DeepPerfusion: Camera-based Blood Volume Pulse Extraction Using a 3D Convolutional Neural Network

Matthieu Scherpf[1], Hannes Ernst[1], Hagen Malberg[1], Martin Schmidt[1]

[1] Institute of Biomedical Engineering, TU Dresden, Dresden, Germany

Abstract

Imaging photoplethysmography (iPPG) is a camera-based approach for the remote extraction of the blood volume pulse (BVP) most commonly applied to facial video recordings. The major challenges of this promising technique are the low amplitude of BVP signals and their superposition with artifacts as well as physiological and non-physiological movement induced distortions. We addressed this complexity with a 3D convolutional neural network, which we called DeepPerfusion, to improve BVP extraction from iPPG. Our approach is based on the idea of enabling DeepPerfusion to learn the extraction of the BVP from videos by understanding their relation to the ground truth signals. First results show that DeepPerfusion outperforms state-of-the-art algorithms for remote BVP extraction demonstrating a mean absolute error of 0.66 beats per minute (up to $60\,\%$ *improvement) regarding the BVP based pulse rate estimation for 21 randomly chosen held out test subjects of the UBFC dataset.*

1. Motivation

Imaging photoplethysmography (iPPG) enables the remote measurement of the blood volume pulse (BVP) using RGB cameras. Usually facial regions are recorded because of the high superficial blood circulation. Due to its simplicity, this technique offers high potential to become an easy-to-use and widely available diagnostic tool. So far, mainly the pulse rate and breathing rate were measured [1,2]. The most challenging problem is the low signal amplitude often affected by poor or altering lighting conditions and artifacts (e.g. caused by head movement relative to the camera). Additionally, the signal is superimposed by complex movements caused by the BVP and the contraction of the heart itself. Therefore, it is reasonable to encounter the given complexity with a neural network based approach.

As of today, several deep learning based approaches exist to extract the BVP signal from facial video recordings. McDuff et al. proposed a 2D convolutional neural network (2D-CNN) trained on the extraction of the BVP signal's first derivative from the temporal normalized frame difference [2]. A first approach using a simple 3D-CNN was introduced by Bousefsaf et al. [3] where the network was trained purely on synthetic data.

We assumed a specialized 3D-CNN to be particularly suitable for capturing the complexity which we explain in more detail in section 2.3. Thus, we designed a new neural network architecture, which we called *DeepPerfusion*, and compared its performance to state-of-the-art algorithmic approaches for remote BVP extraction.

2. Methodology

2.1. Datasets

We used three datasets for training, validating and testing our approach comprising of two datasets acquired at our institution (named CardioVisioIBMT and ColdStressStudy) and a publicly available dataset (UBFC dataset [4]). The datasets of our institution were used for the training procedure whereas the UBFC dataset was primarily used to test our approach. Each dataset comprises of uncompressed facial RGB video recordings and a synchronized BVP signal, i.e. the ground truth. The specifications of the datasets are summarized in Table 1.

2.2. Preprocessing

At first, every video and the according ground truth signal were resampled at 30 samples per second to simplify the subsequent processing steps. Then, we extracted the region of interest (ROI), i.e. the whole facial region, with the use of a state-of-the-art face detection algorithm (from [5]). The next preprocessing elements were essentially based on the dichromatic reflection model (DRM) used by Wang et al. [1], which defines the time-varying color components (red, green and blue) contained in $\vec{C}_k(t)$ of the k-th skin pixel in an image as

$$\vec{C}_k(t) = I(t) \cdot (\vec{v}_s(t) + \vec{v}_d(t)) + \vec{v}_n(t). \qquad (1)$$

Where $\vec{v}_s(t)$ and $\vec{v}_d(t)$ describe the specular and diffuse reflection proportions, respectively. Sensor quanti-

Table 1. Specifications of the datasets that were used for training, validation and testing. Ground truth signal was aquired by PPG sensors measuring the blood volume pulse. The CardioVisioIBMT dataset comprises of two recordings per subject with different resolution and frames per second. Due to different cardiovascular stressors, the used datasets cover a wide range of signal variations. The split proportions of the last column refer to the subjects, i.e. none of the test subjects were seen by the network during training or validation.

