# Deep Learning for Remote Heart Rate Estimation: A Reproducible and Optimal State-of-the-Art Framework

Nelida Mirabet-Herranz[1], Khawla Mallat[2], and Jean-Luc Dugelay[1]

[1] EURECOM, 450 Route des Chappes, 06410 Biot, France
{mirabet,dugelay}@eurecom.fr
[2] SAP Labs, 805 Avenue du Dr Donat– Font de l'Orme, 06259 Mougins, France
mallat@eurecom.fr

**Abstract.** Accurate remote pulse rate measurement from RGB face videos has gained a lot of attention in the past years since it allows for a non-invasive contactless monitoring of a subject's heart rate, useful in numerous potential applications. Nowadays, there is a global trend to monitor e-health parameters without the use of physical devices enabling at-home daily monitoring and telehealth. This paper includes a comprehensive state-of-the-art on remote heart rate estimation from face images. We extensively tested a new framework to better understand several open questions in the domain that are: which areas of the face are the most relevant, how to manage video color components and which performances are possible to reach on a public relevant dataset. From this study, we extract key elements to design an optimal, up-to-date and reproducible framework that can be used as a baseline for accurately estimating the heart rate of a human subject, in particular from the cheek area using the green (G) channel of a RGB video. The results obtained in the public database COHFACE support our input data choices and our 3D-CNN structure as optimal for a remote HR estimation.

**Keywords:** Remote HR estimation · 3D-CNN · G channel · ROI

## 1 Introduction

Heart rate (HR) is an important physiological signal that reflects the physical and emotional status of an individual. Monitoring physiological parameters, such as heart rate is of great importance to address an individuals' health status and it is beneficial not only for patients in critical situations, but also for high-risk patients in home-care and outdoor areas [2]. Photoplethysmography (PPG) is a low-cost and noninvasive means of sensing the cardiovascular blood volume pulse through subtle color variations in reflected light of human skin [1]. Although PPG is typically implemented using dedicated light sources, Verkruysse *et al.* [32] showed that using ambient light as illumination source it is sufficient to capture a person's vital signs from RGB videos. Remote PPG technologies (rPPG), allow for non-intrusive measurements, highly relevant when contact has to be prevented (e.g. skin-damage) or users' cooperation cannot be required (e.g. surveillance). Some studies showed that a laptop camera is enough to capture the subtle changes on skin color that lead to a successful HR estimation[21, 22, 24], making it accessible to every individual with a webcam-equipped laptop or a mobile phone.

In the past years, there has been a growing number of studies dedicated to remote HR estimation via rPPG using data extracted from face videos [5, 12, 21, 34]. Most of those rPPG algorithms are based on handcrafted features and consist of a two-stage pipeline which first extract the rPPG signals from the face, and then perform a frequency analysis to estimate the corresponding average HR from a peak detection algorithm. They also require different preprocessing techniques such as skin segmentation, color space transformation, signal decomposition and filtering steps among others. Some filters require parameter adjustment and tuning according to the data that is being used, making those approaches nearly impossible to replicate as shown in [7]. Nowadays deep learning is successfully used in many tasks related to computer vision and medical analysis, such as body mass index (BMI) estimation from face images [25]. End-to-end deep neural models have out-performed traditional multi-stage methods that require hand-crafted feature manipulation being as well possible to replicate. Therefore, recent works have been focused on implementing deep learning techniques for the rPPG extraction when a large amount of labeled data is available [4, 17, 20, 28, 29]. Its performance can be also improved by the increasing of the training set size, unlike previous hand-crafted methods.

The main contributions of this work are the following: 1) We aim to respond the most common unanswered questions on remote HR estimation by proposing the first study, to our knowledge, of the influence of different inputs on a 3D-CNN based HR estimation network, particularly the selection of face region and the channel choice of a video source. 2) We propose a benchmark for HR estimation by assembling an optimal and reproducible 3D-CNN that directly estimates the HR from face videos. 3) The method is evaluated on the publicly available database COHFACE, allowing comparability with future works, and evaluated against other learning-based HR estimators.

The rest of this paper is organized as follows. In Section 2 a review of the state-of-the-art methods for HR estimation is presented. Section 3 describes the selected neural network that extracts the heart rate from facial videos. The database description, experimental setup and results are presented in Section 4. Finally, we conclude with future research directions in Section 5.

## 2 Related work

In this Section, we give an overview of the state-of-the-art methods for pulse rate estimation presented in two categories: Hand-crafted approaches and learning-based models. The hand-crafted methods aim to estimate the rPPG signal from which the HR is later extracted, while the learning-based models are able to recover the rPPG signal as well as to directly estimate the HR value. In earlier studies, some claims regarding the optimal input unit to use in order to estimate the HR from skin patches were made. Those claims were proved for hand-crafted approaches but not exhaustive study was done for the latest works that include deep learning structures. We aim to identify the most important not-verified statements and to provide an answer to those open questions in Section 4.

