Computers in Biology and Medicine 154 (2023) 106592

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

X-iPPGNet: A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation

Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf *, Choubeila Maaoui

Université de Lorraine, LCOMS, F-57000 Metz, France

ARTICLE INFO

Keywords:
Pulse rate estimation
Convolutional neural networks
End-to-end learning
Imaging photoplethysmography
Xception network

ABSTRACT

Pulse rate (PR) is one of the most important markers for assessing a person's health. With the increasing demand for long-term health monitoring, much attention is being paid to contactless PR estimation using imaging photoplethysmography (iPPG). This non-invasive technique is based on the analysis of subtle changes in skin color. Despite efforts to improve iPPG, the existing algorithms are vulnerable to less-constrained scenarios (i.e., head movements, facial expressions, and environmental conditions). In this article, we propose a novel end-to-end spatio-temporal network, namely **X-iPPGNet**, for instantaneous PR estimation directly from facial video recordings. Unlike most existing systems, our model learns the iPPG concept from scratch without incorporating any prior knowledge or going through the extraction of blood volume pulse signals. Inspired by the Xception network architecture, color channel decoupling is used to learn additional photoplethysmographic information and to effectively reduce the computational cost and memory requirements. Moreover, X-iPPGNet predicts the pulse rate from a short time window (2 s), which has advantages with high and sharply fluctuating pulse rates. The experimental results revealed high performance under all conditions including head motions, facial expressions, and skin tone. Our approach significantly outperforms all current state-of-the-art methods on three benchmark datasets: MMSE-HR ($MAE = 4.10$; $RMSE = 5.32$; $r = 0.85$), UBFC-rPPG ($MAE = 4.99$; $RMSE = 6.26$; $r = 0.67$), MAHNOB-HCI ($MAE = 3.17$; $RMSE = 3.93$; $r = 0.88$).

1. Introduction

Pulse rate (PR) is one of the important indicators of a person's health that needs to be monitored routinely to identify a range of health issues. Electrocardiography and Photoplethysmography (PPG) are the main ways of measuring heart rate activity. Both techniques use contact sensors that need to be attached to body parts. Despite the high accuracy and robustness provided by these devices, specific conditions are required to acquire accurate measurements. Moreover, contact with skin can be inconvenient or even infeasible in some critical cases such as burns, skin ulcers, or contagious diseases [1]. These constraints limit their use in realistic scenarios. Over the last decade, great progress has been made in non-contact pulse rate estimation using imaging photoplethysmography, due to its wide application domains [2–8]. iPPG is an optical technique allowing a remote assessment of the pulse rate by observing the blood-volume variations on a person's face using a simple camera.

Conventional iPPG algorithms are based on hand-crafted features approaches, which generally involve multi-stage pipelines and require multiple image and signal processing steps [2–6,9]. Most of these methods have been carried out under constrained environments and rely on certain assumptions regarding light-skin interaction and head motions. Therefore, they perform reasonably well under controlled conditions. However, their performance degrades significantly under challenging scenarios such as large head movement, poor lighting conditions, and very dark skin [8,10].

Inspired by the recent breakthroughs in computer vision tasks [11–14], current state-of-the-art algorithms incorporate deep learning architectures in different stages of the conventional imaging photoplethysmography pipeline. Deep neural networks have been used to accurately extract the iPPG signal [7,8,15,16]. However, several limitations remain to be resolved. These systems are not end-to-end, so they still require pre-processing or post-processing steps as well as a larger time-span window to estimate pulse rate. Furthermore, heart rate activity should be measured even in unconstrained scenarios. Many factors can affect the measurement: the person may move his head or express emotions, his face can be partially occluded or light conditions may be changing continuously. These situations affect the quality of the extracted iPPG signal, thus degrading the accuracy of the predicted PR values.

* Corresponding author.
E-mail address: frederic.bousefsaf@univ-lorraine.fr (F. Bousefsaf).

https://doi.org/10.1016/j.compbiomed.2023.106592
Received 22 June 2022; Received in revised form 7 December 2022; Accepted 22 January 2023
Available online 24 January 2023

To address these drawbacks, we developed an end-to-end deep learning model (X-iPPGNet) for instantaneous pulse rate estimation directly from raw facial videos. The architecture is fully automatic and does not require any prior knowledge or special pre-processing or post-processing. This work is an extension and improvement of the method proposed as part of the Vision for Vitals Challenge [17]. We propose a new efficient architecture and evaluate its effectiveness using public databases. We also examine the impact of challenging conditions on performance.

The main contributions of this study are summarized as follows:

1. We propose a novel one-stage approach based on an end-to-end trainable neural network.
2. X-iPPGNet predicts pulse rate from short-time video excerpts (2 s), which is particularly relevant in the case of high and sharply fluctuating pulse rates.
3. Color channels decoupling is used to extract additional photoplethysmographic information.
4. The first use of the BP4D+ database in conjunction with data augmentation.
5. Extensive evaluations on multiple public databases to analyze the effectiveness and generalizability of the proposed method against a range of challenging factors.

The remainder of the article is organized as follows: related works are briefly exposed in Section 2. Section 3 presents the materials and methods. Experimental results are presented and discussed in Sections 4 and 5 respectively. Finally, conclusions and future works are given in Section 6.

2. Related works

By surveying existing research on contactless pulse rate estimation using iPPG, we can identify the existence of two major approaches according to the way of iPPG signal extraction, either manually using conventional methods [2–6,9], or automatically using deep learning models [7,8,15,16]. Earlier works on iPPG relied on hand-crafted features approaches that generally include image and signal processing operations. The image processing techniques are first applied to locate the skin regions containing relevant information about the subtle color changes associated with blood flow. Different color spaces and different regions of interest (ROI) were exploited to constitute raw iPPG signals using a spatial averaging operation. Verkruysse et al. [9] have initially computed raw iPPG signals from the green channel using a set of predefined ROI. Several face detectors and trackers have been used to extract the entire face or sub-regions from the face such as the forehead or cheeks [18–22]. Bousefsaf et al. [23] proposed to select only the pixels of interest using a custom skin segmentation, while Tulyakov et al. [24] developed an approach to choose dynamically the ROI using self-adaptive matrix completion. Furthermore, different color spaces have been studied besides the standard RGB. For example, the u* component from the CIE L*u*v* color space [23] and V from YUV have been exploited [25].

In the second step, signal processing algorithms are performed to increase the signal-to-noise ratio and remove the noise from iPPG signal. Some of the popular studies include blind source separation methods, such as independent component analysis [18] and principal components analysis [19]. On the other hand, Haan and his group achieved further improvements by proposing model-based approaches [3–5]. They developed different color subspace transformations to overcome motion artifacts and improve the quality of iPPG signal.

With the great success of deep learning and more specifically convolutional neural networks for medical imaging and computer vision tasks [12,26,27], several groups developed deep learning-based methods for iPPG estimation. According to the recent review of Ni et al. [28], existing methods are built using VGG-style CNN [7,29,30], or combine CNN and LSTM to take into account the temporal information [8,31,32], or use 3D-CNN directly to simultaneously learn spatial and temporal features [15,33–36]. To name some of the promising works, Chen and McDuff [7] proposed a convolutional attention network named DeepPhys, which consists of two-stream CNN to extract blood volume pulse waveform from facial video under varying lighting and significant head motions. They used an appearance model based on an attention mechanism to find the appropriate regions of interest (ROI) and to guide the motion representation model. Radim et al. [29] 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 pulse rate values. Niu et al. [8] generated spatial–temporal maps from multiple ROI over the face and then trained a CNN-RNN network to regress the average PR value. Yu et al. [15] introduced a spatial–temporal deep neural network (PhysNet) to extract iPPG signals from raw facial videos, and then measure the averaged PR and HRV features. AutoHR is a recent contribution proposed by Yu et al. [16]. The authors used temporal difference convolution beside a strong backbone discovered via neural architecture search to estimate accurately the iPPG signal from image sequences.

All the methods mentioned above are based on several processing stages. They mainly use deep learning to recover iPPG signals from facial videos. However, some works have adopted deep neural networks to pulse rate estimation in an end-to-end manner without passing by iPPG signal extraction. Bousefsaf et al. [34] were the first to demonstrate the possibility of pulse rate estimation from a face video without any additional processing. They put forward a 3D CNN trained purely on synthetic data. Huang et al. [32] developed a one-stage spatio-temporal network that combines 3D convolutional and LSTM modules to extract spatial and temporal features and a Dense layer to pulse rate value estimation. Ouzar et al. [37] proposed an efficient model built on a linear stack of depthwise separable convolution layers concatenated with residual connections. This method has advantages in terms of speed and simplicity and can run in real-time both on CPUs and GPUs. Existing iPPG-based PR measurement approaches are summarized in Table 1.

3. Materials and methods

3.1. Datasets

The availability of huge databases and advanced neural architectures have underpinned the great success of deep learning approaches in computer vision tasks. In the field of remote PR estimation, the lack of large-scale heart rate (HR) datasets has limited the use of deep learning models [8]. Existing public domain HR databases are quite limited not only in data size but also in diversity. Head motion, facial expressions, occlusion, and skin tone correspond to the main challenging conditions that affect the performance of contactless pulse rate measurement from facial videos. However, previous works had not addressed all of these problems due to the quality and scale of the aforementioned databases.

For this study, we used four public datasets for pulse rate estimation to evaluate the performance of the proposed method. We trained X-iPPGNet on BP4D+ [38], a public large-scale database, while MAHNOB-HCI [39], UBFC-rPPG [40], and MMSE-HR [38] were used for testing. We briefly describe each of these three datasets in the subsequent paragraphs while we present in detail the BP4D+ database as we are the first to use it for training deep neural networks. Table 2 gives detailed comparisons between the different databases used in our experiments.

3.1.1. MMSE-HR

MMSE-HR [38] was collected for contactless pulse rate estimation under challenging conditions. It consists of 102 RGB facial videos recorded at 25 frames-per-second (fps) from 40 subjects (17 males and 23 females) with various ethnic/racial ancestries. The corresponding average pulse rates were gathered using a contact BVP sensor (sampling frequency: 1K HZ).

Table 1
A brief summary of existing iPPG-based PR estimation approaches and their pros & cons.

	Multiple stage		One stage
	Conventional	Deep learning	
Input	Thermal [41] Monochromatic [43]	Thermal [42]	R.G.B [32]
	R.G.B [24,44] Five band [45]	R.G.B [7,16]	Synthetic Data [34]
Preprocessing	Face ROI detection & tracking [18,24]	Face ROI detection & tracking [8]	Face ROI detection & tracking [32]
	Color space transformation [23,25]	Spatial-temporal maps [8]	
	Signal decomposition [2]	Video magnification [46]	
Postprocessing	Filtering [3,18,19] FFT [2,48] Peaks detection [18,45]	FFT [47] Peaks detection [15] Deep learning model [8,29]	–
iPPG signal extraction	Spatial average [1,6,18]	Deep regression model [7,15]	–
Pros	Allows pulse wave features extraction	Good generezability	Good generezability
		Allows pulse wave features extraction	Easy to deploy Short time span window
Cons	Hard to deploy	Hard to deploy	Does not allow pulse wave features extraction
	Require pre-processing or post-processing steps	Require pre-processing or post-processing steps	
	Large time span window Poor generezability	Large time span window	

Table 2
Summary of the public-domain databases used in our experiments.

Database	Nb of participants	Nb videos	FPS	Ethnicity	Task/Condition
MMSE-HR [38]	40	102	25	Latino/Hispanic, White, African American, Asian, and Others	Emotion elicitation
MAHNOB-HCI [39]	27	527	61	Caucasian and Asian	Emotion elicitation
UBFC-rPPG [40]	42	42	30	–	Interaction
BP4D+ [38]	140	1400	25	Latino/Hispanic, White, African American, Asian, and Others	Emotion elicitation

3.1.2. MAHNOB-HCI

MAHNOB-HCI [39] is a commonly used benchmark to assess the effectiveness and generalizability of non-contact pulse rate estimation methods. It includes 527 videos from 27 subjects (12 males and 15 females) along with their corresponding physiological signals. All videos are recorded at 61 fps with a resolution of 780 × 580 pixels. ECG signal has been used to calculate the ground truth pulse rate values.

3.1.3. UBFC-rPPG

UBFC-rPPG [40] consists of 42 videos from 42 subjects. The videos were recorded using a low-cost webcam at 30 fps and a resolution of 640 × 480 pixels. The duration of each recording varies between 50 and 90 s. A Contec Medical CMS50E finger pulse oximeter is synchronized with the video recordings to establish the ground truth PPG signal.

3.1.4. BP4D+

BP4D+ [38] is a large-scale public database mainly dedicated to multimodal spontaneous emotion recognition based on facial expressions and physiological parameters. It includes several physiological signals such as heart rate, respiratory rate, and blood pressure. Compared to existing pulse rate databases, BP4D+ is significantly larger in terms of data amount and ethnic diversity (including Black, White, Asian, and Hispanic/Latino). Additionally, it was collected under challenging scenarios such as significant head motions, wild pulse rate range, facial expressions, and occlusions. 140 subjects (82 females and 58 males) participated in ten sessions set up to elicit different emotions. 1400 RGB videos lasting 30 s to 1 min were recorded at 25 fps. The resolution of each video is 1040 × 1392 pixels. Pulse rate and other physiological signals were collected with contact sensors at 1K Hz. Fig. 2 shows the histogram of ground truth pulse rate distribution in BP4D+. Pulse rate values vary from 47 to 139 beats per minute (bpm), which almost covers the typical pulse rate range. The histogram forms an inverse Gaussian distribution because most healthy and relaxed adults have a resting heart rate comprised between 70 and 90 beats per minute (see Fig. 2). On the other hand, due to a large amount of corrupted ground truth signals (see a typical example in Fig. 3), we recalculated the pulse rates from the blood pressure signals available in the database. We also removed segments where facial regions are outside the image.

3.2. Proposed framework

The general framework for pulse rate estimation from facial videos is illustrated in Fig. 1. We treat this task as a one-stage regression problem that takes batches of 50 frames (corresponding to 2 s) as input and regresses the pulse rate value as output. First, face segmentation is performed to eliminate the background and non-skin areas [49]. Then the face region is cropped from the segmented face image according to the coordinates of the first non-zero pixel on each side of the image. Finally, the face image sequences are scaled and fed to a 3D fully convolutional neural network. We assume that the proposed architecture can automatically focus on the most vascularized areas of the face. It then learns the spatio-temporal features associated with iPPG.

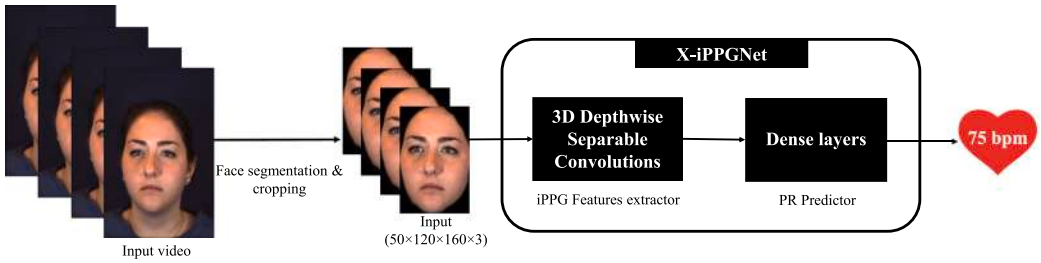


Fig. 1. Overview of the proposed framework for visual pulse rate estimation. Face segmentation and cropping are performed first on the input video to get rid of non-skin areas. Then the facial image sequences are fed to a deep neural network (X-iPPGNet) consisting of 3D Depthwise Separable Convolutions for spatial and temporal features extraction, and Dense layers for pulse rate prediction.

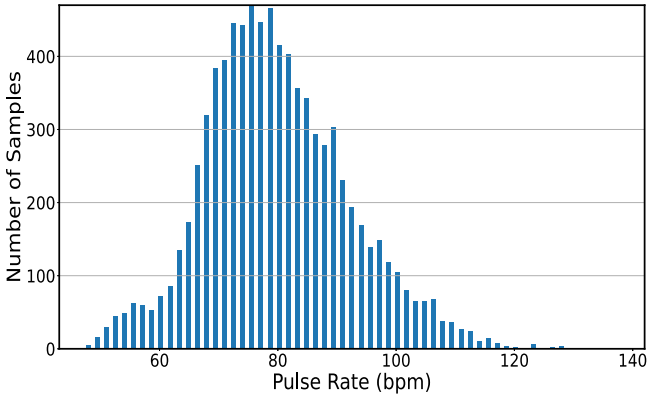


Fig. 2. Distribution of the ground truth pulse rates in BP4D+.

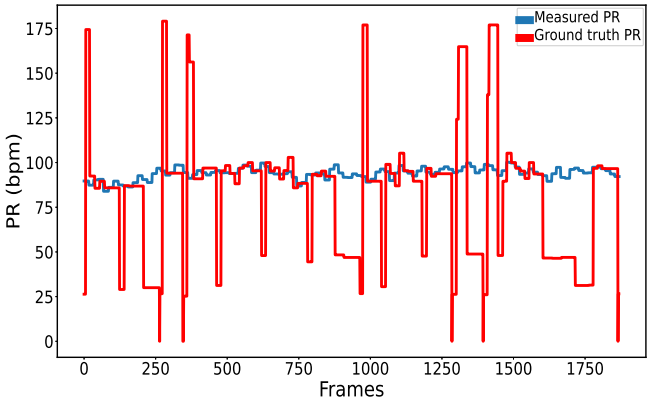


Fig. 3. Example of ground truth pulse rates (participant F005) showing strong inconsistencies. Red curve: ground truth pulse rate provided by the database; Blue curve: pulse rate computed from the raw blood pressure signal.

3.2.1. Face segmentation

The extraction of regions of interest (ROI) is the first step of almost all video-based pulse rate estimation [8,15,18,32,50]. It aims to maximize the signal-to-noise ratio by only keeping the skin pixels that carry the iPPG information. Several face and facial landmarks detectors have been employed to locate ROI. However, these techniques often fail in situations involving head movement, occlusion, or facial expressions. Many other factors can also affect ROI extraction, such as lighting and background. We compared the performance of the three most popular face detectors used for iPPG extraction in terms of efficiency, i.e., Viola&Jones [51], Dlib [52], and MTCNN [53]. Table 3 illustrates the number of missed images on the MMSE-HR dataset [38] presented in Section 3.1. MMSE-HR has been widely used as a test set in several works and contains about 108117 images. The results show

Table 3
Number of missed images according to the most popular face detection algorithms.

Face detector	Number of missed frames
Viola–Jones [51]	1375
Dlib [52]	227
MTCNN [53]	48
Face segmentation [49]	**0**

that the three face detectors mentioned above fail to perform well in unconstrained scenes.

To overcome the limitations of face detectors, especially in unconstrained scenarios, we performed face segmentation using one of the state-of-the-art algorithms [49] (see Table 3). This method, originally proposed for face-swapping ideally works in all conditions without missing any frames. Faces are properly segmented from backgrounds and occlusions with high accuracy. Some processed images extracted from the MMSE-HR database are shown in Fig. 4.

3.2.2. Pulse rate estimation neural network

Most of the existing video-based PR estimation approaches that integrate a deep learning model rely on a VGG-style CNN. Temporal information is processed using recurrent networks [8,32], spatio-temporal convolutions [15,34], or by incorporating another temporal branch in parallel [7]. The VGG-style CNN is a basic architecture that uses a standard convolution stack with no residual blocks [54]. Despite its simplicity, it is more prone to overfitting. It also performs worse than other deep learning architectures on many computer vision tasks [55]. In addition, standard convolution considers all spatial and color channel information together. However, previous studies showed that color channels have different physiological properties and that pulsatile activity varies from one color to another [56]. Although the green channel featuring the strongest plethysmographic signal and carries more PPG information compared to the other channels, the red and blue channels also contained useful and complementary plethysmographic information that should not be neglected [18]. Nevertheless, and to the best of our knowledge, all deep learning-based approaches have combined RGB channels. This can lead to loss of useful features across channels, affecting measurement accuracy.

In this study, we designed an end-to-end deep regression framework based on a modified Xception network [57]. This architecture outperforms other deep learning models in several computer vision tasks [55,58]. Furthermore, it relies on depthwise separable convolution instead of standard convolution operations that require larger amounts of memory and computational cost. A depthwise separable convolution extension for 3D volumes is used[1] to learn the relevant features associated with the cardiac rhythm of each color channel separately.

The idea behind the depthwise separable convolution is that the depth and spatial dimension of a filter can be decoupled within a

[1] https://github.com/alexandrosstergiou/keras-DepthwiseConv3D.

Fig. 4. Examples showing the ability of the face segmentation model to work in difficult scenarios. Top figures: raw images, bottom figures: corresponding segmentations.

convolutional layer. First, the video embedding dimensions are separated and an independent spatio-temporal convolution is performed for each color channel. This operation is called depthwise convolution. It aims to extract local features from each color channel of the input image sequences separately and to capture the temporal relationships among the spatial feature sequences. Then, a pointwise convolution is performed on the convoluted tensor to merge the feature maps across channels in the embedding dimension. This effectively reduces computational costs and memory requirements.

Fig. 5 presents the overall architecture of the proposed X-iPPGNet, which consists of three blocks (entry, middle, and exit). It includes 36 convolutional layers structured in 14 modules, all linked with shortcuts as in the ResNet architecture, except for the first and last modules. Since the network is very deep, these residual connections allow reducing the impact of gradient vanishing. Each convolutional layer is followed by a batch-normalization to stabilize the training process and accelerate the convergence. ReLU activation functions are also used to perform nonlinear mapping. The features extraction output is flattened and fed into two dense layers of 1024 and 1 neurons, respectively, to estimate the pulse rate value.

In summary, the proposed non-contact pulse rate estimation framework is a one-stage pipeline that predicts the average pulse rate in only 2 s video fragments. The input is represented as a 5-dimensional tensor ($Nbacth \times Nbframes \times ImHeight \times ImWeight \times Channel$) (where $Nbatch$ is the batch-size; $Nbframes$ is the length of face video clip; $ImHeight$, $ImWeight$, and $Channel$ are the size of each frame) and the output is the estimated pulse rate in beats per minute.

We consider pulse rate prediction as a one-step regression problem. Training is fully supervised where each 2-seconds video fragment takes a ground truth pulse rate obtained with a contact device as a training label. In the training phase, the network learns to associate the ground truth pulse rate value with each facial video sequence by constructing a mapping relationship between inputs and outputs, i.e., mapping of a three-dimensional tensor (video data) to a single scalar (pulse rate). After the training phase, the network would be able to estimate pulse rate within the trained pulse rate range.

3.2.3. Implementation details

3.2.3.1. Training.

The proposed architecture is implemented with Keras and Tensorflow frameworks and trained with two Nvidia Quadro P6000s. The videos have been cut into sequences of 50 frames (corresponding to 2 s). The size of each frame is $160 \times 120 \times 3$ ($ImHeight \times ImWeight \times Channel$). The total number of sequences is 39762. Inspired by the SWATS optimization procedure [59], we started training with a Rectified Adam (RAdam) optimizer [60] before switching to Stochastic Gradient Descent (SGD) [61] when the validation accuracy stops improving. The learning rate was initially set to 10^{-4}, and then decreased to 10^{-6}. We train the network for about 25 epochs with a batch size of 64 ($Nbacth = 64$) and using the mean-squared-error loss function. In addition, a dropout technique [62] is applied before the final dense layer of the network (the dropout rate is set to 0.4). L1 and L2 regularization strategies are employed as well, which help to overcome overfitting issues and improve the model generalizability to new data.

3.2.3.2. Training set augmentation.

A common problem with limited and imbalanced datasets when training a neural network is overfitting and poor predictive performance, specifically for minority label samples.

X-iPPGNet was first trained without data augmentation. However, several problems that hinder the accuracy of pulse rate predictions have caught our attention. They are mainly caused by the highly imbalanced pulse rate samples in the BP4D+ database and also by the subjects skin tone [38]. Therefore, high and low pulse rate values and the skin color type with fewer samples are more difficult to predict. It is very challenging for a deep model to learn relevant features on poorly represented data. Neural networks tend to focus on targets with large numbers of samples. To address this issue, a data augmentation technique was applied to increase the size of the training set. Since more samples are available in the mid-pulse rates range (70, 90) bpm and less outside this range (see Fig. 2), we performed threefold offline data augmentation on the video sequences associated with pulse rate values greater than 90 bpm or lower than 70 bpm.

Following the same strategy presented in [63], we performed standard geometric augmentation and video magnification to increase the training set size and improve the robustness of the model. The geometric augmentation involves image transformations such as random clockwise and counterclockwise rotations by up to 20 degrees, scaling (in and out) of up to 20%, and horizontal and vertical video image shifting by 10% of the frame's width and height. The Eulerian video magnification (EVM) technique [64] was used to amplify the subtle colorimetric fluctuations due to iPPG in the videos. The intensity of these fluctuations can be weak for pixels that cover dark skin. The EVM method has been proven effective for PR estimation [46,64,65]. This technique takes a cropped ROI video sequence as input and applies spatial decomposition followed by temporal filtering to the frames. Laplacian pyramid is used for spatial decomposition, while temporal filtering is performed by applying the Fourier transform for each pixel. The amplification factor is fixed to 60 while Frequencies outside the

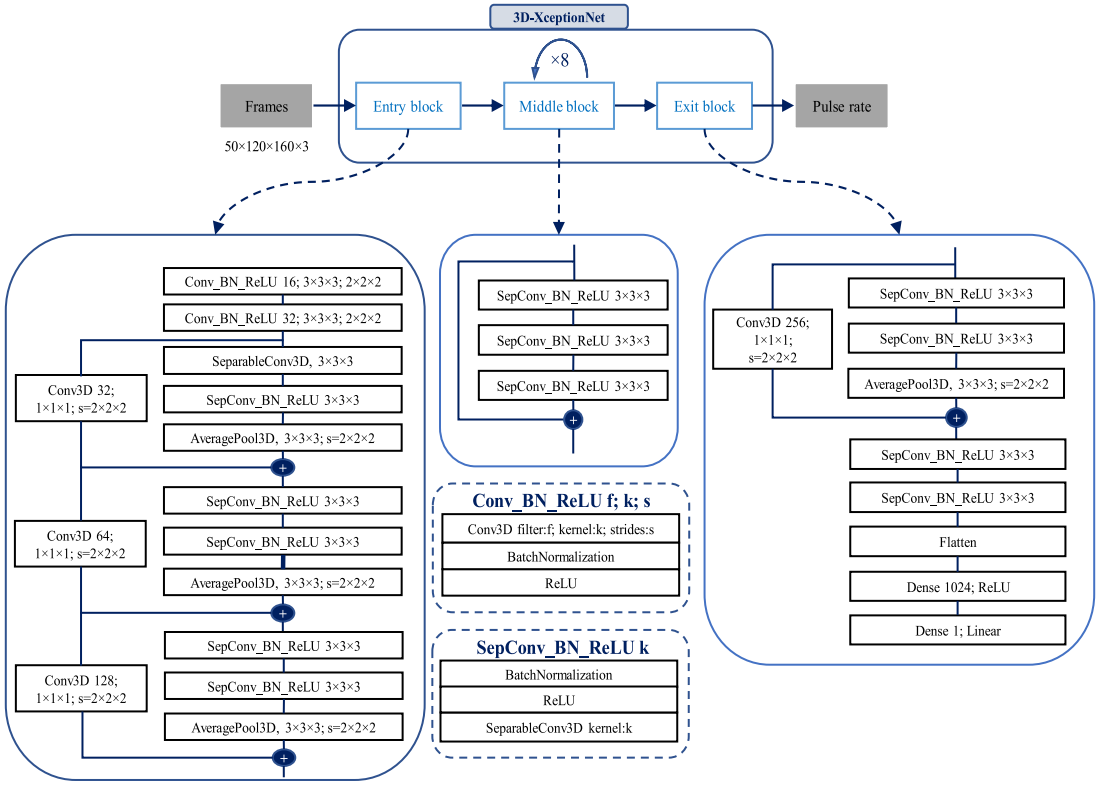


Fig. 5. X-iPPGNet architecture proposed in this work. It corresponds to a modified version of the Xception network. 2D depthwise separable convolution layers are replaced by 3D depthwise separable convolution to capture both spatial and temporal features across video frames. A Dense layer is used instead of a Global Average Pooling layer. The input video fragment first passes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow which ends with a dense layer of 1 neuron, to estimate the corresponding pulse rate.

cutoff (45–240 bpm) are set to zero. Finally, the inverse Fourier transform is applied to reconstruct the frames. The resulting video is then amplified and reveals hidden subtle changes in the skin color instigated by blood flow in facial vessels.

4. Experiments

We aim to achieve several goals in the conducted experiments. First, we prove the possibility of measuring pulse rate with high accuracy without going through the commonly used iPPG signal extraction step. Secondly, we provide a performance comparison with various developed baseline systems as well as other deep learning approaches recently proposed for contactless pulse rate estimation using iPPG. Thirdly, we demonstrate the generalization ability of our method under challenging conditions to illustrate the proposed framework's efficiency.

In order to study the generalizability and the effectiveness of the proposed X-iPPGNet presented in Section 3.2, three widely used public-domain databases are employed namely MMSE-HR [38], MAHNOB-HCI [39], and UBFC-rPPG [40]. MMSE-HR is directly used for testing without any additional processing since it was collected under the same conditions as BP4D+ (the training dataset). UBFC-rPPG and MAHNOB-HCI are downsampled from 30 fps and 61 fps to 25 fps in order to harmonize the fps of training and testing videos. For each experiment, we do not use videos of the same subject in both training and testing. We evaluate and compare the performance with other state-of-the-art techniques using different metrics: the standard deviation (SD), the mean absolute error (MAE, see Eq. (1)), the root mean square error (RMSE, see Eq. (2)), and the Pearson's correlation coefficient (r, see Eq. (3)). PR_i and $\widehat{PR}_i$ represent the ground truth and estimated pulse rate, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|PR_i - \widehat{PR}_i| \tag{1}$$

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

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

4.1. Results

4.1.1. Evaluation on MMSE-HR

We first evaluate the generalization ability of X-iPPGNet by training the network on BP4D+ and testing it on MMSE-HR (see Section 3.1).

Table 4 gives detailed comparisons with several state-of-the-art approaches including hand-crafted methods (Li2014 [66], CHROM [48], SAMC [24]) and deep learning-based methods (EVM-CNN [46], PhysNet [15], RhythmNet [8] and Auto-HR [16]). The X-iPPGNet proposed in this study achieves the best performance (SD = 5.34 bpm; MAE = 4.10 bpm; RMSE = 5.32 bpm and r = 0.85), outperforming all competing methods. Comparison with the other state-of-the-art methods are taken from [16].

4.1.2. Evaluation on UBFC-rPPG

In this experiment, we followed the same strategy presented in [32]. 25 videos were randomly selected to fine-tune the model pre-trained on BP4D+. The remaining videos were reserved for testing. Since the UBFC-rPPG dataset contains very limited facial videos (only one video is recorded for each subject), we used a three-fold subject-independent cross-validation strategy. Performance comparison results with other state-of-the-art techniques are taken from [67] and presented in Table 5. The proposed X-iPPGNet achieves good results and generalizes well in unseen domains. It should be noted that we achieved the best SD (6.25 bpm) and RMSE (6.26 bpm) among the existing methods.

Table 4
PR estimation results by the proposed approach and several state-of-the-art methods on MMSE-HR.

Approach	Method	SD (bpm)	RMSE (bpm)	r
Multiple stage hand-crafted	Li2014	20.02	19.95	0.37
	CHROM	14.08	13.97	0.55
	SAMC	12.24	11.37	0.71
Multiple stage deep learning	RhthmNet	6.98	12.76	0.78
	PhysNet	12.76	13.25	0.44
	AutoHR	5.71	5.87	**0.89**
One stage	**X-iPPGNet (Ours)**	**5.34**	**5.32**	0.85

Table 5
PR estimation results by the proposed approach and several state-of-the-art methods on UBFC-RPPG.

Approach	Method	SD (bpm)	MAE (bpm)	RMSE (bpm)	r
Multiple stage hand-crafted	Green	20.2	10.2	20.6	–
	ICA	18.6	8.43	18.8	–
	CHROM	19.1	10.6	20.3	–
	POS	10.4	**4.12**	10.5	–
Multiple stage deep learning	Meta-rPPG	7.12	5.97	7.42	0.53
One stage	3DCNN	8.55	5.45	8.64	–
	PRNet	6.45	5.29	7.24	–
	X-iPPGNet (Ours)	**6.25**	4.99	**6.26**	**0.67**

Table 6
PR estimation results by the proposed approach and several state-of-the-art methods on MAHNOB-HCI.

Approach	Method	SD (bpm)	MAE (bpm)	RMSE (bpm)	r
Multiple stage hand-crafted	Poh 2011	13.5	–	13.6	0.36
	CHROM	–	13.49	22.36	0.21
	Li 2014	6.88	–	7.62	0.81
	SAMC	5.81	4.96	6.23	0.83
Multiple stage deep learning	SynRhythm	10.88	–	11.08	–
	DeepPhys	–	4.57	–	–
	HR-CNN	–	7.25	9.24	0.51
	rPPGNet	7.82	5.51	7.82	0.78
	RhythmNet	3.99	–	3.99	0.87
	PhysNet	7.84	5.96	7.88	0.76
	AutoHR	4.73	3.78	5.10	0.86
	PulseGAN	–	4.15	6.53	0.71
One stage	**X-iPPGNet (Ours)**	**3.93**	**3.17**	**3.93**	**0.88**

4.1.3. Evaluation on MAHNOB-HCI

We further verify the efficiency and generalizability of X-iPPGNet on MAHNOB-HCI [39], which is the most commonly used dataset for non-contact PR estimation. The high compression rate and spontaneous movements caused by emotional stimulation make PR estimation challenging. We used the same three-fold subject-independent cross-validation protocol as for UBFC-rPPG (see Section 4.1.2). We randomized 66% of the videos to fine-tune the model pre-trained on BP4D+ and used the remaining videos for testing. Table 6 compares the performance of X-iPPGNet with state-of-the-art techniques, including hand-crafted and deep learning-based methods. From the results, we can observe that the X-iPPGNet ranks first on all metrics (SD = 3.93; MAE = 3.17; RMSE = 3.93 and r = 0.88). It is clear that our model performs very well under various image acquisition conditions and highly compressed videos.

4.2. Key components analysis

We also provide additional analysis to examine the impact of challenging factors, i.e., pulse rate distribution values, skin tone, gender, and head movements. All experiments have been conducted on the MMSE-HR dataset.

4.2.1. Impact of pulse rate distribution values

To further analyze the impact of PR distribution values on the performance of X-iPPGNet, we plot the differences between estimated and ground-truth pulse rate versus ground-truth estimation. This Bland–Altman plot (see Fig. 6) shows that the distribution is concentrated inside the 95% limits of agreement (1.96 SD) for low (<70) and mid (70, 90) pulse rates range. However, predictions of high pulse rates exhibit some outliers (>90). We suppose that this observation is connected to the imbalanced training set (see Fig. 2). Furthermore, the error rate increases significantly for higher pulse rates than for mid and low pulse rates due to their fluctuations over the time window [32].

Moreover, the Bland–Altman exhibits a marked negative trend. The model tends to over-estimate low PR and under-estimate high PR because low and high pulse rates are under-represented in the training dataset. We suppose that this observation is a direct consequence of the dataset imbalance. The model tends to produce predictions oriented towards mid-PR values. The PR difference is therefore positive for low PR and negative for high PR.

4.2.2. Impact of skin tone and gender

MMSE-HR was selected to assess the generalizability of our method to different skin tones. This dataset is more diverse in terms of ethnicity (including black, white, Asian, and Hispanic/Latino) compared to UBFC-rPPG [40] and MAHNOB-HCI [39], which are highly biased towards lighter skin. Following the protocol employed by the authors of [68], which is based on the Fitzpatrick scale [69], we divided the database into 4 categories according to skin tone type. In addition to types III and IV, we grouped skin types I + II and V + VI together as there were relatively few subjects in these categories. The predictions of X-iPPGNet for different skin tones are reported in Table 7. The proposed technique exhibits great performance for all skin types and relatively less for dark skin, considering that participants with darker skin tones are underrepresented in the training set.

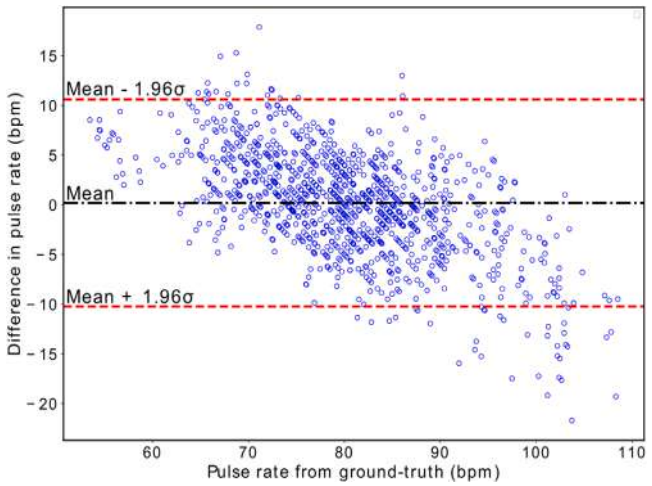
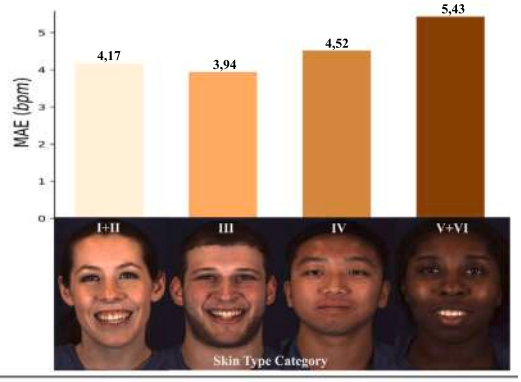


Fig. 6. Bland–Altman plot showing the differences in pulse rate between ground-truth and estimated values plotted against the ground-truth measurements for the MMSE-HR dataset (see Section 3.1). Mean values are represented by black dash–dot lines and 95% limits of agreement (1.96 SD) by red dashed lines.

Table 7
PR MAE, RMSE and r for our method by skin type on MMSE-HR.



Fitzpatrick Skin Types	I+II	III	IV	V+VI
MAE (bpm)	4.17	3.94	4.52	5.43
RMSE (bpm)	5.31	5.18	5.76	6.82
r	0.87	0.81	0.84	0.40

Table 8
Performance of our method on MMSE-HR by gender.

