# LCOMS Lab's approach to the Vision For Vitals (V4V) Challenge

Yassine Ouzar
Université de Lorraine
yassine.ouzar@univ-lorraine.fr

Djamaleddine Djeldjli
Université de Lorraine
djamaleddine.djeldjli@univ-lorraine.fr

Frédéric Bousefsaf
Université de Lorraine
frederic.bousefsaf@univ-lorraine.fr

Choubeila Maaoui
Université de Lorraine
choubeila.maaoui@univ-lorraine.fr

## Abstract

*We present in this paper the LCOMS Lab's approach to the 1st Vision For Vitals (V4V) Challenge organized within ICCV2021. The V4V challenge was focused on computer vision methods for vitals signs measurement from facial videos, including pulse rate (PR) and respiratory rate.*

*We propose a novel end-to-end architecture based on a deep spatiotemporal network for pulse rate estimation from facial video recordings. Unlike existing methods, we predict the pulse rate value directly without passing by iPPG signal extraction and without incorporating any prior knowledge or additional processing steps. We built our network using 3D Depthwise Separable Convolution layers with residual connections to extract spatial and temporal features simultaneously. This is very suitable for real-time measurement because it requires a reduced number of parameters and a short video fragment. The obtained results seem very satisfactory and promising, especially since the experiments were conducted in challenging dataset collected in uncontrolled conditions.*

## 1. Introduction

The measurement of vital parameters including heart rate, respiratory rate, blood pressure and body temperature, is one of the first gestures most practiced in daily clinic [9]. Vital signs are primarily critical indicators that can inform healthcare professionals about a person's physical or psychological well-being. They therefore allow the screening and initial medical treatment of several diseases. Physiological parameters are often measured using invasive or non-invasive sensors in direct contact with the human body. Despite all the advantages of contact technologies, they remain psychologically stressful and often uncomfortable due to the use of contact sensors with the body [1]. In addition, their use is almost impossible in cases of trauma, skin ulcer, burns, congenital and contagious diseases [5]. Therefore, these different limits, together with the strong demand for reliable, comfortable, simple, portable, non-stressful and low-cost technology, has prompted researchers to develop new techniques for non-contact measurement of physiological signals. Imaging PhotoPlethysmoGraphic (iPPG) has been able to gain more attention over the past decade through its various qualities by overcoming the drawbacks of contact measurements mentioned above [17]. Thus, it reduces wiring and increases the safety of patients and medical personnel by minimizing the risk of contamination in case of a contagious disease [5] .

All the studies carried out on Photoplethysmographic imaging have greatly improved its performance in terms of reliability and robustness in case of controlled condition (good lighting and motionless subject) [12, 17, 4, 20, 18]. However, at present most of these methods present a weakness in the case of uncontrolled measurement conditions, in particular the subject's motions and low lighting conditions as well as very dark skin (phototype 6) [1, 14]. In this field, deep learning based methods show better performance than conventional state-of-the-art algorithms based on image and signal processing [11, 21]. Recently, several deep learning architectures have been proposed to extract the iPPG signal from a video stream. the resulting signal is then processed to estimate pulse rate. These methods are not one stage. They still require pre-processing or post-processing steps as well as long-term recording for measurements. In addition, they employ private or public databases collected in a controlled environment. However, this makes the study less realistic as the experiences have to be carried out under unconstrained scenarios.

## 2. Related works

The commonly adopted methods for contactless pulse rate measurement using iPPG consist of two-stage pipelines which divide the prediction process into iPPG signal ex-