### 2.1 Hand-crafted rPPG signal estimation

Traditional measurement approaches for extracting human physiological signals such as HR involve devices that require physical contact. In 2008, Verkruysse *et al.* [32] showed that natural light photo-plethysmography could be used for medical purposes such as remote sensing of vital signs for triage or sport purposes. They also claimed that the G channel of a video contains the strongest plethysmographic signal due to the fact that hemoglobin absorbs green light better than red or blue.

Since small variations in reflected light from the skin can be used to estimate the HR, in the past years, traditional methods studied rPPG measurement from videos taken with digital cameras by analyzing the color changes on facial regions of interest (ROI). In 2010, Poh *et al.* [22] presented a non-contact low-cost method for remotely measuring the HR of a subject using a basic webcam. They extracted the blood volume pulse from the selected facial ROI by spatially averaging the value of the ROI for each color channel and then applying independent component analysis to recover the underlying PPG signal. To compute the average HR value of a video, they applied Fast Fourier Transform (FFT) on the estimated signal to find the highest power spectrum. In [21], they extended their work by adding several temporal filters to prune the PPG recovered signal.

Due to the promising results that previous researchers have obtained, several studies focused on overcoming the problems on rPPG signal recovery. Hann *et al.* [5] highlighted the limitations of blind source separation when motion problems are present in the videos and propose a chrominance-based method that combines two orthogonal projections of the RGB space. Li *et al.* [12] approached the problem of rigid movements by implementing face tracking techniques using facial landmarks. Their research focused as well on the illumination variation problem, which influence was rectified with adaptive filters and by comparing background and foreground illumination. Other authors claimed that the state-of-the-art approaches were not robust enough in natural conditions and tried to improve the quality of the rPPG signal. Tulyakov *et al.* [30] divided the face into multiple ROI regions and introduced a matrix completion approach to prune rPPG signals. Wang *et al.* [34] proposed a projection plane orthogonal to the skin tone for rPPG pulse extraction and afterwards they expanded their research in [38] proposing a joint face detection and alignment model followed by an adaptive patch selection method which chooses the best size-variable triangular patches to exclude undesired facial motions. Later, Niu *et al.* presented one of the first real-time rPPG method for continuous HR measurement which included a multi-patch region selection to remove outlier signals and a distribution-based model to link the rPPG signal to their best HR estimation [16]. Recent works have persisted in the use of hand-crafted methods trying to improve the recovered rPPG signal with band-pass filters [11], adding intermediate steps such as feature points generation for optimum masking and Variational Mode Decomposition (VMD) based filtering [14] or by combining Ensemble Empirical Mode Decomposition (EEMD) with Multiset Canonical Correlation Analysis (MCCA) [27].

The main drawback of the presented methods is that they are partly based on denoising algorithms that do not require any type of training but a complex parameter tuning, making them extremely difficult to reproduce as pointed out in previous researches [7]. They include a spatially averaging of the image color values per ROI which helps in reducing Gaussian noise but fails when the different pixels in the ROI do not follow the same distribution. This average operation is also highly sensitive to different types of noise; motion, lightning and/or sensor artifacts.

### 2.2 HR estimation via learning-based models

The first research, to our knowledge, that introduces machine learning techniques for pulse estimation was presented by Monkaresi *et al.* [15] in 2013. They proposed a modification of [22] to improve the accuracy of HR detection by adding machine learning classification techniques in the last step of the pipeline. After a power spectrum analysis, they explore machine learning techniques to find the cardiovascular pulse frequency. In 2018, Qiu *et al.* approached the problem in a similar way. They applied spatial and temporal filtering to extract the rPPG signal and then they estimated the corresponding HR using a Convolutional Neural Network (CNN) [23].

In the recent years, deep-learning models, especially convolutional networks, have gained more importance in the task of HR estimation. Some of those research works have focused on extracting the rPPG signal from face videos, similarly to the traditional methods. In 2018, Chen *et al.* [4] proposed DeepPhys, the first end-to-end system for recovering physiological signals using a CNN. DeepPhys was trained to learn at the same time the most appropriate mask for ROI selection and to recover the Blood-Volume Pulse (BVP) signals. Yu *et al.* [39] proposed the first known approach that includes the use of 3D-CNN for reconstructing rPPG signals from raw facial videos. In their first research, the whole video frame is passed as an input of the network and the output is expected to be the rPPG estimated signal. In a more recent publication [40], they proposed a two-stage method to extract the rPPG signal in which the 3D-CNN is used for video enhancement to counter video compression loss. Similar to [39], Perepelkina *et al.* [20] developed HeartTrack, a two-stage method that uses a 3D-CNN that recovers the rPPG signal from face frames and a 1D-CNN to map the signals to their corresponding HR values.