Gender	Male	Female
MAE (bpm)	3.74	4.53
RMSE (bpm)	4.76	5.84
r	0.79	0.85

We further evaluated the impact of gender on pulse rate estimation. The results obtained show differences in performance between males and females (see Table 8). This confirms the results of previous study showing a slightly lower error rate for males than for females [8].

4.2.3. Impact of head movement

Visual pulse rate estimation in unconstrained environments remains a challenging task. Besides skin color and environmental conditions, head movements and facial expressions should be considered to build a robust pulse rate measurement system. Pulse rate estimation error for videos with stable subjects and those that include facial expressions and head movements has been computed in order to assess how rigid movements (e.g., head tilt and posture changes) and non-rigid movements (e.g., facial expressions) affect the performance of X-iPPGNet. The results are presented in Table 9. We observe a performance degradation for large movements compared to stable videos but the error remains acceptable.

Table 9
Performance of our method on MMSE-HR under different head movement conditions.

Head movement conditions	Stable	Large movement
MAE (bpm)	3.88	4.44
RMSE (bpm)	4.91	5.74
r	0.86	0.82

Table 10
The time window size of the input video fragment in state-of-the-art methods.

Method	Time window size
DeepPhys [7]	30 s
Siamese-rPPG [50]	20 s
CHROM [52]	10 s
POS [5]	10 s
SynRhythm [70]	10 s
RhythmNet [8]	10 s
2SR [4]	6 s
EVM-CNN [46]	4/6/8 s
PhysNet [15]	2/4(best)/8 s
rPPGNet [33]	2 s (64 frames)
PRNet [32]	2 s (60 frames)
3DCNN [34]	2 s (60 frames)
X-iPPGNet (Ours)	**2 s (50 frames)**

4.2.4. Time window size

The time window size is an important parameter for video-based pulse rate estimation. Previous studies have reported that a longer window size leads to better performance, especially when using band-pass filter operation or power spectral density [15,46]. However, this increases the computational cost which is not suitable for real-time applications. Indeed, there is a trade-off in the size of the time window. If the time window is too large, the predicted pulse rate loses instantaneous information as we average pulse rates in the concerned video fragment. Conversely, the input video fragment may not contain a full cycle of two consecutive beats, resulting in an inaccurate pulse rate estimate. Table 10 presents the window size selected in this work in addition with state-of-the-art methods. All previous studies present much longer time windows than our method, except PRNet [32], 3DCNN [34], and rPPGNet [33]. These methods used a 2-seconds video fragment to estimate pulse rate, but with a higher number of frames.

Table 11 presents computation time and accuracy by window size. It is clear that increasing the window size implies more input images and more trainable parameters, thus increasing computation time. The same applies to accuracy where MAE and RMSE raise with increasing time windows, except for 1-second window which does not cover the low-frequency interval. For this reason, the 2-seconds window has been carefully selected to have a complete cardiac cycle and to cover the entire pulse rate range. Computation times of the methods that use a 2-seconds window is reported in Table 12. X-iPPGNet achieves 140 ms inference time behind PRNet [32], which runs the fastest among the six methods. X-iPPGNet is however deeper and outperforms PRNet in terms of accuracy.

5. Discussion

This work has been undertaken to optimize and improve iPPG-based systems for pulse rate estimation. Most existing studies extract the iPPG signal using either conventional approaches [2,4–6,48,66] or deep

Table 11
Performance and computation time of our method on MMSE-HR using different time window sizes.

Window size	1 s	2 s	3 s	4 s	6 s
MAE (bpm)	10.21	4.10	6.41	7.75	8.13
RMSE (bpm)	12.89	5.32	7.98	9.77	10.02
Computation time (ms)	120	140	160	180	220

Table 12
Computation time of our approach compared to state-of-the-art methods that use a 2-s input window size.

Method	Computation time (ms)
rPPGNet [33]	230
PhysNet [15]	200
3DCNN [34]	155
LCOMS [37]	150
PRNet [32]	**130**
X-iPPGNet (Ours)	140

learning-based methods [7,8,15,16]. Pulse rate is usually computed as the inverse of the average time difference between consecutive beats in the time domain, or as the frequency with the highest power spectrum energy in the frequency domain. Therefore, additional processing steps such as peak detection, Fast Fourier Transform, or Power Spectral Density are required. Moreover, the accuracy depends on the quality of the iPPG waveform and on the accuracy of the main peaks detection. Since publicly available databases are challenging and provide a large number of corrupted and poor-quality PPG signals [38,39,71], this directly affects the main peak location and consequently decreases the accuracy.

The proposed approach corresponds to an end-to-end trainable neural network where pulse rate is directly predicted from facial video recordings without separate iPPG signal recovery and with no prior knowledge. X-iPPGNet merges iPPG signal extraction and pulse rate prediction in one step. We rely on the ability of deep learning models to implicitly learn useful information directly from raw data. The training is fully supervised where each 2-seconds video fragment takes a ground truth pulse rate obtained with a contact device as a training label.

The main advantages of the proposed approach lie in its simplicity and low processing latency. A short time window is used to estimate pulse rate (2 s, 50 video frames). The size of the time window has a direct impact on performances. The larger it is, the higher the error, especially when dealing with higher and sharply fluctuating pulse rates (see Table 11). This is due to the loss of instantaneous information since the pulse rate is estimated by the averaging operation over the time window (As shown in Table 11). Moreover, our approach is more suitable for real-time measurement. The architecture is based on the Xception backbone that significantly reduces the number of parameters and computational costs without any performance degradation.

Since the most important factor when dealing with deep learning-based approaches is data, X-iPPGNet has been trained on BP4D+ to operate accurately in challenging scenarios and enable more robust training. BP4D+ provides a large amount of data and ethnic diversity, as well as challenging conditions. Furthermore, data augmentation is applied to increase the amount of under-represented samples at high and low frequencies. Using such a database in conjunction with data augmentation allows automatic learning of iPPG without hand-crafted features. Additionally, advanced deep learning optimization techniques as well as regularization strategies used in our work help to overcome overfitting issues and improve the model generalizability to new data.

The above experimental results verify the effectiveness of the proposed method and prove the possibility of measuring pulse rate directly from facial videos without going through iPPG signal recovery. Test results on three benchmark databases outperform existing methods and reveal the generalization ability to new data. We also examined the impact of various factors on prediction errors. The evaluation shows good performance in less-constrained scenarios such as head movement, illumination, video compression, and for different skin tones.

5.1. Limitations

The main limitation of our method concerns the way the pulse rate is measured. Although the framework is end-to-end trainable and superior in terms of speed and simplicity, pulse rate prediction without going through iPPG signal extraction does not allow pulse wave features extraction which is useful in medical applications [1] or for affective state recognition [72]. Furthermore, we have identified several issues that can be improved in future studies. First, most publicly available databases are very limited in terms of amount of data [40,73,74]. This lack of data makes training deep learning models more difficult and therefore increases the probability of overfitting and decreases the ability to generalize to new data. Although a few large-scale databases are available [38,39,71], they are not very diverse and are highly skewed towards light skin tones and mid-pulse rates. This leads to a lack of generalization and poor performance for under-represented samples. Using synthetic data [34,70,75,76] or combining multiple datasets [77] can solve the problem of the limited amount of data while applying advanced data augmentation strategies can improve performances for under-represented samples by creating additional and different training instances. Secondly, we noticed a high rate of corruption and poor quality ground truth PPG signals in the databases we used [38–40]. Data preparation and cleaning are essential to properly train the network and avoid overfitting problems. Finally, existing networks often consist of a large number of parameters and require high computational costs, which greatly hampers their application on resource-limited devices such as mobile phones.

6. Conclusion and future works

In this paper, we proposed a novel one-stage approach (X-iPPGNet) for contactless pulse rate estimation from facial video recordings using a deep spatio-temporal network. This approach is an efficient and elegant way to predict pulse rate without separate iPPG signal extraction and with no prior knowledge. X-iPPGNet is inspired by the Xception network architecture, which has proven to be efficient for general-purpose 2D image tasks in terms of accuracy, fast convergence speed, and low computational cost. Our extensive experiments showed the effectiveness of the proposed architecture, which achieves higher accuracy and outperforms existing methods on three popular benchmark datasets such as MMSE-HR, UBFC-rPPG, and MAHNOB-HCI. The results of this study demonstrated that pulse rate can be estimated remotely from facial videos without the need for complicated hand-crafted features or iPPG signal extraction.

Looking forward to our future work, we intend to compare the performance between our one-stage-based approach and two-stage-based methods. We will further analyze the effect of combining real and synthetic data on performance. Furthermore, we envisage investigating lightweight networks to develop a faster and more suitable model for real-time applications. We would also like to investigate the effectiveness of the proposed approach for measuring other physiological parameters, such as blood pressure, respiratory rate, and oxygen saturation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partly funded by the Contrat Plan État Région (CPER) Innovations Technologiques, France, Modélisation et Médecine Personnalisée (IT2MP), France and Fonds Européen de Développement Régional (FEDER), France.

References

[1] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, A. Pruski, Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera, Biomed. Signal Process. Control 64 (2021) 102242.

[2] M.-Z. Poh, D.J. McDuff, R.W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE Trans. Biomed. Eng. 58 (1) (2011) 7–11, http://dx.doi.org/10.1109/TBME.2010.2086456, URL http://ieeexplore.ieee.org/document/5599853/.

[3] G. De Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Trans. Biomed. Eng. 60 (10) (2013) 2878–2886.

[4] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE Trans. Biomed. Eng. 63 (9) (2015) 1974–1984.

[5] W. Wang, A.C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote PPG, IEEE Trans. Biomed. Eng. 64 (7) (2016) 1479–1491.

[6] F. Bousefsaf, C. Maaoui, A. Pruski, Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate, Biomed. Signal Process. Control 8 (6) (2013) 568–574.

[7] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 349–365.

[8] X. Niu, S. Shan, H. Han, X. Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, IEEE Trans. Image Process. 29 (2019) 2409–2423.

[9] W. Verkruysse, L.O. Svaasand, J.S. Nelson, Remote plethysmographic imaging using ambient light, Opt. Express 16 (26) (2008) 21434–21445.

[10] E. Nowara, D. McDuff, A. Veeraraghavan, A meta-analysis of the impact of skin type and gender on non-contact photoplethysmography measurements, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1148–1155.

[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[12] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep learning for computer vision: A brief review, Comput. Intell. Neurosci. 2018 (2018).

[13] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: International Workshop on Human Behavior Understanding, Springer, 2011, pp. 29–39.

[14] K. Suzuki, Overview of deep learning in medical imaging, Radiol. Phys. Technol. 10 (3) (2017) 257–273.

[15] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, 2019, arXiv preprint arXiv:1905.02419.

[16] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Process. Lett. 27 (2020) 1245–1249.

[17] A. Revanur, Z. Li, U.A. Cifti, L. Yin, L.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.

[18] M.-Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Opt. Express 18 (10) (2010) 10762–10774.

[19] M. Lewandowska, J. Rumiński, T. Kocejko, J. Nowak, Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity, in: 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, pp. 405–410.

[20] T. Blöcher, J. Schneider, M. Schinle, W. Stork, An online PPGI approach for camera based heart rate monitoring using beat-to-beat detection, in: 2017 IEEE Sensors Applications Symposium (SAS), 2017, pp. 1–6, http://dx.doi.org/10.1109/SAS.2017.7894052.

[21] M. Kumar, A. Veeraraghavan, A. Sabharwal, DistancePPG: Robust non-contact vital signs monitoring using a camera, Biomed. Opt. Express 6 (5) (2015) 1565–1588.

[22] S. Kwon, J. Kim, D. Lee, K. Park, ROI analysis for remote photoplethysmography on facial video, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 4938–4941.

[23] F. Bousefsaf, C. Maaoui, A. Pruski, Automatic selection of webcam photoplethysmographic pixels based on lightness criteria, J. Med. Biol. Eng. 37 (3) (2017) 374–385, http://dx.doi.org/10.1007/s40846-017-0229-1, URL http://link.springer.com/10.1007/s40846-017-0229-1.

[24] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J.F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2396–2404.

[25] J. Rumiński, Reliability of pulse measurements in videoplethysmography, Metrol. Meas. Syst. 23 (3) (2016).

[26] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Z. Medizinische Phys. 29 (2) (2019) 102–127, http://dx.doi.org/10.1016/j.zemedi.2018.11.002, URL https://www.sciencedirect.com/science/article/pii/S0939388918301181, Special Issue: Deep Learning in Medical Physics.

[27] E. Goceri, N. Goceri, Deep learning in medical image analysis: Recent advances and future trends, 2017, pp. 305–310.

[28] A. Ni, A. Azarang, N. Kehtarnavaz, A review of deep learning-based contactless heart rate measurement methods, Sensors 21 (2021) 3719, http://dx.doi.org/10.3390/s21113719.

[29] R. Špetlík, V. Franc, J. Matas, Visual heart rate estimation with convolutional neural network, in: Proceedings of the British Machine Vision Conference, Newcastle, UK, 2018, pp. 3–6.

[30] A. Reiss, I. Indlekofer, P. Schmidt, K. Van Laerhoven, Deep PPG: Large-scale heart rate estimation with convolutional neural networks, Sensors 19 (14) (2019) http://dx.doi.org/10.3390/s19143079, URL https://www.mdpi.com/1424-8220/19/14/3079.

[31] E. Lee, E. Chen, C.-Y. Lee, Meta-rPPG: Remote heart rate estimation using a transductive meta-learner, in: ECCV, 2020.

[32] B. Huang, C.-L. Lin, W. Chen, C.-F. Juang, X. Wu, A novel one-stage framework for visual pulse rate estimation using deep neural networks, Biomed. Signal Process. Control 66 (2021) 102387, http://dx.doi.org/10.1016/j.bspc.2020.102387, URL https://www.sciencedirect.com/science/article/pii/S1746809420304936.

[33] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, 2019, arXiv:1907.11921.

[34] F. Bousefsaf, A. Pruski, C. Maaoui, 3D convolutional neural networks for remote pulse rate measurement and mapping from facial video, Appl. Sci. 9 (2019) 4364, http://dx.doi.org/10.3390/app9204364.

[35] T. Luguev, D. Seuß, J.-U. Garbas, Deep learning based affective sensing with remote photoplethysmography, in: 2020 54th Annual Conference on Information Sciences and Systems (CISS), 2020, pp. 1–4, http://dx.doi.org/10.1109/CISS48834.2020.1570617362.

[36] O. Perepelkina, M. Artemyev, M. Churikova, M. Grinenko, HeartTrack: Convolutional neural network for remote video-based heart rate monitoring, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1163–1171, http://dx.doi.org/10.1109/CVPRW50498.2020.00152.

[37] Y. Ouzar, D. Djeldjli, F. Bousefsaf, C. Maaoui, LCOMS lab's approach to the vision for vitals (V4V) challenge, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 2750–2754.

[38] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U.A. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, L. Yin, Multimodal spontaneous emotion corpus for human behavior analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3438–3446.

[39] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2012) 42–55, http://dx.doi.org/10.1109/T-AFFC.2011.25.

[40] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognit. Lett. 124 (2019) 82–90, http://dx.doi.org/10.1016/j.patrec.2017.10.017, URL https://www.sciencedirect.com/science/article/pii/S0167865517303860, Award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[41] S. Bennett, T.N. El Harake, R. Goubran, F. Knoefel, Adaptive Eulerian video processing of thermal video: An experimental analysis, IEEE Trans. Instrum. Meas. 66 (10) (2017) 2516–2524, http://dx.doi.org/10.1109/TIM.2017.2684518.

[42] D.-Y. Chen, H.-S. Zou, A.-T. Hsieh, Thermal image based remote heart rate measurement on dynamic subjects using deep learning, in: 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 2020, pp. 1–2, http://dx.doi.org/10.1109/ICCE-Taiwan49838.2020.9258129.

[43] K. Humphreys, T. Ward, C. Markham, Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry, Rev. Sci. Instrum. 78 (4) (2007) 044304.

[44] W. Wang, A.D. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote PPG, IEEE Trans. Biomed. Eng. 64 (2017) 1479–1491.

[45] D.J. McDuff, S. Gontarek, R.W. Picard, Improvements in remote cardiopulmonary measurement using a five band digital camera, IEEE Trans. Biomed. Eng. 61 (2014) 2593–2601.

[46] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A.E. Saddik, EVM-CNN: Real-time contactless heart rate estimation from facial video, IEEE Trans. Multimed. 21 (2019) 1778–1787.

[47] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Adv. Neural Inf. Process. Syst. 33 (2020) 19400–19411.

[48] G. Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Trans. Bio-Med. Eng. 60 (2013) http://dx.doi.org/10.1109/TBME.2013.2266196.
[49] Y. Nirkin, I. Masi, A.T. Tran, T. Hassner, G.G. Medioni, On face segmentation, face swapping, and face perception, 2017, CoRR abs/1704.06729, arXiv:1704.06729, URL http://arxiv.org/abs/1704.06729.
[50] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, S.-H. Chang, Siamese-RPPG network: Remote photoplethysmography signal estimation from face videos, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, SAC '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2066–2073, http://dx.doi.org/10.1145/3341105.3373905.
[51] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, 2001, p. I, http://dx.doi.org/10.1109/CVPR.2001.990517.
[52] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[53] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (2016) 1499–1503.
[54] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[55] P. Zhao, C. Li, M.M. Rahaman, H. Yang, T. Jiang, M. Grzegorzek, A comparison of deep learning classification methods on small-scale image data set: from convolutional neural networks to visual transformers, 2021, arXiv preprint arXiv:2107.07699.
[56] A.V. Moço, S. Stuijk, G. de Haan, Motion robust PPG-imaging through color channel mapping, Biomed. Opt. Express 7 (5) (2016) 1737–1754.
[57] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.
[58] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah, X. Zhang, DS-CNN: A pre-trained xception model based on depth-wise separable convolutional neural network for finger vein recognition, Expert Syst. Appl. 191 (C) (2022) http://dx.doi.org/10.1016/j.eswa.2021.116288.
[59] N. Keskar, R. Socher, Improving generalization performance by switching from adam to SGD, 2017, ArXiv abs/1712.07628.
[60] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the variance of the adaptive learning rate and beyond, in: Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), 2020.
[61] S. Ruder, An overview of gradient descent optimization algorithms, 2016, arXiv preprint arXiv:1609.04747.
[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (56) (2014) 1929–1958, URL http://jmlr.org/papers/v15/srivastava14a.html.
[63] E.M. Nowara, D. McDuff, A. Veeraraghavan, Combining magnification and measurement for non-contact cardiac monitoring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 3810–3819.
[64] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. 31 (4) (2012) 1–8.
[65] N. Miljković, D. Trifunović, Pulse rate assessment: Eulerian video magnification vs. electrocardiography recordings, in: 12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL), IEEE, 2014, pp. 17–20.
[66] X. Li, J. Chen, G. Zhao, M. Pietikäinen, Remote heart rate measurement from face videos under realistic situations, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4264–4271, http://dx.doi.org/10.1109/CVPR.2014.543.
[67] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: European Conference on Computer Vision, Springer, 2020, pp. 392–409.
[68] X. Liu, Z. Jiang, J. Fromm, X. Xu, S.N. Patel, D. McDuff, MetaPhys: Unsupervised few-shot adaptation for non-contact physiological measurement, 2020, ArXiv abs/2010.01773.
[69] T. Fitzpatrick, The validity and practicality of sun-reactive skin types I through VI, Arch. Dermatol. 124 6 (1988) 869–871.
[70] X. Niu, H. Han, S. Shan, X. Chen, SynRhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3580–3585, http://dx.doi.org/10.1109/ICPR.2018.8546321.
[71] X. Niu, H. Han, S. Shan, X. Chen, VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video, in: ACCV, 2018.
[72] Y. Ouzar, F. Bousefsaf, D. Djeldjli, C. Maaoui, Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2460–2469.
[73] G. Heusch, A. Anjos, S. Marcel, A reproducible study on remote heart rate measurement, 2017, arXiv preprint arXiv:1709.00962.
[74] R. Stricker, S. Müller, H.-M. Groß, Non-contact video-based pulse rate measurement on a mobile service robot, 2014, pp. 1056–1062.
[75] D. McDuff, X. Liu, J. Hernandez, E. Wood, T. Baltrusaitis, Synthetic data for multi-parameter camera-based physiological sensing, 2021, arXiv preprint arXiv:2110.04902.
[76] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, X. Chen, PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography, IEEE J. Biomed. Health Inf. 25 (2021) 1373–1384.
[77] B.L. Hill, X. Liu, D. McDuff, Beat-to-beat cardiac pulse rate measurement from video, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 2739–2742.

Estimation of blood pressure waveform from facial video using a deep U-shaped network and the wavelet representation of imaging photoplethysmographic signals

Frédéric Bousefsaf, Théo Desquins, Djamaleddine Djeldjli, Choubeila Maaoui, Alain Pruski

Estimation of blood pressure waveform from facial video using a deep U-shaped network and the wavelet representation of imaging photoplethysmographic signals

Frédéric Bousefsaf[*,1], Théo Desquins[1,2], Djamaleddine Djeldjli[1], Yassine Ouzar[1], Choubeila Maaoui[1], Alain Pruski[1]

1. Université de Lorraine, LCOMS, F-57000 Metz, France

2. i-Virtual, F-57000 Metz, France

Abstract

BACKGROUND. The remote measurement of physiological signals from video has gained a particular attention over the last past years. Estimating cardiovascular parameters like oxygen saturation and arterial blood pressure (BP) is covered by a limited volume of studies and remain a very challenging issue. Recent attempts demonstrated that BP can be estimated from facial video but under very controlled scenarios or with moderate performances. The data used in these works have not been publicly released or were gathered in a clinical setting. **METHODS**. We, in contrast, propose a framework for estimating BP from publicly available data in order to allow replication and to facilitate fair comparison. We developed and trained a deep U-shaped neural network to recover the blood pressure waveform from its imaging photoplethysmographic (iPPG) signal counterpart. The model predicts the continuous wavelet transform (CWT) representation of a BP signal from the CWT of an iPPG signal. Inverse CWT transform is ultimately computed to recover the BP time series. **RESULTS**. The proposed framework has been evaluated on 57 participants using international standards developed by the AAMI and the BHS. Results exhibit close agreement with ground truth BP values. The method satisfies all standards in the estimation of mean and diastolic BP (grade A) and nearly all standards in the estimation of systolic BP (grade B). **CONCLUSIONS**. This is, to the best of our knowledge, the first demonstration of a deep learning-oriented framework that manages to

predict the continuous blood pressure waveform from facial video analysis. Codes developed during the study are publicly available (`https://github.com/frederic-bousefsaf/ippg2bp`).

Keywords: imaging photoplethysmography, blood pressure, continuous wavelet transform, deep learning, U-Net

---

1. Introduction

Research on the remote measurement of physiological signals and cardiovascular parameters from facial video has made significant progress the last past years. The field is booming and supported by several significant studies [1]. The principle, termed imaging (or remote) photoplethysmography (iPPG), consists in measuring the subtle fluctuations of skin color. These fluctuations reflect complex light-tissue interactions. The simplest cameras (webcams) to the most advanced ones (professional, laboratory or industrial cameras) can be employed to reliably recover iPPG signals. Different regions of interest (ROI) have been studied over time but the face remains the most frequently observed area [2]. Several studies demonstrated that pulse rate and its variability can be robustly and precisely estimated with conventional image processing techniques and, more recently, with deep learning solutions [3, 4].

Current research in this field is now directed towards the measurement of new physiological parameters such as oxygen saturation [5] and blood pressure [6]. Estimating arterial blood pressure (BP) from video is covered by a limited volume of studies and remain a very challenging issue.

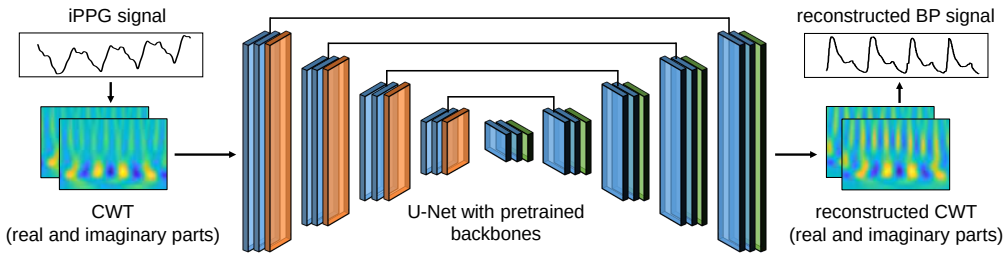


Figure 1: General overview of the method.

Two research directions are considered. First, measurement of the pulse transit time (PTT) on single [7] or several [8] ROI. PTT is a parameter considered to be correlated with blood pressure. Secondly, analysis of the iPPG signal waveform [6, 9]. To our knowledge, deep learning techniques have only been considered by Schrumpf et al. for the estimation of blood pressure from iPPG signals [10]. The model includes 5 layers and exhibit moderate performances, i.e. high mean average error and no compliance with international standards. These recent attempts demonstrated that BP can be estimated from facial video but under very controlled scenarios or with moderate performances. In addition, the data used in these works have not been publicly released or were gathered in a clinical setting. Only Schrumpf et al. released a sub-part of the data employed in their study. At the time of writing, this subset includes small excerpts of iPPG signals and discrete BP values from 17 over 50 participants (see `https://github.com/Fabian-Sc85/non-invasive-bp-estimation-using-deep-learning`). To conclude on this point, training an artificial neural network that accurately estimates blood pressure from video is constrained by the amount of available data because few public databases exist.

We propose, in this article, a framework for estimating BP from publicly available data. The dataset, namely BP4D+, includes video streams of moving participants. Video analysis dedicated to remote physiological sensing is therefore very challenging. A deep learning-oriented method (see figure 1) has been specifically developed to recover the blood pressure waveform from its imaging photoplethysmographic (iPPG) signal counterpart. The deep U-shaped model presented in this work has already been applied for translating iPPG to contact PPG signals in a previous work [11]. The full pipeline includes several stages. Skin pixels are first extracted using a recent segmentation techniques that relies on fully convolutional networks. iPPG signal is computed by averaging all the skin pixels from the green channel. We then employed the continuous wavelet transform (CWT) of iPPG (and respectively BP) signals to train the aforementioned neural architecture. The model therefore predicts a CWT representation of a BP signal from the CWT of an iPPG signal. Inverse CWT transform is ultimately computed to recover the BP time series.

The article includes five additional sections. Section 2 presents the background and related works. Section 3 introduces the used data and the developed methodologies. The full processing pipeline is detailed in this section. The metrics and results of the proposed approach are presented and discussed

in sections 4 and 5, respectively. We present the future works and a summary of the contributions in section 6.

This is, to the best of our knowledge, the first demonstration of a deep learning-oriented framework that manages to predict the continuous blood pressure waveform from iPPG signals computed using publicly released data. Several avenues of interest are envisaged to improve this research that, in its current state, exhibits very encouraging results. Two out of three estimated measures (i.e. diastolic BP and mean BP) already satisfy metrics defined by international standards.

2. Related works

A survey related to blood pressure estimation from video has recently been proposed by Lu et al. [12]. Several studies of interest have nevertheless been proposed since its publication. We therefore, and in the two first subsections, propose to review the studies that exploit iPPG for blood pressure assessment using both conventional and deep learning approaches. The estimation of blood pressure from contact PPG is closely related to this topic. We therefore dedicate the last subsection to this part.

2.1. iPPG for blood pressure estimation from propagation time

Systolic and diastolic blood pressures have been estimated using the propagation time of pulse waves from two different skin areas (typically hand and face) in video recordings [13, 14, 15, 8]. The positional of the two skin areas must be maintained during the measurement. This approach is therefore very restrictive. In this context, the time delay must be robustly assessed. Dedicated techniques were proposed for this purpose the last past years. Shao et al. compared peak locations from iPPG signals measured from two sites [16]. To improve accuracy, the peaks were estimated with two linear curves fitted on the edges of the rising and falling parts of the signal. Fan and Tjahjadib [17] analyzed the wave peaks with a custom signal quality index. Peaks of low confidence are removed using a Kalman filter to improve performances. Sugita et al. proposed to analyze videos of human hands recorded at different heights from the heart [18]. They analyze the difference in amplitude of iPPG pulse waves to build a model that estimates SBP.

2.2. iPPG for blood pressure estimation from single facial region

The estimation of BP from a single facial region is covered by very few studies in the scientific literature. The general approach, inspired from the

contact PPG field [19, 20], consists in computing waveform features that are correlated to BP. In this direction, Djeldjli et al. recently showed that temporal, derivative and area features computed from iPPG and cPPG waveform evolve similarly [21].

Jain et al. developed a simple regression framework that analyzes 21 waveform features computed on the iPPG signal to estimate BP [22]. Sugita et al. proposed to quantify the degree of distortion of iPPG signals [7]. They showed that this quantity exhibits correlation with BP close to correlations computed between BP and propagation times. Viejo et al. estimated BP from video using handcrafted features and machine learning models [23]. They studied the evolution of BP using a shallow neural network in the context of food sensory responses but no direct BP assessment is presented in their article.

The seminal work from Luo et al. [6] presents for the first time a pipeline that includes an artificial intelligence model. A multilayer perceptron has been fed with 30 features computed from iPPG waves. Their results show that iPPG waveform extracted from video exhibits information that are correlated to BP. Combining handcrafted features from iPPG signals with a machine learning approach to estimate systolic and diastolic BP has also been investigated by Rong and Li [9]. Deep learning architectures were recently studied by Schrumpf et al. [10]. The authors fine-tuned a network that integrates convolutional, long short-term memory and dense layers. They conclude that iPPG signals computed from standard RGB video streams may not be suitable to reliably estimate BP. All these studies pointed out the feasibility of remote BP monitoring from facial video but showed that there is still room for improvements and that the estimation remains a very challenging issue. A synthetic overview of the existing studies is presented in table 1. An important disparity in the number of subjects as well as overall low performances can be observed from this table. In addition, all the results presented in these studies have been tested on data that has not been released. To the best of our knowledge, no research dedicated to the estimation of blood pressure from iPPG has yet been conducted with public datasets.

2.3. Blood pressure estimation from contact PPG

Estimating absolute BP values from contact PPG (cPPG) remains a challenging problem even if there is clear evidence that the fluctuations in BP are reflected in cPPG signals [19, 20].

Deep learning techniques have recently been investigated [26] and recent developments show that these frameworks can effectively be deployed to convert BP waveform from cPPG signals. Different type of artificial neural architectures have been proposed the last past years. They combine fully connected [27] or convolutional layers [28] with long short-term memory. Simultaneous estimation of systolic and diastolic BP is ensured by these networks. Demographic features (e.g. weight and height) have additionally been included in machine learning algorithms to improve BP estimation from cPPG signals [29]. Time, frequency and time-frequency features were computed from the PPG and their derivative signals. Feature selection techniques were used for reducing the computational complexity and simultaneously decreasing the chance of over-fitting the machine learning algorithms.

Number of subjects	**Sampling freq. (fps)**	**iPPG signal extraction**	**Features**	**Model**	**Performances**		**Ref.**
					SBP	**DBP**	
17	140	Green	T_{BH} index	–	-0.6†	–	[7]
45	50	PCA	21 time and frequency features	regression	3.90‡ $\pm$ 5.37	3.72‡ $\pm$ 5.08	[22]
45	15	Green	amplitude, freq. and pulse rate)	shallow ANN	–	–	[23]
1328	30	TOI	155 features (30 after PCA)	ANN (MLP)	0.67† 0.39* $\pm$ 7.30	0.63† -0.2* $\pm$ 6.00	[6]
189	30	Green	26 features (16 after feature selection)	SVR	9.97‡ 2.1* $\pm$ 3.35	7.59‡ 0.79* $\pm$ 2.58	[9]
25	32	POS	–	CNN-LSTM-Dense (transfer learning using MIMIC III)	13.6‡	10.3‡	[10]

Table 1: Overview of the existing studies in the field of BP estimation from single facial region in video streams.

*: bias

†: correlation coefficient

‡: Mean Absolute Error (MAE)

ANN: Artificial Neural Network

CNN: Convolutional Neural Network

Green: iPPG signal formed using only the green channel [24]

LSTM: Long Short-Term Memory

MLP: MultiLayer Perceptron

PCA: Principal Component Analysis

POS: Plane-Orthogonal-to-Skin method [25]

SVR: Support Vector Regression

TOI: Transdermal Optical Imaging [6]

A similar framework but with a deep architecture with residual connections has been proposed by Slapnicar et al. [30]. A part of the network is dedicated to the analysis of the spectral representation of the signal using gated recurrent units. Deep learning networks that manage to predict the continuous BP waveform from cPPG signals have recently been proposed [26]. An approximation network learns a rough approximation of the BP waveform while a refinement network further enhances the preliminary estimate. The approximation and refinement networks are based on a U-Net architecture [31].

3. Methods

3.1. Database

BP4D+ is a multimodal dataset publicly available to the research community[1]. The database initially includes the physiological, thermal, 2D video, 3D and different metadata and annotations of 140 participants [32]. Ten tasks were proposed to elicit different emotions in a lab environment.

Because of the nature of the tasks, strong motion artifacts are present alongside an ensemble of videos, leading to difficult iPPG signal extraction. Video analysis for remote physiological sensing is therefore very challenging. We conducted a first selection process where only videos presenting clear iPPG signals have been kept. The procedure relies on a conventional signal-to-noise ratio (SNR). The index is defined using the Fourier transform of iPPG signal in 15-second windowed intervals so that sub-parts of partially impacted videos can be selected. The SNR has already been used in the field of iPPG [33, 34]. All the selected video parts have been manually controlled after this first automatic preselection. A subset of 57 subjects (21 females, 36 males), leading to a total of 157 videos, has been built. We additionally removed samples where the reference continuous blood pressure signal was improperly constituted or flawed (negative values). Details about the selected participants and tasks are available on a dedicated file in the website hosting the project (`https://github.com/frederic-bousefsaf/ippg2bp`). This subset has been employed for training and testing the neural architecture presented in this study.

[1] `http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html`

Each signal (for each participant and for each task) has been processed using the different techniques detailed in section 3.2. Each full-length signal has been split in excerpts of 2.56 seconds defined over 256 values. This constituted a dataset of 4123 portions of signal. About 70% of the data (2887 randomly selected excerpts) has been reserved for training, 15% (618 randomly selected excerpts) for validation and the remaining 15% (618 randomly selected excerpts) for the testing phase. The different sets contain a balanced portfolio of the participants and tasks.

We computed systolic BP (SBP) by averaging the intensities of the max peaks over the entire excerpt. Diastolic BP (DBP) has been computed with a similar strategy but using the min peaks intensities instead of the max ones. Mean arterial pressure (MAP) is the average value computed over all the excerpt samples. The distribution of SBP, DBP and MAP values for the training, validation and test sets are presented in figure 2. The distributions share similar properties and ranges.

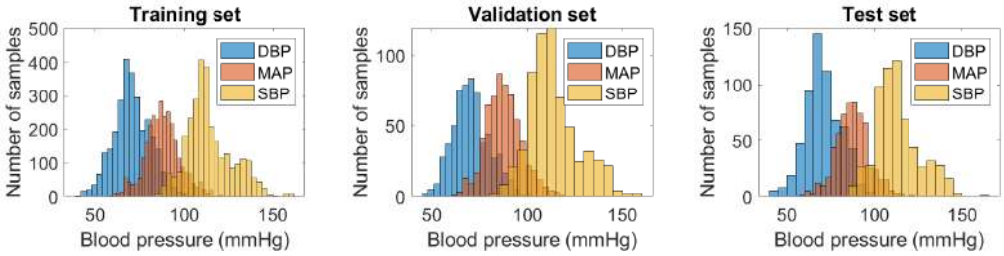


Figure 2: Distribution of DBP, MAP and SBP for the different sets. All the samples were extracted from the BP4D+ dataset.

3.2. iPPG signal constitution

The overall processing pipeline is quite similar to the one presented in [11]. This method (called iPPG 2 cPPG) consists in employing the continuous wavelet representation (real and imaginary parts) of an iPPG signal to reconstruct the wavelet representation of a contact PPG (cPPG) signal. Inverse transform is then computed to recover the cPPG time series.

First, we employed a recent face segmentation technique that relies on fully convolutional networks [37]. The approach robustly removed the background and non-skin areas. The method has recently been employed in the field of imaging photoplethysmography [38].

iPPG signal has been computed by averaging all the remaining skin pixels from the green channel. Figure 3a exhibits a raw iPPG signal computed from one of the BP4D+ video stream. Raw iPPG signals are then interpolated at a sampling frequency of 100 Hz and detrended using a specific low-pass filter [35] based on a smoothness priors that attenuates low frequencies [36].