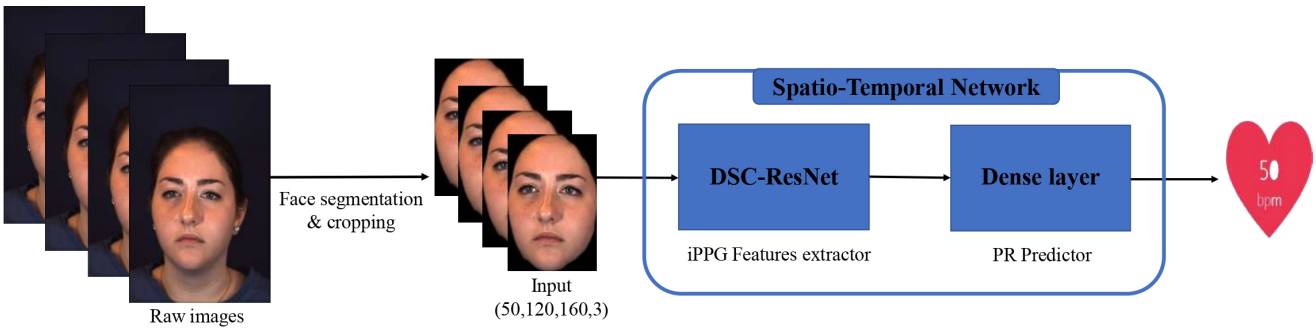


Figure 1. LCOMS Lab's solution pipeline

traction and pulse rate estimation. According to the way of iPPG signal extraction, we can divide the existing works into two major approaches either conventional based methods using image and signal processing algorithms [12, 17, 4, 20, 18, 1], or deep learning based methods that extract the iPPG signal automatically [3, 11, 21, 2]. In this section, we review mainly the state-of-the-art deep learning based methods for contactless pulse rate measurement.

There have been several CNN-based methods for iPPG based contactless pulse rate measurement. Chen and McDuff [3] proposed a two-stream 2D CNN architecture, including one stream of an appearance model to find the appropriate regions of interest (ROI) and the other of motion representation model. The two streams are trained to extract BVP waveform under heterogeneous lighting and significant head motions. Radim et al. [15] proposed a two-stage convolutional neural network method composed of 2D CNN and 1D CNN respectively. The first one extracts the iPPG signal while the second regresses the pulse rate value.

As the 2D CNN cannot directly exploit the temporal features, spatial-temporal modeling techniques were involved in a more explicit way. 3D CNN were used to learn spatial-temporal features for reconstructing precise rPPG signals or estimating pulse rate directly [22, 2]. Niu et al. [11] combined a CNN with gated recurrent units to train spatial-temporal maps generated from multiple ROI. Neural architecture search (NAS) were also proposed to discover a well-performing model with good generalization capacity in less-constrained scenarios [21].

## 3. Our method

The general framework is illustrated in Figure. 1. we consider the task of pulse rate estimation from facial videos as a one stage regression task. We perform first face segmentation [10] to get rid of the background and the non-skin areas. Then, without any additional preprocessing or post processing steps, batches of 50 frames (corresponding to 2 seconds) are fed to a 3D fully convolutional network to learn spatiotemporal features associated with the subtle color changes on these regions to finally estimate the corresponding pulse rate. This section describes each step in detail.

### 3.1. Face segmentation

The commonly used face and facial landmarks detectors often fail in cases of large head motions, occlusions, facial expressions, and black skin. As the dataset used for the challenge is collected under challenging conditions, we perform face segmentation to get rid of non-skin regions that don't hold any color changes associated with cardiac rhythm [10]. The employed algorithm is proposed initially for face swapping and works ideally in challenging scenarios. All the images of the segmented faces are cropped according to the coordinates of the non-zero pixels and then scaled to $160 \times 120 \times 3$ pixels.

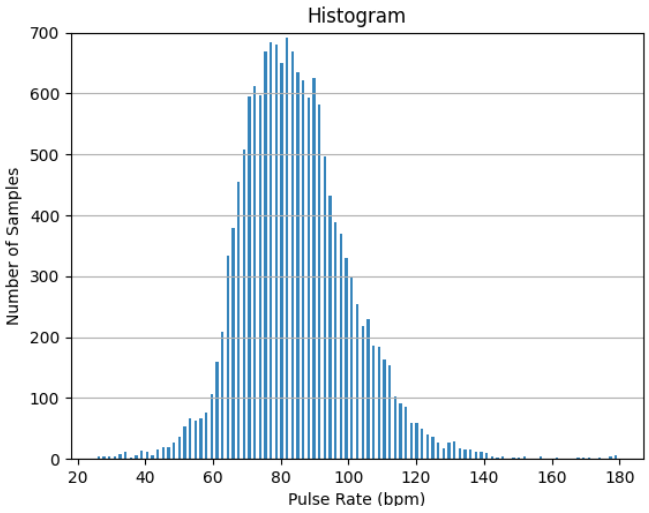


Figure 2. Distribution of the ground truth pulse rates in the V4V database.

### 3.2. Training set augmentation

The ground-truth pulse rates (in beats per minute) of v4v dataset [13] has an inverse Gaussian distribution with more examples for mid pulse rate range [70, 90 Bpm] and less for very high and very low pulse rates (see Figure 2). To avoid the poor predictions for the minority samples, we have performed offline data augmentation on video sequences with pulse rate values larger than 90 Bpm or smaller than 70 Bpm. We have randomly applied image transformations as well (slight rotation, scale, brightness) for each batch to avoid data redundancy and to add robustness of data variation to the network.