Other works have focused on the task of estimating the HR in beats per minute (bpm) from face videos, without an intermediate signal estimation step. In 2018, Spetlik *et al.* [29] proposed their two-step CNN to directly estimate a heart rate from a face video. This network consisted on a pipeline of two CNNs, the first one extracted a 1D embedding from face images and the second one mapped this embedding to the estimated HR of a subject. Later on, Wang *et al.* [37] proposed a double feature extraction stream by adopting first a low-rank constraint to guide the network to learn a robust feature representation and second a rPPG extraction stream. Combining both, they were able to develop a unified neural network to learn the feature extraction and to estimate HR simultaneously. Niu *et al.* [17] introduced a new data transformation to represent both, the temporal and the spatial information in a 2D manner from face videos as input of a deep heart rate estimator. In future researches, they refined this approach by using multiple ROI volumes as its input [18] and by performing data augmentation [19]. Song *et al.* created their own version of spatio-temporal maps constructed from pulse signals extracted from existing rPPG methods to feed their CNN [28]. In [8], Hu *et al.* compared the effectiveness of extracting spatial-temporal facial features using 2D-CNN against 3D-CNN and in [13] Lokendra *et al.* experimented with the utilization of Action Units (AUs) and Temporal Convolutional Networks (TCN) for denoising temporal signals and improve the HR estimation. Other deep learning methods considering the temporal information of a video for direct HR estimation have been explored. Huang *et al.* [9] proposed a deep neural network consisting in 2D convolutional layers and long short-term memory (LSTM) operations. The first to propose a 3D deep learning architecture were Bousefsaf *et al.*, who presented in [3] a method relying on 3D networks with embedded synthetic signals in real videos. This 3D network outputs values recorded in a histogram composed by intervals of 2.5 bpm producing HR predictions per intervals. The model also ensures concurrent mapping by producing a prediction for each local group of pixels which, as already highlighted by some users and confirmed by the authors in their git repository, makes the framework slow since the processing time of one testing sample is on the order of days. As a way to decrease the number of parameters to leverage the tasks of the network, the authors used as input of the 3D structure a random shuffle of the group of pixels in the selected regions of a single channeled frame, and the G-frame was chosen without any further study supporting this choice.

As stated in this Section many researches have work on the task of extracting the rPPG signal and/or the HR of a person from facial videos but little attention has been given on stating an unified criteria for input data choice, specially for the learning based approaches. We aim to verify some claims by providing a study on some choices that an author has to make when implementing a deep learning HR estimator. For this purpose, an overview of recent and relevant works that use deep learning structures for a direct HR estimation from face videos is presented in Table 1. The table presents the model structure chosen in each approach, the type of input data passed to the network, the ROI selected and the public database (if any) in which their results were reported. As reported in Table 1 no comparison between input data or ROI is done when the selection of those needs to be made. The only comparative study made, to the knowledge of the authors, concerned the performance of a HR when a 2D or a 3D CNN is selected as network [8]. We aim to enlarge the state of the art by covering studies such as comparison of different facial areas to be selected as ROI (full face, cheeks and forehead) and which channels of the input video give the more valuable information for a CNN-based HR estimator. We will also report our results in one of the most popular public database for direct HR estimation from face videos using deep learning based approaches, the COHFACE [7] dataset.

## 3 Method

### 3.1 ROI selection

A region of interest (ROI) is a subset of a dataset particularly relevant for a specific purpose. In our study, a ROI is a part of a video frame that contains relevant information for our HR task. The right selection of the ROI on a subject's face is critical to perform an efficient HR estimation. Despite several research works have analyzed the facial region leading to the most accurate estimation, these regions have been tested only with hand-crafted methods [10, 33]. In Section 4, we aim to contribute to this choice by performing an evaluation comparison between the most commonly used ROIs for remote HR estimation.

We extract the cheek area from the face videos as described in Fig. 1. First we divide our videos in 5 second sequences of images creating sub-videos. We detect from every frame in the sub-video the location of the 68 (x,y)-coordinates of the dlib landmark detector to map the shape of the face on the image. Then, we obtain the average landmark points per sub-video and based on those, we compute a 40×40 pixels sized region of each of the two cheek areas for every frame of the sub-video. Most of the HR measurement methods tend to average the color values in the entire ROI and use them as the original rPPG signal. By performing this step, we loose the local information within each ROI, therefore, we choose to pass it entirely as input to our neural network.