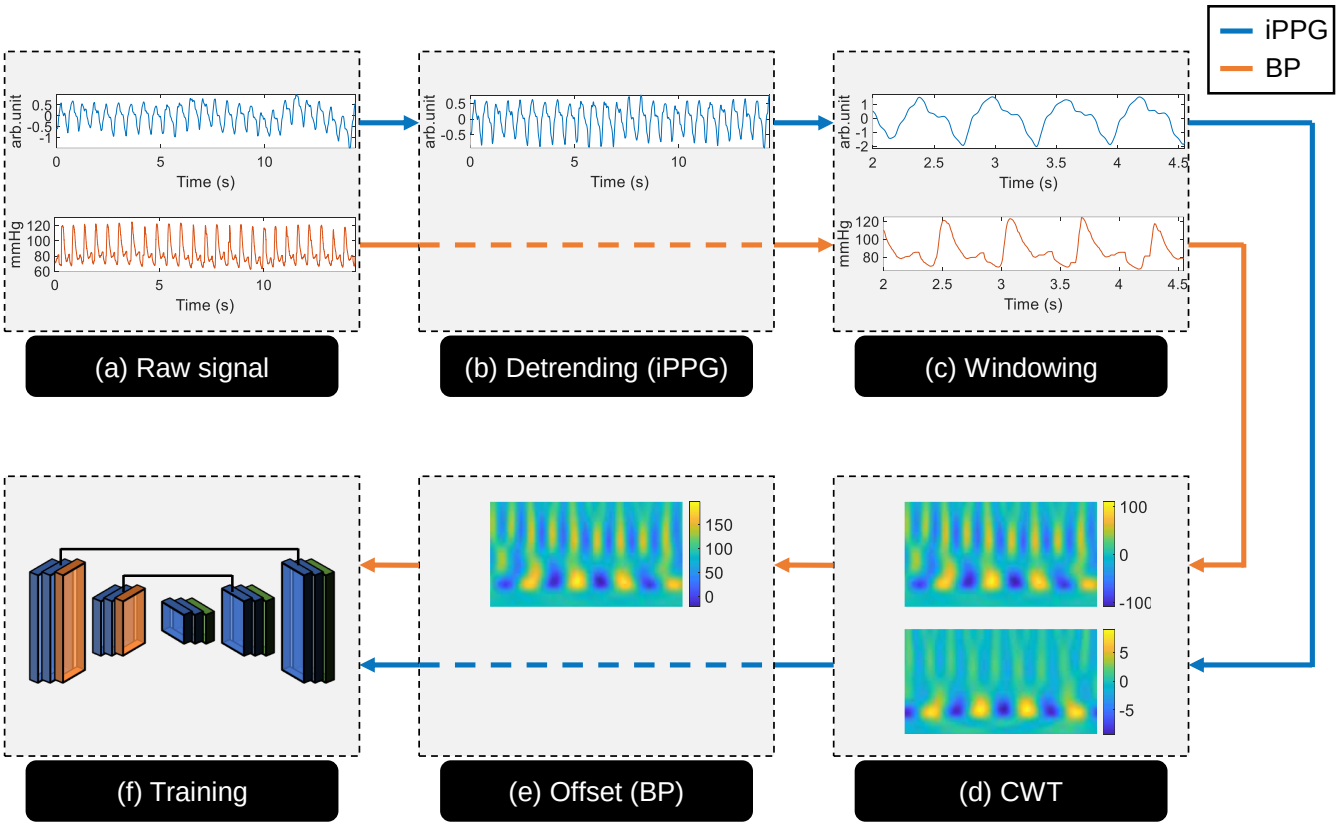


Figure 3: Signal processing before CWT computation. (a) Example of a raw iPPG signal that contains noise and trends (top illustration) and of a BP signal that has been simultaneously recorded using a continuous non-invasive sensor (bottom illustration). (b) iPPG trends removal is ensured by a method [35] that has already been used in this field [36]. (c) Small excerpts of 2.56 seconds are extracted for further processing. (d) The CWT (real part) of both iPPG and BP signals is computed in the frequency range [0.6, 4.5] Hz. (e) The average value is lost when computing the CWT in the aforementioned frequency range. This information is therefore directly encoded in the CWT of the BP signal by adding the mean value to every CWT coefficient. See the the difference in the ranges of the colorbars between subfigures (d) and (e). (f) The CWT (real and imaginary parts) are used for training the neural architecture presented in section 3.3.

Figure 3b shows the impact of the detrending operation on the iPPG signal. We then extract small excerpts for both the iPPG and the ground truth BP signals (see figure 3c for a typical example). An overlapping sliding window scheme has been selected to increase the volume of data employed during training. The sampling frequency of the interpolated iPPG signal being set to 100 Hz, 2.56 seconds are necessary to form time-frequency representations of 256 pixels in width. The window length has therefore been set to 2.56 seconds with an empirically defined step size of 0.5 seconds (50 samples). All the iPPG excerpts have been standardized using the z-score formula (so that $\mu = 0$ and $\sigma = 1$). Training, validation and testing sets were then constituted from this ensemble of excerpts (see section 3.1).

Like in [11], we employed the continuous wavelet transform (CWT) representation to train the neural architecture presented in section 3.3. The global approach is depicted in figure 3. The CWT (equation 1) of a signal $x(t)$ corresponds to a time-frequency representation computed from a prototype function commonly called mother wavelet. Unlike the Fourier transform, the wavelet transform can detect abrupt changes in frequency using a family of wavelets $\psi_{\tau,s}$ (equation 2) computed from the mother wavelet ψ.

$$CWT_x^{\psi}(\tau, s) = \int_{-\infty}^{\infty} x(t)\, \psi_{\tau,s}(t) dt \tag{1}$$

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right) \tag{2}$$

$\psi_{\tau,s}$ corresponds to the mother wavelet dilated by s and translated by τ. Dilating the wavelet allows the transform to analyze larger portions of signal in the time domain, thus covering lower frequencies. Different mother wavelets have been developed and the choice depends mainly on the application and the properties of the signal. The Morlet mother wavelet used in this study was already used in previous work related to the analysis of PPG signals by camera [39, 40, 11].

The original signal $x(t)$ can be reconstructed by the inverse transform:

$$x(t) = \frac{1}{C_{\psi}} \int_0^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} CWT_x^{\psi}(\tau, s) \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right) d\tau\, ds \tag{3}$$

$$C_{\psi} = \int_0^{\infty} \frac{\left|\hat{\psi}(\zeta)\right|^2}{|\zeta|} d\zeta < \infty \tag{4}$$

C_ψ is the admissibility condition and $\hat{\psi}$ is the Fourier transform of ψ.

The continuous wavelet transform was computed on each iPPG and BP signal in the frequency range [0.6, 4.5] Hz, which corresponds to the physiological range of the human heart rate [2]. Typical iPPG signal, BP signal and their respective wavelet representations (real part) are presented in figure 4. As it was presented before, the iPPG signals have been standardized ($\mu = 0$ and $\sigma = 1$, see top-left illustration in figure 4 for a typical example). This type of process has not been applied to the BP signals because we need to recover both the average, systolic and diastolic values (see top-mid illustration in figure 4). The average value being lost when computing the CWT in the frequency range [0.6, 4.5] Hz, we chose to directly encode this information in the CWT of BP signals by adding the mean value to every CWT coefficients (see figure 3e):

$$CWT_{BP} = CWT_{BP} + \mu_{BP} \tag{5}$$

Here, μ_{BP} corresponds to the average value of a BP signal (top-mid illustration in figure 4 for a typical BP signal example) and CWT_{BP} to its

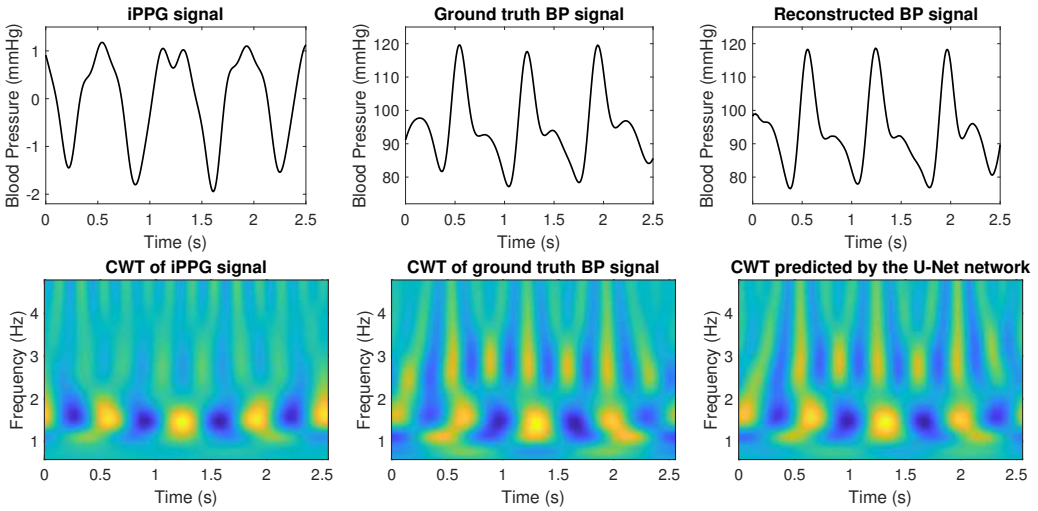


Figure 4: An iPPG and its corresponding ground truth BP are respectively presented in the bottom-left and bottom-mid figures. Their corresponding CWT (real part) are presented below. The transform (a complex image with a real and imaginary part) is computed in the frequency range [0.6, 4.5] Hz. Figures on the right present the CWT predicted by the neural network and the corresponding reconstructed BP signal, computed using the inverse CWT transform.

corresponding CWT (see bottom-mid illustration in figure 4).

The produced wavelet representations have a dimension of $256 \times 256 \times 2$ pixels. They are used to train the neural architectures (figure 3f) presented in the next section.

3.3. Neural architectures

The neural architecture has already been developed and tested in previous work [11]. Briefly, it consists in a U-Net architecture, which was initially proposed by Ronneberger et al. [31], enhanced by a backbone. This type of network has been widely used for segmentation of medical images [41]. Its architecture consists of a descending (encoder) branch completed by an ascending (decoder) branch, giving a U-shape to the network. The descending branch contains an ensemble of convolution and pooling layers. The ascending branch integrates upsampling layers connected to the convolutions of the descending branch. Connections help to restore the spatial information. A schematic representation of the network is provided in figure 5. Each convolutional layer are coupled with a Rectified Linear Unit (ReLU) activation function.

A Backbone (e.g. VGG16) can be integrated into the encoder part of the U-Net network. Its internal parameters can be blocked during training, meaning that the weights of the network remain the same. In practice, a backbone correspond to a model subpart pre-trained on ImageNet, a database deployed for object recognition tasks in images [42]. Training a U-Net network supported by a backbone consists, in this case, in optimizing the internal parameters of the decoder part. This approach can be associated to a transfer learning strategy. In this work, we initialized the U-Net architecture with a ResNeXt101 backbone [43]. The encoder parameters were not blocked during training, meaning that they were optimized during the learning phase. The number of variables to be trained (weights and biases) is 52 million. We chose ResNeXt101 because it performed better than other standard backbones on the reconstruction of contact PPG signals from non contact ones through their continuous wavelet representation, a problem that is in fact quite similar [11].

Conventional regularization techniques (e.g. dropout) have not been introduced while a normalization scheme (i.e. batch normalization) has been employed. Linear activation function was specified because the targeted task corresponds to a regression in the form of a pixel-to-pixel reconstruction of a two-channel wavelet representation.

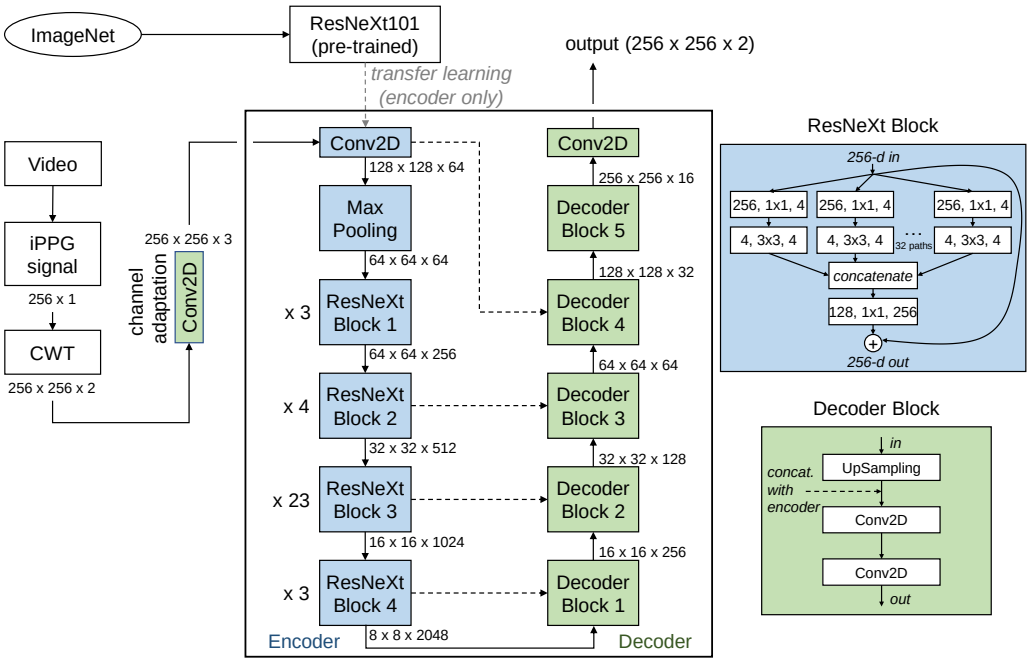


Figure 5: Overview of the U-Net [31] proposed in this study, which includes encoder (down-sampling) and decoder (upsampling) portions. The encoder is replaced by a ResNeXt101 backbone [43]. ResNeXt and decoder blocks are detailed on the right-side of the figure. The input of a ResNeXt block (256 dimensions in the example depicted in the figure) is split into 32 lower dimensional branches (or paths) that will next be merged through concatenation. This architecture exploits Inception's split-transform-merge strategy but with a uniform topology. The parameters of each stage inside this ResNeXt block example are respectively the number of input filters, the filter size and the number of output filters. Each ResNeXt block present different parameters. They are specified in [43].

The input dimensions of a U-Net network supported by a backbone are fixed by the data used for their training (256×256 pixels RGB images from the ImageNet database). The inputs being in our case a two-channels wavelet representation, an adaptation strategy must be introduced. We employed an additional 2D convolutional layer with a $(1, 1)$ kernel that has been placed between the input layer and the encoder part of the network. The neurons of this layer allow conversion of the input from N to 3 channels. The weights of all the networks have randomly been initialized by the method proposed by Glorot and Bengio [44]. Biases are initialized to zero. The Mean Squared Error (MSE) has been selected as loss for training all the models:

$$MSE = \frac{1}{n}\sum_{i,j}\left(CWT_{i,j} - \widehat{CWT}_{i,j}\right)^2 \tag{6}$$

CWT corresponds to the wavelet transform (see figure 3) of the ground truth BP signal. $\widehat{CWT}$ is the wavelet representation predicted by the neural network starting from the wavelet representation of the iPPG signal.

The architecture implementation was carried out under Python using Keras API and Tensorflow library. The Segmentation Models library [45] proposed by P. Yakubovskiy was used to develop the neural network. The training sessions were launched over 500 epochs through batches of 16 images. We used, in this study, the Adam optimization algorithm [46] with a learning rate of 0.001. A dedicated computer equipped with a dual Intel Xeon Silver 4114 and two Nvidia Quadro P6000s was used to carry out network learning.

4. Results

The proposed U-Net architecture transforms an iPPG signal to a continuous BP signal through their wavelet representation. Figure 4 illustrates a typical example of BP estimation (top-right figure) from an iPPG wave (top-left figure). The predicted waveform closely follows the ground truth BP wave presented in top-mid figure. The shape and magnitude, which were initially different, have been preserved. We can notice small phase differences in the wavelet representations of the iPPG signal (bottom-left figure)

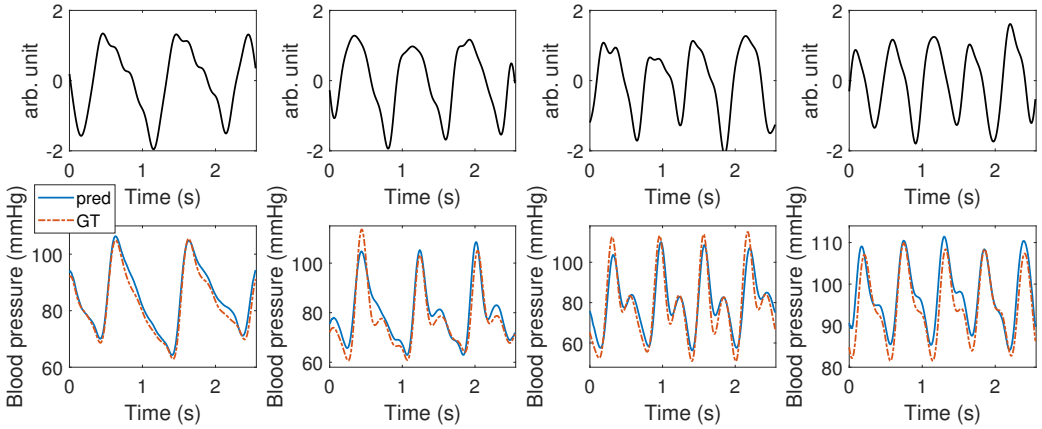


Figure 6: Typical BP signals reconstruction for several pulse rate values. Top figures: iPPG signals. Bottom figures: predicted and ground truth (GT) BP.

and the ground truth BP signal (bottom-mid figure). The neural network learned this specificity, the reconstructed wavelet representation (bottom-right figure) being in phase with the ground truth one (bottom-mid figure). The phase has therefore been properly recovered. This follows previous observations that we made when testing this U-Net to transform contact PPG to iPPG signals [11] and observations from other authors that employed deep learning to convert contact PPG to BP waves [26].

Figure 6 illustrates several examples of blood pressure estimation from iPPG signals. We evaluated the performances of the proposed technique with international standards [47, 48] from the Association for the Advancement of Medical Instrumentation (AAMI) and from the British Hypertension Society (BHS). We, however, emphasize that BP4D+ contains videos and physiological data that have not been recorded in a clinical setting. Also, the constituted subset integrates 57 participants while the AAMI recommends to evaluate BP estimation techniques on a minimum of 85 subjects.

4.1. General metrics and Bland-Altman plots

The Mean Absolute Error (MAE, equation 7) and the Root Mean Square Error ($RMSE$, equation 8) have been used to quantify the level of agreement between the predicted ($\widehat{BP}$) and the ground truth blood pressure (BP). We computed these metrics for DBP, MAP and SBP over all the test set (see section 3.1).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|BP_i - \widehat{BP_i}| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(BP_i - \widehat{BP_i}\right)^2} \tag{8}$$

Table 2 presents a comparative analysis of results taken from similar works. Bland-Altman representations have been computed for DBP, MAP and SBP over all the test data. The average between the estimated and ground truth BP values is depicted on the x-axis while the differences between the estimated and ground truth BP values are depicted on the y-axis. The resulting plots are presented in figure 7. Means are represented by dash-dot lines and 95% limits of agreement ($\pm$ 1.96 SD) by dashed lines. The ranges of these limits are [-12.3 14.3], [-12.0 11.6] and [-19.6 16.6] for DBP, MAP and SBP respectively.

		MAE (mmHg)	RMSE (mmHg)
Rong and Li [9]	DBP	7.59	–
	SBP	9.97	–
Schrumpf et al. [10]	DBP	10.3	–
	SBP	13.6	–
iPPG2BP (our results)	DBP	5.1	6.85
	MAP	4.47	6.01
	SBP	6.73	9.34

Table 2: Blood pressure estimation errors. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have been computed between the estimated and ground truth DBP, MAP and SBP. Results from similar studies are also reported.

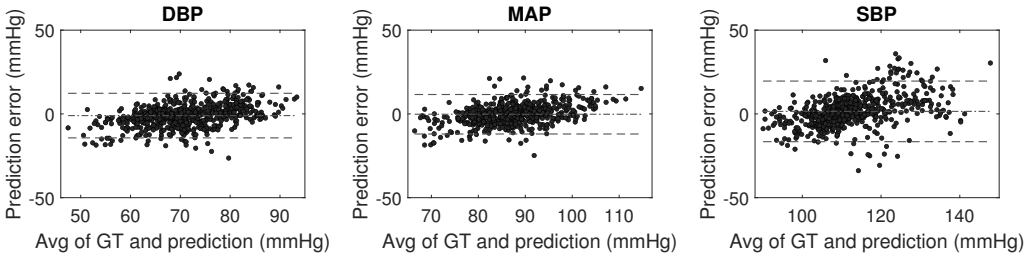


Figure 7: Bland-Altman plots for DBP, MAP and SBP prediction. Means are represented by dash-dot lines and 95% limits of agreement ($\pm$ 1.96 SD) by dashed lines.

4.2. BHS standards

The BHS assesses blood pressure estimation techniques by their cumulative percentage of errors [47]. Different grades are provided (see table 3) according to the percentage of the predictions on the test samples that fall under three empiric thresholds, i.e. 5, 10 and 15 mmHg.

Table 3 presents a comparative analysis of the BHS evaluation on our results. We reported the values provided by Rong and Li [9] as it appears to be the only study that computed BHS metrics. Our results exhibit good overall performances with more than 60%, 87% and 95% of the test samples having estimation errors less than, respectively, 5, 10 and 15 mmHg for both DBP and MAP (grade A). More than 50% and 79% of SBP predictions fall under 5 and 10 mmHg respectively (grade B) while 89.6% of SBP predictions fall under 15 mmHg, which is slightly under the 90% threshold.

The conclusions drawn from the analysis of the results presented in table 3 are graphically presented in figure 8.

		Cumulative Error Percentage		
		$\leq$ 5 mmHg	$\leq$ 10 mmHg	$\leq$ 15 mmHg
Rong and Li [9]	DBP	55.4%	85.7%	98.2%
	SBP	48.2%	78.6%	94.6%
iPPG2BP (our results)	DBP	60.2%	87.1%	95.8%
	MAP	66.8%	90.9%	96.4%
	SBP	50.2%	79.0%	89.6%
BHS	grade A	60%	85%	95%
	grade B	50%	75%	90%
	grade C	40%	65%	85%

Table 3: BHS metrics for DBP, MAP and SBP prediction.

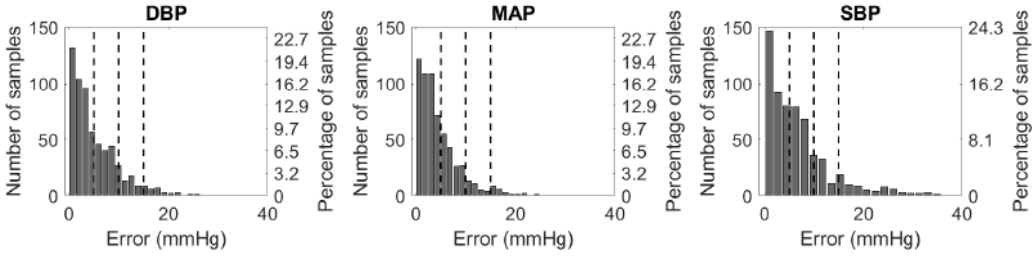


Figure 8: Absolute error in DBP, MAP and SBP predictions. Dashed lines represent the 5, 10 and 15 mmHg thresholds recommended by the BHS.

4.3. AAMI standards

The AAMI proposes to assess blood pressure estimation techniques by analyzing the mean error (ME) and the standard deviation of errors (SDE) on the test set [48]. The former must be lower than 5 mmHg while the latter must be lower than 8 mmHg to fully respect the recommendation.

Table 4 presents a comparative analysis of the AAMI evaluation on our results. We additionally reported the values provided by Luo et al. [6] and Rong and Li [9]. Our results exhibit good overall performances. Both DBP and MAP satisfy the AAMI standards. They exhibit a small ME and a SDE lower than 8 mmHg. Regarding SBP estimations, the ME condition is fulfilled but the SDE is a bit higher (1.2 mmHg over the 8 mmHg threshold defined by the AAMI).

The histograms of prediction errors for DBP, MAP and SBP are presented in figure 9. The spread of these histograms gives a graphical picture of the different SDE presented in 4 (narrower for MAP, wider for SBP).

		ME (mmHg)	SDE (mmHg)
Luo et al. [6]	DBP	-0.20	6.00
	SBP	0.39	7.30
Rong and Li [9]	DBP	0.79	2.58
	SBP	2.1	3.35
iPPG2BP (our results)	DBP	-1.001	6.781
	MAP	-0.205	6.007
	SBP	1.51	9.221
AAMI standard		≤ 5	≤ 8

Table 4: AAMI metrics for DBP, MAP and SBP prediction. ME: Mean Error; SDE: Standard Deviation of Errors.

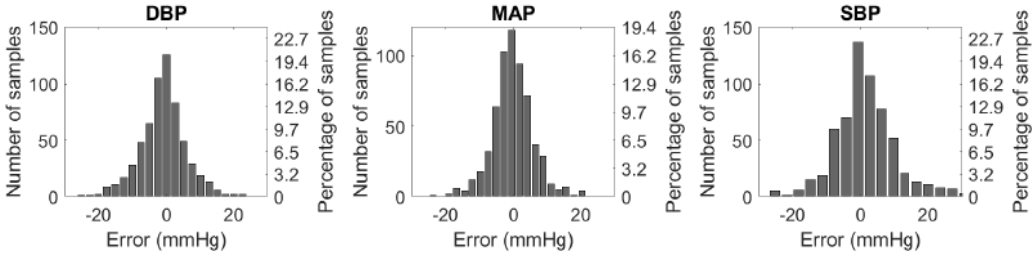


Figure 9: Error in DBP, MAP and DBP predictions.

5. Discussion

The method presented in this paper corresponds to one of the few proposals that relies on deep learning to estimate blood pressure from facial video. We propose, in the next subsection, to discuss and compare our results with related works. Section 5.2 presents the limitations of this study. We ultimately present and discuss the results of a leave-one-patient-out cross-validation procedure (section 5.3).

5.1. About the results presented in this study

Regarding previous works, and to the best of our knowledge, only Rong and Li presented Bland-Altman representations to assess their results. The technique proposed by the authors seems to underestimate low BP values and overestimate high BP values, both for DBP and SBP [9]. Our results depict a similar tendency but with lesser impact, the Bland-Altman plots presented in figure 7 being quite consistent across all the BP range. Table 2 presents a comparative analysis of results taken from similar works. The technique proposed in this study performs better than the other methods in

terms of MAE and RMSE. We, however, emphasize that the results reported from other studies were computed from data of different nature. To the best of our knowledge, these data are not publicly available.

Results presented in sections, 4.2 and 4.3 exhibit a relevant level of agreement between predicted and ground truth BP values. It can however be observed that several BP predictions exceed the 15 mmHg threshold, in particular for SBP (see table 3). We emphasize that no other techniques focusing on the analysis of BP from a single facial video have obtained grade B in SBP prediction, in particular from challenging data. Techniques dedicated to the conversion of contact PPG signals to the BP waveform [26] or from contact PPG signals to DBP and SBP values [30, 28, 29] also produce SBP estimations that are less relevant than DBP estimations. We do not report the AAMI and BHS analysis from Schrumpf et al. because none of their results seems to satisfy the requirements [10].

Integrating the wavelet representation of iPPG signals instead of raw iPPG signals in the network is a key-point of the method presented in this study. We here take advantage of transfer learning through a ResNeXt backbone pre-trained on large databases [11]. U-Nets have been widely used for segmentation of medical images and can be trained with a low volume of data [41].

5.2. Limitations

Figure 10 presents a prediction of lesser quality where the mean BP value is approximately estimated by the model. Apart from the mean error, DBP and SBP seem to be properly estimated. Adding more data during the learning phase of the network may solve, or at least minimize, this mean error. Balancing the distribution of ground truth BP values while varying the iPPG and BP waveform (shape of the signals) may be a relevant approach to tackle this issue.

All the presented results are limited by the current dataset: a low percentage of subjects (<85) has been used to derive the results presented in section 4. We point out that the reference blood pressure, gathered using a continuous non-invasive sensor, has not been recorded in a clinical setting. There might be irrelevant ground truth values, ultimately leading to improper BP learning by the U-Net model presented in section 3.3. We also emphasize that only videos presenting clear iPPG signals have been included in the dataset. Videos with motion can lead to iPPG signals that contain strong artifacts. This particular source of noise can negatively impact the

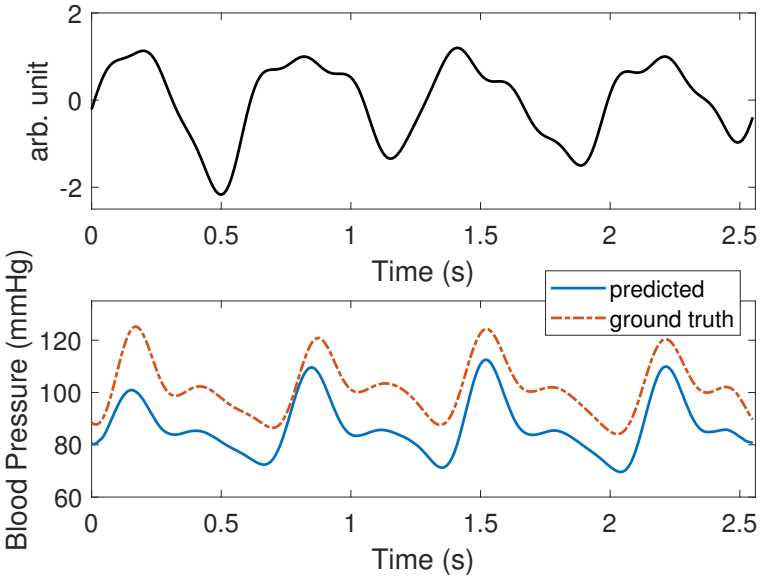


Figure 10: Predictions of lesser quality. Top figure: iPPG signal. Bottom figure: predicted and ground truth BP.

CWT coefficients. Including noisy iPPG signals into the dataset will be the objective of future works. Broaden the currently limited dataset is necessary so that all types of noise are represented.

The data distributions presented in figure 2 are not well-balanced across all the BP range. This can drastically impact training, in particular by negating the generalization power of the model (see next subsection). To tackle this issue, the development of a smart overlapping selection could be a potential approach. It would consist in automatically increasing the overlapping to produce more signals in the underrepresented BP ranges. We also emphasize that data augmentation strategies were recently proposed in the field of pulse rate estimation from video to improve the models performances [49]. These approaches are however not conceivable in the case of BP estimation because removing frames or augmenting the videos with conventional transformations may directly impact the shape of iPPG waveforms. Developing an augmentation strategy towards the wavelet representations, by for example adding random noise to the CWT coefficients, can here be an approach of interest.

The distributions of the training, validation and test sets presented in figure 2 contain a mix of all the participants data. In the next subsection,

we analyze the impact of a leave-one-patient-out cross-validation procedure on the performances.

5.3. Leave-one-patient-out cross-validation

Table 5 presents the assessment of BP using the method proposed in this study (section 3) but under a leave-one-patient-out cross-validation procedure (three folds). We can observe a decrease in performances over all the folds, even if some values are close from the international standards recommendations. The Bland-Altman representations for DBP, MAP and SBP over all data from the first fold are presented in figure 11. They exhibit wider point clouds than those computed from the randomly distributed subsets (see the Bland-Altman plots presented in figure 7) where each set includes a balanced portfolio of participants and tasks (details in section 3.1). We can also observe that SBP predictions depicted in figure 11 follow an inverse trend than those displayed in figure 7. Here, the trained model overestimates SBP in low BP values and underestimates SBP in high BP values. All these results exhibit a limitation in the generalization power of the network but are, in contrast, encouraging because the model has been trained with limited data.

It can also be observed, from table 5, that the model performed poorly for SBP estimations of fold 2. After a closer look on the iPPG signals and ground truth BP, we remarked that this decrease in performance was due to a patient who presents the highest SBP values. All these patient signals were included in the test and were therefore totally missing from the training set. We therefore believe that the network did not learn the features relative to these specific samples. As stated in the previous subsection, broadening the dataset is a necessary step to improve generalization.

6. Conclusion and future works

We proposed, in this article, a deep learning-oriented solution dedicated to the recovering of blood pressure from facial video. The reconstruction is carried out using a U-shaped network supported by a ResNeXt backbone from the time-frequency representation of the iPPG signal. To the best of our knowledge, this study presents the first demonstration of an automatic framework that manages to estimate the continuous BP waveform from facial video. The approach corresponds to an efficient way for predicting BP without a prior extraction of complicated hand-crafted waveform features from

Fold	Errors (mmHg)		BHS			AAMI (mmHg)		
	MAE	RMSE	≤ 5	≤ 10	≤ 15	ME	SDE	
1	8.28	11.78	49%	73%	83%	4.23	10.99	DBP
	7.52	10.66	50%	76%	86%	3.39	10.11	MAP
	9.79	12.64	33%	61%	76%	4.56	11.79	SBP
2	5.83	7.12	48%	85%	97%	0	7.12	DBP
	8.03	10.24	43%	65%	87%	3.97	9.45	MAP
	16.41	21.61	26%	46%	57%	12.99	17.27	SBP
3	11.43	14.12	28%	51%	68%	-4.77	13.29	DBP
	8.11	10.21	38%	69%	86%	-3.81	9.47	MAP
	8.87	11.33	38%	62%	81%	-4.77	10.28	SBP

Table 5: Assessment of the proposed solution under a leave-one-patient-out cross-validation procedure.

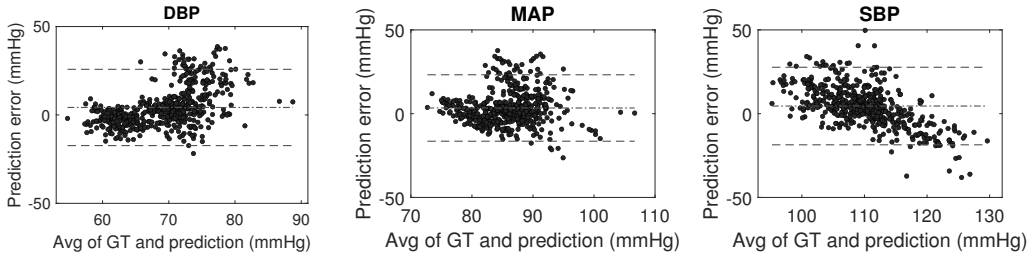


Figure 11: Bland-Altman plots for DBP, MAP and SBP prediction under a leave-one-patient-out, three folds, cross-validation procedure (only the results from the first fold are presented here).

the iPPG signal. Our extensive experiments showed the effectiveness of the proposed method, which achieves high accuracy and satisfies all international standards in the estimation of mean and diastolic BP (grade A) and nearly all international standards in the estimation of systolic BP (grade B).

Several ways of improvement for this work are considered. We first propose expanding the currently limited volume of data by increasing the number of included recordings and participants. We, in this study, conducted a manual selection of videos that presented well-defined iPPG signals. This step can be automatized using a quality index [17]. Also, it has recently been shown that data augmentation strategies can significantly improve the performances of deep learning models dedicated to pulse rate estimation from video [49]. Producing more overlapped signals in the range of low represented BP values might be a first considered approach for re-balancing the dataset distribution.

The Morlet wavelet has been used as a prototype function for the computation of the CWT. We propose evaluating the impact on performances with different mother wavelets as well as investigating different time-frequency representations like short-time Fourier and constant-Q transforms.

Inputting directly the video stream in an end-to-end architecture rather than the time-frequency representation of iPPG signal will be the subject of long-term research. We also envisage to extend this work in the context of blood oxygen saturation using a similar approach (inputting CWT representations of iPPG signals to a deep U-Net model).

7. Acknowledgments

This work has been partly funded by the Contrat Plan État Région (CPER) Innovations Technologiques, Modélisation et Médecine Personnalisée (IT2MP) and Fonds Européen de Développement Régional (FEDER).

References

[1] D. McDuff, Camera measurement of physiological vital signs, arXiv preprint arXiv:2111.11547 (2021).

[2] S. Zaunseder, A. Trumpp, D. Wedekind, H. Malberg, Cardiovascular assessment by imaging photoplethysmography–a review, Biomedical Engineering/Biomedizinische Technik (2018).

[3] A. Ni, A. Azarang, N. Kehtarnavaz, A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods, Sensors 21 (2021) 3719. URL: https://www.mdpi.com/1424-8220/21/11/3719. doi:10.3390/s21113719.

[4] C.-H. Cheng, K.-L. Wong, J.-W. Chin, T.-T. Chan, R. H. Y. So, Deep Learning Methods for Remote Heart Rate Measurement: A Review and Future Research Agenda, Sensors 21 (2021) 6296. URL: https://www.mdpi.com/1424-8220/21/18/6296. doi:10.3390/s21186296.

[5] A. Al-Naji, G. A. Khalid, J. F. Mahdi, J. Chahl, Non-Contact SpO2 Prediction System Based on a Digital Camera, Applied Sciences 11 (2021) 4255. URL: https://www.mdpi.com/2076-3417/11/9/4255. doi:10.3390/app11094255.

[6] H. Luo, D. Yang, A. Barszczyk, N. Vempala, J. Wei, S. J. Wu, P. P. Zheng, G. Fu, K. Lee, Z.-P. Feng, Smartphone-based blood pressure measurement using transdermal optical imaging technology, Circulation: Cardiovascular Imaging 12 (2019) e008857.

[7] N. Sugita, M. Yoshizawa, M. Abe, A. Tanaka, N. Homma, T. Yambe, Contactless Technique for Measuring Blood-Pressure Variability from One Region in Video Plethysmography, Journal of Medical and Biological Engineering (2018) 1–10.

[8] X. Fan, Q. Ye, X. Yang, S. D. Choudhury, Robust blood pressure estimation using an RGB camera, Journal of Ambient Intelligence and Humanized Computing (2018) 1–8.

[9] M. Rong, K. Li, A Blood Pressure Prediction Method Based on Imaging Photoplethysmography in combination with Machine Learning, Biomedical Signal Processing and Control 64 (2021) 102328. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809420304444. doi:10.1016/j.bspc.2020.102328.

[10] F. Schrumpf, P. Frenzel, C. Aust, G. Osterhoff, M. Fuchs, Assessment of Non-Invasive Blood Pressure Prediction from PPG and rPPG Signals Using Deep Learning, Sensors 21 (2021) 6022. URL: https://www.mdpi.com/1424-8220/21/18/6022. doi:10.3390/s21186022.

[11] F. Bousefsaf, D. Djeldjli, Y. Ouzar, C. Maaoui, A. Pruski, iPPG 2 cPPG: reconstructing contact from imaging photoplethysmographic signals using U-Net architectures, Computers in Biology and Medicine 138 (2021) 104860. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482521006545. doi:10.1016/j.compbiomed.2021.104860.

[12] Y. Lu, C. Wang, M. Q.-H. Meng, Video-based Contactless Blood Pressure Estimation: A Review, in: 2020 IEEE International Conference on Real-time Computing and Robotics (RCAR), IEEE, Asahikawa, Japan, 2020, pp. 62–67. URL: https://ieeexplore.ieee.org/document/9303040/. doi:10.1109/RCAR49640.2020.9303040.

[13] N. Sugita, K. Obara, M. Yoshizawa, M. Abe, A. Tanaka, N. Homma, Techniques for estimating blood pressure variation using video images, in: Engineering in Medicine and Biology Society (EMBC), 2015 37th

Annual International Conference of the IEEE, IEEE, 2015, pp. 4218–4221.

[14] I. C. Jeong, J. Finkelstein, Introducing contactless blood pressure assessment using a high speed video camera, Journal of medical systems 40 (2016) 77.

[15] P.-W. Huang, C.-H. Lin, M.-L. Chung, T.-M. Lin, B.-F. Wu, Image based contactless blood pressure assessment using Pulse Transit Time, in: Automatic Control Conference (CACS), 2017 International, IEEE, 2017, pp. 1–6.

[16] D. Shao, Y. Yang, C. Liu, F. Tsow, H. Yu, N. Tao, Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time, IEEE Transactions on Biomedical Engineering 61 (2014) 2760–2767.

[17] X. Fan, T. Tjahjadi, Robust contactless pulse transit time estimation based on signal quality metric, Pattern Recognition Letters 137 (2020) 12–16.

[18] N. Sugita, T. Noro, M. Yoshizawa, K. Ichiji, S. Yamaki, N. Homma, Estimation of Absolute Blood Pressure Using Video Images Captured at Different Heights from the Heart, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 4458–4461.

