Skin Segmentation for Imaging Photoplethysmography Using a Specialized Deep Learning Approach

Matthieu Scherpf, Hannes Ernst, Leo Misera, Hagen Malberg, Martin Schmidt

Institute of Biomedical Engineering, TU Dresden, Dresden, Germany

Abstract

Imaging photoplethysmography (iPPG) is a camera-based approach for the remote measurement of superficial tissue perfusion most commonly applied to facial video recordings. Since only tissue contains information about perfusion, skin detection is a necessary processing step. Several approaches for the detection of skin pixels in video recordings have been developed, e.g. using color thresholds. Within this work we present a deep learning based approach capable of combining color and morphology information, which makes the skin detection robust against different illumination conditions. We evaluated our new approach using two datasets with 182 individuals of different gender, age, skin tone and illumination conditions. Our approach outperformed state-of-the-art algorithms or yielded at least comparable results (mean absolute error of estimated pulse rate improved by up to 68 %). The method presented allows more accurate assessment of superficial tissue perfusion with iPPG.

1. Motivation

Imaging photoplethysmography (iPPG) enables the remote measurement of the blood volume pulse using RGB cameras. Usually facial regions are recorded because of the high superficial perfusion. Due to its simplicity, iPPG offers high potential to become an easy-to-use and widely available diagnostic tool. Because only skin pixels contain relevant information, skin detection is a necessary processing step for the application of iPPG. Several approaches for the detection of skin pixels in video recordings have been developed. Some methods rely solely on color thresholding which is based on the spectral characteristics of human skin [1]. Other approaches use additional processing steps to include morphological information [2].

In this work, we present a new deep learning based approach which uses the *DeepLabV3+* [3] network architecture and is trained on the specific task of skin detection. The capability of DeepLabV3+ to process complex information enables the combination of color and morphology features for robust skin detection under various conditions (e.g. illumination and skin color). For evaluation purposes, we compared the network with two state-of-the-art algorithms for skin detection.

2. Methodology

2.1. DeepLabV3+ Implementation

Originally introduced for semantic image segmentation, DeepLabV3+ is a deep neural network architecture capable of reliably segmenting the content of an image [3]. A task-specific training procedure tunes this general ability of image segmentation towards more specific purposes such as skin detection. We implemented such a training procedure using suitable datasets alongside image augmentation. We provide our source code and model as open source[1].

Our implementation is based on the official Deeplab repository available on Github[2]. We used the DeepLabV3+ architecture with the *Xception*-Backbone. The network was trained using four datasets commonly used for skin segmentation. To summurize, these datasets consist of 7049 images (see Table 1) showing one or more people within various settings (e.g. outdoor and indoor) and exhibit a large diversity regarding the recording parameters (e.g. illumination and resolution). For improved readability, DeepLabV3+ will be called *Deeplab* in the following sections.

2.2. Evaluation Datasets

We evaluated our approach with the *Multimodal Spontaneous Emotion database (BP4D+)* [10] and the *Univ. Bourgogne Franche-Comté Remote PhotoPlethysmoGraphy (UBFC)* dataset [11]. The BP4D+ consists of 140 individuals of different sex, age and skin color. For every subject, 10 RGB videos of 10 to 60 seconds length were recorded. The UBFC consists of 42 individuals, with one

[1] https://github.com/BeCuriousS/ippg-toolbox Release: v1.0

[2] https://github.com/tensorflow/models/tree/master/research/deeplab Release: v2.4.0

 ISSN: 2325-887X DOI: 10.22489/CinC.2021.195

Table 1: Deeplab datasets. Abbr.: ECU Face and Skin Detection dataset (ECU), dataset based on FERET and AR Facial Images (SFA), dataset for Hand Gesture Recognition (HGR), dataset from Schmugge et al. (SCH).

Dataset	ECU	SFA	HGR	SCH	Overall
Images	3999	1118	1558	374	7049
Source	[4]	[5]	[6–8]	[9]	-