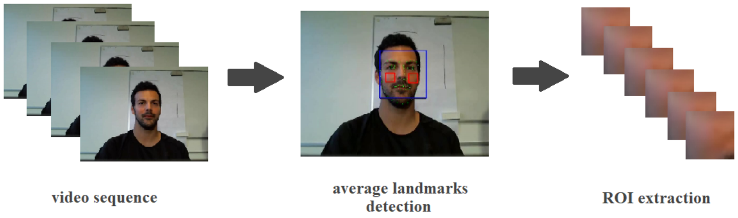

**Fig. 1.** Diagram of the proposed ROI extraction approach from a video sequence.

### 3.2 Green channel selection

In early studies, the strength of the plethysmographic signal in the G channel of a face video was proved sufficent [32]. This is consistent with the fact that

**Table 1.** Overview of the most relevant deep learning structures aiming for a direct HR estimation from face videos. The table includes the model structure, the type of input data, the ROI selected, and the public database in which the results were reported.

| Paper | Year | Structure | Input data | ROI | Databases | Metrics | Code available |
|---|---|---|---|---|---|---|---|
| HR-CNN [29] | 2018 | CNN | RGB | Full frame | COHFACE<br>MAHNOB<br>PURE<br>ECG-Fitness | MAE<br>RMSE<br>$\rho$ | Yes |
| SynRhythm [17] | 2018 | ResNet18 | Spatial-temporal maps | Nose and cheeks | MAHNOB<br>MMSE-HR | Me<br>STDe<br>RMSE<br>MER | No |
| 2-stream CNN [37] | 2019 | Two layer LSTM | Spatial-temporal maps | Full frame | COHFACE<br>PURE | MAE<br>RMSE<br>$\rho$ | No |
| RhythmNet [18] | 2019 | ResNet18 | Spatial-temporal maps | Full face | MAHNOB<br>MMSE-HR | Me<br>STDe<br>MAE<br>RSME<br>MER<br>$\rho$ | Yes but trained model not shared. |
| 3D-Mapping [3] | 2019 | 3DCNN | Shuffled G pixels | Full frame | UBFC-RPPG | Me<br>STDe<br>MAE<br>RMSE | Yes but trained model not shared. |
| Visual-CNN [9] | 2020 | CONV2D with LSTM | RGB | Full face | - | STDe<br>MAE<br>RSME<br>MER<br>$\rho$ | No |
| Robust-CNN [19] | 2020 | CNN | Spatial-temporal maps | Full face | MMSE-HR | STDe<br>MAE<br>RSME<br>MER<br>$\rho$ | No |
| HR-CNN [28] | 2020 | ResNet18 | Spatial-temporal maps | Nose and cheeks | MAHNOB<br>ECG-FITNESS | Me<br>STDe<br>MAE<br>RSME<br>MER<br>$\rho$ | No |
| AND-rPPG [13] | 2021 | Temporal Convolution Networks | RGB | Full face | COHFACE<br>UBFC-rPPG | STDe<br>MAE<br>RSME<br>$\rho$ | No |
| rPPGNet [8] | 2021 | **2DCNN vs 3DCNN** | RGB | Full frame | COHFACE<br>PURE | Me<br>STDe<br>MAE<br>RSME | No |
| Ours | 2022 | 3D-CNN | **RGB vs R vs G vs B** | **Full face vs cheeks vs forehead** | COHFACE | Me<br>STDe<br>MAE<br>RSME<br>MER<br>$\rho$ | Yes, upon request |

hemoglobin absorbs green light better than red and blue [31] light. However, in [32] the fact that the R and B channels may contain complementary information is highlighted, this is why in Section 4 we perform an evaluation of the effectiveness of the G channel selection for our method compared to the choice of R and B channels and the use of the three RGB channels as originally provided in the video. Other deep learning approaches [3] used as input of their structures just the G channel of the face videos although our approach differs from theirs in a crucial point: they consider the selection of the G channel as a way to leverage the tasks of a CNN, reducing the number of parameters of the network without any study that justifies the selection.

### 3.3 Neural network

Convolutional neural networks (CNN) are a type of deep learning models that usually act directly on raw inputs such as images to extract patterns for various tasks. CNNs have been proved very efficient, particularly for classification tasks with which we are dealing in this paper. Those models are often limited to handling 2D inputs. A three dimensional CNN is a network of processing layers used to reduce three dimensional data to its key features so that it can be more easily classified. We model our input data in a three dimensional representation, where the first two dimensions correspond to the 2D images while the third dimension represents time.