[19] M. Elgendi, On the analysis of fingertip photoplethysmogram signals, Current cardiology reviews 8 (2012) 14–25.

[20] E. von Wowern, G. Östling, P. M. Nilsson, P. Olofsson, Digital photoplethysmography for assessment of arterial stiffness: repeatability and comparison with applanation tonometry, PloS one 10 (2015) e0135659.

[21] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, A. Pruski, Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera, Biomedical Signal Processing and Control 64 (2021) 102242.

[22] M. Jain, S. Deb, A. Subramanyam, Face video based touchless blood pressure and heart rate estimation, in: Multimedia Signal Processing

(MMSP), 2016 IEEE 18th International Workshop on, IEEE, 2016, pp. 1–5.

[23] C. G. Viejo, S. Fuentes, D. D. Torrico, F. R. Dunshea, Non-Contact Heart Rate and Blood Pressure Estimations from Video Analysis and Machine Learning Modelling Applied to Food Sensory Responses: A Case Study for Chocolate, Sensors 18 (2018) 1802.

[24] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Optics express 16 (2008) 21434–21445.

[25] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic Principles of Remote PPG, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491.

[26] N. Ibtehaz, M. S. Rahman, PPG2ABP: Translating Photoplethysmogram (PPG) Signals to Arterial Blood Pressure (ABP) Waveforms using Fully Convolutional Neural Networks, arXiv preprint arXiv:2005.01669 (2020).

[27] M. S. Tanveer, M. K. Hasan, Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ANN-LSTM network, Biomedical Signal Processing and Control 51 (2019) 382–392.

[28] M. Panwar, A. Gautam, D. Biswas, A. Acharyya, PP-Net: A Deep Learning Framework for PPG based Blood Pressure and Heart Rate Estimation, IEEE Sensors Journal (2020). Publisher: IEEE.

[29] M. H. Chowdhury, M. N. I. Shuzan, M. E. Chowdhury, Z. B. Mahbub, M. M. Uddin, A. Khandakar, M. B. I. Reaz, Estimating Blood Pressure from the Photoplethysmogram Signal and Demographic Features Using Machine Learning Techniques, Sensors 20 (2020) 3127. Publisher: Multidisciplinary Digital Publishing Institute.

[30] G. Slapničar, N. Mlakar, M. Luštrek, Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network, Sensors 19 (2019) 3420. Publisher: Multidisciplinary Digital Publishing Institute.

[31] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[32] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, others, Multimodal spontaneous emotion corpus for human behavior analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3438–3446.

[33] G. de Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886.

[34] A. Hammer, M. Scherpf, M. Schmidt, H. Ernst, H. Malberg, K. Matschke, A. Dragu, J. Martin, O. Bota, Camera-based assessment of cutaneous perfusion strength in a clinical setting, Physiological Measurement (2022). URL: http://iopscience.iop.org/article/10.1088/1361-6579/ac557d.

[35] M. P. Tarvainen, P. O. Ranta-Aho, P. A. Karjalainen, An advanced detrending method with application to HRV analysis, IEEE transactions on biomedical engineering 49 (2002) 172–175. Publisher: IEEE.

[36] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE transactions on biomedical engineering 58 (2011) 7–11.

[37] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, G. Medioni, On face segmentation, face swapping, and face perception, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 98–105.

[38] Y. Ouzar, D. Djeldjli, F. Bousefsaf, C. Maaoui, LCOMS Lab's Approach to the Vision for Vitals (V4V) Challenge, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2750–2754.

[39] F. Bousefsaf, C. Maaoui, A. Pruski, Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate, Biomedical Signal Processing and Control 8 (2013) 568–574.

[40] F. Bousefsaf, C. Maaoui, A. Pruski, Peripheral vasomotor activity assessment using a continuous wavelet analysis on webcam photoplethysmographic signals, Bio-medical materials and engineering 27 (2016) 527–538.

[41] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, others, Deep learning for segmentation using an open large-scale dataset in 2d echocardiography, IEEE transactions on medical imaging (2019).

[42] E. C. Too, L. Yujian, S. Njuki, L. Yingchun, A comparative study of fine-tuning deep learning models for plant disease identification, Computers and Electronics in Agriculture 161 (2019) 272–279. Publisher: Elsevier.

[43] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[44] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[45] P. Yakubovskiy, Segmentation Models, GitHub, 2019. URL: https://github.com/qubvel/segmentation_models, publication Title: GitHub repository.

[46] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[47] E. O'Brien, J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, K. O'Malley, M. Jamieson, D. Altman, M. Bland, N. Atkins, The british hypertension society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems., Journal of hypertension 8 (1990) 607–619.

[48] G. S. Stergiou, B. Alpert, S. Mieke, R. Asmar, N. Atkins, S. Eckert, G. Frick, B. Friedman, T. Graßl, T. Ichikawa, others, A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement, Hypertension 71 (2018) 368–374. Publisher: Am Heart Assoc.

[49] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching, IEEE Signal Processing Letters 27 (2020) 1245–1249. URL: https://ieeexplore.ieee.org/document/9133501/. doi:10.1109/LSP.2020.3007086.

iPPG 2 cPPG: Reconstructing contact from imaging photoplethysmographic signals using U-Net architectures

Frédéric Bousefsaf, Djamaleddine Djeldjli, Yassine Ouzar, Choubeila Maaoui, Alain Pruski

iPPG 2 cPPG: reconstructing contact from imaging photoplethysmographic signals using U-Net architectures

Frédéric Bousefsaf*, Djamaleddine Djeldjli, Yassine Ouzar, Choubeila Maaoui and Alain Pruski

Université de Lorraine, LCOMS, F-57000 Metz, France

Abstract

Imaging photoplethysmography (iPPG) is an optical technique dedicated to the assessment of several vital functions using a simple camera. Significant efforts have been made to reliably estimate heart and respiratory rates. Currently, research is focusing on the remote estimation of oxygen saturation and blood pressure (BP). The limited number of publicly available data tends to restrict the advancements related to BP estimation. To overcome this limit, we propose to split the problem in a two-stage processing chain: (i) converting iPPG to contact PPG (cPPG) signals using available video dataset and (ii) estimate BP from converted cPPG signals by exploiting large existing databases (e.g. MIMIC). This article presents the first developments where a method for converting iPPG signals measured using a camera into cPPG signals measured by contact sensors is proposed. Real and imaginary parts of the continuous wavelet transform (CWT) of cPPG and iPPG signals are passed to various deep pre-trained U-shaped architectures. Conventional metrics and specific waveform estimators have been implemented to validate the relevance of the predictions. The results exhibit good agreements towards a large portion of metrics, showing that the neural architectures properly estimated cPPG from iPPG signals through their CWT representations. The performance indicates that BP estimation from iPPG signals converted to cPPG signals can now be envisaged. Consequently, future work will focus on the integration of models dedicated to BP estimation trained on MIMIC. This is the first demonstration of a method for accurate reconstruction of cPPG from iPPG signals satisfying pulse waveform criteria.

Keywords: imaging photoplethysmography, U-Net, blood volume pulse,

1. Introduction

In the recent years, research on contactless technologies dedicated to physiological signals measurement have made significant progress [1]. Photoplethysmography (PPG) can be remotely measured by observing the subtle fluctuations of skin color. These fluctuations reflect complex light-tissue interactions, from which their origin is not fully agreed [2]. The simplest cameras (webcams) to the most advanced ones (professional, laboratory or industrial cameras) can be used to reliably measure PPG signals [3]. Different regions of interest (ROI) have been studied over time but the face remains the most frequently observed area [4].

The field is booming and supported by several significant studies. Computer vision, image processing and artificial intelligence (AI) methods have been used or developed specifically to reliably transform input video into biomedical parameters [4]. Numerous studies have shown that pulse rate and its variability can be estimated with high robustness. In this context, artificial intelligence is playing an increasingly important role [5] where the most efficient pulse rate measurement methods are now based on deep neural models [6]. These architectures are often based on convolutional layers [7] and can be trained with synthetic data [8] reinforced by real data [9].

Current research in this field is now directed towards the measurement of new physiological parameters such as oxygen saturation [10] and blood pressure [11]. They impact the amplitude and waveform of PPG signals over

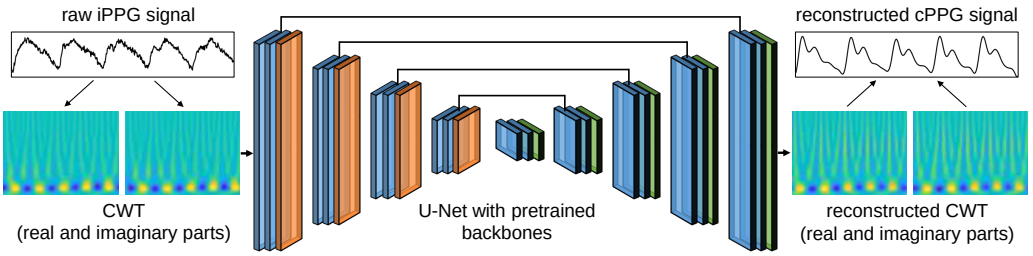


Figure 1: General overview of the method.

different wavelength ranges. Blood pressure estimation based on video analysis is complex and very few works show its feasibility. Two research directions are considered. First, measurement of the pulse transit time (PTT) on single [12] or several [13] ROI. PTT is a parameter considered to be correlated with blood pressure. Secondly, analysis of the PPG signal waveform [11]. To our knowledge, deep learning techniques based on video analysis have not been considered for the estimation of blood pressure yet.

Training an artificial neural network that accurately estimates blood pressure from video is constrained by the amount of available data because few public databases exist. Djeldjli et al. recently showed that temporal, derivative and area features computed from imaging PPG (iPPG) waveform and contact sensor (placed on the finger or the ear) evolve similarly [14]. This point is important because it motivates the present study. We envisage estimating BP with a two-stage processing chain. A model dedicated to the conversion of iPPG signals to contact PPG (cPPG) signals using available video dataset corresponds to the first part of the processing chain. The second stage consists in constituting a deep learning model dedicated to blood pressure estimation from these converted signals by exploiting large existing databases (e.g. MIMIC [15]).

The developments related to the first stage are presented in this study. To add more details, we propose to train a deep U-shaped neural architecture (U-Net) dedicated to the conversion of contact PPG signals from imaging PPG signals simultaneously measured on the face by conventional video analysis. Continuous wavelet representation of the signals is employed to take advantage of transfer learning through pretrained backbones on large databases. To the best of our knowledge, this is the first demonstration of a method for accurate reconstruction of cPPG from iPPG signals.

The article includes five additional sections. Section 2 presents the background and related works. Section 3 introduces the used data and the developed methodologies. The metrics and results of the proposed approach are presented and discussed in section 4. We present the future works and a summary of the contributions in sections 5 and 6, respectively.

2. Related works

This section reviews the studies that exploit deep learning for iPPG analysis as well as conventional and deep learning approaches for blood pressure assessment from both iPPG and cPPG.

2.1. Deep Learning for iPPG signal and pulse rate estimation

Relevant surveys in the imaging PPG field of research have been proposed the last past years [1, 3, 4]. They cover conventional techniques that generally include both image and signal processing approaches to improve PPG signal-to-noise ratio and therefore the estimation of biomedical parameters like pulse and breathing rates. Video and image processing operations like face detection, tracking of region(s) of interest and skin segmentation have been employed [16, 17, 18]. Constituting an iPPG signal from a sequence of frames is usually carried out with a spatial averaging operation [19]). Standard signal processing techniques include blind source separation approaches [20], Fourier and Wavelet transforms [21]. The impact of color space on pulse rate assessment has also been investigated in previous research [22, 17].

The most recent studies present artificial intelligence through deep learning methods to automatically estimate the pulse signal or directly the pulse rate. These approaches currently deliver the best performances and present root mean squared errors between 2.7 and 3.8 beats per minute [5] on public datasets like UBFC-RPPG [23], MAHNOB-HCI [24] and PURE [25]. Both hybrid and end-to-end approaches have been investigated. Hybrid strategies take either processed frames or iPPG signals as input and output the biomedical parameters of interest. End-to-end models takes a video (sequence of frames) as input and output the biomedical parameters.

Hybrid strategies combine conventional with deep learning methods. For instance, Qiu et al. developed a three-stage pipeline including face tracking, features extraction and finally pulse rate estimation based on a convolutional neural network (CNN) [26]. Hsu et al. proposed a deep CNN trained to predict pulse rate based on the time–frequency representation of processed iPPG signals [27]. Chen et al. proposed DeepPhys [28] and DeepMag [29], deep CNN trained to respectively predict pulse wave and magnify color variations produced by the periodic changes in blood flow. Inputs are transformed using a skin reflection model while the convolutional layers are guided using attention masks to ensure the robust estimation of PPG signals under lighting fluctuation and motion. They used a modified version of VGG, a model dedicated to object recognition in images [30].

End-to-end strategies were recently investigated through different neural architectures: CNN-based extractor and estimator [31], 3D CNN [8, 32, 33], combination of CNN and long short-term memory [32, 34], CNN and gated recurrent unit [35], Siamese network including two branches with identical structure that analyze two different facial regions [36] and temporal difference

convolution [6]. These models have been trained with synthetic data [8] reinforced by real data [9, 33]. They estimate the pulse signal [32] or directly the pulse rate from a sequence of images.

Few studies investigated the interpretability and behavior of the models to understand the representations learned by the features. Zhan et al. studied this aspect by analyzing that CNN properly learn PPG during training [7]. They conclude that color variations produced by blood flow fluctuations are correctly exploited by the neural networks.

2.2. Blood pressure assessment from iPPG

Both systolic and diastolic blood pressures (BP) have been estimated using the propagation time of pulse waves from two different skin areas (typically hand and face) in video recordings [37, 38]. The positional of the two skin areas must be maintained during the measurement. This approach is therefore very restrictive. The scientific literature covers few studies dedicated to the estimation of BP from a single facial region [39, 12, 40, 41]. To the best of our knowledge, only the seminal work from Luo et al. [11] presents a pipeline that includes an artificial intelligence model. They feed a multilayer perceptron with 155 features (reduced to 30 after principal component analysis) computed from iPPG waves. Their results show that PPG waveform extracted from video exhibits information that relates to BP. All these studies pointed out the feasibility of remote BP monitoring from facial video but showed that there is still room for improvements.

2.3. Blood pressure assessment from cPPG

Based on the current literature, there is clear evidence that the fluctuations in BP are reflected in cPPG signals [42, 43] even if estimating absolute BP values from cPPG remains a challenging problem. The changes in morphological contours due to interaction of other physiological systems make the extraction of features, and thus the estimation of BP, challenging but achievable [44]. Exploration of deep learning techniques is here particularly interesting because it allows overriding of handcrafted features [45]. These features are somewhat restricted because the cPPG waveform fluctuates from subject to subject and also because the filtering procedure can change its morphology [46].

Several recent studies show that deep learning frameworks can effectively be deployed to translate BP from cPPG signals. Tanveer and Hasan proposed to associate artificial neural network (ANN) with long short-term memory

for BP estimation [47]. A similar network structure was proposed by Panwar et al. in 2020 [48]. 1D CNN replace the ANN part from Tanveer and Hasan architecture. The network concurrently estimates diastolic BP, systolic BP and heart rate from a single cPPG signal. Chowdhury et al. then proposed to employ machine learning algorithms dedicated to BP estimation using cPPG signal and demographic features (e.g. weight and height) [49]. Time, frequency and time-frequency features were extracted from the PPG and their derivative signals. Feature selection techniques were used for reducing the computational complexity and simultaneously decreasing the chance of over-fitting the machine learning algorithms. Slapnicar et al. introduced a similar framework but with a deep neural network architecture with residual connections [50]. A part of the network is dedicated to the analysis of the signal spectral representation using gated recurrent units. Ibtehaz et Raman employed a deep learning based method that manages to predict the continuous BP waveform from cPPG signals. An approximation network learns a rough approximation of the BP waveform while a refinement network further enhances the preliminary estimate. The approximation and refinement networks are based on U-Net [51].

3. Methods

3.1. Database and experimental protocol

The data used to learn the neural models (section 3.3) have been presented in a previously published article [14]. 12 volunteers aged between 20 and 35 years participated to the study. The experiments were conducted in a dark room where the only source of light was two Neewer LED panels (NL480) set to 2700 lux / m with a color temperature of 3750 K (neutral white light). During the experiments, they were asked to seat at approximately 1 meter from a fast camera (16mm C Series Lens mounted on a EO-2223C Color camera from Edmund Optics). The recorded sequences of RGB images were save without compression at resolution 640 $\times$ 480 pixels (24 bits per pixel) and with a frame rate of 125 frames per second. Autoexposure and white balance have been disabled.

The ground truth cPPG signals were recorded using approved contact probes (BVP-Flex / Pro. By Thought Technologies Ltd.) placed on the finger and the ear. Two 60-second videos were recorded for every participant. First video: participants were asked to stay calm and breathe normally. Second video: participants were asked to hold their breath as much as possible, the

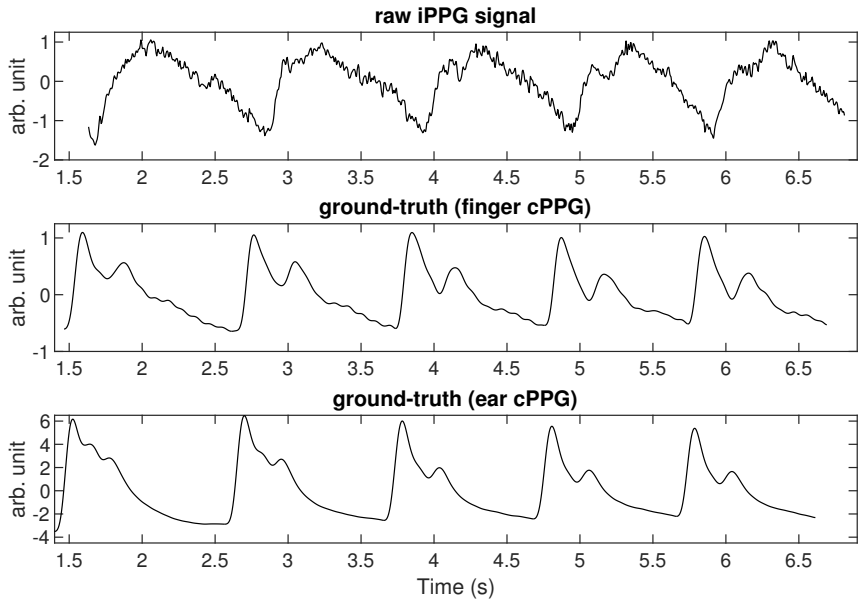


Figure 2: Excerpts of participant #1 (collected during breath holding experiment). Top figure: raw iPPG signals computed with a spatial averaging operation over the forehead region [19]. Video recordings have been collected using a fast camera (125 frames per second). Reference cPPG signals have been recorded with contact probes placed on the finger (middle figure) and the ear (bottom figure).

objective being to cause physiological variations that modify blood pressure and impact the recorded PPG signals. We refer the reader to the original publication for more details concerning the procedure and the material used [14].

The database contains 724 signals. Each of them contains 5 PPG waves (more details in section 3.2) defined over 256 values. About 80% of the data (600 signals) were reserved for training and 20% (124 signals) for testing. The sets contain a balanced portfolio of the different participants and tasks. We evaluated the models relevance through k-fold cross-validation (k=5). A fold contains 120 signals that are reserved for validation. The 4 remaining folds include 480 signals that are employed for training the neural models.

3.2. Image and signal processing

The forehead corresponds to a relevant area of interest in terms of signal-to-noise ratio [17]. This region has been automatically detected with a model

composed of 68 points positioned on the main shapes of the face [52]. These different points are tracked along the video. Some of them are used to find the position of the forehead. In practice, algorithms for face and facial landmarks detection included in OpenCV [1] and Dlib [2] libraries have been employed.

iPPG signals are computed by averaging all the forehead pixels from the green channel. This technique has been used since the very first publications related to the measurement of contactless PPG signals by camera [19]. The raw iPPG signals are then detrended using a specific low-pass filter [53] based on a smoothness priors that attenuates low frequencies [20]. We then robustly detect the valleys to extract each PPG signal wave. Each signal is ultimately sampled over 256 points and contains 5 successive iPPG waves. An excerpt is presented in figure 2. The ground truth cPPG signals measured at the finger and the ear are also presented in this figure. All the signals have been standardize ($\mu = 0$ and $\sigma = 1$).

In this article, we propose to exploit the wavelet representation of PPG signals to train the different neural architectures presented in section 3.3 (figure 1). The continuous wavelet transform (equation 1) of a signal $x(t)$ corresponds to a time-frequency representation computed from a prototype function commonly called mother wavelet. Unlike the Fourier transform, the wavelet transform can detect abrupt changes in frequency using a family of wavelets $\psi_{\tau,s}$ (equation 2) computed from the mother wavelet ψ.

$$CWT_x^{\psi}(\tau, s) = \int_{-\infty}^{\infty} x(t)\, \psi_{\tau,s}(t) dt \tag{1}$$

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right) \tag{2}$$

$\psi_{\tau,s}$ corresponds to the mother wavelet dilated by s and translated by τ. Dilating the wavelet allows the transform to analyze larger portions of signal in the time domain, thus covering lower frequencies. Different mother wavelets have been developed and the choice depends mainly on the application and the properties of the signal. The Morlet mother wavelet used in this study was already used in previous work related to the analysis of PPG signals by camera [54].

[1] https://opencv.org/
[2] http://dlib.net/

The original signal $x(t)$ can be reconstructed by the inverse transform:

$$x(t) = \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty \frac{1}{s^2} CWT_x^\psi(\tau, s) \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-\tau}{s}\right) d\tau \, ds \tag{3}$$

$$C_\psi = \int_0^\infty \frac{\left|\hat{\psi}(\zeta)\right|^2}{|\zeta|} d\zeta < \infty \tag{4}$$

C_ψ is the admissibility condition and $\hat{\psi}$ is the Fourier transform of ψ.

The continuous wavelet transform was computed on each PPG signal in the frequency range [0.6, 4.5] Hz, which corresponds to the physiological range of the human heart rate [4]. Wavelet representations of dimension 256×256 will be used to train the neural architectures presented in section 3.3.

Typical iPPG signal, cPPG signal and their respective wavelet representations (real, imaginary and absolute part) are presented in figure 3. A typical difference in shape between both signals and in phase between their wavelet representations can be noted: the real part of the iPPG signal starts with a series of low intensity coefficients (blue pseudo-ellipse) while the real part of the cPPG signal starts with strong intensity coefficients (yellow pseudo-

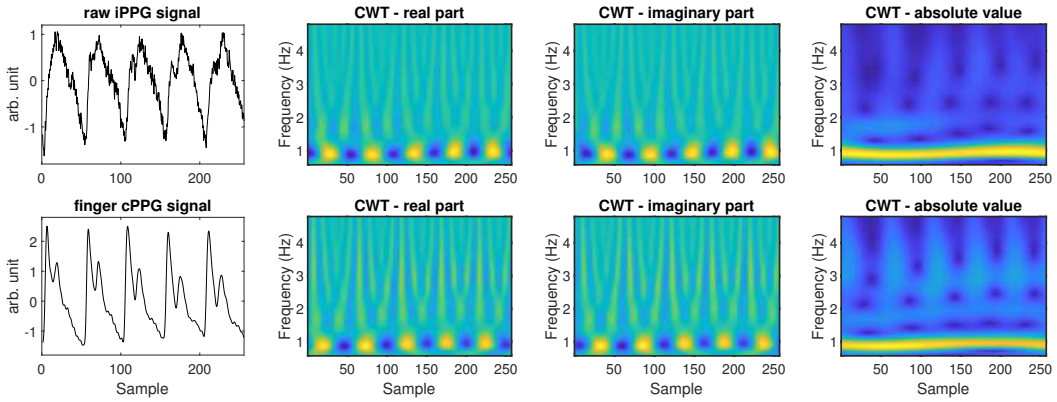


Figure 3: The continuous wavelet transform of the iPPG signal (top figure) and cPPG signal (ear or finger, see bottom figure for a finger cPPG signal) is computed in the frequency range [0.6, 4.5] Hz. The wavelet representation of the iPPG signal (a complex image with a real and imaginary part) serves as input for training the neural networks presented in section 3.3. The absolute of the continuous wavelet transform is depicted for information and is not learned by the model.

ellipse). The neural network will learn this specificity during the training phase.

3.3. Neural architectures

The U-Net neural architecture was initially proposed by Ronneberger et al. [51]. This network has been used for segmentation of medical images [55]. Its architecture consists of a descending (encoder) branch completed by an ascending (decoder) branch, giving a U-shape to the network. The descending branch contains an ensemble of convolution and pooling layers. The ascending branch integrates deconvolution layers connected to the convolutions of the descending branch. Connections help to restore the spatial information. A schematic representation of the network is given in figure 1. In this study, we employ the U-Net1 version proposed by Leclerc et al. [55]. The model hyperparameters vary slightly compared to the original version proposed by Ronneberger et al. Details are presented in table 1. The number of filters is given for the first and for the last convolutional block as well as at the center of the network, where the spatial information is most compressed. Each convolutional layer integrates a core $(3, 3)$ coupled to a Rectified Linear Unit (ReLU) activation function.

A Backbone (e.g. VGG16) can be integrated into the encoder part of the U-Net network (figure 4). Its internal parameter are blocked during training (the weights of the network remain fixed). In practice, a backbone correspond to a model subpart pre-trained on ImageNet, a database deployed for object

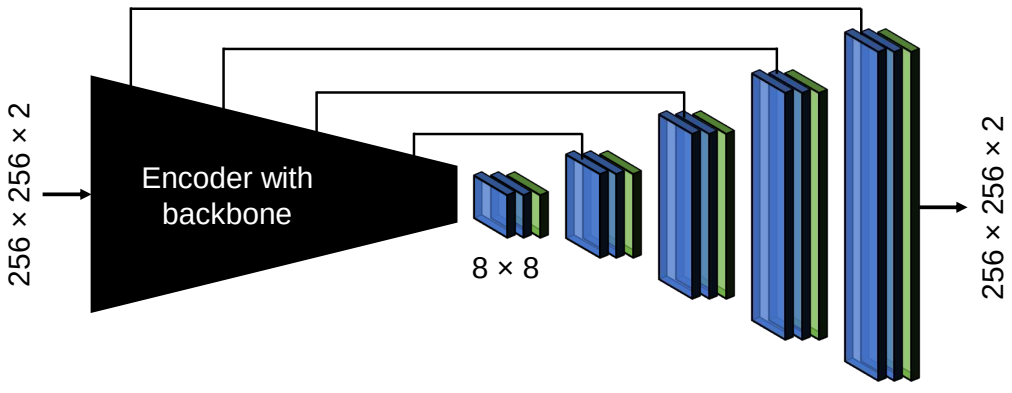


Figure 4: A backbone corresponds to a pre-trained network included in the encoder part of U-Net.

recognition tasks in images [61]. Training a U-Net network supported by a backbone consists in optimizing the internal parameters of the decoder part. This approach can be associated to a transfer learning strategy.

The various backbones tested and their main characteristics are summarized in table 1. VGG [30] is a model composed of $(3, 3)$ convolutional layers and pooling layers. The 16-layer version (VGG16) was used in this study. ResNet [56] are neural modules nested in a larger network (network-in-network) through residual units composed of convolutional filters. The architecture is about 8 times deeper than VGG. ResNet models at different depth levels (18, 34, 50, 101 and 152 layers) were trained on the ImageNet database but only the 101 layers was used in this study. DenseNet networks [60] include Dense blocks that are densely connected together: each layer is directly connected with the following ones. Thus, the input vector of a given layer integrates all the characteristics of those that precede it. The 201-layer version was chosen. Inception networks [59] contain modules composed of convolution and pooling layers of different sizes. The InceptionV3 and InceptionResNetV2 versions (with residual connections) were used in this work.

Conventional regularization techniques (e.g. dropout) have not been introduced while a normalization scheme (i.e. batch normalization) is used in networks having a backbone. These details are summarized in table 1. No output activation function was specified because the targeted task corresponds to a regression in the form of a pixel-to-pixel reconstruction of a two-channel wavelet representation. The number of variables to be trained

Network	**Number of conv. filters**	**Lowest resolution**	**Normalization**	**Number of parameters**
U-Net1 [55]	$32 \downarrow 128 \uparrow 16$	8×8	$\varnothing$	2M
U-Net$_{\text{VGG16}}$ [30]	$64 \downarrow 512 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{VGG19}}$ [30]	$64 \downarrow 512 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{ResNet101}}$ [56]	$64 \downarrow 2048 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{ResNeXt101}}$ [57]	$64 \downarrow 2048 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{SE-ResNet101}}$ [58]	$64 \downarrow 2048 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{SE-ResNeXt101}}$ [58]	$64 \downarrow 2048 \uparrow 16$	8×8	BatchNorm	9M
U-Net$_{\text{InceptionResNetV2}}$ [59]	$32 \downarrow 2080 \uparrow 16$	8×8	BatchNorm	7.5M
U-Net$_{\text{InceptionV3}}$ [59]	$32 \downarrow 448 \uparrow 16$	8×8	BatchNorm	8M
U-Net$_{\text{DenseNet201}}$ [60]	$64 \downarrow 128 \uparrow 16$	8×8	BatchNorm	8.5M

Table 1: Main properties of the U-Net networks used in this study.

(weights and biases) is comprised between 2 and 9 million (table 1).

The input dimensions of networks with backbones are fixed by the data used for their training (256 × 256 pixels RGB images from the ImageNet database). The inputs being in our case two-channels wavelet representations, it is necessary to introduce an adaptation strategy. An additional 2D convolutional layer with a $(1, 1)$ kernel has therefore been placed between the input layer and the encoder part of the network. The neurons of this layer allow conversion of the input from N to 3 channels. The weights of all the networks have randomly been initialized by the method proposed by Glorot and Bengio [62]. Biases are initialized to zero. The Mean Squared Error (MSE) has been selected as loss for training all the models:

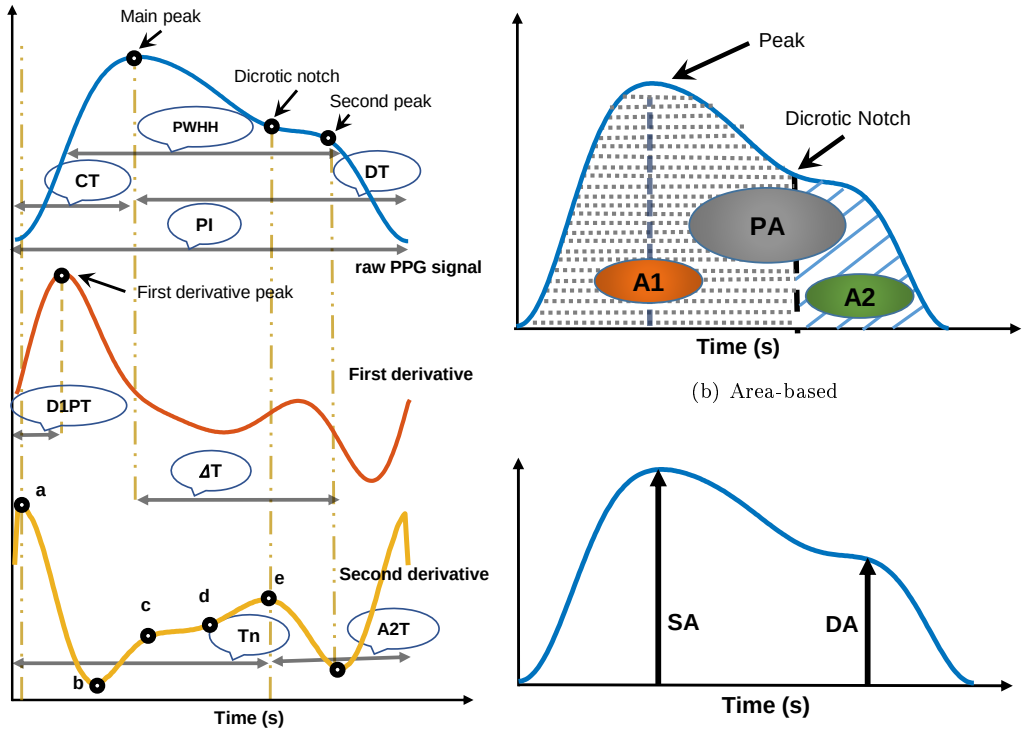
$$MSE = \frac{1}{n} \sum_{i,j} \left(CWT_{i,j} - \widehat{CWT}_{i,j} \right)^2 \tag{5}$$

CWT corresponds to the wavelet transform (see section 3.2) of the ground truth cPPG signal. $\widehat{CWT}$ is the wavelet representation predicted by the neural network starting from the wavelet representation of the iPPG signal.

The architecture implementation was carried out under Python using Keras API and Tensorflow library. The Segmentation Models library [63] proposed by P. Yakubovskiy was used to develop the neural networks presented in table 1. The training sessions were launched over 5000 epochs through batches of 16 images. We used, in this study, the Adam optimization algorithm [64] with a learning rate of 0.0001. A dedicated computer equipped with a dual Intel Xeon Silver 4114 and two Nvidia Quadro P6000s was used to carry out network learning.

3.4. Waveform estimators

Different features have been proposed to characterize the waveform of a PPG signal [42]. In order to validate the predictions of the neural architectures presented in the previous section, we propose to compare the estimates of the most commonly observed waveform features [42] [43] between the reconstructed PPG signal (computed using the inverse transform of the predicted wavelet representation) and the ground truth cPPG signal. It has recently been shown that some of these features can properly be estimated on iPPG signals [14], the contact and contactless waveform features evolving in a same way.



(a) Temporal and derivative-based

(b) Area-based

(c) Amplitude-based

Figure 5: Presentation of the features computed from a PPG wave. These parameters have been categorized in four groups. Temporal: Pulse Interval (PI), Crest Time (CT), Diastolic Time (DT), time between the main peak and the secondary peak (ΔT), Dicrotic Notch Time (Tn), Pulse Width at Half Height (PWHH), time between the dicrotic notch and the end of the wave (A2T) and First Derivative Peak Time (D1PT). Derivatives: a, b, c, d and e correspond to specific points that are detected on the second derivative. Area: Pulse Area (PA) and area computed between the start of the wave and the inflection point (A1) and between the inflection point and the end of the wave (A2). Amplitude: Systolic Amplitude (SA) and Diastolic Amplitude (DA).

Waveform features can be categorized into 4 families: temporal, amplitude-based, area-based, and (first and second) derivative-based. All features are presented in figure 5. We refer the reader to the article of Elgendi et al. [42] which details the PPG waveform features and their physiological interpretation.

3.4.1. Temporal features

The Pulse Interval (PI) corresponds to the total time of the wave, which is measured between two successive valleys. This feature is used to estimate the pulse rate. The Crest Time (CT) corresponds to the time between the start (first valley) and the main peak of the wave. The Diastolic Time (DT) corresponds to the time between the main peak and the end of the wave. ΔT corresponds to the time between the main peak and the secondary peak. Dicrotic Notch Time (Tn) is the time between the start of the wave and the dicrotic notch. A2T corresponds to the time between the dicrotic notch and the end of the wave. Pulse Width at Half Height (PWHH) is the time equal to the width of the wave at half height. The First Derivative Peak Time (D1PT) parameter corresponds to the time between the start of the wave and its first derivative peak.

3.4.2. Features based on first and second derivatives

The points a, b, c, d and e (figure 5a) are detected on the second derivative of the PPG signal. These points reflect the wave inflections. They are used to compute all the ratios presented in figures 9, 10 and 11. These ratios change with age and reflect arterial stiffness [43].

3.4.3. Area-based features

The area-based features are shown in figure 5b. The Pulse Area (PA) parameter corresponds to the total area of the PPG wave. Area 1 (A1) is computed between the start of the wave and the inflection point (systolic phase). Area 2 (A2) is computed between the inflection point and the end of the wave (diastolic phase). The Inflection Point Area ratio (IPA) corresponds to the ratio between A2 and A1.

3.4.4. Amplitude-based features

The systolic (SA) and diastolic (DA) amplitudes are calculated from the main and the secondary peaks (figure 5c). The Reflection Index (RI) is the ratio between DA and SA while the Augmentation Index (AI) is the difference between SA and DA divided by SA.

3.5. Metrics

In this section, we detail the different metrics employed for evaluating the performances of the models. The Root Mean Squared Error ($RMSE$, equation 6) has been computed between the PPG traces obtained after inverse

wavelet transform (equation 3). Because the amplitudes are arbitrary and normalized, we also propose the Mean Absolute Percentage Error ($MAPE$, see equation 7). Both metrics along with scatter plots and Pearson correlation coefficients have been used to quantify the level of agreement between the predicted ($\widehat{PPG}$) and the ground truth signals (PPG).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}\left(\widehat{PPG_i} - PPG_i\right)^2} \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{i}\left|\frac{\widehat{PPG_i} - PPG_i}{PPG_i}\right| \tag{7}$$

4. Results and discussion

4.1. Learning performance

k-fold cross-validation results for each model are presented in table 2. The MSE correspond to the minimum validation loss (equation 5) observed during training. Each value presented in the table corresponds to the average and standard deviation computed for a specific U-Net network from the lowest MSE of each fold.

Network	MSE_{finger}	MSE_{ear}
U-Net1	0.382 ± 0.054	0.266 ± 0.024
U-Net$_{\text{VGG16}}$	0.319 ± 0.029	0.224 ± 0.032
U-Net$_{\text{VGG19}}$	0.322 ± 0.033	0.232 ± 0.031
U-Net$_{\text{ResNet101}}$	0.341 ± 0.037	0.244 ± 0.022
U-Net$_{\text{ResNeXt101}}$	**0.316 ± 0.036**	**0.222 ± 0.022**
U-Net$_{\text{SE-ResNet101}}$	0.367 ± 0.031	0.249 ± 0.021
U-Net$_{\text{SE-ResNeXt101}}$	0.368 ± 0.042	0.259 ± 0.024
U-Net$_{\text{InceptionResNetV2}}$	0.385 ± 0.041	0.268 ± 0.030
U-Net$_{\text{InceptionV3}}$	0.386 ± 0.036	0.271 ± 0.026
U-Net$_{\text{DenseNet201}}$	0.317 ± 0.036	0.234 ± 0.027

Table 2: k-fold cross-validation results for each model presented in table 2. The MSE (see equation 5) is computed between predicted and ground truth CWT transforms (real and imaginary parts). U-Net1 corresponds to the neural network proposed by Leclerc et al. [55], which does not include a pre-trained backbone. All the other neural networks are U-shaped architectures supported by a backbone.