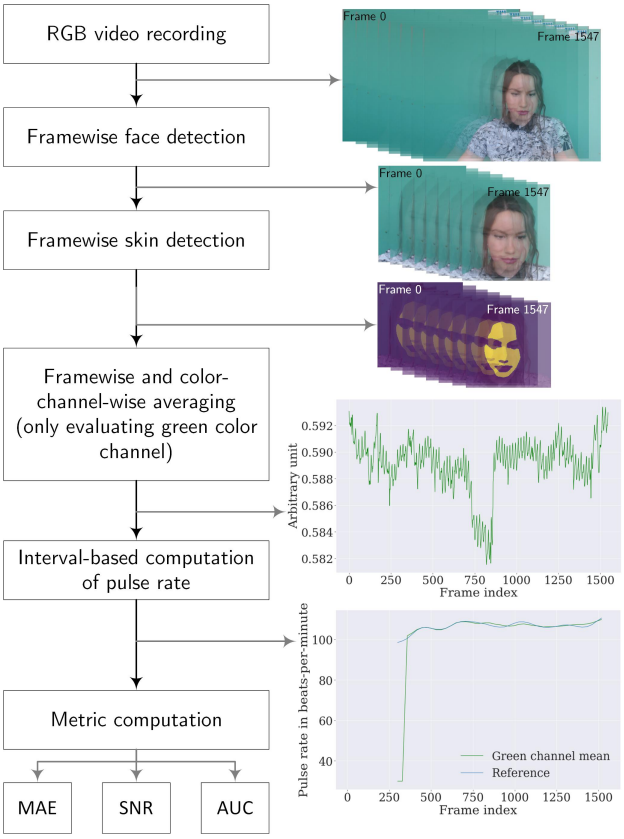


Figure 1: Processing pipeline for the iPPG performance comparison. The frames are taken from the UBFC dataset [11]. Abbr.: mean absolute error (MAE), signal-to-noise ratio (SNR), area under curve (AUC).

RGB video of approx. 60 seconds length per recording. Both datasets use a constant illumination setting but differ regarding the illumination characteristics.

2.3. Processing Pipeline

Figure 1 illustrates our processing pipeline which consisted of the five steps described in the following.

Framewise face detection was implemented to roughly focus on the region of interest (ROI) and therefore to discard unnecessary information, i.e. most of the recorded background.

Framewise skin detection was performed by the three different approaches to compare Deeplab with two state-of-the-art approaches called *Cheref* [1] and *Levelset* [2]. Cheref combines thresholding in several color spaces exploiting the specific color properties of human skin. This represents the simplest of the three approaches and was implemented according to the official repository[3]. Levelset uses the combination of color thresholding with level set segmentation and tracking to create the final skin ROI.

Framewise green color channel averaging yielded one scalar value per frame representing the mean value of the green color channel for the skin mask. Finally, a 1D signal for each video recording was created this way. Within this signal, information about the pulsation of blood can be expected. We used this simple iPPG extraction method to focus on the skin detection and avoid influences of any other processing steps.

Interval-based computation of the pulse rate was subsequently performed for each of the 1D signals. Each signal was split into intervals using a sliding window with a length of 10 seconds and an overlap of 9 seconds. The pulse rate of an interval was defined as the maximum frequency between 0.5 Hz (equals 30 bpm) and 3.33 Hz (equals 200 bpm) extracted from the periodogram. These frequency bounds for the pulse rate extraction were chosen according to *IEC 60601-2-27*. The pulse rate extracted from video recordings is referred to as *estimated pulse rate*. The ground truth pulse rate was either computed from a PPG finger probe in the same way (UBFC) or included in the dataset (BP4D+).

Within the **metric computation**, the mean absolute error (MAE) between the estimated and the ground truth pulse rate, the signal-to-noise ratio (SNR) and the area under the accuracy curve (AUC) were computed. The AUC represents an accuracy measure and ranges between 0 and 1 [12]. The SNR was computed similar to [13] using 0.1 Hz (equals 6 bpm) and 0.2 Hz (equals 12 bpm) around the fundamental oscillation, i.e. the ground truth pulse rate and its first harmonic, respectively. To reveal significant differences for the computed metrics, we applied the *Wilcoxon signed-rank test*.

3. Results and Discussion

The results for the three approaches are displayed in Figure 2. **UBFC** (see Fig. 2a): Deeplab outperformed the two state of the art approaches. For Deeplab, the MAE decreased in the median by 68 % and 55 % compared to Cheref and Levelset, respectively. The median AUC increased by 48 % and 16 %. The median of the SNR of

[3]https://github.com/CHEREF-Mehdi/SkinDetection Release: No official release made

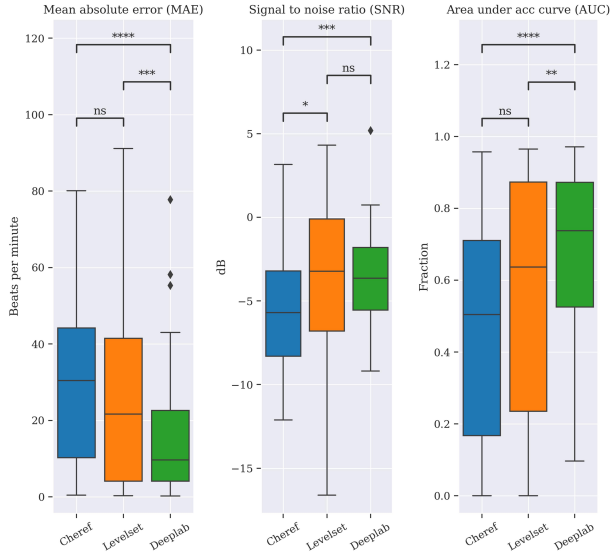
Levelset was slightly higher (0.4 dB) but the variation of Deeplab, measured by the interquartile range, was 28 % lower.