Recent works on the literature have proved that 3D-CNN structures successfully handle 3D data such as videos [35, 36]. We believe in the potential of 3D-CNN for extracting the rPPG information embedded in human faces in the same way that we suspected that this type of network has not been exploited yet. Two other works have intended a HR estimation using a 3D-CNN but in our view major drawbacks from those approaches encourage us to propose our optimal and reproducible option. In [3] the authors presented a 3D-CNN that produces a prediction for every pixel present in a video stream leading to a heavy network that leads to a processing time of days for one test video. In [8] a comparison between 2D-CNN and 3D-CNN is performed being the rough implementation of 3D-CNN proved to be more suitable for the HR estimation task. In this work, Hu et al. apply several techniques to improve the performance of both networks but their main focus lays on adding modules to the 2D-CNN structure leading to a lack of exploitation of their 3D-CNN.

The architecture of the selected 3D-CNN is shown in Fig. 2. The input video patch samples are of the size $(300, 40, 40, 1)$ being 300 the number of frames, $(40, 40)$ the ROI size defined in the Section 3.1 and 1 representing the G channel. This input data is passed to the first *convolution layer*, where the video patch is transformed by kernels, sets of learnable filters. The *convolution layers* are followed by *pooling layers*, where filters evaluate small sections at a time to abstract the values to maps. We use *maxpooling layers*, that act as noise suppressant by taking the highest value of an area. After an alternated use of *convolution layers* and *pooling layers* our network has two *dense layers*, resulting from flattening the last *maxpooling layer*. Our last *dense layer* implements a *softmax function* which assigns decimal probabilities to each class to solve the multi-classification problem. Those decimal probabilities add up to 1 for a faster convergence. We decided to exploit the softmax function at the output layer of our network as a way to handle outliers for a better estimation of the HR. By leveraging classification over regression, our network is more resilient to outliers.

When the probabilities are identified and analyzed, the output is assigned to a value, in our case, a *one hot encoding* representation of the HR. The output of our network is then a vector of length $l$, being $l$ the number of classes. In this case, $l = 52$ classes from 48 to 100 bpm with a step of one. Finally, after a conversion from *one hot encoding* vectors to scalar, we perform an average for all the predictions per sub-video for both cheeks, computing the final HR prediction.

All the results in this paper are reproducible using open source tools. The trained model will be publicly available upon request to the authors.

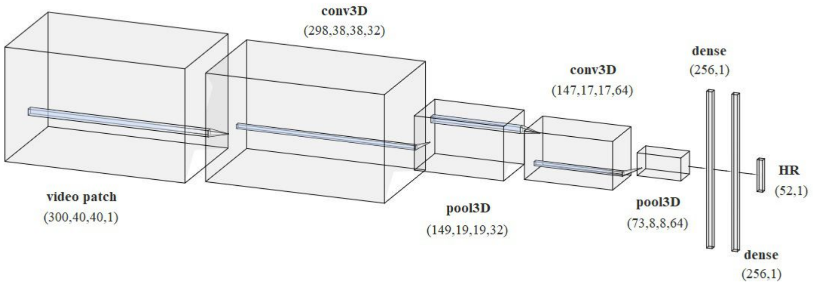

**Fig. 2.** Model architecture. The network takes the data as a 3D input, then alternates between 3D Convolutional layers and 3D MaxPool layers, ending with two fully connected layers that output the estimated HR.

### 3.4 Implementation details

The 3D-CNN structure was implemented in TensorFlow and Keras using a standard chain of conv3D layers, maxpool3D layers and activation functions. After each maxpool3D layer, a batch normalization was applied. Batch normalization was initialized with weights randomly sampled from a Gaussian and their values were scaled with a value $\gamma$ and shifted with a value $\beta$, parameters learnt during training. This was performed to avoid a linear activation of the inputs. A dropout of 0.5 was applied after each batch normalization to ensure a good training process by preventing model overfitting. Rectified linear activation functions were used in every conv3D and dense layer.

The size of the kernels was set to $3 \times 3 \times 3$ for the convolutional layers and to $2 \times 2 \times 2$ for the max pooling layers. The weights of the kernels were initialized sampled from a normal distribution with a mean of zero and a standard deviation of $\sqrt{\frac{\alpha}{n}}$ with $n$ equal to the number of input samples. The model was trained for 10 epochs, Adam optimizer was selected with learning rate set to 0.001 and the loss function chosen was categorical cross-entropy.