Independently of the measurement site, the network supported by ResNeXt101 presents the lowest MSE, thus indicating the best performance in terms of wavelet transform reconstruction (real and imaginary parts). We note that performances of architectures supported by VGG16 and DenseNet101 are close from ResNeXt101. Backbones based on ResNet and ResNeXt structure with squeeze and excitation are less efficient. U-Net1 presents higher MSE values than the other models. This observation probably reflects the fact that the network contains between 4 to 5 times less trainable parameters. Models supported by a backbone performed generally better. This translates a real impact of pre-trained convolutional layers on very large databases. As a reminder, the backbone layers are blocked during the training phase. Inception-based backbones also present degraded performances.

Regarding the two sites, ear measurements deliver better general performances (lower MSE) than finger measurements. We assume that this gap

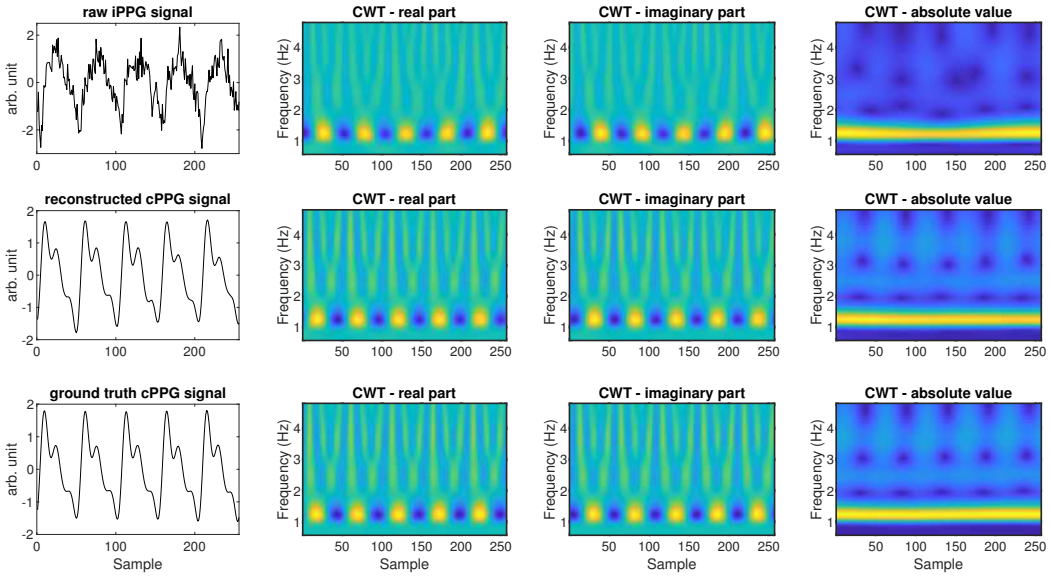


Figure 6: The real and imaginary parts of the reconstructed wavelet transform by the U-Net network supported by ResNeXt101 (middle figures) are similar to those computed from the finger ground truth signal (bottom figures). We can notice a small phase difference in the wavelet representations of the raw iPPG signal (top figures) and the ground truth cPPG signal (bottom figures). The neural network learned this specificity, the reconstructed wavelet transform being in phase with the ground truth one. The absolute representations are depicted for information.

reflects the differences between signal waveform: a PPG signal measured at the forehead surface is generally closer to a PPG signal measured at the ear than measured at the finger [65].

4.2. Point-to-point validation of reconstructed PPG signals

This section is dedicated to the evaluation of PPG signals produced by the neural architectures presented in table 1.

The trained neural models deliver a two-channel wavelet representation (a real part and an imaginary part). The temporal PPG signal is then reconstructed from the inverse transform (equation 3). An example is presented in figure 6, where we can appreciate the prediction quality of the real and imaginary parts of the wavelet transform produced by the U-Net$_{\text{ResNeXt101}}$ network. The phase has been properly recovered. We can also observe that the dicrotic notch is well reproduced whereas it was almost absent on the raw iPPG signal. The reconstructed PPG signal is smooth and its width is smaller. This shows that the network properly corrects the high frequency

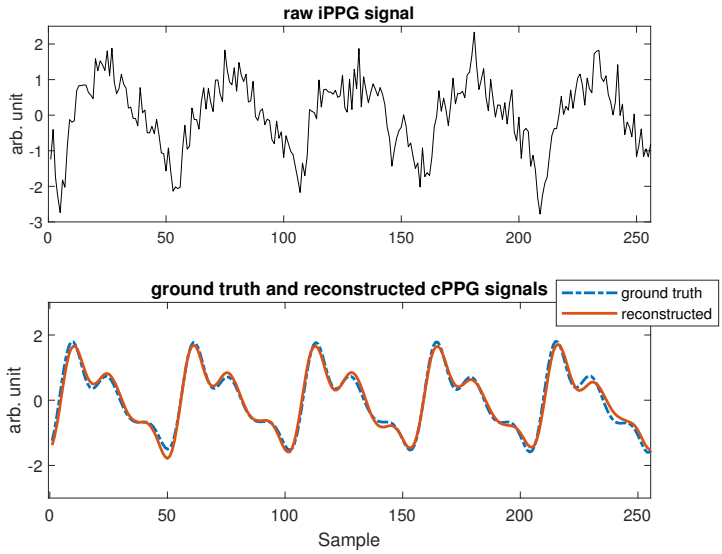


Figure 7: PPG signal prediction (bottom figure) from an iPPG signal (top figure). U-Net supported by ResNeXt101 and trained on finger cPPG signals produced wavelet coefficients that gave, after inverse transform, the reconstructed PPG signal. Ground truth and reconstructed signals are quite similar even if small discrepancies can be noticed.

coefficients, which transcribe the noise, as well as the central frequency coefficients, which determine the pulse signal.

In order to better appreciate the quality of the reconstruction, we present, in figure 7, a superposition of a reference finger cPPG signal and the PPG signal predicted by the U-Net network supported by ResNeXt101 (after computation of the inverse wavelet transform). The $RMSE$ and $MAPE$ have been computed between the two signals (equations 6 and 7). The results after cross-validation on k-fold are presented in table 3. The predictions delivered by the neural models present good overall performance.

The error on the U-Net network supported by ResNeXt101 is slightly lower, which is consistent with the results presented in section 4.1 and table 2. This particular network was therefore selected for further analysis. Table 4 presents the same results but across the test set. Additional comparisons, in particular raw iPPG against ground truth cPPG signals, are presented for information. The errors are here much more important, the $MAPE$ being higher than 50%. The last row of the table is given for comparison and indicates the error between the cPPG signals recorded on the two measurement sites.

Network	$\mathbf{cPPG_{finger}}$ ***vs*** $\widehat{\mathbf{cPPG_{finger}}}$		$\mathbf{cPPG_{ear}}$ ***vs*** $\widehat{\mathbf{cPPG_{ear}}}$	
	$RMSE$	$MAPE$	$RMSE$	$MAPE$
U-Net1	0.260 ± 0.018	0.064 ± 0.010	0.210 ± 0.010	0.033 ± 0.007
$\text{U-Net}_{\text{VGG16}}$	0.245 ± 0.013	0.053 ± 0.014	0.196 ± 0.014	0.031 ± 0.010
$\text{U-Net}_{\text{VGG19}}$	0.248 ± 0.011	0.055 ± 0.012	0.197 ± 0.014	0.034 ± 0.006
$\text{U-Net}_{\text{ResNet101}}$	0.251 ± 0.013	0.058 ± 0.009	0.205 ± 0.010	0.032 ± 0.008
$\mathbf{U\text{-}Net_{ResNeXt101}}$	**0.244 ± 0.014**	**0.045 ± 0.008**	**0.196 ± 0.009**	**0.032 ± 0.009**
$\text{U-Net}_{\text{SE-ResNet101}}$	0.260 ± 0.010	0.058 ± 0.008	0.207 ± 0.012	0.032 ± 0.005
$\text{U-Net}_{\text{SE-ResNeXt101}}$	0.261 ± 0.014	0.060 ± 0.003	0.211 ± 0.012	0.037 ± 0.009
$\text{U-Net}_{\text{InceptionResNetV2}}$	0.265 ± 0.012	0.063 ± 0.008	0.213 ± 0.013	0.038 ± 0.007
$\text{U-Net}_{\text{InceptionV3}}$	0.266 ± 0.011	0.061 ± 0.010	0.213 ± 0.011	0.032 ± 0.004
$\text{U-Net}_{\text{DenseNet101}}$	0.245 ± 0.012	0.052 ± 0.007	0.201 ± 0.012	0.033 ± 0.004

Table 3: k-fold cross-validation for $RMSE$ and $MAPE$ (see equations 6 and 7) computed between reconstructed PPG signals and ground truth cPPG signals. $\text{cPPG}_{\text{finger}}$ and cPPG_{ear} correspond to ground truth cPPG signals measured at finger and ear respectively (see signal depicted in blue in figure 7 for a typical example). $\widehat{\text{cPPG}}_{\text{finger}}$ and $\widehat{\text{cPPG}}_{\text{ear}}$ correspond to reconstructed PPG signals computed by inverse transform on the CWT predicted by the different neural architectures (see signal depicted in orange in figure 7 for a typical example).

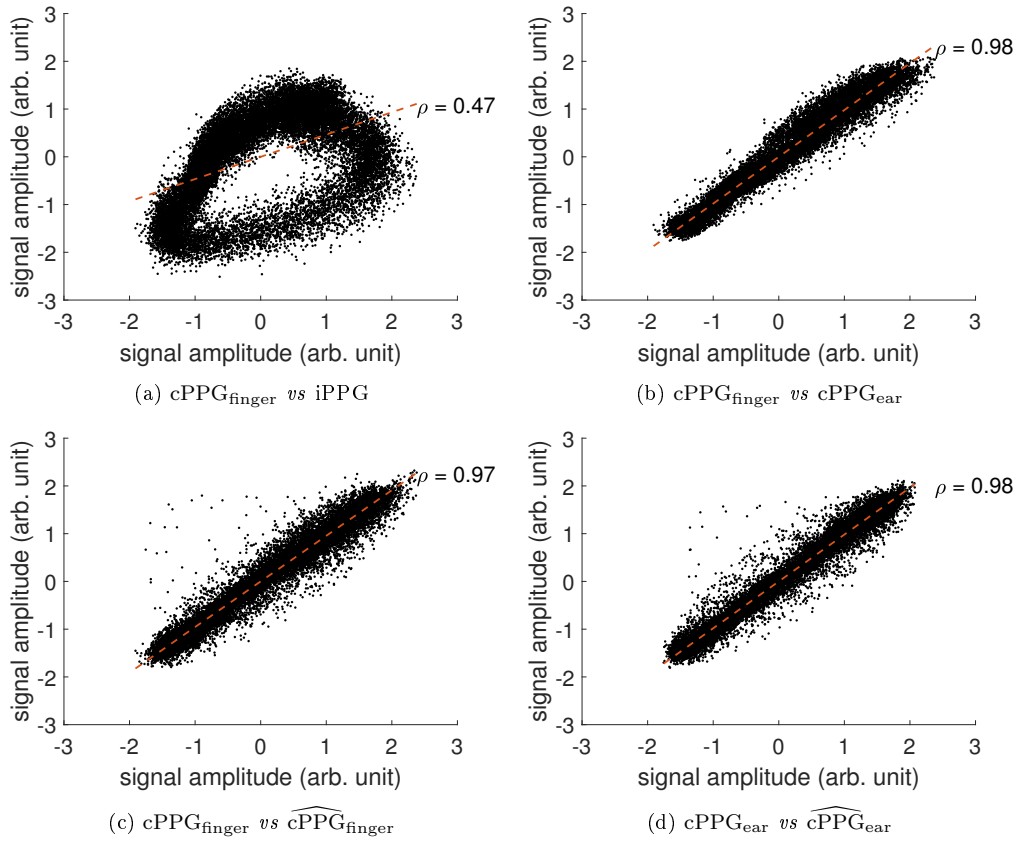
Comparison	*RMSE*	*MAPE*	ρ
$\text{cPPG}_{\text{finger}}$ *vs* $\widehat{\text{cPPG}}_{\text{finger}}$	0.219	0.045	0.97
cPPG_{ear} *vs* $\widehat{\text{cPPG}}_{\text{ear}}$	0.185	0.0187	0.98
$\text{cPPG}_{\text{finger}}$ *vs* iPPG	0.985	0.534	0.47
cPPG_{ear} *vs* iPPG	0.994	0.543	0.46
$\text{cPPG}_{\text{finger}}$ *vs* cPPG_{ear}	0.198	0.020	0.98

Table 4: *RMSE*, *MAPE* and Pearson correlation (ρ) computed across samples included in the test set for ground truth cPPG signals, predicted cPPG signals and raw iPPG signals. An illustration of an iPPG signal is presented in black in figure 7. Predicted signals ($\widehat{\text{cPPG}}$) are produced by the selected $\text{U-Net}_{\text{ResNeXt101}}$ model (see signal depicted in orange in figure 7 for a typical example). All correlations presented p-values lower than 0.001.

Figure 8 presents scatter plots coupled with Pearson correlation coefficients. These representations aim to assess and compare the amplitudes of iPPG, ground truth cPPG and reconstructed cPPG signals over the test set. The graph representing cPPG_{ear} against iPPG signals is not presented in this figure because of its close similarity with the graph presented in figure 8a. The concentric shape of the points distribution reflects the natural waveform difference between raw iPPG signals and cPPG signals. This specificity is mainly due to the dicrotic notch which is generally prominent on cPPG signals and, in contrast, not perceptible on iPPG signals (see figure 2 for a typical example). The inherent pulse width difference between cPPG and iPPG signals also impacts the scatter plot representation presented in figure 8a. Figure 8b depicts finger and ear cPPG measurements and is provided for information.

Figures 8c and 8d illustrate the quality of cPPG signal reconstruction by the $\text{U-Net}_{\text{ResNeXt101}}$ network on the test set. The Pearson correlations coupled with the statistical results presented in table 4 (in particular the low *MAPE*) show that the PPG waveform is suitably reconstructed through its wavelet representation. This conclusion is valid for both finger (figure 8c) and ear (figure 8d) cPPG signals.

We propose, in the next subsection, an in-depth analysis of these results by studding pulse waveform features, whose values are originally very different between iPPG and cPPG signals.



(a) $\text{cPPG}_{\text{finger}}$ *vs* iPPG

(b) $\text{cPPG}_{\text{finger}}$ *vs* cPPG_{ear}

(c) $\text{cPPG}_{\text{finger}}$ *vs* $\widehat{\text{cPPG}}_{\text{finger}}$

(d) cPPG_{ear} *vs* $\widehat{\text{cPPG}}_{\text{ear}}$

Figure 8: Scatter plots along with their respective Pearson correlation (ρ). All the p-values are lower than 0.001. The concentric shape observed in figure (a) reflects the natural waveform difference between raw iPPG signals and cPPG signals. Figure (b) depicts finger and ear cPPG measurements. Bottom row figures present the cPPG signals reconstructed by the $\text{U-Net}_{\text{ResNeXt101}}$ network for both finger (c) and ear (d) measurement sites.

4.3. Waveform features

The point-to-point evaluation presented in the previous subsection provides an overall vision of the predictions quality made by the neural architecture presented in table 1. Here, we propose an evaluation of the reconstructed PPG waves through specific waveform features across the test set. The studied features have briefly been presented in section 3.4. They are divided into 4 categories: temporal, area-based, amplitude-based and based on first and second derivatives.

Scatter plots along with their correlation coefficients are presented for each feature in figures 9 and 10. We focus this specific evaluation on the U-Net$_{\text{ResNeXt101}}$ network. A good general performance on each feature can be observed on each subfigure, showing that the neural network (that take as input CWT of iPPG waves) reliably recovered the shape of finger and ear cPPG waves. As a reminder, iPPG signals computed from video on the forehead region are quite noisy, include artifacts and present a signature that is very different from cPPG signals measured on other sites [65] (see figure 2).

Several temporal features like PI (total width of the pulse wave) show high correlations. PI directly reflects the pulse rate, a parameter estimated from iPPG signals with reliability and precision. Crest time (CT) presents better correlation than DT (diastole time), which seems to be in accordance with studies focusing on arterial pressure estimation based on PPG waveform analysis [45]. In contrast, the temporal parameter ΔT exhibits low correlation. We assume that the specific points associated with the detection of ΔT, in particular the secondary peak, are less accurately recovered. Its estimation is therefore potentially less reliable. It is however interesting to note that this weak correlation is also observed in figure 11 that presents a scatter plot computed between finger cPPG and ear cPPG signals for each waveform feature.

The parameters related to the amplitudes (SA, DA, RI and AI) present more or less high scores. The arbitrary nature of the PPG signals amplitudes makes their estimation very complex. The amplitude of cPPG signals is mainly modulated by the pressure applied between the sensor and the measurement site, by the light absorption of the tissues as well as by the optical properties of the skin. The iPPG signal amplitude also depends on the emitted and reflected quantity of light, the distance as well as internal camera parameters. In general, the predictions produced from finger cPPG signals (figure 9) exhibit higher correlations for the amplitude features than for the predictions computed from ear cPPG signals (figure 10).

Waveform features related to areas and derivatives are relatively well transcribed by the neural model. The correlations presented in figures 9 and 10 are close to the correlations between finger cPPG and ear cPPG signals presented in figure 11.

Overall, the reconstructions of cPPG signals measured on the ear (figure 10) exhibit features that are slightly better correlated with the corresponding ground truth than those measured on the finger (figure 9). This conclusion

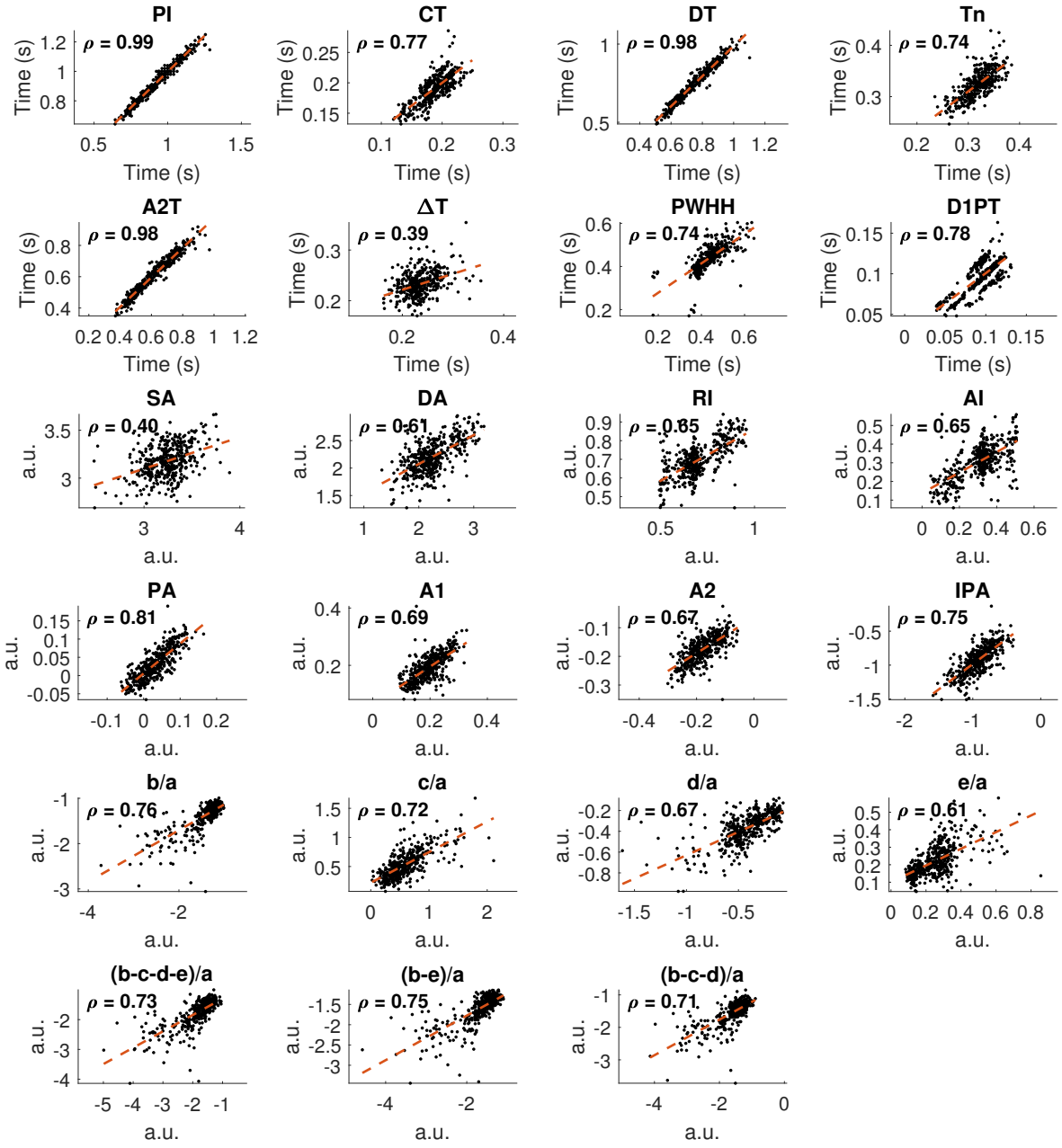


Figure 9: Scatter plots showing the different waveform features computed from ground truth finger cPPG signals ($\text{cPPG}_{\text{finger}}$, x-axis) against the waveform features computed from signals reconstructed by U-$\text{Net}_{\text{ResNeXt101}}$ network ($\widehat{\text{cPPG}}_{\text{finger}}$, y-axis). Associated Pearson correlation coefficients are presented for each feature (on each sub-figure). p-values are all lower than 0.001.

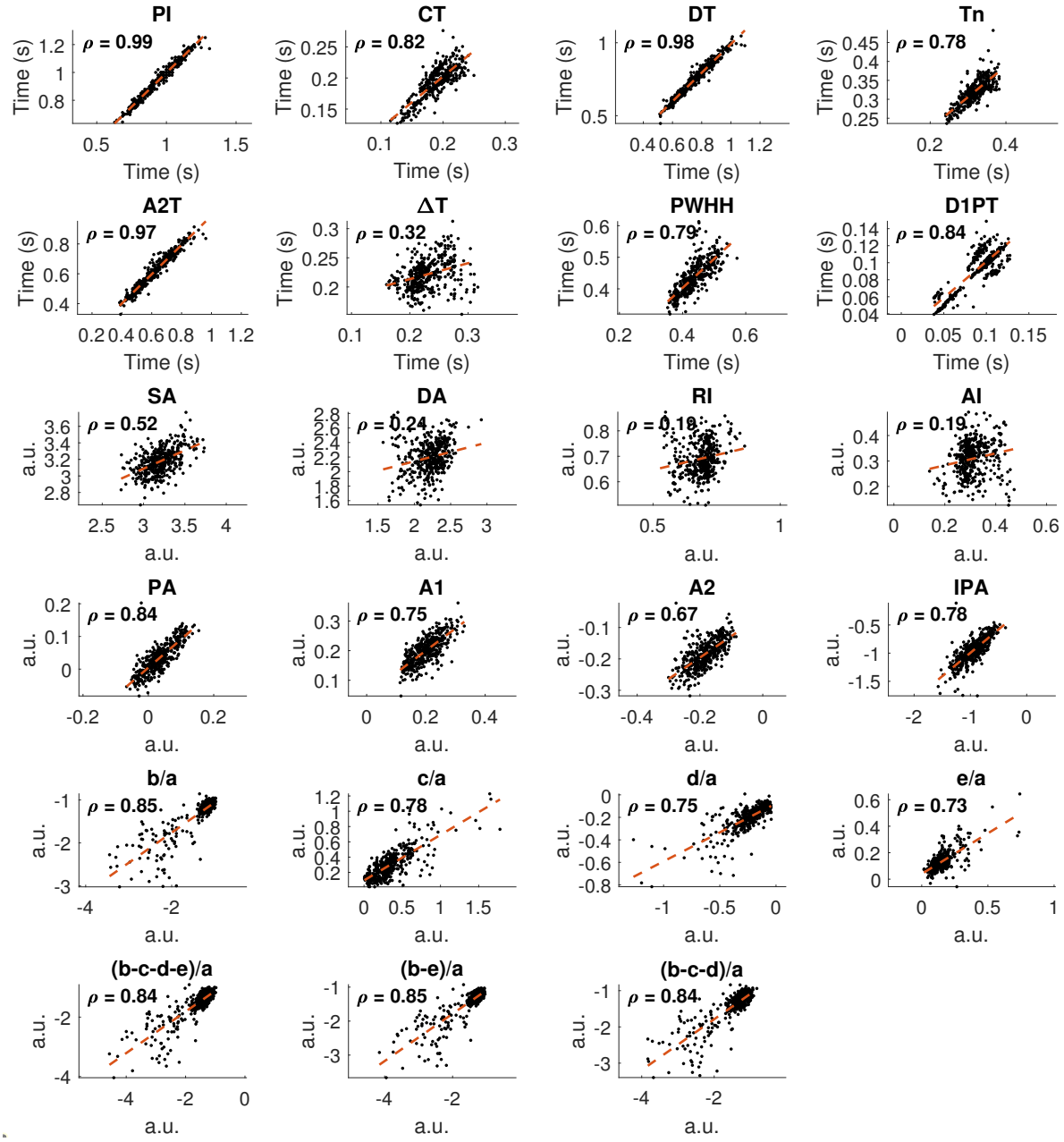


Figure 10: Scatter plots showing the different waveform features computed from ground truth ear cPPG signals (cPPG_{ear}, x-axis) against the waveform features computed from signals reconstructed by U-$\text{Net}_{\text{ResNeXt101}}$ network ($\widehat{\text{cPPG}}_{\text{ear}}$, y-axis). Associated Pearson correlation coefficients are presented for each feature (on each sub-figure). p-values are all lower than 0.001.

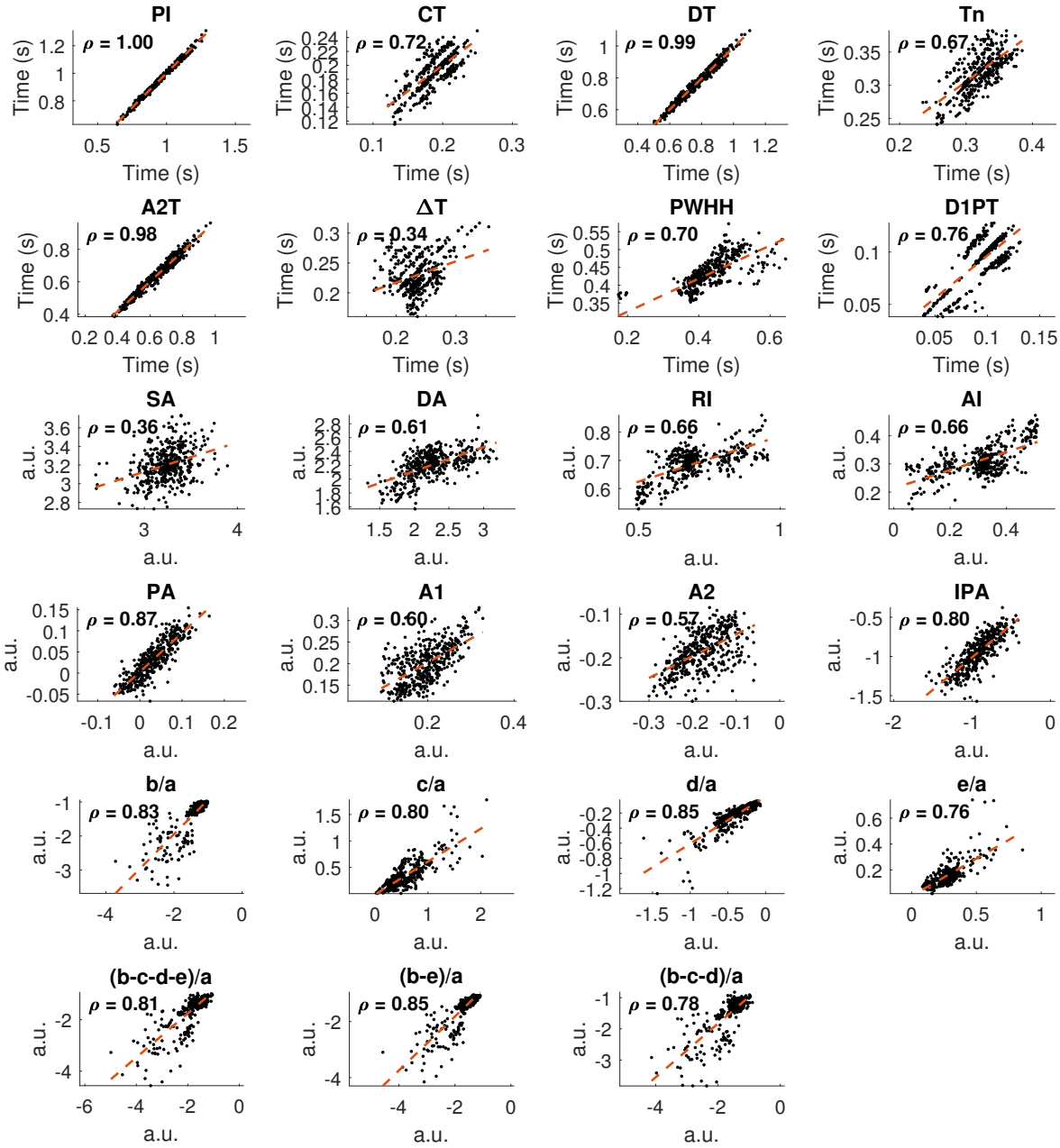


Figure 11: Scatter plots showing the different waveform features computed from ground truth finger cPPG signals ($\text{cPPG}_{\text{finger}}$, x-axis) against the waveform features computed from ground truth ear cPPG signals (cPPG_{ear}, y-axis). Associated Pearson correlation coefficients are presented for each feature (on each sub-figure). p-values are all lower than 0.001.

is in accordance with what we presented in sections 4.1 and 4.2, in particular in tables 2 and 3. We assume that this difference in performance is due to the recovering of the dicrotic notch and the secondary peak that characterize PPG signals. The notch is much more prominent on finger cPPG signals than on ear cPPG signals. It directly impacts the profile of the wave by considerably modifying the inflections and therefore the features linked to the second derivatives. The neural models trained on the wavelet representations computed from finger cPPG signals must therefore recover the coefficients describing the dicrotic notch with more difficulty because this trait is rarely apparent on raw iPPG signals. The top illustration presented in figure 12 shows a prediction of lesser quality where the successive dicrotic notches are approximately reconstructed by the model. The bottom illustration exhibits phase discrepancies. These differences do not systematically impact the shape of the waves but can create unwanted fluctuations in several temporal features, the number one factor being the pulse interval.

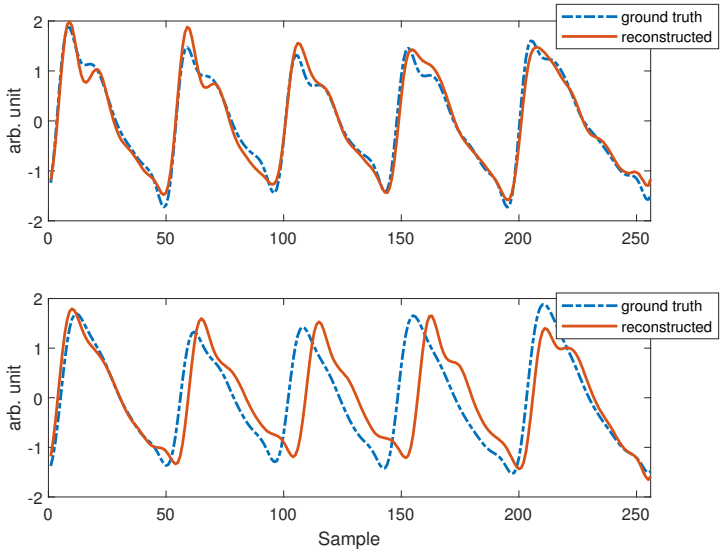


Figure 12: Predictions of lesser quality include approximate dicrotic notch reconstruction (top figure) or phase discrepancies (bottom figure). The signals presented in the two subfigures correspond to finger PPG signals.

5. Limitations and future works

The principal limitation of this study corresponds to the small number of volunteers that participated to the experiments. First validation of the concept on well-formed signals validated this choice. Thus, the videos we employed present a high frame rate which, after processing, results in highly sampled iPPG signals. These signals do not completely reflect those constituted from frames delivered by conventional cameras or webcams. In addition, participants were asked to remain still even during the breath holding experiment.

Several ways of improvement for this work are therefore considered. We first propose expanding the currently limited database by increasing the number of recordings and participants. We also envisage studying the impact of skin color, which directly affects the quality of PPG signals, on performances by assessing the evolution of waveform features against skin phototype.

Continuous wavelet transform using Morlet's wavelet has been employed in this work. We propose evaluating the impact on performances with different mother wavelets as well as investigating different time-frequency representations like short-time Fourier and constant-Q transforms. Modification of the internal parameters of the U-Net architectures (e.g. the number of layers and number of neurons by layer) will also be assessed. Moreover, we propose to study the impact of convolutional attention networks [28] and temporal difference convolution [6] on performances. Currently, the wavelet transform of 5 consecutive waves sampled over 256 values are inputted to the neural network. We envisage varying the number of consecutive waves but with particular consideration for small values (e.g. a single wave) that can produce inconsistencies in the time-frequency representations.

As stated at the beginning of this section, the videos used in this research were acquired by a fast (125 fps) camera. We plan to study in future work iPPG signals computed from recordings delivered by conventional (30 fps) cameras. The waves present, in this context, less details and are therefore more complex to analyze. Training models with larger volume of data can however be envisaged because many databases dedicated to the study of PPG signals measured by conventional cameras are now publicly available.

Inputting video in an U-Net architecture rather than time-frequency representation will be the subject of long-term research. We propose to test 3D U-Net architectures coupled with custom loss function that will constrain reconstruction of cPPG signals through their waveform features. This specific

loss function will be directly integrated into the training phase. The neural network will thus try to minimize an overall error regarding the shape of the pulse waves. Compliance with these criteria could thus allow high quality reconstruction of cPPG from iPPG waves.

6. Summary of contributions

We proposed, in this article, neural architectures that allow accurate recovering of cPPG signals from iPPG signals estimated in video recordings. The reconstruction is carried out using the time-frequency representation of the signals via the continuous wavelet transform. The proposed neural networks correspond to U-Net architectures supported by specific backbones. The recovered signals present waveform features close to those computed on ground truth finger and ear cPPG signals. To the best of our knowledge, this is the first demonstration of a method for accurate reconstruction of cPPG from iPPG signals.

The main motivation behind this work corresponds to the possibility of proposing an estimation of arterial blood pressure from video by analyzing iPPG signals. The next step towards this direction is therefore the integration of the recovered cPPG signals into AI models dedicated to the estimation of blood pressure using contact signals collected from large public databases [50, 48, 45].

7. Acknowledgments

This work has been partly funded by the Contrat Plan État Région (CPER) Innovations Technologiques, Modélisation et Médecine Personnalisée (IT2MP) and Fonds Européen de Développement Régional (FEDER).

References

[1] A. Al-Naji, K. Gibson, S.-H. Lee, J. Chahl, Monitoring of Cardiorespiratory Signal: Principles of Remote Measurements and Review of Methods, IEEE Access (2017).

[2] M. V. Volkov, N. B. Margaryants, A. V. Potemkin, M. A. Volynsky, I. P. Gurov, O. V. Mamontov, A. A. Kamshilin, Video capillaroscopy clarifies mechanism of the photoplethysmographic waveform appearance, Scientific reports 7 (2017) 13298.

[3] M. Hassan, A. Malik, D. Fofi, N. Saad, B. Karasfi, Y. Ali, F. Meriaudeau, Heart rate estimation using facial video: A review, Biomedical Signal Processing and Control 38 (2017) 346–360.

[4] S. Zaunseder, A. Trumpp, D. Wedekind, H. Malberg, Cardiovascular assessment by imaging photoplethysmography–a review, Biomedical Engineering/Biomedizinische Technik (2018).

[5] A. Ni, A. Azarang, N. Kehtarnavaz, A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods, Sensors 21 (2021) 3719. URL: https://www.mdpi.com/1424-8220/21/11/3719. doi:10.3390/s21113719.

[6] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching, IEEE Signal Processing Letters 27 (2020) 1245–1249. URL: https://ieeexplore.ieee.org/document/9133501/. doi:10.1109/LSP.2020.3007086.

[7] Q. Zhan, W. Wang, G. de Haan, Analysis of CNN-based remote-PPG to understand limitations and sensitivities, arXiv preprint arXiv:1911.02736 (2019).

[8] F. Bousefsaf, A. Pruski, C. Maaoui, 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video, Applied Sciences 9 (2019) 4364. URL: https://www.mdpi.com/2076-3417/9/20/4364. doi:10.3390/app9204364.

[9] X. Niu, H. Han, S. Shan, X. Chen, Synrhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3580–3585.

[10] A. Moço, W. Verkruysse, Pulse oximetry based on photoplethysmography imaging with red and green light, Journal of Clinical Monitoring and Computing (2020) 1–11. Publisher: Springer.

[11] H. Luo, D. Yang, A. Barszczyk, N. Vempala, J. Wei, S. J. Wu, P. P. Zheng, G. Fu, K. Lee, Z.-P. Feng, Smartphone-based blood pressure measurement using transdermal optical imaging technology, Circulation: Cardiovascular Imaging 12 (2019) e008857.

[12] N. Sugita, M. Yoshizawa, M. Abe, A. Tanaka, N. Homma, T. Yambe, Contactless Technique for Measuring Blood-Pressure Variability from One Region in Video Plethysmography, Journal of Medical and Biological Engineering (2018) 1–10.

[13] X. Fan, Q. Ye, X. Yang, S. D. Choudhury, Robust blood pressure estimation using an RGB camera, Journal of Ambient Intelligence and Humanized Computing (2018) 1–8.

[14] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, A. Pruski, Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera, Biomedical Signal Processing and Control 64 (2021) 102242.

[15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220.

[16] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, K.-W. Cheung, Block-based adaptive ROI for remote photoplethysmography, Multimedia Tools and Applications (2017) 1–27.

[17] F. Bousefsaf, C. Maaoui, A. Pruski, Automatic Selection of Webcam Photoplethysmographic Pixels Based on Lightness Criteria, Journal of Medical and Biological Engineering 37 (2017) 374–385.

[18] S. Bobbia, D. Luguern, Y. Benezeth, K. Nakamura, R. Gomez, J. Dubois, Real-Time Temporal Superpixels for Unsupervised Remote Photoplethysmography, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1341–1348.

