. If an input

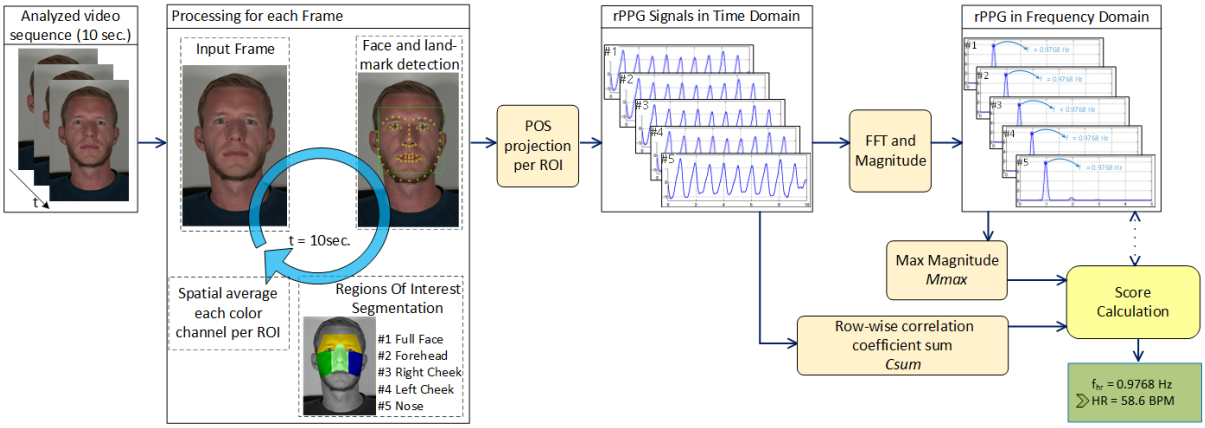

Figure 2. This figure shows the principle workflow of the proposed method. The analyzed input sequence is always 10 s (except when the whole video file is smaller). From each frame of this sequence, the facial landmarks are extracted. The face is segmented into five ROIs (full face, forehead, right cheek, left cheek, and nose) based on these landmarks. Next, the RGB color channels of each ROI are spatially averaged. The resulting temporal signals are then the input for the POS transformation. After an FFT, the maximum magnitude $Mmax$ (peak on the power spectrum density) is selected.The correlation coefficient sum $Csum$ is calculated from the rPPG signals in the time domain. Finally, the ROI that holds the HR is determined based on $Mmax$ and $Csum$.

video is shorter than 10 s, the entire video is analyzed, and the measured heart rate is mapped to every frame.

As shown in [7, 11], measuring the heart rate in the frequency domain is an established procedure. Thereby, for each ROI, the time domain rPPG signal is transformed via FFT into the frequency domain. As illustrated in Figure 2 (rPPG Frequency Domain), the frequency component with the highest magnitude in the range from 0.75 Hz to 4.0 Hz (corresponding to a heart rate between 45 BPM and 240 BPM) represents the heart frequency f_{HR} for the ROI during the analyzed sliding window. The magnitude of f_{HR} is referred to as $Mmax$ (peak on the power spectrum density). The heart rate in beats per minute (BPM) is the result of $60 \cdot f_{HR}$.

In this paper, a spectral frequency resolution of $\Delta f = 0.1\,\mathrm{BPM}/60$ is chosen for the FFT. To reach this resolution, the number of FFT points N_{fft} is calculated using:

$$N_{fft} = \frac{f_s}{\Delta f}, \tag{4}$$

where f_s is the sampling frequency. For the V4V-Challenge data $f_s = 25$ Hz (see Section 4).

As the last step, the maximum magnitude $Mmax$ and the rPPG signals in the time domain are used to determine the HR calculation's reliability for each ROI. The HR of the ROI with the highest reliability score is selected since this region holds the most robust HR information.