# 4 Experiments

## 4.1 Experimental settings

Many existing methods reported their results in private self-collected datasets making difficult a performance comparison of the individual approaches. We want to demonstrate whether our method is capable of performing under different illumination conditions when sometimes part of the subject's face is barely illuminated, i.e. the light source comes from one side and not frontal. To validate those hypothesis and enable a fair comparison with other approaches, we evaluated our method on the public and challenging dataset COHFACE. The COHFACE dataset [7] is composed of 160 facial videos captured at 20 fps with a resolution of 640×480 collected from 40 healthy individuals and their physiological signals. The database includes 12 female and 28 male subjects between 19 and 67 years old. Each video has a length of 60 seconds. Physiological readings were taken by a BVP sensor, which measures changes in skin reflectance to near-infrared lighting caused by the varying oxigen level in the blood due to heart beating. We converted the BVP signals to HR measurement using a function from the Bob package o.db.cohface [26]. The videos in this database have realistic illumination conditions, the subjects are recorded under two different lighting conditions as shown in Fig. 3: (a) Studio, closed blinds, avoiding natural light, and using extra light from a spot to homogeneously illuminate the subject's face, (b) Natural, all the lights turned off and open blinds. The daylight videos (b) represent one of the main challenges of this research since the right side of the subject's face is not well illuminated, being the value of the pixels for every channel close to 0. This will generate dark ROI videos that might act as disturbance to the network in the learning process. But as discussed in [6], a varied training data, that is representative of realistic conditions, enables deep learning models to extract information that is independent of the acquisition scenario. We take advantage of the self-leaning characteristic of neural networks to face the challenges presented in COHFACE.

Different metrics have been used in the literature for reporting the HR estimation performance of an approach. Evaluating a deep learning algorithm with different evaluation metrics is an essential part of its validation because it gives an overall assessment of a model's performance. We present the mean and standard deviation (Me, STDe) in bpm of the HR error, the mean absolute HR error (MAE) in bpm, the root mean squared HR error (RMSE) in bpm, the mean of error rate percentage (MER) in bpm, and Pearson's correlation coefficients ($\rho$).

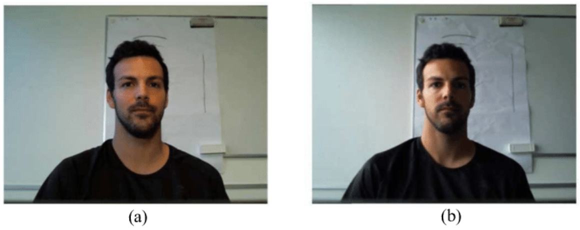

**Fig. 3.** Example video frames of two videos from a subject of the COHFACE dataset. Frame (a) shows the subject's face illuminated with studio light and frame (b) with daylight coming form a left source.

### 4.2 HR estimation results

In this study, we will compare our method with other deep learning based methods for direct HR estimation. Similar to [29], we performed a subject-exclusive split of the videos for training and testing subsets. The training set is composed of 24 subjects and a testing set of the remaining 12.

The results in Table 2 show that the proposed 3D CNN structure presents a competitive performance achieving the lowest STD in the COHFACE dataset. It achieves higher performance compared to [29] confirming that a sequential processing of the spatial and afterwards the temporal information of a video as proposed in [29] cannot capture the HR information as well as our network, which processes both spatial and temporal information simultaneously. It also overperforms, for every metric reported [37] whose double stream cannot beat the power of simultaneous 3D convolutions among all input video patches. Furthermore, by using as input the cheek area of the video we acheive lower MAE and RMSE compared to the denoising patches obtained by the full face in [13]. We prove here that the selection of an optimal face region outperforms the denoising of the full face. Finally, our model surpasses for almost every metric the two networks proposed by [8], which aimed for a CNN-based feature maps extractor from full faces. In their work they implemented a rough version of 2D and 3D-CNN and then they improve both structures by adding aggregation functions. The performance of those networks is presented in Table 2. Their further promote improvements in the 2D model putting on the side the 3D model. The results highlight the optimal performance of our 3D-CNN indicating how an end-to-end 3D CNN can overperform a 2D structure in accurately estimating a subject's HR directly from the cheek area without the need of any other intermediate face representation. This provides a new way to capture the rPPG information without compromising the model accuracy.

In addition, the processing time of one 60s video with 20 fps for our 3D-CNN is of $0,1$ milisecond. The proposed 3D network does not require any extra pre or post processing step, making it highly efficient and convenient for online estimation.

**Table 2.** Comparison between our model and other neural network based approaches on COHFACE.

| **Method** | **STD** | **MAE** | **RMSE** | **MER** | $\rho$ |
|---|---|---|---|---|---|
| HR-CNN [29] | - | 8.10 | 10.78 | - | 0.29 |
| 2-STREAM CNN [37] | - | 8.09 | 9.96 | - | 0.40 |
| 3D-rPPGNet [8] | 8.98 | 5.86 | 9.12 | - | - |
| 2D-rPPGNet [8] | 8,08 | 5.59 | 8.12 | - | **0.63** |
| AND-rPPG [13] | 7.83 | 6.81 | 8.06 | - | **0.63** |
| Ours | **7.23** | **5.5** | **7.74** | **7.12** | 0.62 |