[19] W. Verkruysse, L. O. Svaasand, J. S. Nelson, Remote plethysmographic imaging using ambient light., Optics express 16 (2008) 21434–21445.

[20] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter physiological measurements using a webcam, IEEE transactions on biomedical engineering 58 (2011) 7–11.

[21] F. Bousefsaf, C. Maaoui, A. Pruski, Peripheral vasomotor activity assessment using a continuous wavelet analysis on webcam photoplethysmographic signals, Bio-medical materials and engineering 27 (2016) 527–538.

[22] W. Wang, A. C. den Brinker, S. Stuijk, G. de Haan, Algorithmic Principles of Remote PPG, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491.

[23] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters (2017).

[24] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Transactions on Affective Computing 3 (2012) 42–55.

[25] R. Stricker, S. Müller, H.-M. Gross, Non-contact video-based pulse rate measurement on a mobile service robot, in: Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on, IEEE, 2014, pp. 1056–1062.

[26] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A. El Saddik, EVM-CNN: Real-time contactless heart rate estimation from facial video, IEEE Transactions on Multimedia 21 (2018) 1778–1787. Publisher: IEEE.

[27] G.-S. Hsu, A. Ambikapathi, M.-S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in: Biometrics (IJCB), 2017 IEEE International Joint Conference on, IEEE, 2017, pp. 383–389.

[28] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 349–365.

[29] W. Chen, D. McDuff, DeepMag: Source Specific Motion Magnification Using Gradient Ascent, arXiv preprint arXiv:1808.03338 (2018).

[30] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San

Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1409.1556.

[31] R. Špetlík, V. Franc, J. Matas, Visual Heart Rate Estimation with Convolutional Neural Network, in: The British Machine Vision Conference (BMVC), 2018.

[32] Z. Yu, X. Li, G. Zhao, Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, in: BMVC, 2019.

[33] O. Perepelkina, M. Artemyev, M. Churikova, M. Grinenko, 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, 2020, pp. 288–289.

[34] E. Lee, E. Chen, C.-Y. Lee, Meta-rppg: Remote heart rate estimation using a transductive meta-learner, in: European Conference on Computer Vision, Springer, 2020, pp. 392–409.

[35] X. Niu, S. Shan, H. Han, X. Chen, RhythmNet: End-to-end Heart Rate Estimation from Face via Spatial-temporal Representation, IEEE Transactions on Image Processing (2019). Publisher: IEEE.

[36] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, S.-H. Chang, Siamese-rPPG network: remote photoplethysmography signal estimation from face videos, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020, pp. 2066–2073.

[37] N. Sugita, K. Obara, M. Yoshizawa, M. Abe, A. Tanaka, N. Homma, Techniques for estimating blood pressure variation using video images, in: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, 2015, pp. 4218–4221.

[38] I. C. Jeong, J. Finkelstein, Introducing contactless blood pressure assessment using a high speed video camera, Journal of medical systems 40 (2016) 77.

[39] M. Jain, S. Deb, A. Subramanyam, Face video based touchless blood pressure and heart rate estimation, in: Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on, IEEE, 2016, pp. 1–5.

[40] C. G. Viejo, S. Fuentes, D. D. Torrico, F. R. Dunshea, Non-Contact Heart Rate and Blood Pressure Estimations from Video Analysis and Machine Learning Modelling Applied to Food Sensory Responses: A Case Study for Chocolate, Sensors 18 (2018) 1802.

[41] N. Sugita, T. Noro, M. Yoshizawa, K. Ichiji, S. Yamaki, N. Homma, Estimation of Absolute Blood Pressure Using Video Images Captured at Different Heights from the Heart, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 4458–4461.

[42] M. Elgendi, On the analysis of fingertip photoplethysmogram signals, Current cardiology reviews 8 (2012) 14–25.

[43] E. von Wowern, G. Östling, P. M. Nilsson, P. Olofsson, Digital photoplethysmography for assessment of arterial stiffness: repeatability and comparison with applanation tonometry, PloS one 10 (2015) e0135659.

[44] S. S. Mousavi, M. Firouzmand, M. Charmi, M. Hemmati, M. Moghadam, Y. Ghorbani, Blood pressure estimation from appropriate and inappropriate ppg signals using a whole-based method, Biomedical Signal Processing and Control 47 (2019) 196–206.

[45] N. Ibtehaz, M. S. Rahman, PPG2ABP: Translating Photoplethysmogram (PPG) Signals to Arterial Blood Pressure (ABP) Waveforms using Fully Convolutional Neural Networks, arXiv preprint arXiv:2005.01669 (2020).

[46] M. Elgendi, R. Fletcher, Y. Liang, N. Howard, N. H. Lovell, D. Abbott, K. Lim, R. Ward, The use of photoplethysmography for assessing hypertension, npj Digital Medicine 2 (2019). URL: http://www.nature.com/articles/s41746-019-0136-7. doi:10.1038/s41746-019-0136-7.

[47] M. S. Tanveer, M. K. Hasan, Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based

ANN-LSTM network, Biomedical Signal Processing and Control 51 (2019) 382–392.

[48] M. Panwar, A. Gautam, D. Biswas, A. Acharyya, PP-Net: A Deep Learning Framework for PPG based Blood Pressure and Heart Rate Estimation, IEEE Sensors Journal (2020). Publisher: IEEE.

[49] M. H. Chowdhury, M. N. I. Shuzan, M. E. Chowdhury, Z. B. Mahbub, M. M. Uddin, A. Khandakar, M. B. I. Reaz, Estimating Blood Pressure from the Photoplethysmogram Signal and Demographic Features Using Machine Learning Techniques, Sensors 20 (2020) 3127. Publisher: Multidisciplinary Digital Publishing Institute.

[50] G. Slapničar, N. Mlakar, M. Luštrek, Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network, Sensors 19 (2019) 3420. Publisher: Multidisciplinary Digital Publishing Institute.

[51] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[52] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1867–1874.

[53] M. Tarvainen, P. Ranta-aho, P. Karjalainen, An advanced detrending method with application to HRV analysis, IEEE Transactions on Biomedical Engineering 49 (2002) 172–175. URL: http://ieeexplore.ieee.org/document/979357/. doi:10.1109/10.979357.

[54] F. Bousefsaf, C. Maaoui, A. Pruski, Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate, Biomedical Signal Processing and Control 8 (2013) 568–574.

[55] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, others, Deep learning for segmentation using an open large-scale dataset in 2d echocardiography, IEEE transactions on medical imaging (2019).

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[57] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

[58] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.

[60] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[61] E. C. Too, L. Yujian, S. Njuki, L. Yingchun, A comparative study of fine-tuning deep learning models for plant disease identification, Computers and Electronics in Agriculture 161 (2019) 272–279.

[62] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[63] P. Yakubovskiy, Segmentation Models, GitHub, 2019. URL: https://github.com/qubvel/segmentation_models, publication Title: GitHub repository.

[64] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[65] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, D. Zheng, Quantitative Comparison of Photoplethysmographic Waveform Characteristics: Effect of Measurement Site, Frontiers in physiology 10 (2019).

Multimodal Stress State Detection from Facial Videos Using Physiological Signals and Facial Features

Yassine Ouzar(✉), Lynda Lagha, Frédéric Bousefsaf, and Choubeila Maaoui

Université de Lorraine, LCOMS, 57000 Metz, France
{yassine.ouzar,frederic.bousefsaf,choubeila.maaoui}@univ-lorraine.fr, laghalynda@outlook.com

Abstract. Stress is a complex phenomenon that affects the body and mind on multiple levels, encompassing both psychological and physiological aspects. Recent studies have used multiple modalities to comprehensively describe stress by exploiting the complementarity of multimodal signals. In this paper, we investigate the feasibility of fusing facial features with physiological cues on human stress state estimation. We adopt a multiple modalities fusion using a camera as a single input source and based on the remote photoplethysmography method for non-contact physiological signals measurement. The frameworks rely on modern AI techniques and the experiments were conducted using the new UBFC-Phys dataset dedicated to multimodal psychophysiological studies of social stress. The experimental results revealed high performance when fusing facial features with remote pulse rate variability with an accuracy of 91.07%.

Keywords: Stress detection · Multimodality · Machine learning · Remote photoplethysmography · Facial features

1 Introduction

Cognitive and mental stress corresponds to an important issue in modern societies. Several studies in this field of research recognize mental stress as a key factor in diseases and pathologies like depression, sleep disorders, stroke and heart attack [9]. These effects are particularly induced by a high and daily mental workload.

Various techniques have been developed for treating or preventing this condition, including stress detection techniques that rely on the processing and the analysis of physiological signals which exhibit a high potential alongside an increasing interest from the scientific community. These methods are based on, among others, heart or pulse rate and its variability, breathing rate, skin temperature, and electrodermal activity (or skin conductance) [3]. These physiological

J.-J. Rousseau and B. Kapralos (Eds.): ICPR 2022 Workshops, LNCS 13646, pp. 139–150, 2023.
https://doi.org/10.1007/978-3-031-37745-7_10

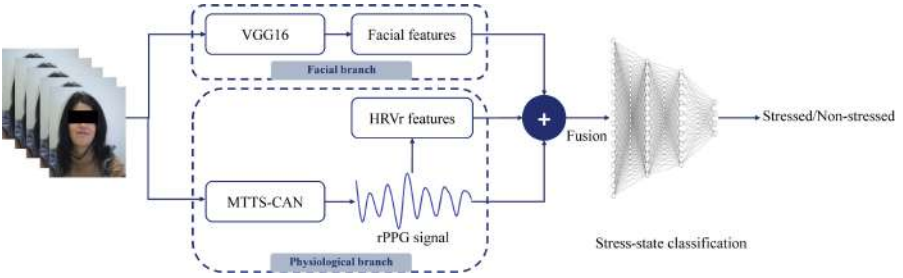


Fig. 1. Overview of the proposed system for multimodal stress state recognition using facial features, rPPG signals, and remote PRV features. It consists of two pipelines. The first one extracts the facial features using a pre-trained VGG16, while the second pipeline extracts physiological signals using a state-of-the-art architecture called MTTS-CAN. The latter recovers the rPPG signal and from which pulse rate variability features can be measured. The extracted features of each modality are then fused and fed to a feed-forward neural network for stressed/non-stressed classification.

signals are also frequently analyzed in the field of affective computing and emotion recognition [26]. The objective consists in automatically processing these signals to predict a stress state or level with a high level of confidence.

Contact sensors are usually employed to record the aforementioned physiological signals. Conventional cameras, through facial video analysis, can also be employed to compute pulse rate, pulse rate variability, peripheral vasomotor activity, and breathing rate to remotely detect mental stress [4,12,14,15], engagement [19] and more generally in applications that relate to the affective computing field of research. Recent techniques include facial features [32], pupil diameter, and blinking rate [20] to strengthen the assessment of stress levels.

According to the recent review of Arsalan et al. [3], most existing studies have examined the use of facial features and physiological cues separately or by combining multiple physiological signals recorded by different sensors. We propose, in this paper, a multimodal video-based method for mental stress state assessment based on facial features and physiological signals. Only a single input source is used to extract features from each modality. Replacing contact intrusive devices with a camera for physiological data measurement may avoid the problems related to asynchrony across modalities, which are usually unaligned. Moreover, it reduces the discomfort caused by the contact sensors that are psychologically stressful. A public dataset, namely UBFC-Phys [17], has been employed to train and evaluate the models proposed in this article. To the best of our knowledge, this is the first study to use this multimodal stress database apart from the original paper.

In the remainder of this paper, stress recognition-related works are presented in Sect. 2. Section 3 details our proposed approach. Then, in Sect. 4, our method is evaluated. Finally, conclusions and future works are given in Sect. 5.

2 Related Works

In recent years, there has been an increasing number of interesting works done in the field of affective computing, including emotion and stress recognition. Despite the distinction between stress and emotion [8], they share some common attributes. Both cause physical and physiological changes in response to a particular stimulus. Just like emotion, various modalities have been used for stress detection either in unimodal [4,6,31,34] or multimodal way [1,10,32]. These modalities can be divided into two classes: external physical cues such as facial expressions, pupil and head movement; internal physiological signals such as heart rate and its variability, breathing rate, skin temperature, and electrodermal activity. Prasetio et al. [23] proposed a stress recognition system based on facial features (such as eyes, nose, and mouth) extracted from face images. Viegas et al. [29] identify the stress state from face videos using 17 action units. An eye tracker device was used by Pedrotti et al. [22] to analyze the correlation between stress and pupil diameter. Physiological signals also have been widely used for stress recognition. They are measured by a contact sensor or remotely using a simple camera. Zubair et al. [35] developed a five-level stress detection system based on PPG signals collected using a pulse sensor on the fingertips. Bousefsaf and al. [4] showed that mental stress can be estimated from pulse rate variability obtained from remote and low-cost devices. Mcduff et al. [16] recently proposed a cognitive stress estimation system based on peripheral hemodynamics and vasomotion power extracted from rPPG amplitudes.

Recent studies have shown that multimodal stress detection systems exceed the performance of unimodal systems [3]. Existing multimodal stress detection schemes can be divided into a fusion of physiological signals only and a fusion of remote modalities with physiological signals acquired from contact sensors. Despite the results obtained, they follow a constrained experimental setup under laboratory conditions due to the use of intrusive and sensitive equipment that is psychologically stressful. In addition, while dealing with multiple signals of different natures gathered from different sources, they may conflict with each other due to asynchrony across modalities and thus lead to misestimation.

3 Materials and Method

3.1 Dataset

We explored the UBFC-Phys [17], a public multimodal database dedicated to psycho-physiological studies. The UBFC-Phys dataset provides data collected from 56 undergraduate psychology student participants, including 46 females and 10 males, all between the ages of 19 and 38 (with a mean age of 21.8 years). Participants underwent a social stress-inducing experiment in three stages: a resting task T1, a speaking task T2, and an arithmetic task T3 (T2 and T3 being the stressful tasks), during which participants were filmed and wore a wristband that allowed the measurement of their blood volume pulse (BVP) and electrodermal activity (EDA) signals. A form for calculating the level of stress

(anxiety score) is presented to participants before and after the experiment. For each participant, three videos (one video per task) of 3 min duration were recorded at a frame rate of 35 fps and with a resolution of 1024 × 1024 pixels. BVP and EDA signals for each task as well as their anxiety scores calculated before and after the experiment are publicly available.

3.2 Data Preparation

The conducted experiments include two types of physiological features measured in contact using a wristband or remotely from video recordings. For contact-based physiological features, BVP signals and their derivative contact pulse rate variability (PRVc) features are used. A similar procedure has been conducted for video-based physiological features, rPPG signal, and its derivative remote pulse rate variability (PRVr) are employed. Adding to that the exploitation of the facial features extracted from the video recordings by transfer learning.

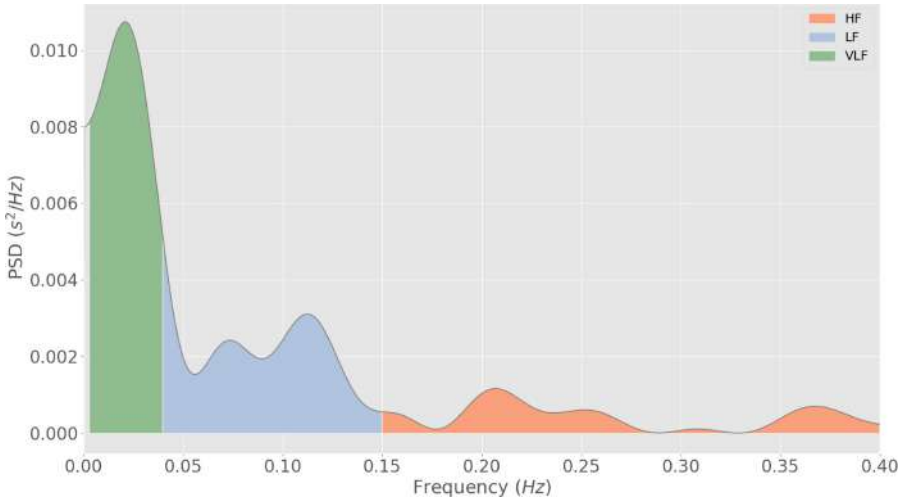


Fig. 2. A representative power spectral density for IBI series showing The areas of VLF, LF, and HF powers of the PRV.

3.2.1 Contact-based Features

First, the ground-truth BVP signals are resampled to the sampling rate of the camera (35 fps). Then, detrending is performed using a smoothness priors approach [28]. After that, we applied a 2nd-order Butterworth band-pass filter with a cutoff frequency of 0.75 and 2.5 Hz to keep only the information related to the pulse waveform. From the filtered BVP signals, 8 contact pulse rate variability

(PRVc) features have been computed. Peak detection is first performed to locate the instant of time at which the heartbeat occurs. The contact pulse rate is computed in the time domain by the inverse of the interbeat interval (IBI) divided by 60 to get the frequency in beats per minute. From the pulse rate variations during the video recording session, we computed the mean (meanHR), standard deviation (stdHR), maximum (maxHR) and minimum (minHR) of the pulse rate series. The root mean square of successive interval differences (RMSSD) is also calculated (see Eq. 1). This parameter gives an evaluation of the vagal activity reflected in pulse variability [25].

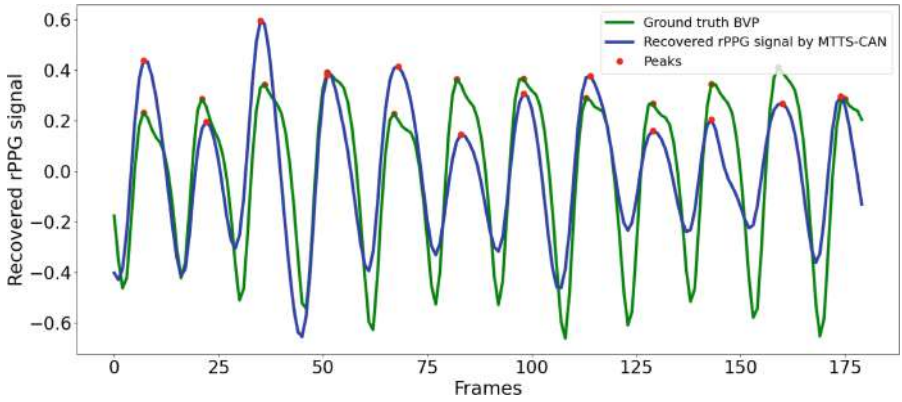


Fig. 3. Comparison between a predicted signal by MTTS-CAN and the ground-truth BVP signal taken from the UBFC-Phys dataset. The amplitudes are different but the peak location seems relevant which is important for IBIs measurement.

Three pulse rate variability features were extracted in the frequency domain. The IBI series were interpolated with cubic Hermite and the power spectra were obtained by employing Welch's method [30]. From the different oscillatory components of the power spectral density (PSD), low frequency (LF) and high frequency (HF) components were computed. The LF component is modulated by baroreflex activity and contains both sympathetic and parasympathetic activity, while the HF component reflects the parasympathetic branch of the autonomic nervous system [2]. The LF and HF powers of the pulse rate variability were computed as the area under the PSD curve corresponding to 0.04–0.15 Hz and 0.15–0.4 Hz respectively (see Fig. 2). The LF/HF, which represents the sympatho-vagal balance [5] has also been computed. The very low frequency (VLF) components were not employed in our experiments.

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2} \tag{1}$$

3.2.2 Video-based Physiological Features

Remote photoplethysmography (rPPG) and ballistocardiography (BCG) are the two main methods for measuring physiological signals by camera [13]. In our experiment, we used the rPPG technique for the estimation of pulse signals. Despite the advantages and limitations of each method, the BCG has been less exploited in recent years for different reasons. Compared to rPPG method, BCG is more difficult to implement because of its morphology which varies according to the subjects and the sensor used adding to that its sensitivity to noise and artifacts of movements.

rPPG is an optical technique that captures cardiac signals by observing the variation of blood volume on the person's face using a camera. The captured light reflected from the skin is translated into a variation of the rPPG signal. Same to BVP signal, several characteristics can be derived from the rPPG signal such as pulse rate, breathing rate, and remote pulse rate variability (PRVr).

Among the popular methods in the state-of-the-art, we used the Multi-Task Sequential Shift Convolutional Network (MTTS-CAN) proposed by Liu et al. [11] for rPPG signals extraction. MTTS-CAN is an end-to-end deep neural network that combines a convolutional attention mechanism with a time-shifting module. For a better appreciation of the quality of the rPPG signal, we present in Fig. 3 an overlay of a BVP ground truth signal in contact and an rPPG signal predicted by the MTTS-CAN network. We clearly observe a correlation between the estimated rPPG signal and the ground truth, moreover, the peaks are very close which is important for IBIs measurement.

For PRVr features extraction, we processed the raw rPPG signals and extracted the same parameters as for contact signals (see Sect. 3.2.3).

3.2.3 Facial Features

Deep learning models have proven efficient for general-purpose 2D image tasks compared to traditional machine learning algorithms. However, a large amount of data is required to train the model properly in order to achieve high performance. Due to data scarcity, we looked at the transfer learning approach as a viable alternative and to reduce the development effort at the same time. A pre-trained VGG16 model [27] is adopted as facial features extractor. VGG16 is very popular and has proved to be very efficient and achieve high recognition accuracy in computer vision tasks [7]. The network consists of a features extraction block based on convolution layers and the classifier block that consists of dense layers. The features extraction block is frozen, while the classifier block is modified by replacing the upper dense layers compatible with stressed/non-stressed classification. Afterward, the network is fine-tuned with UBFC-Phys data dedicated to stress recognition.

4 Results and Discussion

The experiments were carried out using the same specifications presented in the original article of the UBFC-Phys dataset [17]. Using the supplementary material II provided with the paper[1], 101 of 168 tasks were selected. The 67 removed tasks were eliminated by testing the correlation between BVP and rPPG signals to detect the corrupted signals. We used 7-fold subject-independent cross-validation strategy on both separate and fused modalities. We randomly created 7-fold using 85% of the data for training and the remaining 15% for testing. The average accuracy across each fold is reported in Table 1 and 2.

Three different experiments were performed for stress state detection: a) using physiological modalities only (contact and non-contact), b) using facial features only, and c) merging physiological signals and facial features. The non-stress is represented by task T1, while T2 and T3 represent the stress state.

Table 1. Non-stress vs stress state classification results based on physiological signals

Features	Classifiers	Accuracy (%)
BVP	SVM RBF Kernel	69.72
	SVM Poly Kernel	58.58
	NB	**72.61**
	RF	66.96
	KNN	44.22
rPPG	SVM RBF Kernel	57.81
	SVM Poly Kernel	57.81
	NB	61.82
	RF	**62.40**
	KNN	59.96
PRVc	SVM RBF Kernel	72.74
	SVM Poly Kernel	74.55
	NB	**78.16**
	RF	58.58
	KNN	73.64
PRVr	SVM RBF Kernel	58.58
	SVM Poly Kernel	57.61
	NB	56.92
	RF	58.58
	KNN	**72.22**

[1] https://ieeexplore.ieee.org/ielx7/5165369/10056372/9346017/supp2-3056960.pdf?arnumber=9346017.

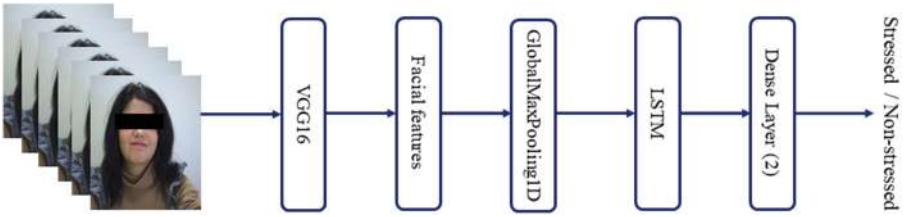


Fig. 4. stress state recognition system using facial features.

4.1 Stress Recognition from Physiological Signals

We used machine learning algorithms and ideas proposed in the original article of the UBFC-Phys dataset to compare the performance of the different features [17]. Five classifiers were considered: Support Vector Machine (SVM) with a polynomial kernel, SVM with Radial Basis Function (RBF), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). We conducted the same 7-fold cross-validation on the five algorithms. Each classifier was trained with 85% of the signals, and the remaining 15% were used for testing.

Table 1 provides the recognition accuracy using contact (BVP and PRVc) and non-contact (rPPG and PRVr) features. In this experiment, the best result was achieved by contact-based physiological features. PRVc features reached the highest accuracy at 78.16%, followed by BVP signals with an accuracy of 72.61%. The best performance for contact-based physiological features was obtained with the Naive Bayes classifier. By comparing the obtained results, we note that the stress state recognition accuracy with the BVP signal outperforms the rPPG one. A similar observation can be drawn for the contact and non-contact PRV features, better accuracy was achieved with the contact features compared to the remote ones. These observations are in the line with what has been reported in previous studies [21] but in contradiction with the results presented in the article that introduced the UBFC-Phys [17]. The authors reported higher accuracies with video-based physiological modalities than with contact-based physiological features. We suppose that the performance of each modality depends on the type of classifier and its parameters, as well as the rPPG signal recovering method. In their experiments [17], a conventional framework consisting of several signal and image processing steps was used. Here, we choose to adopt a novel end-to-end deep learning approach that extracts the rPPG waveform automatically without any additional pre-processing or post-processing steps [11].

4.2 Stress State Recognition from Facial Features

A transfer learning strategy is adopted in this experiment to leverage the knowledge from the object recognition domain to stress recognition by replacing the upper dense layers and fine-tuning the network with UBFC-Phys data dedicated to stress recognition. The proposed system is illustrated in Fig. 4. First, each

frame of the videos is resized to ($224 \times 224 \times 3$ and then passed to the pre-trained VGG16 model [27], which is initially trained with the ImageNet dataset for object recognition. The output features of VGG16 (before the dense layers) are extracted and then vectorized using GlobalMaxPooling1D to obtain the facial feature vector. This vector is passed to an LSTM layer to consider the temporal dimension. Finally, it is passed to a dense layer composed of 2 neurons for classification. For this purpose, a sigmoid activation function is applied to the dense layer, enabling binary stress classification (stress/non-stress).

The result presented in Table 2 shows that facial features-based stress state recognition outperforms physiological features either measured in contact or non-contact. This confirms the results reported in previous studies where recognition accuracy of affects/emotions using visual features (e.g. facial expressions) outperforms physiological modalities [21].

4.3 Stress State Recognition from Facial Features and Physiological Signals

Figure 1 presents the overall architecture of the proposed multimodal stress recognition system. It includes two pipelines to extract the features of each modality from facial video recordings. Each video of the UBFC-Phys dataset is fed to the facial features network and to the rPPG extractor network (MTTS-CAN). The first pipeline extracts the features vector after the flatten layer using the pre-trained weights of VGG16 (See Fig. 4), while the second pipeline returns either the rPPG signal recovered through the MTTS-CAN network [11] or PRVr features. We conducted two experiments on our multimodal stress recognition system. The first one combined the facial features with the PRVr features only, then with the rPPG signal only. The concatenation result vector of the two modalities is passed into two dense layers with 256 and 2 neurons respectively. The first layer takes the rectified linear units as the hidden units while the second one uses the sigmoid activation function to predict the corresponding stress class either stress or non-stress state.

The average accuracies of fusing the facial features with the PRVr features and with the rPPG signals are shown in Table 2. As we can see, combining facial features with PRVr features improve significantly the classification accuracy and deliver better accuracy (91.07%) compared to using facial or PRVr features separately. The fusion of facial features and the rPPG signals slightly improves performance, achieving an accuracy of 83.12%.

Table 2. Non-stress vs stress state classification results based on facial features only and on a fusion between facial features and remote physiological signals

Method	Accuracy (%)
Facial features	82.48
Facial features + rPPG	83.12
Facial features + PRVr	**91.07**

5 Conclusion

A multimodal approach to stress state recognition through video-based physiological signals and facial features has been proposed. Physiological cues are measured remotely from facial video recordings using the rPPG technique, while facial features were extracted by transfer learning. In such a manner, only a single input source was utilized to extract features from each modality. Both unimodal and multimodal experiments were performed. Analysis has shown that facial features are more relevant and allow for the highest level of accuracy. Compared to performance using only facial features, merging facial features with physiological signals provided a more accurate estimation, indicating the effectiveness of multimodal analysis.

Future tasks are to further improve the method's accuracy and to use other physiological modalities such as electrodermal activity and respiratory rate and rPPG waveform-based features linked to vasomotor activity and blood pressure. We also aim to use the stress score provided by the UBFC-Phys dataset and move to other approaches for facial features extraction using action units and facial landmarks. Furthermore, we intend to extend our work to other datasets such as RECOLA [24], AMIGOS [18] and BP4D+ [33] databases. The RECOLA dataset [24] can directly be used as it is annotated with stress labels and provides videos and the corresponding physiological signals. We plan to annotate the other two databases by exploiting other emotion label classes or by using electrodermal activity signals that correlate strongly with stress. We also plan to improve and search for the best features extractor model by comparing the most commonly employed neural architectures (e.g. ResNet, Xception, Inception).

References

1. Almeida., J., Rodrigues., F.: Facial expression recognition system for stress detection with deep learning. In: Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS, pp. 256–263. INSTICC, SciTePress (2021). https://doi.org/10.5220/0010474202560263
2. Appelhans, B., Luecken, L.: Heart rate variability as an index of regulated emotional responding. Rev. Gen. Psychol. **10**(3), 229–240 (2006). https://doi.org/10.1037/1089-2680.10.3.229
3. Arsalan, A., Anwar, S.M., Majid, M.: Mental stress detection using data from wearable and non-wearable sensors: a review. arXiv preprint arXiv:2202.03033 (2022)
4. Bousefsaf, F., Maaoui, C., Pruski, A.: Remote detection of mental workload changes using cardiac parameters assessed with a low-cost webcam. Comput. Biol. Med. **53**, 154–163 (2014). https://doi.org/10.1016/j.compbiomed.2014.07.014
5. Burr, R.L.: Interpretation of normalized spectral heart rate variability indices in sleep research: a critical review. Sleep **30**(7), 913–919 (2007)
6. Cinaz, B., Arnrich, B., Marca, R.L., Tröster, G.: Monitoring of mental workload levels during an everyday life office-work scenario. Pers. Ubiquit. Comput. **17**, 229–239 (2011)
7. Dubey, A.K., Jain, V.: Automatic facial recognition using vgg16 based transfer learning model. J. Inf. Optim. Sci. **41**(7), 1589–1596 (2020). https://doi.org/10.1080/02522667.2020.1809126

8. Epel, E.S., et al.: More than a feeling: a unified view of stress measurement for population science. Front. Neuroendocrinol. **49**, 146–169 (2018). https://doi.org/10.1016/j.yfrne.2018.03.001, https://www.sciencedirect.com/science/article/pii/S0091302218300219, stress and the Brain
9. Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M.: Review on psychological stress detection using biosignals. IEEE Trans. Affect. Comput. **13**(1), 440–460 (2022). https://doi.org/10.1109/TAFFC.2019.2927337
10. Kurniawan, H., Maslov, A.V., Pechenizkiy, M.: Stress detection from speech and galvanic skin response signals. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 209–214 (2013). https://doi.org/10.1109/CBMS.2013.6627790
11. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. Adv. Neural Inf. Process. Syst. **33**, 19400–19411 (2020)
12. Maaoui, C., Bousefsaf, F., Pruski, A.: Automatic human stress detection based on webcam photoplethysmographic signals. J. Mech. Med. Biol. **16**, 1650039 (2016)
13. McDuff, D.: Camera measurement of physiological vital signs. arXiv preprint arXiv:2111.11547 (2021)
14. McDuff, D.J., Gontarek, S., Picard, R.W.: Remote measurement of cognitive stress via heart rate variability. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2957–2960 (2014)
15. McDuff, D.J., Hernández, J., Gontarek, S., Picard, R.W.: Cogcam: contact-free measurement of cognitive stress during computer tasks with a digital camera. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016)
16. McDuff, D.J., et al.: Non-contact imaging of peripheral hemodynamics during cognitive and psychological stressors. Sci. Rep. **10**, 10887 (2020)
17. Meziati Sabour, R., Benezeth, Y., De Oliveira, P., Chappe, J., Yang, F.: Ubfc-phys: a multimodal database for psychophysiological studies of social stress. IEEE Trans. Affect. Comput. 1 (2021). https://doi.org/10.1109/TAFFC.2021.3056960
18. Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I.: Amigos: a dataset for affect, personality and mood research on individuals and groups. IEEE Trans. Affect. Comput. **12**(2), 479–493 (2021). https://doi.org/10.1109/TAFFC.2018.2884461
19. Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans. Affect. Comput. **8**(1), 15–28 (2017). https://doi.org/10.1109/TAFFC.2016.2515084
20. Nagasawa, T., Takahashi, R., Koopipat, C., Tsumura, N.: Stress estimation using multimodal biosignal information from RGB facial video. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1181–1187 (2020). https://doi.org/10.1109/CVPRW50498.2020.00154
21. Ouzar, Y., Bousefsaf, F., Djeldjli, D., Maaoui, C.: Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2460–2469, June 2022
22. Pedrotti, M., et al.: Automatic stress classification with pupil diameter analysis. Int. J. Hum.-Comput. Interact. **30**, 220–236 (2014)

23. Prasetio, B.H., Tamura, H., Tanno, K.: The facial stress recognition based on multi-histogram features and convolutional neural network. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 881–887 (2018). https://doi.org/10.1109/SMC.2018.00157
24. Ringeval, F., Sonderegger, A., Sauer, J.S., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8 (2013)
25. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. Front. Public Health **5**, 258 (2017)
26. Shu, L., et al.: A review of emotion recognition using physiological signals. Sensors **18**(7) (2018). https://doi.org/10.3390/s18072074, https://www.mdpi.com/1424-8220/18/7/2074
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Tarvainen, M.P., Ranta-aho, P.O., Karjalainen, P.A.: An advanced detrending method with application to HRV analysis. IEEE Trans. Biomed. Eng. **49**, 172–175 (2002)
29. Viegas, C., Lau, S.H., Maxion, R., Hauptmann, A.: Towards independent stress detection: a dependent model using facial action units. In: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2018). https://doi.org/10.1109/CBMI.2018.8516497
30. Welch, P.: The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967)
31. Zhang, H., Zhu, Y., Maniyeri, J., Guan, C.: Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2985–2988 (2014). https://doi.org/10.1109/EMBC.2014.6944250
32. Zhang, H., Feng, L., Li, N., Jin, Z., Cao, L.: Video-based stress detection through deep learning. Sensors **20**(19) (2020). https://www.mdpi.com/1424-8220/20/19/5552
33. Zhang, Z., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3438–3446 (2016)
34. Zhou, G., Hansen, J., Kaiser, J.: Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **9**(3), 201–216 (2001). https://doi.org/10.1109/89.905995
35. Zubair, M., Yoon, C.: Multilevel mental stress detection using ultra-short pulse rate variability series. Biomed. Signal Process. Control. **57**, 101736 (2020)

Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals

Yassine Ouzar, Frédéric Bousefsaf, Djamaleddine Djeldjli, Choubeila Maaoui
Université de Lorraine, LCOMS, F-57000, Metz, France
{yassine.ouzar, frederic.bousefsaf, djamaleddine.djeldjli, choubeila.maaoui} @univ-lorraine.fr

Abstract

Human's affective state recognition remains a challenging topic due to the complexity of emotions, which involves experiential, behavioral, and physiological elements. Since it is difficult to comprehensively describe emotion in terms of single modalities, recent studies have focused on fusion strategy to exploit the complementarity of multimodal signals. In this article, we study the feasibility of fusing facial expressions with physiological cues on human emotion recognition accuracy. The contributions of this work are threefold: 1) We propose a new spatiotemporal network for facial expression recognition using a 3D squeeze and exitation based 3D Xception architecture (squeeze and exitation Xception network). 2) We adopt the first multiple modalities fusion using single input source which, to the best of our knowledge, no existing multimodal emotion recognition system has attempted to identify emotional state from only facial videos using facial expressions and physiological signals features. 3) We compare the performance of the unimodal approach using only facial expressions or physiological data, to multimodal systems fusing facial expressions with video-based physiological cues. In our experiments, physiological signals such as the iPPG signal and features of heart rate variability measured remotely using the imaging photoplethysmography (iPPG) method are used. The preliminary results show that the multimodal fusion model improves the accuracy of emotion recognition, and merging facial expressions features with iPPG signal gives the best accuracy with 71.90 %.

1. Introduction

Human faces are a rich source of information. They are characterized by a great expressive richness to convey emotions, which makes them widely used to identify a person's emotional state through facial expressions. Despite the impressive results achieved by facial expressions recognition systems on acted databases with controlled conditions [31, 49, 54, 55], they are rarely faced with real situations. In a natural environment, reliability cannot be guaranteed and performance degrades considerably [20, 32, 43]. In addition to environmental conditions (camera angles, lighting conditions and occlusion of multiple parts of the face) and the ability to control and fake emotions by people, facial expressions are also more affected by social and cultural differences. Human expressiveness can vary among individuals and can be expressed differently. Additionally, facial expressions can be a mix of different emotion status that occur at the same time or may not be expressed at all. Consequently, using facial expressions to identify person's emotional state can lead to wrong inferences.

Recently, few studies have proposed emotion recognition systems that use physiological cues extracted from the face using the imaging photoplethysmography method [2, 30]. The advantage of using physiological parameters to assess emotion compared to facial expressions is : physiological data are a response to the autonomic nervous system (ANS), which is involuntarily activated and therefore uncontrollable.