3. Reliability Determination

Two features are used as input for the reliability determination, cf. Figure 2. The first feature is the maximum magnitude $Mmax$ (corresponding to f_{HR}) of each ROI. As the rPPG signal of each individual ROI should be linear dependent to the others, the second feature is using the correlation coefficient of the five different rPPG signals. The correlation coefficient ρ between each of these waveforms is calculated to quantify this dependence using

$$\rho(v_j, v_k) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{v_{j,i} - \mu(v_j)}{\sigma(v_j)} \right) \left(\frac{v_{k,i} - \mu(v_k)}{\sigma(v_k)} \right), \tag{5}$$

where v_j and v_k represent the vital sign signals (rPPG signal) of two different ROIs, i is the sample index in the time domain rPPG signal, μ is the mean and σ is the standard deviation. Generally, all five rPPG signals should be similar. However, each signal is unique due to noise, and the stronger the influence of noise, the smaller ρ to the other waveforms. Therefore, the calculated ρ are arranged into a correlation coefficient matrix

$$\mathbf{R} = \begin{pmatrix} \rho(v_1, v_1) & \rho(v_1, v_2) & \dots & \rho(v_1, v_5) \\ \rho(v_2, v_1) & \rho(v_2, v_2) & \dots & \rho(v_2, v_5) \\ \vdots & \vdots & \ddots & \\ \rho(v_5, v_1) & & & \rho(v_5, v_5) \end{pmatrix}, \tag{6}$$

where v_i is the rPPG signal of ROI i. The sum $Csum$ is calculated for each row of $\mathbf{R}$. The highest sum shows the signal with the strongest agreement.

The goal is to select the region where the frequency component of the $Mmax$ corresponds best to the HR. However, a high $Mmax$ does not necessarily represent the actual HR but another periodicity in the signal v. This periodicity can be caused by motion, respiration, illumination changes, noise from the camera sensor, partial face coverings, or poor rPPG signal strength. Such an rPGG signal is then referred to as noisy.

ROI	$Mmax_{norm}$	$Csum_{norm}$	reliability	HR
#1	0.427	1.0	1.427	123.4
#2	0.714	0.405	1.119	71.9
#3	0.564	0.257	0.821	105.8
#4	0.0	0.0	0.0	229.5
#5	1.0	0.562	1.562	**74.3**

Table 1. Example for the reliability calculation. These values are the results for the first 10 s of the randomly chosen video file *File M034_T1.mkv*. In this example, the nose region (#5) has the highest reliability and corresponds to an average HR of 74.3 BPM during the analyzed period. The HR column is in BPM. Therefore, the ground truth for these 10 s varies between 74.1 BPM and 90.6 BPM and has an average of 81.7 BPM.

In our region-based approach, several regions (in the best case, all of them) can be expected to hold the correct HR. Thus, the waveform of v is supposed to be similar in several ROIs. In order to exclude noisy ROIs with high $Mmax$, the correlation coefficients ρ between the individual regions are calculated.

Suppose many regions are noisy and only one region holds the correct heart rate. In that case, this correct region has a high $Mmax$ and a higher $Csum$ to the other signals than the noisy signals to the others. The region with the accurate HR has a similarity to all regions since all are rPPG signals. Thus, the noisy signals are more likely to be present in the correct signal than among themselves. On the other hand, the noise in the signals is assumed to be random, so the ρ of all single noisy ROIs is relatively small and is the respective $Csum$.

Thus, the sum of the correlation coefficient $Csum$ is a suitable complement to the $Mmax$ for the reliability determination. Both inputs are min-max-scaled (to $Mmax_{norm}$ and $Csum_{norm}$) and then added region-wise. Table 1 shows an example of the reliability determination. Finally, the highest reliability proposes the region that holds the HR.

4. Dataset

The dataset [13] provided by the V4V-Challenge group is divided into three sets (training, validation, and test). In total, 1358 videos are provided. The length of these recordings varies between 5 s and 206 s. The physiological data were collected with the BIOPAC MP150 data acquisition system (see [21] for more information). This system captures physiological signals with a sample rate of 1000 Hz. First, the blood pressure signal is captured with the blood pressure monitoring system Biopac NIBP100D from the arm and fingers of the subject. Then, with the BIOPAC AcqKnowledge software, the heart rate labels are extracted from the blood pressure signal. Thus, the published ground truth data include the raw blood pressure signal and the heart rate labels for each video frame ($f_s = 25$ Hz).

5. Results

For each video file, the standard mean-absolute-error (MAE), root-mean-square-error (RMSE), and Pearson Correlation Coefficient ρ are calculated between the measured heart rate and the ground truth labels. Table 2 shows the result for the training (divided into two sub-sets) and validation set. In this table, the means of all evaluation values are calculated ($\overline{MAE}$, $\overline{RMSE}$, $\overline{|\rho|}$). The mean of the MAE over all video file is 8.324 BPM.