### 4.3 Effectiveness of input choice selection

We also perform a study of the effectiveness of different video input choices: ROI and input channel selection. As a baseline experiments, we trained and tested our 3D-CNN on full face 3 channeled videos. For the consequent experiment, each of the RGB channels was used to train the network on the full face videos, and the results are reported in Table 3. The experiments suggest that even though no clear choice between the use of RGB vs G as input can be made, the selection of just the R or B channels clearly decreases the network's performance. As a next step, we trained and tested the 3D-CNN by feeding it with a 40×40 and 30×80 ROI for the cheeks and forehead experiments respectively. Those areas detected and cropped using the 68 (x,y)-coordinates of the dlib landmark detector as explained on Section 3.1. The results presented in Table 3 indicate that taking as input a smaller and more specific area than the full face is particularly beneficial to perform an accurate HR estimation especially in the case of the cheeks region. The cheek area is less affected by nonrigid motion such as smiling or talking and can yield better results since in some cases the forehead can be occluded by hair or other monitoring devices. However they can be equally affected by difficult illumination conditions explaining why in both areas, the results are specially promising for the G channel, highlighting how in adversarial illumination conditions (e.g. natural light sources that do not distribute the light equally on the face skin areas) the proposed 3D-CNN predicts successfully the HR only passing the G channel.

## 5 CONCLUSION AND FUTURE WORKS

Remote HR estimation allows a pulse rate extraction from the skin regions in face videos without any type of physical contact with the subject. In this study, we presented a review on the most relevant SoA methods for a remote rPPG and HR estimation from RGB face videos. For the learning based approaches, we summarized some of the choices regarding the structure and the data that those models use and we extract some key experiments that, in our view, are lacking from the current literature. More specifically, we perform a comparison of the most common ROI for remote HR estimation obtaining best results using the cheek area and we evaluate the choice of the G channel as input against

**Table 3.** Evaluation of the proposed network using different input video channels and ROI on the COHFACE dataset.

| **Method** | **Me** | **STD** | **MAE** | **RMSE** | **MER** | $\rho$ |
|---|---|---|---|---|---|---|
| Full face RGB | 2.43 | 9.55 | 8.22 | 9.86 | 11.32 | 0.28 |
| Full face R | 4.05 | 11.08 | 9.82 | 11.80 | 12.87 | 0.11 |
| Full face G | 1.44 | 10.35 | 8.23 | 10.45 | 11.54 | 0.29 |
| Full face B | -0.29 | 10.47 | 8.95 | 10.47 | 12.79 | 0.23 |
| Forehead RGB | -1.22 | 10.61 | 8.35 | 10.68 | 12.13 | 0.42 |
| Forehead G | -3.17 | 8.80 | 7.84 | 9.35 | 11.71 | 0.52 |
| Cheeks RGB | **0.01** | 7.99 | 5.78 | 7.99 | 7.99 | 0.46 |
| Cheeks G | 2.75 | **7.23** | **5.5** | **7.74** | **7.12** | **0.62** |

using the three channels of RGB videos. We highlight how some other deep learning based models, require a pipeline of different techniques that can be costly in terms of memory and time, therefore, non suitable for real-time usage with affordable devices such as mobile phones. Our 3DCNN has a processing time of 0, 1 milisecond for a 60s video with 20 fps. We propose a competitive, fast and reproducible HR estimation method based on a 3D-CNN structure and we evaluate the network against similar deep learning state-of-the-art structures on the publicly available dataset COHFACE.

Future challenges include the evaluation of the G channel as input against using the three channels of RGB videos. A 3 stream 3D-CNN for the R, G and B channels will be further explored as well as an adaptive ROI selection for forehead skin areas that can be more equally illuminated in natural light conditions.

## References

1. Allen, J.: Photoplethysmography and its application in clinical physiological measurement. Physiological measurement **28**(3) (2007)
2. Blazek, V.: Ambient and unobtrusive cardiorespiratory monitoring. In: 2016 ELEKTRO. IEEE (2016)
3. Bousefsaf, F., Pruski, A., Maaoui, C.: 3d convolutional neural networks for remote pulse rate measurement and mapping from facial video. Applied Sciences **9**(20) (2019)
4. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
5. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60**(10), 2878–2886 (2013)
6. Hernandez-Ortega, J., Fierrez, J., Morales, A., Diaz, D.: A comparative evaluation of heart rate estimation methods using face videos. In: 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE (2020)
7. Heusch, G., Anjos, A., Marcel, S.: A reproducible study on remote heart rate measurement. arXiv preprint arXiv:1709.00962 (2017)