Most existing studies have examined the use of facial expressions and physiological cues separately [12, 24, 31, 38, 49]. However, little attention has been paid to a fusion between these two modalities [8, 18, 51]. Combining the two can improve recognition accuracy and provide greater reliability by continuously gathering information about the person's emotional state despite missing acquisition or misleading information that may occur when using a single modality, operating in a noisy environment or in the case of falsified expression. Additionally, fusion of multiple modalities can help to compensate errors and resolve ambiguities by learning useful representations of data of different nature. However, The main limitation is related to asynchrony across modalities, which are usually unaligned. In addition, physiological data are collected through intrusive devices that are psychologically stressful and this can modify the measurement results of physiological signals. Therefore, this will certainly affect the accuracy of emotion scoring [9]. In this work, we propose the first video-based

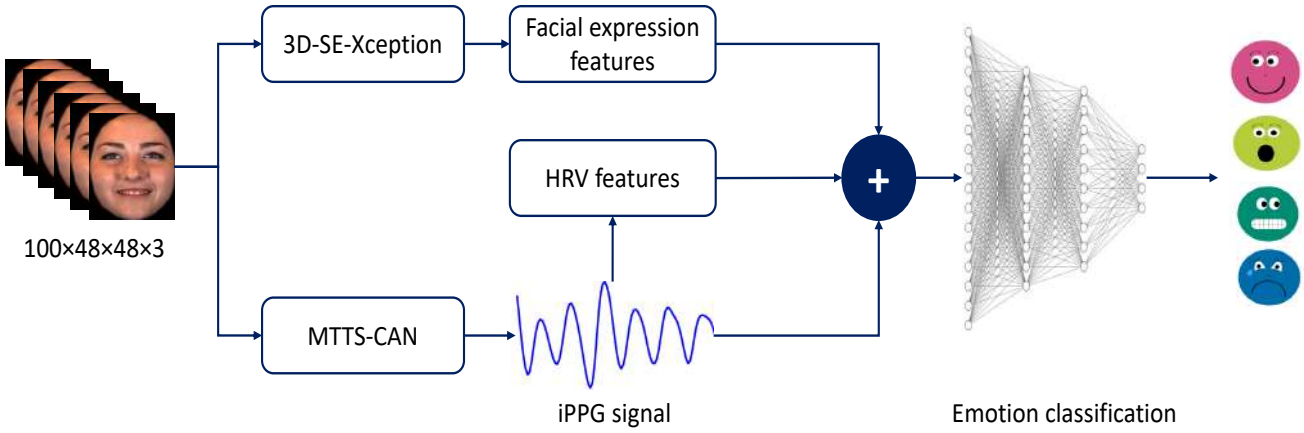


Figure 1. Proposed system for multimodal emotion recognition using facial expressions, iPPG signals and HRV features.

multimodal spontaneous emotion recognition that combines facial expressions with physiological data to derive the advantages of each modality.

In this paper the physiological parameters are measured from facial video recordings based on imaging photoplethysmography principal [6]. While, facial expressions features are extracted using a new spatio-temporal network that combines 3D squeeze and exitation module with 3D Xception architecture. The features vector of facial expressions is then merged with the physiological signals to ultimately estimate the corresponding emotion. In the remainder of this paper, human emotion recognition related works are presented in Section 2. Section 3 details our proposed approach. Then, in Section 4, our method is evaluated. Finally, conclusions and future works are given in Section 5.

2. Related works

In literature, various modalities have been used to recognize emotion either in unimodal [12, 36, 45, 47] or multimodal way [5, 18, 39]. Initial research on unimodal emotion recognition systems have focused on the expressiveness of the face because it is visible and it is easier to collect a large set of facial data. The commonly adopted methods for facial expression recognition are either deep learning or handcrafted based approaches [22, 25]. However, deep learning techniques have made a great success due to their high generalizability for new data and their ability to automatically extract robust features and learn complex nonlinear representations. Today, the state of the art deep learning methods allow to achieve a categorization of facial expressions with a reliability of around 98% in controlled situations [23]. Nevertheless, several real environment issues can degrade recognition accuracy such as lighting variations or background appear [28]. Additionally, deep learning algorithms often fail in the case of expressionless faces or falsified expressions.

To address this issue, some attempts have been made to identify emotion through physiological data that are managed by the autonomous nervous system (ANS) which is involuntarily activated and therefore can not be controlled [12]. Physiological signals such as electroencephalography, electrocardiography, skin temperature and electromyography are reliable data for quantifying emotions [10]. However, they are acquired by intrusive contact sensors that can interfere with the subjects and modify their emotional state. Moreover, the complexity of measurement and the sensitivity of the electrodes of these devices strongly limit their scope of application, since they cannot be used outside of the laboratory. Therefore, recent studies have focused on wearable devices that provide various biosignals such as blood volume pulse (BVP) and electrodermal activity and their derivatives to explore new application fields. Going even further, recent works have used heart rate variability measured by the camera to detect emotional state [3, 30]. They rely on imaging photoplethysmography method, which allows non-contact extraction of the blood volume pulse signal from facial video recording, making it more interesting and promising among the other physiological signals that require contact devices and the presence of a specialist to monitor them.

Numerous literature studies show that multimodal emotion recognition systems outperform unimodal approaches [11, 33]. For this reason, several works have merged facial expressions with physiological data to develop reliable systems [8, 18, 21]. Despite the obtained results , they follow a constrained experimental setup under laboratory conditions due to the use of intrusive and sensitive equipment. In addition, dealing with multiple signals of different nature gathered from different sources, may conflict with each other due to asynchrony across modalities and thus lead to misestimation.

3. Materials and Methods

3.1. Dataset

Although many multimodal emotion databases are available, few of them provide physiological signals. The existing datasets for multimodal emotion recognition from facial expressions and physiological signals are quite limited not only in data size but also in diversity. In this study we explore a new multimodal sponatneous emotions database named BP4D+ [53]. Compared to existing datasets such as MAHNOB [42] and DEAP [19], BP4D+ is a large scale dataset that includes annotated action units (AUs) and discrete emotion categories. In addition, it contains numerous challenging conditions and diversity in terms of significant head motion and ethnic diversity, making it more interesting and challenging. Since its creation, BP4D+ has been widely used in several works related to affective computing and vital signs measurement [26, 46, 50].

This dataset consists of RGB and thermal images, 2D and 3D facial landmarks, actions units and 8 physiological signals collected with contact sensor. 140 subjects (82 females and 58 males) of different ethnic ancestry participated in 10 sessions designed to induce the following emotions : Happiness (T1), Surprise (T2), Sadness (T3), Startle (T4), Skeptical (T5), Embarrassment (T6), Fear (T7), Pain (T8), Anger (T9), and Disgust (T10). 1400 RGB videos lasting 30 seconds to 1 minute were recorded at a frame rate of 25 fps. The resolution of each image is 1040×1392 pixels. Among the 10 tasks, only four emotions are used in our experiments, corresponding to happiness, embarrassment, fear and pain. These emotion tasks are provided with manually coded action units (33 in total) that were computed only for the most expressive frames of each task.

3.2. Data preparation

First, the most expressive frames are extracted from each emotion task using action units code provided in the database. Then, we follow the same protocol used in [37]. A robust face swapping-based segmentation method is used to get rid of non-skin regions that do not hold any color changes associated with cardiac activity [35]. This step improves imaging photoplethysmographic signal extraction from face skin. All the images of the segmented faces are cropped according to the coordinates of the non-zero pixels and then scaled to $48 \times 48 \times 3$. Besides, data augmentation strategy is applied for the training set to create additional and different training instances. Several image transformations such as rotating the image by varying degrees, translating it and flipping it horizontally and vertically, cropping, zooming in, or changing the contrast of the image have been randomly applied on video fragments. It helps to reduce overfitting and improve the generalizability of the model.

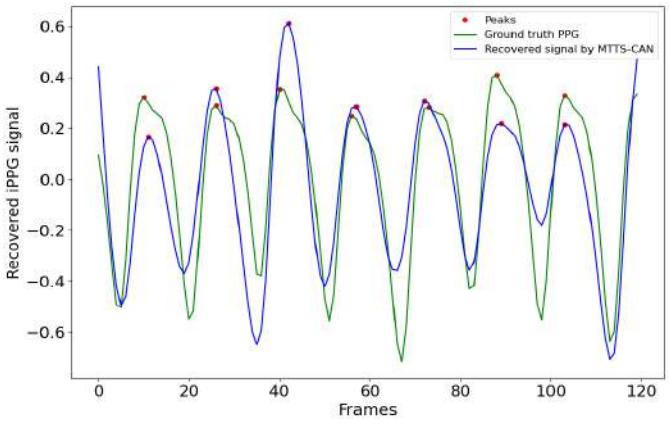


Figure 2. Comparison between a predicted signal by MTTS-CAN and the ground-truth PPG signal taken from BP4D+ dataset.

3.3. Video-based physiological signals measurement

In this study, physiological parameters are measured remotely using imaging photoplethysmography method [6]. iPPG is an optical technique for capturing cardiac signals by observing the blood-volume variations on a person's face using a simple camera. The captured light reflected by the skin is translated to a variation of the iPPG signal. Several important vital signs can be derived from the iPPG waveform such as pulse rate, respiration rate and heart rate variability (HRV). However, among these physiological features, only iPPG signal and its derivative HRV features have been used in our experiment. It was reported in several studies that heart rate variability is one of the most important physiological characteristic that reflects affective states of a person [3, 30]. HRV features can be derived from time interval variation between consecutive heartbeats in iPPG signal. [14].

iPPG extraction algorithms can be divided to hand-crafted based algorithms [52] that use signal/image processing steps and deep learning based approaches [34]. In this work, we used a multi-task sequential shift convolutional attention network (MTTS-CAN) proposed by Liu et al. to extract the iPPG signal [29]. MTTS-CAN is one of the recent popular state-of-the-art deep learning based method that provides good performance in terms of heart and respiratory rates measurement. In order to better appreciate the quality of the recovered iPPG signal, we present, in Figure 2, a superposition of a ground truth PPG signal recorded by contact sensor and the iPPG signal predicted by the MTTS-CAN network. It is clear that the estimated iPPG signal is strongly correlated with the ground truth and the location of the peaks is very close.

The core module of MTTS-CAN is a hybrid network that uses the attention mechanism in conjunction with Tem-

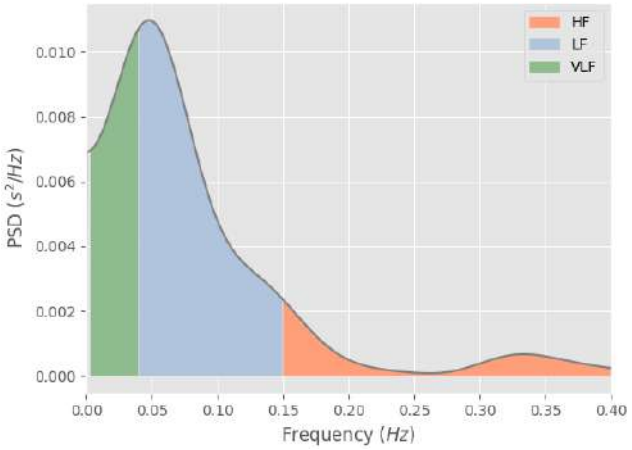


Figure 3. A representative PSD for IBI signal showing The areas of VLF, LF and HF powers of the HRV.

poral Shift Modules [29]. The recovered iPPG signals by MTTS-CAN allow HRV features extraction both in the time-domain and in the frequency-domain. For both time and frequency analysis, peak detection is performed to locate the instant of time at which heartbeat occurs (which allows to compute HRV features).

In time domain, heart rate is calculated as the inverse of the of the interbeat interval (IBI) divided by 60 to get the frequency in beats per minute. From the heart rate variations in the selected window, we computed the mean (meanHR) and standard deviation (stdHR) of the heart rate series. The root mean square of successive interval differences (RMSSD) is also calculated (see Equation 1). This parameter allows assessing vagal activity reflected in heart variability [40].

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2} \quad (1)$$

In frequency domain, the IBI series were interpolated with cubic Hermite and the power spectra were obtained by employing Welch's method [48]. The power spectral density (PSD) of a signal makes it possible to analyze its different oscillatory components such as HRV low frequency (LF) and high frequency (HF) components. The LF component is modulated by baroreflex activity and contains both sympathetic and parasympathetic activity, while the HF component reflects parasympathetic branch of the ANS [1]. The LF and HF powers of the HRV were computed as the area under the PSD curve corresponding to 0.04-0.15Hz and 0.15-0.4Hz respectively (see Figure 3). We also computed the ratio LF/HF, which represents the sympatho-vagal balance [4]. The very low frequency (VLF) components were not employed in our experiments.

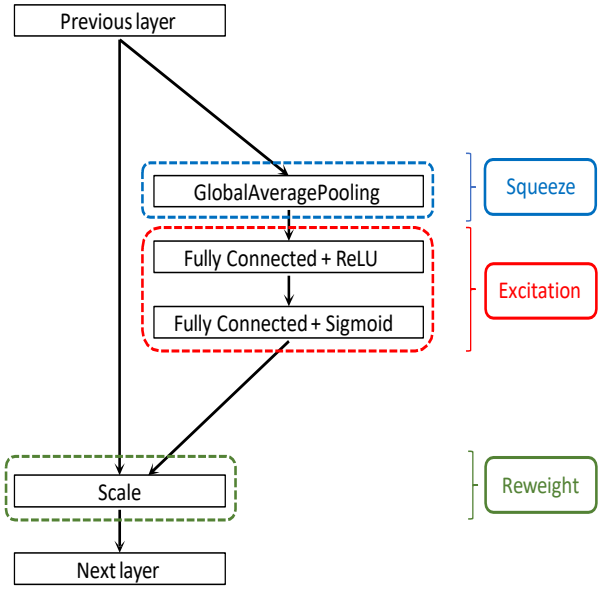


Figure 4. The Squeeze-and-Excitation module consists of global average pooling as a Squeeze operation. The two Fully Connected layers are then used to to learn the feature weights. We first reduce the feature dimension with a shrinkage parameter r, then we recover the dimension with the same r in the next fully connected layer. After the excitation operation, the SE block use the scale operation to re-weight the input layers, by element-wise multiplying the raw input by the excitation output.

3.4. Facial expressions recognition network

Xception network is one of the state-of-the-art methods that has proven efficient for general purpose 2D image tasks in terms of accuracy, fast convergence speed and low computational costs [7]. Xception is a derivative of Inception network [44]. It replaces Inception modules with depthwise separable convolution layers and adds residual connections. This modification, compared to Inception architecture, greatly reduces the computational cost and memory requirements, while maintaining similar (or slightly better) performance. The depth-wise separable convolution performs spatial convolution by channel separately without considering the relationship between different channels, while conventional convolution considers all spatial and channel information together. Exploiting channel dependency is an important way to improve convolutional neural network. Therefore, we fuse Xception network with Squeeze and Excitation (SE) [16] module to achieve channel weighting and maintain or improve classification ac-

curacy while reducing the number of parameters and the amount of computation. The SE block aims to explicitly model the interdependency between the channels of the image, in order to recalibrate the channel-wise feature maps in a computationally efficient manner.

The structure of the SE block is depicted in Figure 4. The SE processing blocks are composed of two successive parts: Squeeze and Excitation. The squeeze operation uses a global average pooling layer, while the excitation phase consists of two fully-connected layers that take the rectified linear units and sigmoid activation units as the hidden units respectively. In our implementation, 3D version of Xception network and SE block are used instead of the original implementations that only consider the spatial information. In this way, we simultaneously extract spatio-tempoal features without adding additional layers to take into account the temporal features.

Figure 5 presents the overall architecture of the proposed 3D-SE-XceptionNet which consists of three blocks (entry, middle and exit) as the original architecture of Xception network. However, the model structure is simplified by reducing the number of repetitive depthwise separable convolution layers. Our new mini Xception includes 15 convolution layers instead of 36 compared to the original version. These convolutional layers are structured into 14 modules, all linked with shortcuts as in ResNet architecture [15] except the first and last modules. SE blocks are inserted after the residual connections. The output of the features extraction is flattened and passed to two dense layers with 256 and 4 neurons respectively. The first dense layer takes the rectified linear units as the hidden units while the second takes the softmax activation function to predict the corresponding emotion classes.

4. Results and Discussion

The BP4D+ dataset was split to 90 percent training set and 10 percent validation set. Training and validation were performed three times with different samples in order to verify the consistency of the system. Three different experiments were conducted to classify emotions : using (a) facial expressions only, (b) physiological modalities only, and (c) facial expressions and physiological signals together.

4.1. Implementation details

The proposed system is implemented with Keras and tensorflow frameworks and ran on Nvidia Quadro P6000s. As BP4D+ is sampled at 25 fps, the length of face video clip is set to Nbframes = 100 frames (corresponding to 4 seconds) while the size of each image frame is $48 \times 48 \times 3$ ($ImHeight \times ImWidth \times Channel$). We used Rectified Adam (RAdam) optimizer [27] to optimize a categorical crossentropy loss function. We trained the network for 50 epochs with batch size = 16, learning rate 10^-4 and decay = 10^-2. L1 and L2 regularization strategies with coefficient equal 10^-2 are employed which help to overcome overfitting issue and improve the model generalizability to new data.

4.2. Emotion recognition from facial expressions

5 state-of-the-art networks are compared : 3D-VGG [41], 3D-ResNet [15], 3D-DenseNet [17], 3D-Inception [44] and 3D-Xception [7]. We train these architectures using the BP4D+ dataset and then we compare their performance with our proposed model. As shown in Table 1, our 3D-SE-Xception network outperforms the state-of-the-art deep learning architectures. Note that in the conducted experiments, we do not perform any special preprocessing to the input images except face segmentation (See section 3.2). Compared to other architectures, the accuracy improves to the highest value of 63.40% when the Xception network is fused with the SE block. The proposed framework derives more targeted feature information through the SE module, meanwhile using the Xception network to avoid the vanishing gradient problem through residual connections and reduce the computational cost and memory requirements through the depthwise separable convolutions.

Table 1. Comparison of proposed method to state-of-the-art networks on spontaneous data for facial expression recognition.

Method	Accuracy
3D-DenseNet [17]	37.91
3D-Inception [44]	42.48
3D-ResNet [15]	44.44
3D-VGG [41]	49.02
3D-Xception [7]	53.59
3D-SE-Xception (Ours)	**63.40**

Figure 6 shows the confusion matrix for the emotion recognition system based on facial expressions. The overall performance of the proposed network was 63.4%. Happiness and pain are the most recognized emotions with an accuracy of 80% and 81% respectively, while fear is missclassified as happiness and pain. This can partially be explained by the multiple behaviors that may occur during the expression of this emotion.

4.3. Emotion recognition from Physiological signals

Emotion classification from physiological signals is performed using iPPG signals and HRV features. Three different fusion schemes were conducted for emotion recognition

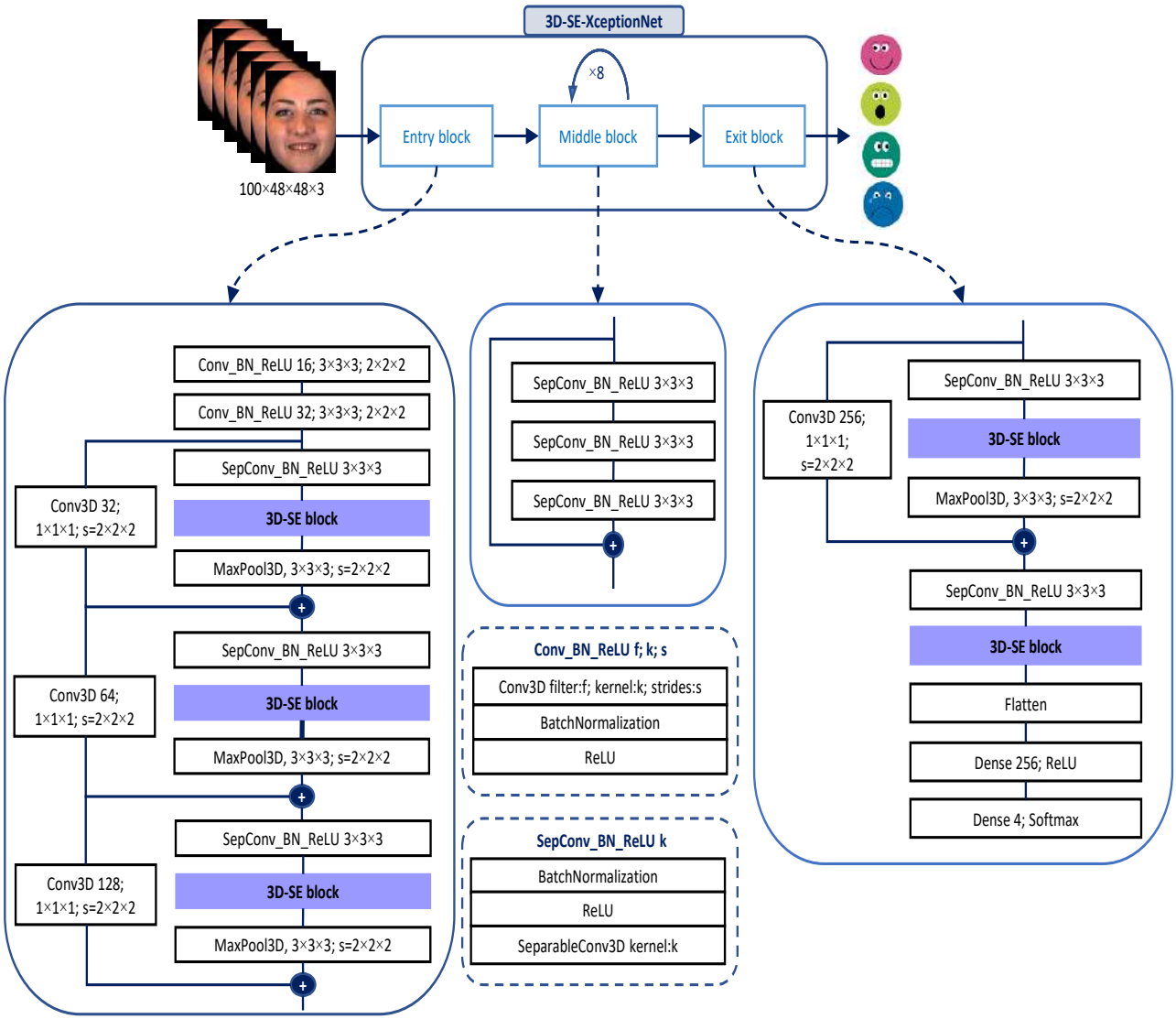


Figure 5. The network structure of 3D-SE-Xception corresponds to a modified version of the Xception network. 2D depthwise separable convolution layers are replaced by 3D depthwise separable convolution to capture both spatial and temporal features across video frames. The SE block was embedded in the model to enhance the useful feature channels and weaken the useless feature channels through channel-wise feature maps recalibration. Two dense layers are used instead of Global Average Pooling.The input video fragment first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow which ends in a dense layer with 4 neuron to classify emotions.

using physiological data. First, iPPG signals and HRV features are used separately to classify emotions. Then, we merge them to see which approach gives the best accuracy.

Inspired by the work of Fabiano et al. [13], a feedforward neural network is used in our experiments. It consists of two layers. The input layer has the same number of neurons as the input length (100 for iPPG modality , 6 for HRV), while the output layer includes the same number of neurons as the number of classes of emotion to predict. The activation function for the input layer is ReLU, while the softmax activation function is employed for the output layer.

Table 2 illustrates the recognition accuracy using iPPG signals and HRV features separately and after fusion between them. As can be seen from table 2, whether physiological signals are used separately or combined, the recognition accuracy is low compared to facial expressions. Besides, the performance when using iPPG signals is better than HRV. This can be justified by the short length of the

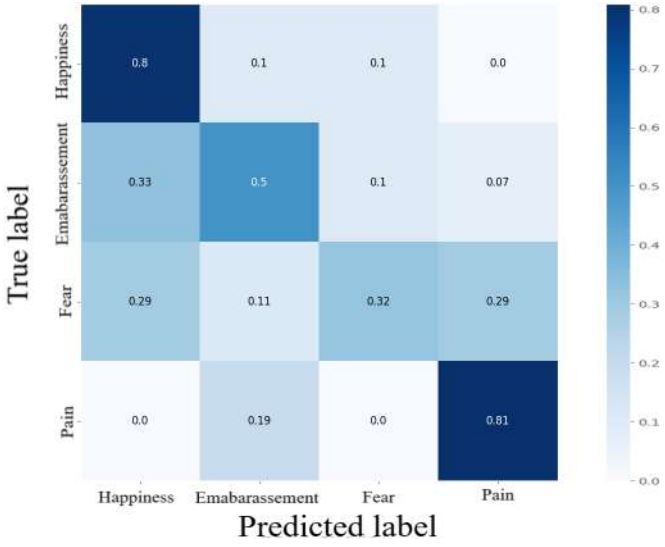


Figure 6. Emotion classification confusion matrix using facial expressions.

iPPG signals used for HRV analysis as well as the signal quality which is prone to noise and artifacts due to movements and lighting conditions. Therefore, it has an impact on the accuracy of HRV characteristics. On the other hand, iPPG and HRV fusion exhibit lower performance. This may be related to the lack of correlation between the iPPG signal and the HRV characteristics.

Table 2. Comparison of emotion recognition accuracy from physiological signals. Abbreviations: (iPPG :Emotion from iPPG signals), (HRV : Emotions from HRV features), (iPPG + HRV : Emotions from the combined iPPG and HRV).

Method	Accuracy
iPPG	55.33
HRV	53.59
iPPG + HRV	44.64

4.4. Multimodal emotion recognition

The architecture of our multimodal emotion recognition system is shown in Figure 1. Basically, the proposed model consists of two pipelines allowing to extract the features of each modality from video streams (See section 3.3 and 3.4). Each video of BP4D+ is fed to the facial expression network (3D-SE-Xception) and to the iPPG signal network (MTTS-CAN). The first pipeline extracts the features vector after the flatten layer (See Figure 1 using the pre-trained weights of our 3D-SE-Xception model, while the second pipeline returns either the iPPG signal recovered through the MTTS-CAN network or HRV features. Hence, three experiments have been carried out for multimodal emotion recognition. First, facial expressions features are combined with only the iPPG signal, then only with HRV vector. Finally, all modalities are fused. The contactenation result vector is then passed to two dense layers with 256 and 4 neurons respectively. The first dense layer takes the rectified linear units as the hidden units while the second takes the softmax activation function to predict the corresponding emotion class.

The recognition accuracy for each experiment is reported in Table 3. The results show that combining facial expression features with physiological parameters improve the performance compared to unimodal approach either using facial expressions or physiological data separately. This confirms previous studies that have obtained the same results where the precision of the fusion exceeds unimodality systems, and the performance of facial expressions modality is always better compared to physiological signals [8, 18]. Furthermore, the lack of correlation between the iPPG signal and HRV features impacts performance, whether merging just these two modalities or their fusion with facial expressions.

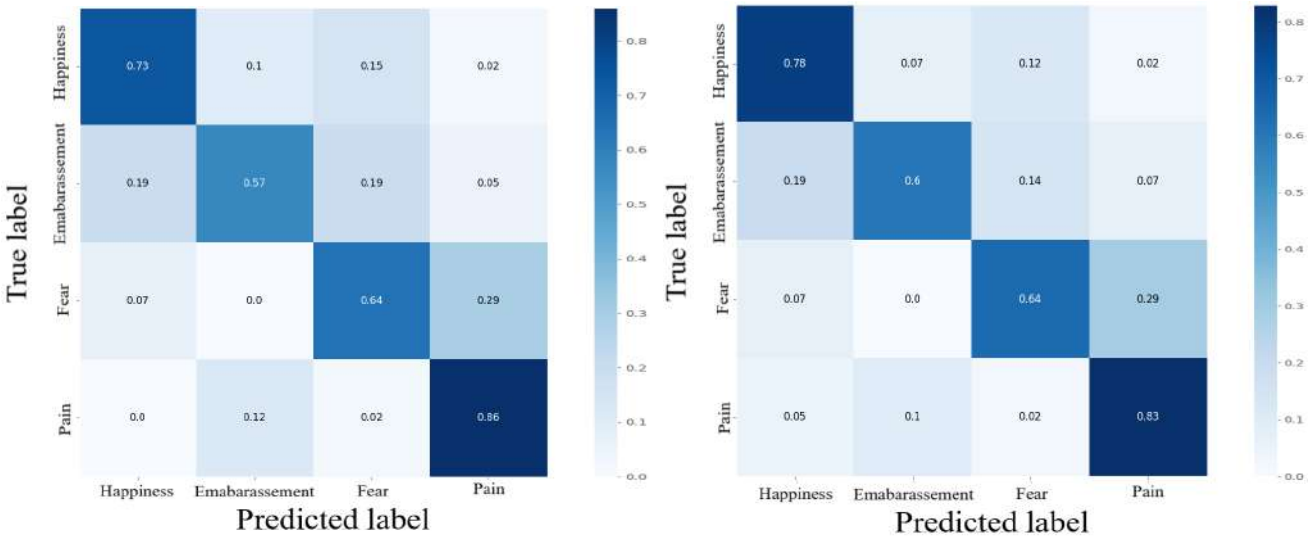
Table 3. Comparison of multimodal emotion recognition accuracy from facial expressions and physiological signals.

Method	Accuracy
Facial expressions + HRV	70.59
Facial expressions + iPPG	71.90
Facial expressions + iPPG + HRV	67.97

Figure 7 shows the confusion matrix for the multimodal emotion recognition system based on facial expressions and HRV features fusion, and facial expressions and iPPG fusion. The overall performance of the proposed network is 70.59% and 71.90 % respectively. Compared to using facial expressions only, the fusion with physiological signals improved significantly the accuracy for missclassified emotions. For example, fear accuracy has been doubled from 32% to 64% for each fusion schemes.

4.5. Discussion

Facial expressions and physiological signals modalities establish superiority to each other. The combination of facial expression features and iPPG signal achieved the highest accuracy of around 72%. This slightly outperforms the fusion between facial expressions and HRV features. However, merging only the iPPG signal and HRV features, or



(a) Facial expressions and HRV features fusion.

(b) Facial expressions and iPPG signals fusion.

Figure 7. Multimodal emotion classification confusion matrix using facial expressions with HRV Features and iPPG signals.

with facial expressions features, gives the lowest accuracy. We hypothesize that these two modalities may interfere with each other, thus impacting the recognition accuracy. In addition, using multiple modalities considerably improved performance for miss-classified emotions such as Fear. Although facial expressions are visible and easy to categorize compared to physiological cues, incorporating with physiological modalities can provide complementary information and further enhance the performance. On the other hand, the results obtained fit perfectly with existing multimodal systems that use multiple input data sources and demonstrate the possibility of using only facial videos to recognize emotions using human physiological and physical cues.

5. Conclusion

This paper proposes a new framework for multimodal emotion recognition through facial expressions and physiological signals. A novel spatiotemporal neural network has been proposed, which fused Squeeze-and-Excitation modules with a 3D Xception network to recalibrate the channel-wise feature maps in a computationally efficient manner. Two physiological parameters were selected, namely the iPPG signal and the HRV features. Unlike existing studies, physiological cues were measured remotely based on imaging photoplethysmography method. This way, only single input source were used to extract features from each modality. It is very interesting and promising to recognize emotions in a multimodal way with a single non-intrusive sensor. using a camera that is integrated on all digital devices used in daily life allows to reduce the cost and to make the system more accessible. Furthermore, video-based physiological signals measurement is more practical and may reduces the discomfort caused by the contact devices. Overall, we have shown that fusion of two modalities (facial expressions with iPPG signals or facial expressions with HRV features) gives significant improvements and offer potential for more accurate recognition of affects and emotions.

As future work, we intend to incorporate other physiological signals and test the performance on other multimodal emotion datasets. We will further explore the complexity of expressions to understand the poor performance of certain emotions.

References

[1] Bradley M. Appelhans and Linda Luecken. Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, 10(3):229–240, Sept. 2006. 4

[2] Yannick Benezeth, Peixi Li, Richard Macwan, Keisuke Nakamura, Randy Gomez, and Fan Yang. Remote heart rate variability for emotional state monitoring. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 153–156, 2018. 1

[3] Yannick Benezeth, Peixi Li, Richard Macwan, Keisuke Nakamura, Randy Gomez, and Fan Yang. Remote heart rate variability for emotional state monitoring. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 153–156. IEEE, 2018. 2, 3

[4] Robert L. Burr. Interpretation of normalized spectral heart rate variability indices in sleep research: a critical review. *Sleep*, 30 7:913–9, 2007. 4

[5] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP Interna-*

tional Conference on Artificial Intelligence Applications and Innovations, pages 375–388. Springer, 2007. 2

[6] A V J Challoner and C A Ramsay. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Physics in Medicine and Biology*, 19(3):317–328, may 1974. 2, 3

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. 4, 5

[8] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, 8:168865–168878, 2020. 1, 2, 7

[9] Djamaleddine Djeldjli, Frédéric Bousefsaf, Choubeila Maaoui, and Fethi Bereksi-Reguig. Imaging photoplethysmography: Signal waveform analysis. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, pages 830–834, 2019. 1

[10] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Buinskas. Human emotion recognition: Review of sensors and methods. *Sensors (Basel, Switzerland)*, 20, 2020. 2

[11] Sidney K. D'Mello and Jacqueline Kory Westlund. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47:1 – 36, 2015. 2

[12] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence. 1, 2

[13] Diego Fabiano and Shaun Canavan. Emotion recognition using fused physiological signals. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 42–48, 2019. 6

[14] Miha Fingar and Primoz Podrzaj. Feasibility of assessing ultra-short-term pulse rate variability from video recordings. *PeerJ*, 8, 2020. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[18] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational Intelligence and Neuroscience*, 2017, 2017. 1, 2, 7

[19] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 3

[20] Dimitrios Kollias and Stefanos Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 12(3):595–606, 2020. 1

[21] Jukka Kortelainen, Suvi Tiinanen, Xiaohua Huang, Xiaobai Li, Seppo Laukka, Matti Pietikäinen, and Tapio Seppänen. Multimodal emotion recognition by combining physiological signals and facial expressions: A preliminary study. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5238–5241, 2012. 2

[22] Jyoti Kumari, R. Rajesh, and K.M. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15). 2

[23] James Ren Lee, Linda Wang, and Alexander Wong. Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition. *Frontiers in Artificial Intelligence*, 3, 2021. 2

[24] Min Seop Lee, Yun Kyu Lee, Dong Sung Pae, Myo Taeg Lim, Dong Won Kim, and Tae-Koo Kang. Fast emotion recognition based on single pulse ppg signal with convolutional neural network. *Applied Sciences*, 2019. 1

[25] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 2

[26] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018. 3

[27] 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. 5

[28] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 2

[29] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 3, 4

[30] Timur Luguev, Dominik Seuß, and Jens-Uwe Garbas. Deep learning based affective sensing with remote photoplethysmography. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–4, 2020. 1, 2, 3

[31] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021. 1

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1

[33] Nazmun Nahid, Arafat Rahman, and Md Atiqur Rahman Ahad. Contactless human emotion analysis across different modalities. 2021. 2

[34] Aoxin Ni, Arian Azarang, and Nasser Kehtarnavaz. A review of deep learning-based contactless heart rate measurement methods. *Sensors*, 21:3719, 05 2021. 3

[35] 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. 3

[36] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2):505–523, 2018. 2

[37] Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf, and Choubeila Maaoui. Lcoms lab's approach to the vision for vitals (v4v) challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2754, 2021. 3

[38] Aasim Raheel, Muhammad Majid, Majdi R. Alnowami, and Syed Muhammad Anwar. Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. *Sensors (Basel, Switzerland)*, 20, 2020. 1

[39] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 2

[40] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017. 4

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[42] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 3

[43] Bo Sun, Liandong Li, Guoyan Zhou, and Jun He. Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6):1 – 8, 2016. 1

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4, 5

[45] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J. Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. 2

[46] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016. 3

[47] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019. 2

[48] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. 4

[49] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018. 1

[50] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[51] Yi Yang, Qiang Gao, Xiaolin Song, Yu Song, Zemin Mao, and Junjie Liu. Facial expression and eeg fusion for investigating continuous emotions of deaf subjects. *IEEE Sensors Journal*, 21(15):16894–16903, 2021. 1

[52] Sebastian Zaunseder, Alexander Trumpp, Daniel Wedekind, and Hagen Malberg. Cardiovascular assessment by imaging photoplethysmography – a review. *Biomedical Engineering / Biomedizinische Technik*, 63:617 – 634, 2018. 3

[53] Zheng Zhang, J. Girard, Yue Wu, X. Zhang, Peng Liu, U. A. Ciftci, S. Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, J. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016. 3

[54] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 1

[55] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016. 1

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-

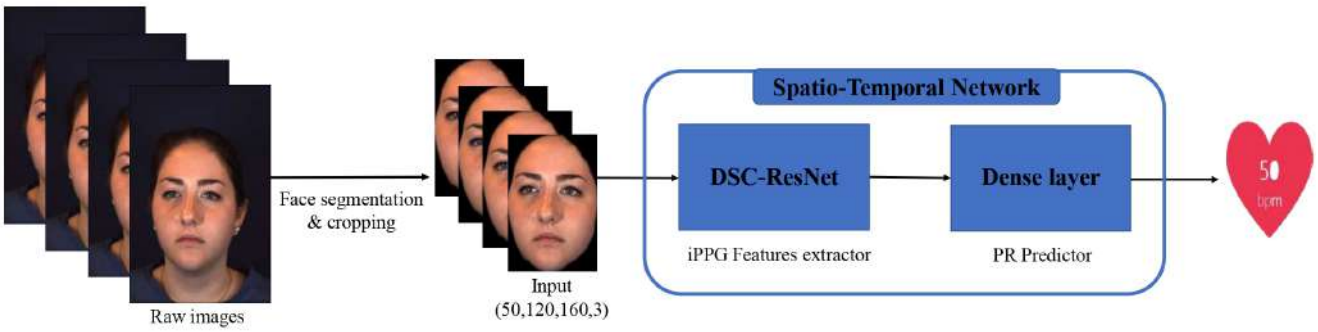


Figure 1. LCOMS Lab's solution pipeline

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.

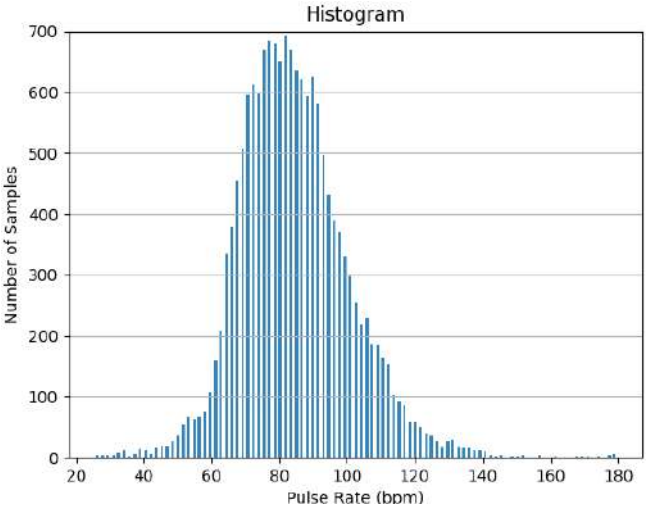


Figure 2. Distribution of the ground truth pulse rates in the V4V database.

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.