Dataset	*Number of subjects*	*Mean video length (minutes)*	*Color channel bit depth*	*Resolution (pixels)*	*Frames per second*	*Ground truth sample rate in Hz*	*Percentage used for training/ validation/ testing*
CardioVisioIBMT	19	30	12	320×420	100	1000	80/20/0
				640×840	30		
ColdStressStudy	41	30	12	320×420	100	1000	80/20/0
UBFC dataset [4]	42	1	8	480×640	30	30	25/25/50

zation noise is taken into account with $\vec{v}_n(t)$. The luminance intensity is expressed by $I(t)$. We assumed all time-dependent components in Equation 1 to have approximately a recording setting specific constant value regarding a sufficient video interval length and small movement of the recorded surface. Therefore, we split each video into overlapping video segments of two seconds length and divided each pixel by its temporal mean value to minimize the dependency to the recording setting specific conditions, e.g. the light source characteristics. We considered the temporal normalization step as being of high relevance regarding the generalization performance of our approach. We chose the proposed interval length to guarantee the presence of at least one cardiac cycle. To reduce the influence of camera sensor induced quantization noise, we applied a gaussian filter and resized the videos to a resolution of 60×60 pixels using bilinear interpolation. Subsequently, each pixel was centered around its mean and scaled by its standard deviation to accelerate the learning procedure.

The ground truth BVP signals were low-pass filtered using an 8th-order Butterworth filter with a cut-off frequency of 10 Hz. Since the ground truth data for each dataset was recorded with different hardware, the signal intensities also vary in different ranges. To facilitate the learning process and improve the generalization ability, we centered each ground truth signal around its mean value and scaled it by its standard deviation.

The last step consisted of the input and output generation for the neural network. From each video segment we extracted seven frames and the according seven sample points from the ground truth. The network was then trained on predicting the fourth ground truth sample point, i.e. the middle sample point, from the seven input frames. Figure 1 summarizes the preprocessing steps.

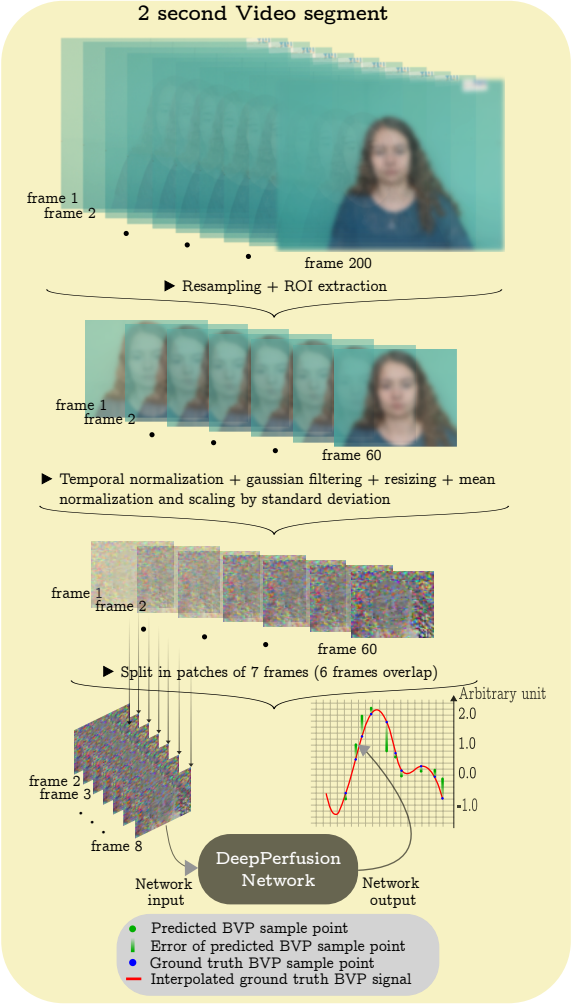

Figure 1. Preprocessing steps to generate the input data for DeepPerfusion. Subject images (taken from [4]) are blurred to preserve privacy.

2.3. DeepPerfusion network architecture

We chose a 3D-CNN because of its superiority regarding the extraction of information from a 3D input-space. Here, two dimensions are used by an image, i.e. a 2D input matrix (ignoring color channels). The third dimension is represented by the time, i.e. a sequence of images. We considered the third dimension being highly relevant because of the following assumptions: (1) it should allow to account for previous and future movement identification in the input sequence and (2) it should enable the search for specific signal patterns. In our case, the network was supposed to search for BVP signal information.