8. Hu, M., Qian, F., Wang, X., He, L., Guo, D., Ren, F.: Robust heart rate estimation with spatial-temporal attention network from facial videos. IEEE Transactions on Cognitive and Developmental Systems (2021)
9. Huang, B., Chang, C.M., Lin, C.L., Chen, W., Juang, C.F., Wu, X.: Visual heart rate estimation from facial video based on cnn. In: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE (2020)
10. Kwon, S., Kim, J., Lee, D., Park, K.: Roi analysis for remote photoplethysmography on facial video. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2015)
11. Lamba, P.S., Virmani, D.: Contactless heart rate estimation from face videos. Journal of Statistics and Management Systems **23**(7), 1275–1284 (2020)
12. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4264–4271 (2014)
13. Lokendra, B., Puneet, G.: And-rppg: A novel denoising-rppg network for improving remote heart rate estimation. Computers in biology and medicine p. 105146 (2021)
14. Mehta, A.D., Sharma, H.: Heart rate estimation from rgb facial videos using robust face demarcation and vmd. In: 2021 National Conference on Communications (NCC). pp. 1–6. IEEE (2021)
15. Monkaresi, H., Calvo, R.A., Yan, H.: A machine learning approach to improve contactless heart rate monitoring using a webcam. IEEE journal of biomedical and health informatics **18**(4), 1153–1160 (2013)
16. Niu, X., Han, H., Shan, S., Chen, X.: Continuous heart rate measurement from face: A robust rppg approach with distribution learning. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 642–650. IEEE (2017)
17. Niu, X., Han, H., Shan, S., Chen, X.: Synrhythm: Learning a deep heart rate estimator from general to specific. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE (2018)
18. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing **29** (2019)
19. Niu, X., Zhao, X., Han, H., Das, A., Dantcheva, A., Shan, S., Chen, X.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (2019)
20. Perepelkina, O., Artemyev, M., Churikova, M., Grinenko, M.: Hearttrack: Convolutional neural network for remote video-based heart rate monitoring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 288–289 (2020)
21. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE transactions on biomedical engineering **58**(1), 7–11 (2010)
22. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18**(10), 10762–10774 (2010)
23. Qiu, Y., Liu, Y., Arteaga-Falconi, J., Dong, H., El Saddik, A.: Evm-cnn: Real-time contactless heart rate estimation from facial video. IEEE transactions on multimedia **21**(7) (2018)
24. Rahman, H., Ahmed, M.U., Begum, S., Funk, P.: Real time heart rate monitoring from facial rgb color video using webcam. In: The 29th Annual Workshop of the

Swedish Artificial Intelligence Society (SAIS), 2–3 June 2016, Malmö, Sweden. No. 129, Linköping University Electronic Press (2016)

25. Siddiqui, H., Rattani, A., Kisku, D.R., Dean, T.: Ai-based bmi inference from facial images: An application to weight monitoring. preprint arXiv:2010.07442 (2020)
26. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. Affective Computing, IEEE Transactions on **3**(1) (2012)
27. Song, R., Li, J., Wang, M., Cheng, J., Li, C., Chen, X.: Remote photoplethysmography with an eemd-mcca method robust against spatially uneven illuminations. IEEE Sensors Journal **21**(12), 13484–13494 (2021)
28. Song, R., Zhang, S., Li, C., Zhang, Y., Cheng, J., Chen, X.: Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks. IEEE Transactions on Instrumentation and Measurement **69**(10) (2020)
29. Špetlík, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: Proceedings of the British Machine Vision Conference, Newcastle, UK (2018)
30. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: 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. pp. 2396–2404 (2016)
31. Van Kampen, E., Zijlstra, W.G.: Determination of hemoglobin and its derivatives. Advances in clinical chemistry **8** (1966)
32. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Optics express **16**(26) (2008)
33. Wang, G.: Influence of roi selection for remote photoplethysmography with singular spectrum analysis. In: 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). pp. 416–420. IEEE (2021)
34. Wang, W., den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering **64**(7) (2016)
35. Wang, X., Xie, W., Song, J.: Learning spatiotemporal features with 3dcnn and convgru for video anomaly detection. In: 2018 14th IEEE International Conference on Signal Processing (ICSP). pp. 474–479. IEEE (2018)
36. Wang, Y., Dantcheva, A.: A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 515–519. IEEE (2020)
37. Wang, Z.K., Kao, Y., Hsu, C.T.: Vision-based heart rate estimation via a two-stream cnn. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 3327–3331. IEEE (2019)
38. Wang, Z., Yang, X., Cheng, K.T.: Accurate face alignment and adaptive patch selection for heart rate estimation from videos under realistic scenarios. PloS one **13**(5), e0197275 (2018)
39. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. arXiv preprint arXiv:1905.02419 (2019)
40. Yu, Z., Peng, W., Li, X., Hong, X., Zhao, G.: Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: Proceedings of the Int. Conference on Computer Vision (2019)