### 3.3. Pulse rate estimation neural network

The most existing methods on contactless pulse rate measurement using iPPG consist of two-stage frameworks which extract first iPPG signal and then estimate PR by peak detection. [3, 22, 11, 15, 12, 6, 19]. This approach can achieve more reliable predictions but increases the computation cost and require a long-time window, hence being less convenient for real-time applications. Unlike the commonly used approach, we treat this task as a one-stage regression problem which predict the average pulse rate in only 2 seconds video fragments (2 seconds or T = 50 frames) (see Figure 3). Inspired by mobilenet architecture [7], we built our network using a linear stack of depthwise separable convolution layers to reduce the computational cost and memory requirements. Residual connections are used as well to avoid vanishing gradient problems. Each depthwise separable convolution layer is followed by a batch normalization and ReLU activation function. The final activations of the last convolution layer are then flattened and passed to two dense layers with 1024 and 1 neurons respectively, to estimate the pulse rate value.

## 4. Experiment

### 4.1. Dataset

V4V dataset provided by the organizers of the V4V Challenge is used for both training and testing [13]. It consists of totally 1400 RGB videos recorded from 140 participants (82 females and 58 males) with diverse ethnic ancestries. Each participant is involved in 10 sessions that aimed at evoking different emotions which makes it more challenging for heart rate estimation. The length of each video is between 30 seconds to 1 minute. The frame rate is 25 fps, and the resolution of each image is 1040 x 1392 pixels. Heart rate is collected by a contact sensor operating at a sample rate of 1 kHz. Since we use in our experiment 2 seconds video fragment to predict the pulse rate value, each 50 frames take the mean of 2000 pulse rate values as label.

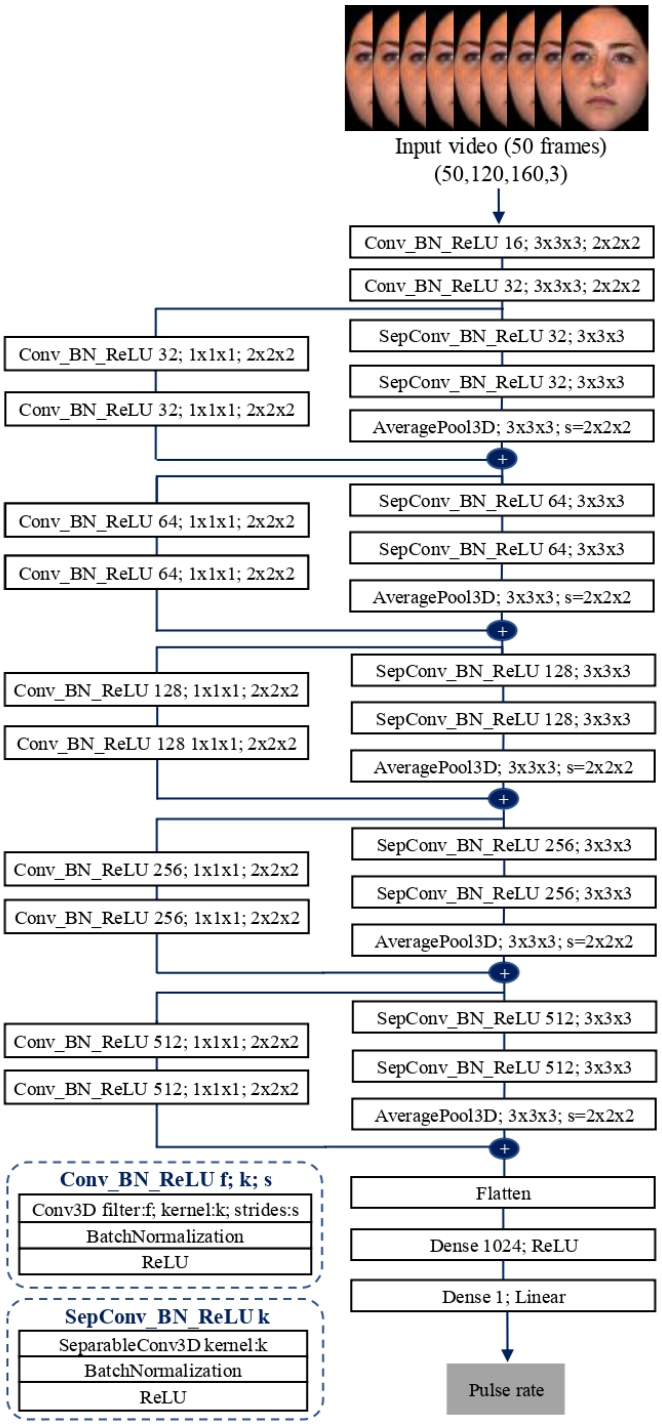


Figure 3. The framework of spatio-temporal networks for pulse rate estimation directly from facial videos recording.