BP4D+ (see Fig. 2b): Deeplab delivers similar performance compared to Cheref regarding all metrics. The median differences between Deeplab and Cheref were 1.6 % and 2.1 % for MAE and AUC, respectively. Levelset achieved the lowest performance with an increased median MAE of 52 % compared to Deeplab.

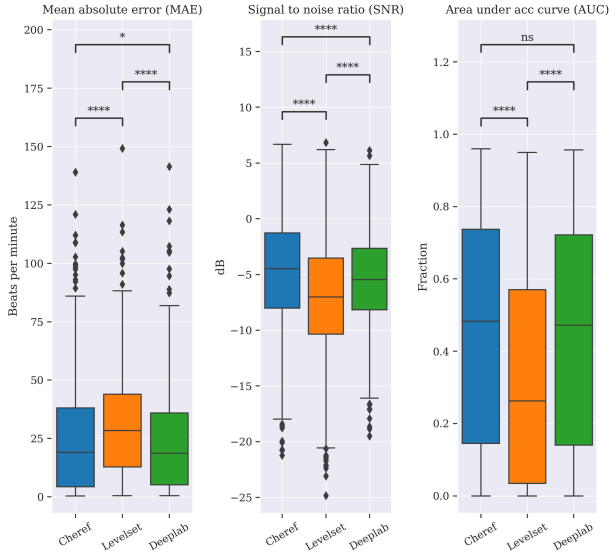
Overall, Deeplab demonstrated a both sensitive and robust performance, i.e. its results were the best or at least as good as the best state of the art approach's results. In comparison, the skin masks detected with Deeplab were more robust (see Fig. 2c and 2d). The Cheref approach performance was lower for the UBFC dataset than for the BP4D+ dataset which we attribute to the different illumination setting. This underlines Deeplab's superior performance in combining color and morphology features.

4. Conclusion

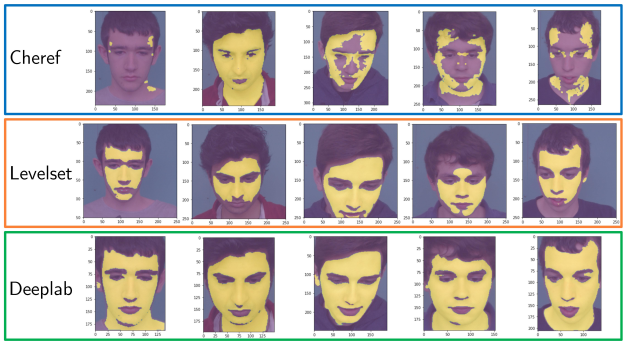
Our Deeplab approach, consisting of the DeepLabV3+ neural network architecture trained for the specific task of skin detection, shows both sensitive and robust perfor-



(a) Metrics for UBFC dataset. Deeplab achieved the best performance for MAE and AUC. Regarding the SNR Deeplab achieved comparable performance to Levelset. Median values for Cheref/Levelset/Deeplab: 30.4/21.7/9.7 (MAE in bpm), -5.7/-3.2/-3.6 (SNR in dB), 0.50/0.64/0.74 (AUC).



(b) Metrics for BP4D+ dataset. Deeplab and Cheref achieved similar performance for all metrics outperforming Levelset. Median values for Cheref/Levelset/Deeplab: 19.0/28.4/18.7 (MAE in bpm), -4.5/-7.0/-5.5 (SNR in dB), 0.48/0.26/0.47 (AUC).



(c) Example skin masks for UBFC dataset. Cheref computed the worst skin masks. Overall, the Deeplab computed the most suitable skin masks.

(d) Example skin masks for BP4D dataset. The Levelset approach computed the worst skin masks. Cheref and Deeplab skin masks delivered comparable results.