The relatively high errors and standard deviations can be explained to a large extent by the very fluctuating ground truth values. In the ground truth, a newly determined HR occurs approx every 10 to 30 frames. This new HR frequently fluctuates largely with jumps of ± 15 BPM. In the example given in Table 1, the HR varies between 74.1 BPM and 90.6 BPM in the analyzed 10 s interval. Such recurring abrupt changes in HR are physiologically abnormal for healthy people at rest. In our method, a mean heart rate is calculated over this 10 s interval and then set for a range of at least three seconds. Using the mean HR prevents such abrupt changes in short time intervals, being a reason for the high MAE deviations shown in Table 2.

	$\overline{MAE}$ $\sigma(MAE)$	$\overline{RMSE}$ $\sigma(RMSE)$	$\overline{\lvert\rho\rvert}$ $\sigma(\lvert\rho\rvert)$
Training set 1	7.307 (9.221)	9.566 (10.272)	0.35 (0.231)
Training set 2	7.673 (7.507)	10.292 (9.486)	0.349 (0.229)
Validation set	10.511 (10.858)	13.648 (13.101)	0.302 (0.233)
All Videos	8.324 (9.241)	10.956 (10.989)	0.336 (0.231)

Table 2. Each of the training sets consists of 362 video files; the validation set of 276 files. For each video file, the MAE, RMSE, and R is calculated. This table shows the mean and standard deviation σ of these values for the corresponding dataset.

6. Conclusion

This paper describes our method for determining the heart rate used in the V4V-Challenge. The proposed region-based approach provides adequate results with a mean MAE of 8.324 BPM across all data. At the same time, the standard deviation (see Table 2) for each data set is relatively high. These high standard deviations and errors are due to the large fluctuations in the ground truth data, making them less meaningful. Furthermore, our model-based method is entirely automatic and does not require large amounts of data for training or time-consuming training sessions; our approach can be applied immediately.

References

[1] J G Betts, P Desaix, E W Johnson, J E Johnson, O Korol, D Kruse, B Poe, OpenStax College, J Wise, M D Womble, and Others. *Anatomy & Physiology*. Open Textbook Library. OpenStax College, Rice University, 2013. 1

[2] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019. 2

[3] Weixuan Chen and Daniel McDuff. DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11206 LNCS, pages 356–373, 2018. 2

[4] Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2

[5] Davis E King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009. 2

[6] Benjamin Kossack, Eric L Wisotzky, Anna Hilsmann, and Peter Eisert. Local Remote Photoplethysmography Signal Analysis for Application in Presentation Attack Detection. *Vision, Modeling and Visualization*, pages 135–142, 2019. 2

[7] Benjamin Kossack, Eric L. Wisotzky, Anna Hilsmann, Peter Eisert, and Ronny Hänsch. Local blood flow analysis and visualization from RGB-video sequences. *Current Directions in Biomedical Engineering*, 5(1):373–375, 2019. 2, 3

[8] Karl H.E. Kroemer, Hiltrud J. Kroemer, and Katrin E. Kroemer-Elbert. *Engineering physiology: Bases of human factors engineering/ergonomics*. 2010. 1

[9] Sungjun Kwon, Jeehoon Kim, Dongseok Lee, and Kwangsuk Park. ROI analysis for remote photoplethysmography on facial video. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4938–4941, 2015. 2

[10] 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, 2018. 2

[11] 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. 1, 2, 3

[12] 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*, pages 1165–1170, 2017. 1, 2

[13] Ambareesh Revanur, Zhihua Li, Umur A Cifti, Lijun Yin, and László A Jeni. The First Vision For Vitals (V4V) Challenge for Non-Contact Video-Based Physiological Estimation. In *IEEE/CVF International Conference on Computer Vision Workshops, 2021*, 2021. 1, 4

[14] Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable Photoplethysmographic Sensors—Past and Present. *Electronics*, 3(2):282–302, 2014. 1

[15] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F. Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[16] 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, 2017. 1, 2

[17] Ze Yang, Haofei Wang, and Feng Lu. Assessment of Deep Learning-based Heart Rate Estimation using Remote Photoplethysmography under Different Illuminations. 2021. 2

[18] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *30th British Machine Vision Conference 2019, BMVC 2019*, 2020. 2

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, oct 2019. 2

[20] Sebastian Zaunseder, Alexander Trumpp, Daniel Wedekind, and Hagen Malberg. Cardiovascular assessment by imaging photoplethysmography - A review. *Biomedizinische Technik*, pages 1–18, 2018. 1

[21] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:3438–3446, 2016. 4