### 4.2. Evaluation Metrics

We evaluate the performance of our approach on the test set of V4V dataset provided for the V4V challenge [13].

Three widely evaluation metrics were used including the mean absolute error (MAE, see equation 1), the root mean square error (RMSE, see equation 2), and the Pearson's correlation coefficient (r, see Equation 3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|PR_i - \widehat{PR_i}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(PR_i - \widehat{PR_i}\right)^2} \quad (2)$$

$$r = \frac{\sum_{i=1}^{n}(PR_i - \overline{PR_i})(\widehat{PR_i} - \overline{\widehat{PR_i}})}{\sqrt{\sum_{i=1}^{n}(PR_i - \overline{PR_i})^2(\widehat{PR_i} - \overline{\widehat{PR_i}})^2}} \quad (3)$$

The MAE and RMSE show the difference between the predicted and the ground truth pulse values. While the pearson correlation coefficiennt R examines the strength and direction of the linear relationship between them on scale of [-1 1]. The smaller value indicates better performnace for MAE and RMSE whilst the larger R indicates better performance.

### 4.3. Implementation details

We implemented our method in keras and Tensorflow frameworks and ran it on Nvidia Quadro P6000s. We used Rectified Adam (RAdam) optimizer [8] to optimize MSE loss. We trained the network for 30 epochs with batch size = 50, learning rate $10^-4$ and decay = $10^-2$. It took approximately 20 minute for each epoch. In addition to a dropout layer [16] of 0.4 ratio that is applied before the final dense layer of the networks, L1 and L2 regularization strategies with coefficient equal $10^-3$ are employed which help to overcome overfitting issue and improve the model generalizability to new data.

## 5. Results

The proposed end to end approach is trained and tested on the V4V dataset without using any external data. It shows good performance with an MAE of 11.60 bpm, an RMSE of 14.90 bpm and a r of 0.20.The obtained results seem very satisfactory and promising, although the training is carried out on an unbalanced data set. Moreover, our approach was initially developed to perform a prediction upon every 2 second recording portion (50 frames). But prediction per frame was instructed in the challenge. Thus, we think that our model was not fully adapted with this requirement, and this may be the reason why the average error over the entire test set was a bit high. Despite that, our model runs in real-time both at GPU ( 150ms) and CPU ( 260ms).

## 6. Conclusion

In this paper, we proposed LCOMS Lab's approach for contactless pulse rate estimation from facial videos. Pulse rate values estimated with this method was submitted for the 1st V4V Challenge [13]. All the experiments were conducted on the challenging V4V dataset provided by the challenge organizers.

The proposed solution is an efficient model built on a linear stack of depthwise seprable convolution layers concatenated with residual connections. This combination significantly reduces the number of parameters and the computational time without any performance degradation. This architecture performs competitively and can serve as a baseline for future robust architecture in real time applications.

## References


[1] 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.

[2] Frédéric Bousefsaf, Alain Pruski, and Choubeila Maaoui. 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. *Applied Sciences*, 9:4364, 10 2019.

[3] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

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

[5] Djamaleddine Djeldjli, Frédéric Bousefsaf, Choubeila Maaoui, Fethi Bereksi-Reguig, and Alain Pruski. Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera. *Biomedical Signal Processing and Control*, 64:102242, 2021.

[6] Gerard Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on bio-medical engineering*, 60, 06 2013.

[7] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, M. Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

[8] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

[9] Craig Lockwood, Tiffany Conroy-Hiller, and Tamara Page. Vital signs. *JBI reports*, 2(6):207–230, 2004.

[10] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard G. Medioni. On face segmentation, face swapping, and face perception. *CoRR*, abs/1704.06729, 2017.

[11] 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.

[12] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, 58(1):7–11, Jan. 2011.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[14] Ruchika Sinhal, Kavita Singh, and Anuraj Shankar. Estimating vital signs through non-contact video-based approaches: A survey. In *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*, pages 139–141. IEEE, 2017.

[15] Radim Špetlík, Vojtech Franc, and Jirí Matas. Visual heart rate estimation with convolutional neural network.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

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

[18] 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, 2016.

[19] Wenjin Wang, A. D. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64:1479–1491, 2017.

[20] Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2015.

[21] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching. *IEEE Signal Processing Letters*, 27:1245–1249, 2020.

[22] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419*, 2019.