While there exist many 2D-CNN architectures, this is not the case for 3D-CNNs. We built our network architecture in the well-known style of VGG-Net which is originally used for image classification [6]. Figure 2 shows the structure of our so-called DeepPerfusion network. We used three convolutional stages followed by a fully connected stage. Overall, DeepPerfusion consists of roughly five million trainable parameters.

Because we expected the BVP signal in small surrounding areas of a maximum convolution output value to be important as well, we chose average pooling instead of the more commonly used maximum pooling layers so that the pooling layer output composes not only of major but also of minor activations. For the second last layer, we implemented a $tanh$ activation function as we expected its smoother non-linearity to allow a higher flexibility regarding the BVP signal construction of the last layer, i.e. the network output.

2.4. Training, validation and testing procedure

We split the datasets into three parts: training, validation and testing partition. The proportions are indicated in Table 1 and are related to the number of subjects. We trained DeepPerfusion with a batch size of 128 using the *ADAM* optimization algorithm combined with a learning rate of 0.0001. The network was trained on the minimization of the mean of squared errors, i.e. the mean of the squared deviations of the network's output to the preprocessed ground truth BVP signal. For our study, we used the *TensorFlow* package in conjunction with the *Keras* backend.

2.5. Evaluation

We compared our results to state-of-the-art remote BVP extraction algorithms *POS* and *CHROM* from Wang et al. [1] and De Haan et al. [7]. Both algorithms require an additional skin detection step. We implemented POS, CHROM and the skin detection based on the *iPhys*-Toolbox [8].

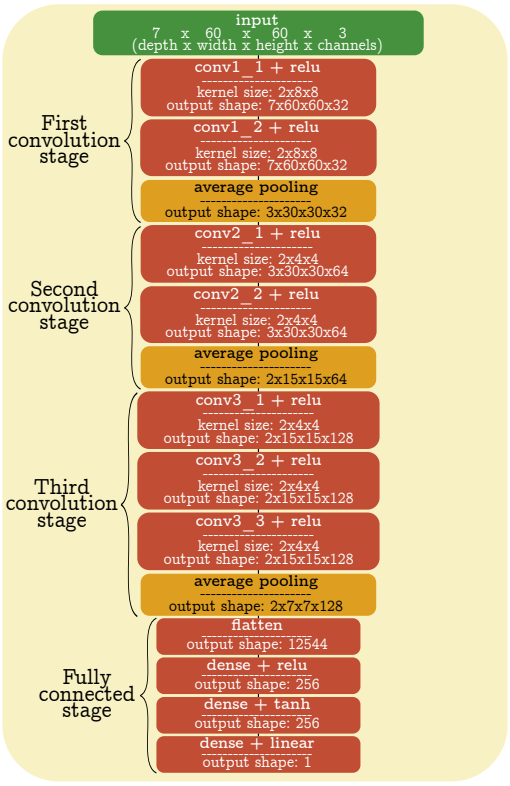

Figure 2. The network architecture of DeepPerfusion. The network has an overall capacity of roughly 5 million parameters. Abbr.: Rectified linear unit (relu), convolution layer (conv).

The computed remote BVP signals passed the same postprocessing which consisted of two steps. At first, the signals were upsampled at $125\,\mathrm{Hz}$ using peace-wise cubic hermite interpolation. Subsequently, a 2nd-order Butterworth filter was applied with lower and upper cut-off frequencies of 0.6 and $8.2\,\mathrm{Hz}$. These boundaries were chosen based on the relevant physiological information we expected in the BVP signal. We extracted the pulse rate of remotely acquired BVP signals and ground truth BVP signals from the frequency domain using a sliding 10 seconds window with a stride of one second.