Figure 2: Results for the two evaluation datasets and the three skin detection approaches. Boxplots of the metrics for a) UBFC and b) BP4D+. Exemplary skin masks for c) UBFC and d) BP4D+. Significance from statistical analysis (Wilcoxon signed-rank test) marked as: ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.00001$.

mance for iPPG on two large and diverse datasets. As there are no postprocessing steps for Deeplab implemented yet, we will investigate such steps to further improve the computed skin masks. Face detection was used for a more efficient calculation of the skin masks. However, Deeplab is capable for the detection of any skin area and not limited to the face. In conclusion, robust skin detection as provided by our Deeplab approach enables the extraction of iPPG signals even under varying conditions and allows for more accurate remote assessment of superficial tissue perfusion with iPPG.

Acknowledgments

The authors are grateful to the Centre for Information Services and High Performance Computing TU Dresden for providing its facilities for high throughput calculations.

Funding

This work was partly supported by grants of the European Regional Development Fund, the Free State of Saxony and the German Research Foundation (ERDF 100278533; DFG 319919706/GRK2323).

References

[1] Dahmani D, Cheref M, Larabi S. Zero-sum game theory model for segmenting skin regions. Image and Vision Computing 2020;(99). ISSN 02628856.

[2] Trumpp A, Rasche S, Wedekind D, Schmidt M, Waldow T, Gaetjen F, Plötze K, Malberg H, Matschke K, Zaunseder S. Skin detection and tracking for camera-based photoplethysmography using a bayesian classifier and level set segmentation. In Bildverarbeitung für die Medizin. Heidelberg: Springer Vieweg, Berlin, Heidelberg, 2017; 43–48. URL http://link.springer.com/10.1007/978-3-662-54345-0_16.

[3] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds.), Computer Vision – ECCV 2018, volume 11211 of Lecture Notes in Computer Science. Cham: Springer International Publishing. ISBN 978-3-030-01233-5, 2018; 833–851.

[4] Phung SL, Bouzerdoum A, Chai D. Skin segmentation using color pixel classification: analysis and comparison. IEEE transactions on pattern analysis and machine intelligence 2005;27(1):148–154.

[5] Casati JPB, Moraes DR, Rodrigues ELL. Sfa: A human skin image database based on feret and ar facial images. In IX workshop de Visao Computational. Rio de Janeiro, 2013; .

[6] Grzejszczak T, Kawulok M, Galuszka A. Hand landmarks detection and localization in color images. Multimedia Tools and Applications 2016;75(23):16363–16387. ISSN 1380-7501.

[7] Kawulok M, Kawulok J, Nalepa J, Smolka B. Self-adaptive algorithm for segmenting skin regions. EURASIP Journal on Advances in Signal Processing 2014; 2014(170):1–22. ISSN 1687-6180. URL http://asp.eurasipjournals.com/content/2014/1/170.

[8] Nalepa J, Kawulok M. Fast and accurate hand shape classification. In Kozielski S, Mrozek D, Kasprowski P, Malysiak-Mrozek B, Kostrzewa D (eds.), Beyond Databases, Architectures, and Structures, volume 424 of Communications in Computer and Information Science. Springer. ISBN 978-3-319-06931-9, 2014; 364–373.

[9] Schmugge SJ, Jayaram S, Shin MC, Tsap LV. Objective evaluation of approaches of skin detection using roc analysis. Computer Vision and Image Understanding 2007;108(1-2):41–51. ISSN 1077-3142. URL https://www.sciencedirect.com/science/article/pii/S1077314206002268.

[10] Zhang Z, Girard JM, Wu Y, Zhang X, Liu P, Ciftci U, Canavan S, Reale M, Horowitz A, Yang H, Cohn JF, Ji Q, Yin L. Multimodal spontaneous emotion corpus for human behavior analysis. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE. ISBN 978-1-4673-8851-1, 2016; 3438–3446. URL http://ieeexplore.ieee.org/document/7780743/.

[11] Bobbia S, Macwan R, Benezeth Y, Mansouri A, Dubois J. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters 2019;124:82–90. ISSN 01678655. URL https://www.sciencedirect.com/science/article/pii/S0167865517303860.

[12] Wang W, Den Brinker AC, Stuijk S, de Haan G. Color-distortion filtering for remote photoplethysmography. In 12th IEEE International Conference on Automatic Face and Gesture Recognition - FG 2017. Piscataway, NJ: IEEE. ISBN 978-1-5090-4023-0, 2017; 71–78.

[13] Scherpf M, Ernst H, Malberg H, Schmidt M. Deepperfusion: Camera-based blood volume pulse extraction using a 3d convolutional neural network. In 2020 Computing in Cardiology. 2020; 1–4.

Address for correspondence:

Matthieu Scherpf
Institute of Biomedical Engineering, TU Dresden
Fetscherstraße 29, 01307 Dresden
Matthieu.Scherpf@tu-dresden.de