To compare our results, we used four different metrics related to the quality assessment of pulse rate extraction: root of the mean of squared errors (RMSE), mean of absolute errors (MAE), pearson correlation coefficient (r) and pulse rate accuracy (PR ACC). PR ACC was defined as the ratio of correctly acquired pulse rates. Following *IEC 60601-2-27* (originally for ECG heart rates), a pulse rate was deemed erroneous if the absolute difference between the remotely acquired pulse rate and ground truth pulse rate exceeds the greater of either $5\,\mathrm{BPM}$ or $10\,\%$ of the ground truth pulse rate. Additionally, we calculated the signal-to-noise ratio (SNR) according to de Haan et al. [7]

Table 2. Results of metrics assessing the remotely acquired pulse rate for POS, CHROM and DeepPerfusion for the 21 held out test subjects. Standard deviations, computed on an inter-subject basis, are provided in brackets. Red color indicates the best result for a given metric. Units: MAE and RMSE in BPM, SNR in dB, r and PR ACC without unit.

Metric	*DeepPerfusion*	*POS*	*CHROM*
MAE	**0.66 (0.57)**	3.56 (9.19)	1.66 (5.34)
RMSE	**1.11 (1.10)**	5.74 (11.84)	2.28 (6.39)
r	**0.94 (0.14)**	0.83 (0.30)	0.93 (0.16)
PR ACC	**0.99 (0.03)**	0.93 (0.22)	0.96 (0.16)
SNR	**3.37 (2.31)**	2.92 (3.79)	3.09 (3.25)

for pulse rates from 36 up to 240 beats per minute (BPM) which equals 0.6 - $4\,\mathrm{Hz}$. In contrast to [7], we enlarged the upper SNR frequency boundary from $4\,\mathrm{Hz}$ to $8.2\,\mathrm{Hz}$ to avoid systematic exclusion of the first harmonic for pulse rates higher than $120\,\mathrm{BPM}$.

3. Results and discussion

Table 2 shows the results of POS, CHROM and DeepPerfusion for the 21 held out test subjects of the UBFC dataset. DeepPerfusion exhibited an improved performance in comparison with POS and CHROM which holds true for all analysed metrics. Compared to CHROM, which performed second best, the MAE and the RMSE are reduced by $1.00\,\mathrm{BPM}$ ($60\,\%$ improvement) and $1.17\,\mathrm{BPM}$ ($51\,\%$ improvement), respectively. According to PR ACC, DeepPerfusion was capable to detect additional $3\,\%$ of the pulse rates correctly compared to CHROM while having a lower standard deviation in the test set.

We observed that POS and CHROM performed poorly on a small number of subjects naturally leading to a worse overall metric which can also be seen in the significantly larger standard deviations compared to DeepPerfusion. The reason for this remains unclear as we could not find significant abnormalities compared to other subjects especially regarding the most critical step of skin detection.

4. Conclusion

Our results demonstrate the considerable enhancement of using a specialized deep learning based approach for BVP extraction from facial video recordings. In the next step, we will test DeepPerfusion on a larger dataset exhibiting a larger variation regarding subject age and skin tone. Further, we will analyse the impact of DeepPerfusion being pre-trained onto the dataset that is also used for testing to answer the question of its generalization performance on new - i.e. never seen - recording settings. Given that our final objective is the beat-to-beat and pulse shape analysis using remotely acquired BVP signals, we will investigate experiments to further improve the network performance especially towards a higher SNR.

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 German Research Foundation, the European Social Found and the Free State of Saxony (ERDF 100278533; DFG 319919706/GRK2323; ESF 100339450, “MEDICOS”).

References

[1] Wang W, Den Brinker AC, Stuijk S, de Haan G. Algorithmic principles of remote ppg. IEEE transactions on bio medical engineering 2017;64(7):1479–1491.

[2] Chen W, McDuff D. Deepphys: Video-based physiological measurement using convolutional attention networks. URL https://arxiv.org/pdf/1805.07888.

[3] Bousefsaf F, Pruski A, Maaoui C. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. Applied Sciences 2019;9(20):4364.

[4] 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.

[5] Bradski G. The opencv library. Dr Dobbs Journal of Software Tools 2000;.

[6] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations. 2015; .

[7] de Haan G, Jeanne V. Robust pulse rate from chrominance-based rppg. IEEE transactions on bio medical engineering 2013;60(10):2878–2886.

[8] McDuff D, Blackford E. iphys: An open non-contact imaging-based physiological measurement toolbox. URL https://arxiv.org/pdf/1901.04366.

Address for correspondence:

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