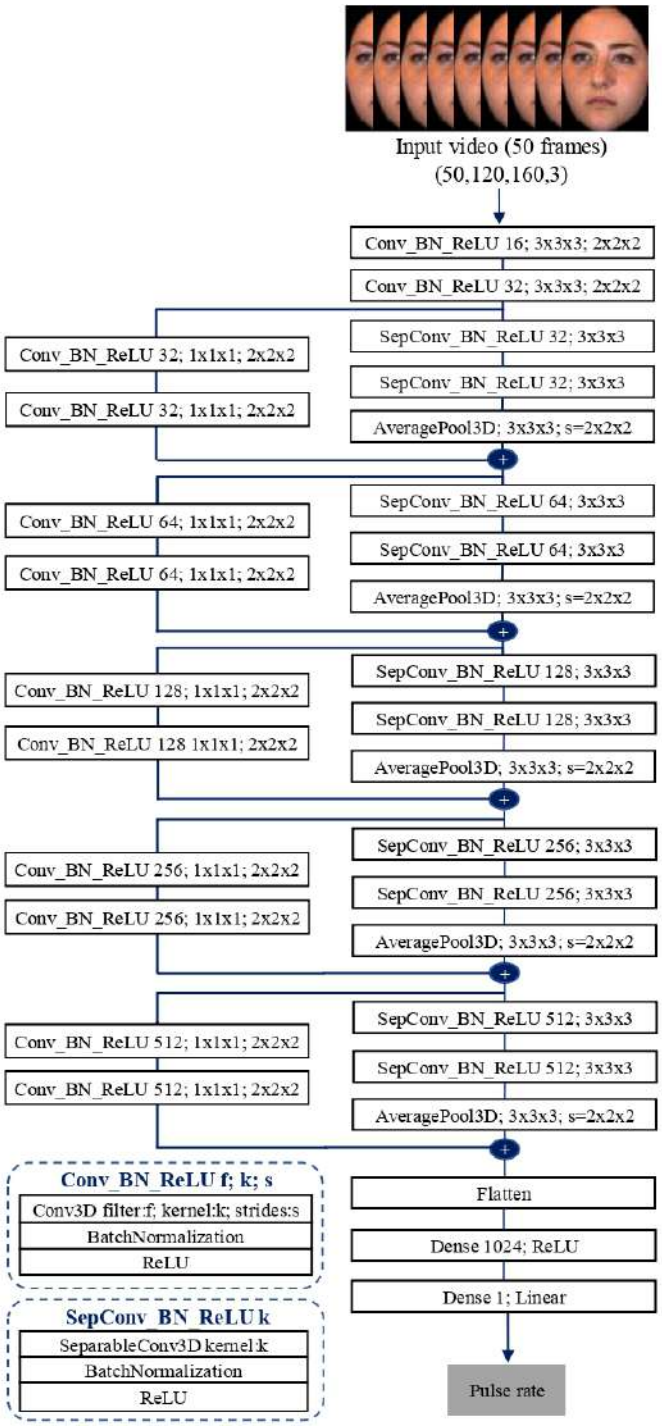


Figure 3. The framework of spatio-temporal networks for pulse rate estimation directly from facial videos recording.

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.

Transformée en ondelettes et IA pour la reconstruction d'un signal PPG en contact à partir de sa version sans contact

Frédéric Bousefsaf, Djamaleddine Djeldjli, Choubeila Maaoui, Alain Pruski

Transformée en ondelettes et IA pour la reconstruction d'un signal PPG en contact à partir de sa version sans contact

Frédéric BOUSEFSAF, Djamaleddine DJELDJLI, Yassine OUZAR, Choubeila MAAOUI, Alain PRUSKI

LCOMS, Université de Lorraine
7 rue Marconi, 57070 Metz, France
frederic.bousefsaf@univ-lorraine.fr

Résumé – La mesure de signaux photopléthysmographiques (PPG) sans contact est une technique de mesure non invasive permettant d'estimer un ensemble de fonctions vitales par analyse vidéo délivrée par caméra. Nous proposons, dans cet article, une méthode permettant de convertir un signal PPG mesuré par caméra en un signal PPG mesuré en contact (cPPG). L'objectif à plus long terme consistera à transformer le signal cPPG en signal de tension artérielle afin de proposer une chaine de traitement permettant d'estimer la tension à partir d'une vidéo. La méthode que nous proposons dans cet article repose sur la transformée en ondelettes et sur des modèles d'IA modernes. Les résultats reflètent la pertinence de l'approche et montrent qu'une estimation de la pression artérielle à partir d'un signal PPG caméra converti en signal en contact est envisageable.

Abstract – Imaging photoplethysmography (iPPG) is an optical technique dedicated to the assessment of several vital functions using a simple camera. We here propose a method for converting iPPG to contact PPG (cPPG) signals for, in future works, translating this cPPG signal to blood pressure. This would allow remote measurement of blood pressure from video. The continuous wavelet transform of cPPG and iPPG signals and deep neural networks are employed in this study. The results exhibit good agreements towards several metrics, showing that the neural architectures properly estimated cPPG from iPPG signals through their CWT representations.

1 Introduction

Les recherches portant sur la mesure de signaux physiologiques par des technologies sans contact ont connu des avancées significatives ces dernières années [1]. La photopléthysmographie (PPG) est mesurable à distance en observant les fines fluctuations de la couleur de la peau d'une personne. Le domaine est en plein essor et est soutenu par un ensemble d'études [2]. Des méthodes issues de la vision par ordinateur, du traitement d'images et de l'intelligence artificielle (IA) ont été utilisées ou développées spécifiquement pour transformer avec fiabilité la vidéo d'entrée en paramètres biomédicaux. Ces méthodes reposent principalement sur des modèles neuronaux [3].

Les recherches dans ce domaine s'orientent désormais vers la mesure de nouveaux paramètres physiologiques tels que la tension artérielle [4]. La mesure de la tension par analyse vidéo est complexe et peu de travaux montrent sa faisabilité. Deux directions sont à l'étude : (i) la mesure du *pulse transit time*, paramètre admis comme étant corrélé avec la pression artérielle [5] ainsi que (ii) l'étude directe de l'onde PPG [4]. Les résultats de ces études sont mitigés.

Des tentatives d'utilisation de modèles d'IA ont récemment été proposées [6]. L'apprentissage des modèles est cependant contraint par les faibles quantités de données actuellement disponibles. L'apprentissage d'un modèle neuronal profond permettant d'estimer avec précision la pression artérielle par analyse vidéo est donc difficilement enviseagable pour le moment. Nous avons récemment montré que les caractéristiques temporelles, de courbure et de surface des signaux PPG évoluent de manière comparable entre les mesures caméra et les mesures en contact pris au doigt ou à l'oreille [7]. Ce point est important car il motive la présente étude. Nous partons de l'hypothèse qu'un signal PPG mesuré par caméra (imaging PPG, iPPG) peut être converti en un signal PPG en contact (cPPG) par le

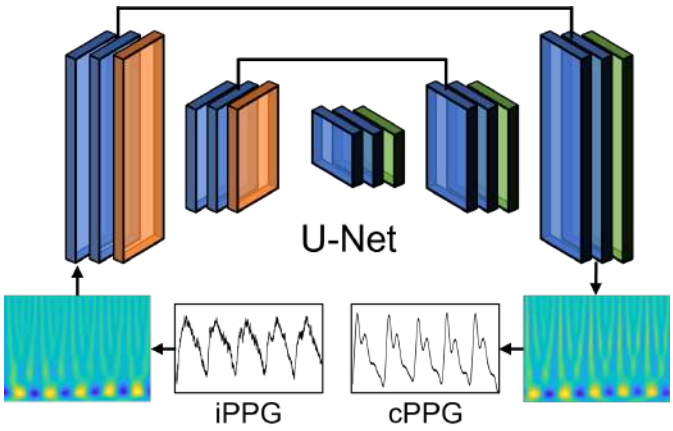


FIGURE 1 – Vue d'ensemble de la méthode proposée. Le signal iPPG est calculé à partir d'une analyse vidéo du visage de la personne. Sa représentation en ondelettes traverse le réseau U-Net. La transformée inverse de la représentation en ondelettes prédite permet de former le signal PPG en contact (cPPG) mesuré traditionnellement via un capteur placé sur le doigt.

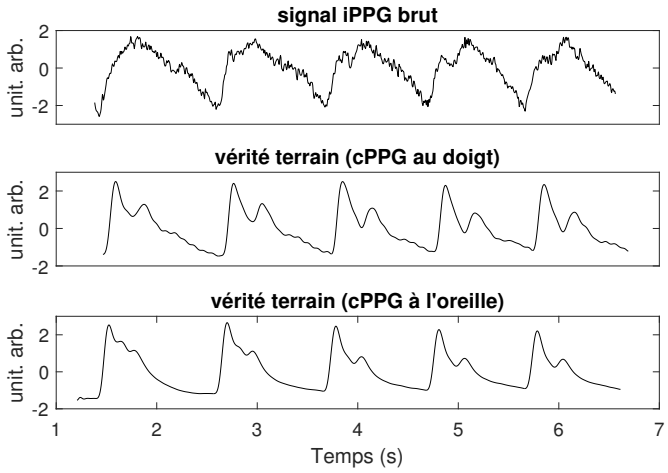


FIGURE 2 – Illustration de signaux extraits du participant # 1 durant une phase de maintien de respiration.

biais d'un modèle d'IA utilisant pour entrée la représentation en ondelettes continue des signaux. L'objectif final et à plus long terme consistera à transformer ce signal cPPG en signal de tension artérielle car ce champ est soutenu par une littérature plus mature, avec des études présentant des méthodes dont les performances respectent les standards internationaux [8].

2 Méthodes

2.1 Base de données et protocole expérimental

Les données utilisées pour apprendre les modèles neuronaux présentés en section 2.3 ont été présentées dans un article publié précédemment. 12 volontaires ont participé à l'étude. L'âge des participants est compris entre 20 et 35 ans. Ils ont été placés à environ 1 mètre d'une caméra rapide (125 fps). Les références ont été acquises grâce à des capteurs PPG en contact placés au doigt et à l'oreille. Deux essais de 60 secondes ont été proposés aux participant de l'étude. Premier test : nous demandions aux participants de rester au calme et de respirer normalement. Second test : il était demandé aux participants de retenir leur respiration autant que faire se peut, l'objectif étant de provoquer des variations physiologiques qui modifient la pression artérielle et impactent les signaux PPG enregistrés. Nous renvoyons le lecteur vers la publication originale pour plus de détails concernant la procédure et le matériel utilisé [7].

La base de données contient 724 signaux échantillonnés sur 256 points. Chaque signal contient 5 ondes PPG. La base de données est aléatoirement séparée en deux jeux : 75 % est dédié à l'entrainement des réseaux (soit 543 signaux) et 25 % à la validation (181 signaux).

2.2 Traitement des images et des signaux

Le front correspond à une région d'intérêt pertinente en matière de rapport signal sur bruit [9]. La région est détectée à partir d'un modèle composé de 68 points épousant les formes principales du visage. Ces différents points sont suivis le long de la vidéo et certains d'entre eux permettent de calculer automatiquement la position du front. En pratique, les algorithmes de détection du visage et des caractéristiques faciales respectivement inclus dans les librairies OpenCV et Dlib ont été utilisés.

Le signal PPG caméra est construit à partir d'une moyenne spatiale sur le canal vert des pixels du front. Cette technique a été utilisée dès les toutes premières publications relatives à la mesure de signaux PPG sans contact par caméra [9]. Les tendances basses fréquences du signal brut sont supprimées par un filtre passe-bas spécifique [10]. Une détection robuste des vallées est ensuite calculée pour extraire chaque onde. In fine, chaque signal de la base de données est échantillonné sur 256 points et contient 5 ondes PPG successives. Un signal calculé à partir d'une des vidéos est présenté en figure 2. Les signaux en contact de référence mesurés au doigt et à l'oreille sont aussi présentés sur cette figure. Tous les signaux ont été centrés (moyenne nulle) et réduits (écart type égal à un).

Il est proposé, dans cet article, d'exploiter la représentation en ondelettes pour entrainer les différentes architectures neuronales présentées en section 2.3. L'utilisation directe du signal iPPG en entrée d'un modèle d'IA est soutenue par une littérature très faible [6] et des essais préliminaires mais non concluants ont été menés par notre équipe de recherche (résultats non publiés).

La transformée en ondelettes continue d'un signal correspond à une représentation temps-fréquence calculée à partir d'une fonction prototype communément appelée ondelette mère. Contrairement à la transformée de Fourier, la transformée en ondelettes permet de détecter des variations abruptes de fréquence dans les signaux. Différentes ondelettes mères ont été développées et le choix dépend principalement de l'application et des propriétés du signal analysé. L'ondelette mère de Morlet, déjà utilisée dans de précédents travaux relatifs à l'analyse de la PPG par caméra [11], a été retenue dans cette étude.

La transformée en ondelettes continue a été calculée sur chaque signal PPG dans la plage de fréquences physiologiques des battements du coeur humain, soit [0.6, 4.5] Hz [2]. La représentation en ondelettes qui servira à entrainer les architectures neuronales est de dimension 256×256. Un signal caméra et en contact au doigt avec leur représentation en ondelettes respective (partie réelle) sont présentés en figure 1. Notons la différence de forme entre les signaux et de phase entre les représentations en ondelettes : la partie réelle du signal caméra démarre sur une série de coefficients de faible intensité (pseudo-ellipse bleue) tandis que la partie réelle du signal en contact démarre sur des coefficients de forte intensité (pseudo-ellipse jaune). Il s'agit d'une particularité que le réseau de neurones apprendra pendant la phase d'entrainement.

2.3 Développement des architectures neuronales

Nous proposons d'exploiter l'architecture U-Net initialement utilisé dans le cadre de la segmentation d'images médicales. Cette architecture est constituée d'une branche descendante (en-

codeur) complétée par une branche ascendante (décodeur), donnant une forme de U au réseau. La branche descendante contient un enchevêtrement de couches de convolution et de *pooling*. La branche ascendante intègre des couches de déconvolution connectées aux convolutions de la branche descendante. Les connexions permettent de restaurer l'information spatiale. Une représentation schématique du réseau est proposée en figure 1.

Des squelettes (*backbones*) peuvent être intégrés dans la partie encodeur du réseau U-Net. Les paramètres internes du squelette sont bloqués pendant l'entrainement (les poids du réseau restent fixes). Il s'agit en pratique de modèles pré-entrainés sur la base de données ImageNet pour des tâches de reconnaissance d'objet dans les images [12]. L'apprentissage d'un réseau U-Net soutenu par un squelette consiste à optimiser les paramètres internes de la partie décodeur. Cette stratégie est similaire à un apprentissage par transfert. Différents squelettes populaires ont été testés : la version 16 couches (VGG–16), la version 101 couches de ResNet [13], la version 201 couches du réseau DenseNet [14] ainsi que les réseaux Inception [15] InceptionV3 et InceptionResNetV2. Les techniques conventionnelles de régularisation n'ont pas été introduites tandis qu'un schéma de normalisation (i.e. *batch normalization*) est utilisé dans les réseaux possédant un squelette. La tâche ne correspond pas à une classification de données mais à une régression sous la forme d'une reconstruction pixel à pixel d'une représentation en ondelettes sur deux canaux. Le nombre de variables à entrainer (poids et biais) est compris entre 2 et 9 millions.

3 Résultats et discussion

3.1 Performances des apprentissages

La fonction de coût (*loss*) des moindres carrés (*mean squared error*) a été utilisée :

$$MSE = \frac{1}{n}\sum_{i,j}\left(CWT_{i,j} - \widehat{CWT}_{i,j}\right)^2 \qquad (1)$$

CWT correspond à la transformée en ondelettes (voir section 2.2) du signal PPG en contact et $\widehat{CWT}$ à celle prédite par le réseau de neurones à partir de la transformée calculée sur le signal caméra.

Les valeurs minimales des courbes d'évolution de la fonction de coût pour chaque réseau sont répertoriées dans le tableau 1. Indépendamment du site de mesure, le réseau utilisant VGG–16 pour squelette présente la plus faible MSE, traduisant ainsi les meilleurs performances en terme de reconstruction de la représentation en ondelettes. Notons tout de même que les valeurs minimales sont proches, en particulier celles calculées à partir des signaux PPG mesurés sur l'oreille.

Réseau	MSE_{doigt}	$MSE_{oreille}$
U-Net1	0.327	0.231
VGG–16	**0.282**	**0.228**
ResNeXt101	0.316	0.227
InceptionResNetV2	0.323	0.238
InceptionV3	0.318	0.233
DenseNet201	0.308	0.229

TABLE 1 – Minimum de la fonction de coût pour chaque modèle. U-Net1 correspond au réseau initial n'intégrant pas de squelette. Les autres réseaux correspondent à des architectures U-Net soutenues par un squelette.

Nous pouvons aussi observer, toujours dans le tableau 1, une meilleure performance générale (plus faible MSE) sur les reconstructions en ondelettes des signaux en contact mesurés à l'oreille par rapport aux signaux en contact mesurés au doigt. Nous supposons que cet écart reflète les différences de forme d'onde mesurée entre les sites, la forme d'une onde PPG caméra étant en général plus proche d'une onde mesurée à l'oreille que d'une onde mesurée au doigt [7].

3.2 Validation point à point des signaux reconstruits

Les modèles neuronaux entrainés délivrent une représentation en ondelettes sur deux plans (une partie réelle et une partie imaginaire). Le signal PPG temporel est reconstruit à partir de la transformée inverse. Un exemple est présenté en figure 3, où il est possible d'apprécier la qualité de la prédiction. L'écart de phase est correctement rectifié par le réseau. Nous pouvons observer que le rebond caractéristique de l'onde PPG est convenablement reproduit alors qu'il est presque toujours absent sur le signal caméra. Nous voyons que le signal a été lissé et que la largeur des ondes est plus faible, montrant que le réseau corrige les coefficients hautes fréquences qui transcrivent les bruits ainsi que les coefficients des fréquences centrales qui déterminent la partie pulsée du signal.

La RMSE (équation 2) a été calculée entre les différentes paires de signaux. Les amplitudes de ces derniers étant arbitraires et normalisées, nous proposons d'observer l'erreur absolue moyenne en pourcentage (*mean absolute percentage error*, voir équation 3).

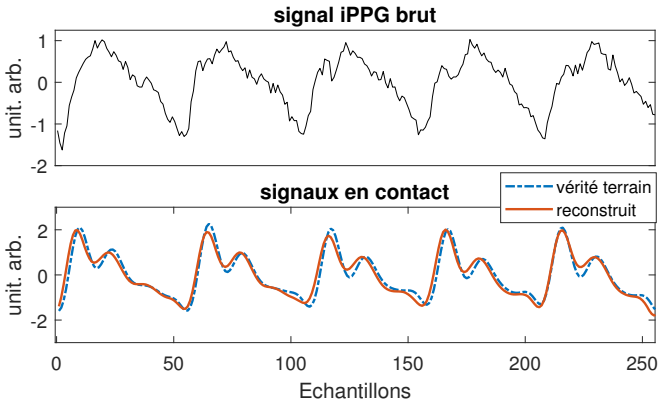


FIGURE 3 – Exemple de reconstruction d'un signal en contact au doigt (figure du bas) à partir du signal caméra (figure du haut). Notons la bonne qualité de la reconstruction même si quelques erreurs sont visuellement perceptibles.

$$RMSE = \sqrt{MSE} \tag{2}$$

$$MAPE = \frac{1}{n}\sum_{i}\left|\frac{p_i - q_i}{p_i}\right| \tag{3}$$

La $MAPE$ est ici calculée entre deux signaux (p et q dans l'éq. 3). Les résultats sont présentés dans le tableau 2 où nous pouvons observer la performance des prédictions délivrées par les modèles neuronaux. L'erreur sur le réseau utilisant le squelette VGG–16 est légèrement plus faible, ce qui cohérent avec les résultats présentés en section 3.1 et dans le tableau 1.

4 Synthèse des contributions et travaux futurs

Nous avons proposé, dans cet article, une architecture neuronale permettant de reconstruire avec précision une onde PPG en contact à partir d'une onde PPG sans contact estimée par analyse vidéo. La reconstruction est effectuée par le biais de la représentation temps-fréquence du signal via sa transformée en ondelettes continue. Les réseaux de neurones proposés correspondent à des architectures U-Net avec et sans squelette. Le signal reconstruit est proche de la vérité terrain en contact.

La motivation principale de ce travail correspond à la possibilité de proposer une estimation de la pression artérielle par l'analyse d'ondes PPG mesurées par caméra. La prochaine étape consistera donc à intégrer les signaux reconstruits dans des modèles d'IA dédiés à l'estimation de la pression artérielle par signaux en contact, ces derniers pouvant être collectés sur de larges bases de données publiques (e.g. MIMIC).

Des pistes d'amélioration de ce travail sont envisagées. Nous proposons dans un premier temps d'étoffer la base de données qui est actuellement limitée en volume et en nombre de participants. Les vidéos exploitées dans cette recherche ont été acquises par une caméra rapide (125 fps). Nous envisageons d'étudier dans des travaux futurs les signaux formés à partir de caméras classiques (30 fps). Les ondes PPG acquises par de tels capteurs sont moins détaillées et donc plus complexes à analyser. Il sera en contrepartie possible d'entrainer les modèles avec un volume plus conséquent de données, de nombreuses bases dédiées à l'étude de signaux PPG mesurés par des caméras classiques étant désormais publiquement disponibles. Une intégration directe de la vidéo plutôt que des représentations temps-fréquence dans l'architecture U-Net fera l'objet de travaux de recherche sur le plus long terme.

Réseau	$\text{cPPG}_{\text{doigt}}$ *vs* $\widehat{\text{cPPG}}_{\text{doigt}}$	$\text{cPPG}_{\text{oreille}}$ *vs* $\widehat{\text{cPPG}}_{\text{oreille}}$
U-Net1	0.25 (0.06)	0.20 (0.04)
$\text{U-Net}_{\text{VGG16}}$	**0.23 (0.05)**	0.20 (0.04)
$\text{U-Net}_{\text{ResNeXt101}}$	0.24 (0.06)	0.20 (0.04)
$\text{U-Net}_{\text{InceptionResNetV2}}$	0.25 (0.06)	0.20 (0.04)
$\text{U-Net}_{\text{InceptionV3}}$	0.25 (0.06)	0.20 (0.04)
$\text{U-Net}_{\text{DenseNet201}}$	0.24 (0.06)	0.20 (0.04)

TABLE 2 – $RMSE$ ($MAPE$) (voir les équations 2 et 3) calculées entre les prédictions délivrées par les différentes architectures neuronales et les vérités terrain. $\text{cPPG}_{\text{doigt}}$ et $\text{cPPG}_{\text{oreille}}$ correspondent aux signaux de vérité terrain mesurés au doigt et à l'oreille respectivement (voir courbe bleue sur la figure 3 pour un exemple typique). $\widehat{\text{cPPG}}_{\text{doigt}}$ et $\widehat{\text{cPPG}}_{\text{oreille}}$ correspondent aux prédictions calculées par la transformée inverse des représentations en ondelettes délivrées par les modèles neuronaux (voir courbe orange sur la figure 3).

Références

[1] D. McDuff, "Camera measurement of physiological vital signs," *arXiv preprint arXiv :2111.11547*, 2021.

[2] S. Zaunseder et al., "Cardiovascular assessment by imaging photoplethysmography-a review," *Bio. Eng.*, 2018.

[3] A. Ni et al., "A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods," *Sensors*, 2021.

[4] H. Luo et al., "Smartphone-based blood pressure measurement using transdermal optical imaging technology," *Circulation : Card. Imag.*, vol. 12, p. e008857, 2019.

[5] N. Sugita et al., "Contactless Technique for Measuring Blood-Pressure Variability from One Region in Video Plethysmography," *Journal of Med. and Bio. Eng.*, 2018.

[6] F. Schrumpf et al., "Assessment of Non-Invasive Blood Pressure Prediction from PPG and rPPG Signals Using Deep Learning," *Sensors*, vol. 21, no. 18, p. 6022, 2021.

[7] D. Djeldjli et al., "Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera," *Biomed. Sig. Proc. and Control*, vol. 64, 2021.

[8] J. Cheng et al., "Prediction of arterial blood pressure waveforms from photoplethysmogram signals via fully convolutional neural nets," *Comp in Bio and Med*, 2021.

[9] W. Verkruysse et al., "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, 2008.

[10] M. Tarvainen et al., "An advanced detrending method with application to HRV analysis," *IEEE Trans. on Biomed. Eng.*, vol. 49, no. 2, pp. 172–175, Feb. 2002.

[11] F. Bousefsaf et al., "Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate," *Bio. Sig. Proc. and Control*, vol. 8, pp. 568–574, 2013.

[12] E. C. Too et al., "A comparative study of fine-tuning deep learning models for plant disease identification," *Comp. and Elec. in Agr.*, vol. 161, pp. 272–279, 2019.

[13] K. He et al., "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[14] G. Huang et al., "Densely connected convolutional networks," in *IEEE CVPR*, 2017, pp. 4700–4708.

[15] C. Szegedy et al., "Inception-v4, inception-resnet and the impact of res. connections on learning," in *AAAI*, 2017.

Estimation sans contact de la tension artérielle par intelligence artificielle

Frédéric Bousefsaf, Théo Desquins, Djamaleddine Djeldjli, Yassine Ouzar, Choubeila Maaoui, Alain Pruski

Estimation sans contact de la tension artérielle par intelligence artificielle

Frédéric Bousefsaf* Théo Desquins*† Djamaleddine Djeldjli* Yassine Ouzar* Choubeila Maaoui* Alain Pruski*
* *LCOMS, Université de Lorraine*
F-57000 Metz, France
frederic.bousefsaf@univ-lorraine.fr
† *i-Virtual*
F-57000 Metz, France
theo.desquins@i-virtual.fr

***Résumé*—Les pathologies cardiovasculaires sont désignées par l'OMS comme étant la première cause de mortalité dans le monde. Elles sont responsables de près de la moitié des maladies invalidantes aujourd'hui. Un ensemble de technologies permettent de mesurer différents signaux physiologiques et fonctions vitales sans qu'aucun contact avec la personne ne soit nécessaire. Les caméras et webcams sont des technologies omniprésentes et accessibles. Elles sont désormais utilisées afin d'évaluer l'état de l'appareil cardiovasculaire en vue du diagnostic de pathologies relatives au cœur ou aux vaisseaux sanguins. Nous proposons, dans cet article, une nouvelle méthode permettant de mesurer la tension artérielle d'une personne à partir de vidéos délivrées par une caméra. L'analyse est effectuée sur le visage de la personne par une observation des fines variations de couleur qui apparaissent à chaque fois que le cœur bat et envoie du sang dans le corps. L'intelligence artificielle, à travers le développement de modèles d'apprentissage profond (deep learning), est ici utilisée. L'estimation déportée de fonctions physiologiques concerne tout autant les personnes saines que malades ou immobilisées, vieillissantes ou en perte d'autonomie ainsi que dépendantes ou en situation de handicap.**

***Mots clés*—fonctions vitales, intelligence artificielle, technologies sans contact, activité cardiovasculaire, tension artérielle**

I. Introduction

L'OMS désigne les maladies cardiovasculaires comme étant la première cause de mortalité dans le monde [1]. Ces pathologies et leurs risques s'accentuent avec le vieillissement en raison de la prévalence élective dans les tranches d'âge les plus élevées de la population. Les pathologies respiratoires ou cardiovasculaires sont responsables de près de la moitié des maladies invalidantes. Les personnes en situation de handicap (en particulier dans le cas de déficience motrice) sont plus sujettes à des pathologies d'origine cardiovasculaire comparées aux personnes sans handicap [2]. Une tension artérielle élevée en est un exemple typique.

Dans ce contexte, la mesure de données physiologiques et médicales à distance correspond à une solution d'intérêt : elle permet aux personnes d'effectuer des mesures fréquentes de leurs fonctions vitales, favorisant ainsi le diagnostic précoce ou un meilleur suivi de la ou des pathologies. Idéalement, les mesures déportées doivent être prises de manière non-invasive ; sans instrumentation supplémentaire ou spécifique ; sans contact, de préférence par le biais des caméras embarquées dans les systèmes mobiles. L'estimation déportée de fonctions physiologiques concerne tout autant les personnes saines (diagnostic précoce) que malades ou immobilisées, vieillissantes ou en perte d'autonomie ainsi que dépendantes ou en situation de handicap (maladies invalidantes notamment).

Différentes technologies ont été développées ou utilisées au fil des années pour mesurer des fonctions et indicateurs biomédicaux à distance [3]. Ces systèmes sont de plus en plus préférés aux capteurs en contact car ils permettent de réduire la gêne occasionnée par l'instrumentation (patchs adhésifs à placer sur la peau, câbles...) tout en améliorant le confort d'utilisation. D'un point de vue médical, une utilisation continue des éléments en contact peut entrainer des irritations voire des infections, notamment sur des peaux sensibles (personnes brulées par exemple).

Les caméras et webcams sont des technologies qui permettent de mesurer un ensemble de signaux physiologiques liés à l'activité cardiaque et vasculaire [4], [5]. Le principe repose sur la photopléthysmographie et consiste à observer les

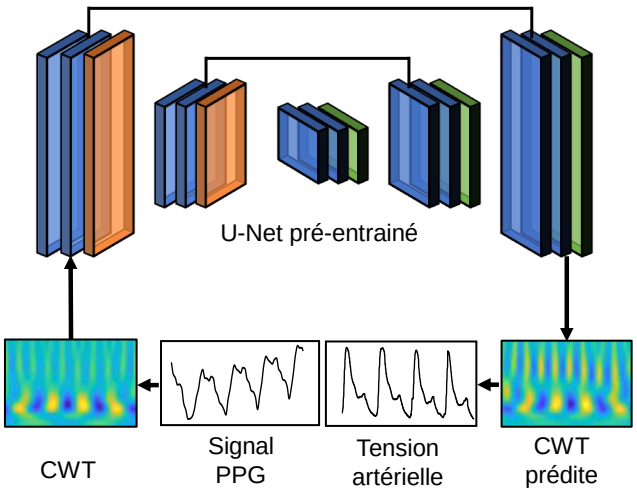


Figure 1: Vue d'ensemble de la méthode proposée dans ce travail. L'analyse vidéo du visage d'une personne permet de former son signal PPG. Le signal, à travers sa représentation en ondelettes (CWT), entre dans un modèle d'IA en forme de U qui se charge de le convertir en signal de tension artérielle.

variations de couleur sur la peau du visage pour en extraire les fluctuations périodiques du volume sanguin, délivrant ainsi un signal proche de l'onde de pouls. Un ensemble de paramètres physiologiques peuvent être calculés à partir de ce signal [6]–[10]. Les techniques d'intelligence artificielle (IA) sont de plus en plus étudiées [11], [12]. Les modèles d'IA délivrent généralement des performances plus intéressantes que les techniques conventionnelles reposant sur des opérations manuelles de traitement des images et du signal.

Nous proposons dans cet article une méthode permettant d'estimer, à partir de vidéos délivrées par une caméra, la tension artérielle d'une personne. L'analyse est effectuée par mesure de la photopléthysmographie sur le visage de la personne. Cela consiste à observer les fines variations de couleur apparaissant sur la peau. Nous présentons les détails de ce principe en section II. La section III est dédiée à la présentation de la méthode. Les étapes de traitement du signal et le modèle d'IA déployé pour répondre à cette problématique sont présentés. Les résultats et perspectives qui se dégagent de cette recherche sont exposées en dernière section. Cette étude s'inscrit dans la continuité de travaux récemment proposés par notre groupe de recherche [13]–[15].

Cette recherche présente l'une des toutes premières démonstrations de mesure de la tension artérielle par IA sur des flux vidéos délivrés par des caméras standards. Les résultats respectent d'ores et déjà un ensemble de métriques définies par les standards internationaux.

II. Mesure de la photopléthysmographie

La PhotoPléthysmoGraphie (PPG) repose sur un principe particulier : le sang absorbe plus de lumière que les tissus physiologiques tels que la peau [4]. Ainsi, la PPG correspond à la mesure des variations du volume sanguin par l'absorption et réflexion de la lumière (Fig. 2). Ces fluctuations de volume sont entrainées à chaque battement cardiaque (le volume croit lors de la contraction et décroit lorsque le muscle cardiaque se relâche).

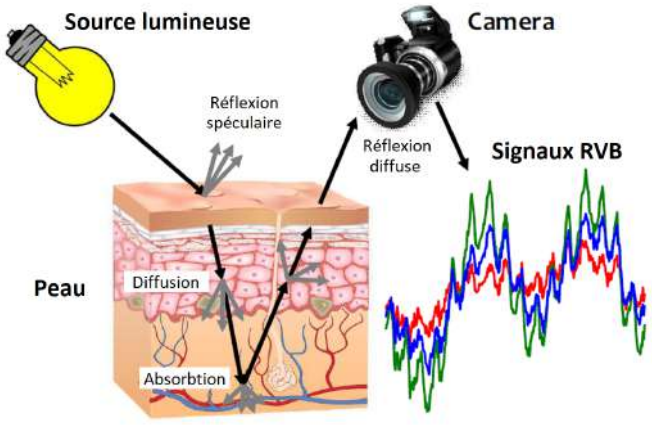


FIGURE 2: La photopléthysmographie consiste à mesurer les variations de l'absorption de lumière par les vaisseaux sanguins via une caméra. Les signaux formés (un par composante colorimétrique du capteur) traduisent les évolutions du volume sanguin à chaque battement cardiaque. Figure extraite de [16].

Les premières études portant sur la mesure de la PPG par caméra ont été introduites en 2008 par Verkruysse et al. [18]. Les chercheurs mesuraient les signaux PPG à une distance d'environ 1 mètre dans une région d'intérêt définie manuellement sur le visage du sujet observé. Les pixels de la région d'intérêt sont moyennés à chaque trame et pour chaque canal chromatique rouge, vert et bleu (RVB) du capteur. Un groupe de pixels est ainsi transformé en un scalaire pour une image donnée.

Ce processus, répété pour chacune des trames, permet de transformer une vidéo RVB en trois vecteurs (Fig. 2) qui contiennent différentes informations physiologiques [7]–[9] dont notamment la fréquence cardiaque, le taux d'oxygène dans le sang, la pression sanguine ainsi que le rythme respiratoire. Les signaux PPG sont la plupart du temps lissés par filtre passe-bande [6] afin de réduire le bruit et les artefacts de mesure les plus marqués.

Le choix des régions d'intérêt du visage retenu pour la mesure de la PPG est un paramètre fondamental [18], [19]. Une étape de pré-segmentation de certaines parties du visage [10] ou de l'ensemble des pixels de la peau [6] peut être introduite. Le calcul de la moyenne spatiale, permettant de transformer les trames de la vidéo en signal, n'est effectué que sur les pixels retenus à l'issue de la pré-segmentation. Le mouvement correspond à la principale limite des méthodes. La PPG par caméra a néanmoins été exploitée de manière très soutenue ces dernières années [4], [5].

III. Méthodes

A. Base de données

BP4D+[1] est une base de données publique et ouverte à la communauté de recherche. La base intègre initialement des signaux physiologiques de référence (dont la tension artérielle continue mesurée par un capteur en contact), des images thermiques, des vidéos et des scans 3D de 140 participants [20]. Dix tâches ont été développées pour induire des émotions en laboratoire. La nature des tâches entraine des mouvements plus ou moins intenses chez les participants. Ces déplacements créent des artefacts dans les vidéos, ce qui complexifie l'extraction du signal PPG. Une première phase de sélection, où seules les vidéos présentant des signaux PPG identifiables et de bonne qualité, a été effectuée. Nous avons ainsi conservé 57 sujets (21 femmes et 36 hommes) pour un total de 157 vidéos (car plusieurs tâches par sujet). Nous avons ensuite supprimé les échantillons présentant des signaux de tension artérielle de référence corrompus on incohérents. Le listing des participants retenus est disponible sur le site web du projet (https://github.com/frederic-bousefsaf/ippg2bp). Ce sous-ensemble a été utilisé pour entrainer les modèles d'IA présentés dans cette étude.

1. http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html

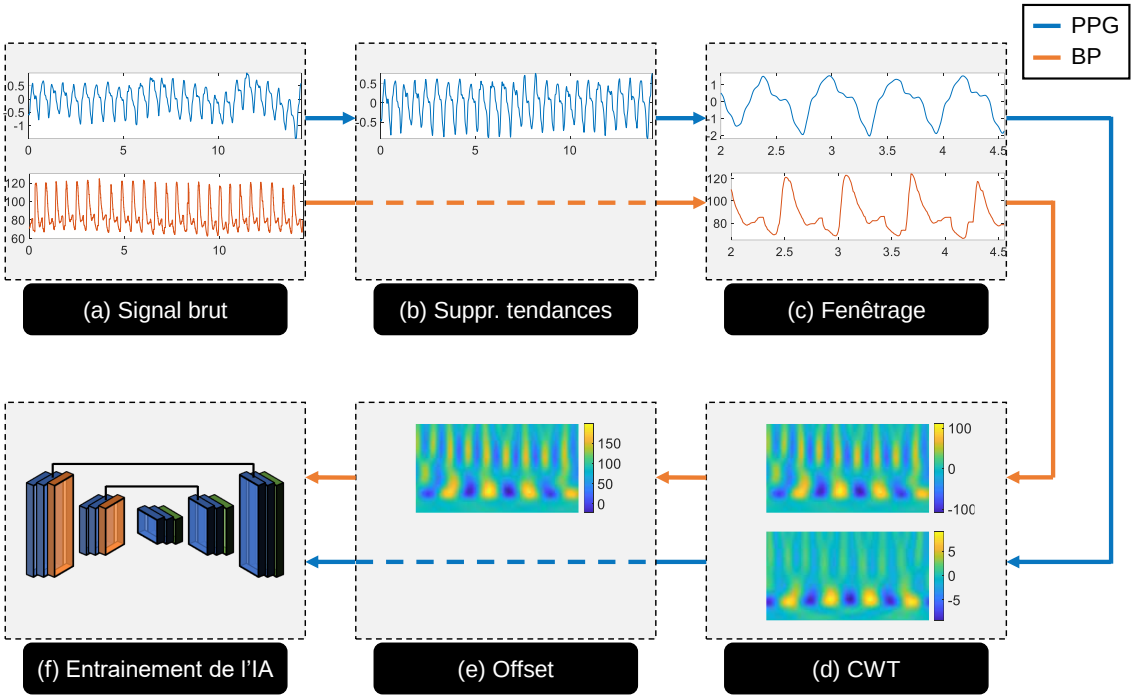


FIGURE 3: (a) Illustration d'un signal PPG calculé à partir d'une vidéo tirée de la base BP4D+ (en bleu) accompagné du signal de tension artérielle (Blood Pressure, BP) de référence (en orange). (b) La tendance basse fréquence du signal PPG est supprimée par une méthode de filtrage spécifique [17]. (c) Le signal est séparé en une collection d'extraits de 2.56 secondes. (d) La transformée en ondelettes continue (CWT) des signaux PPG et de tension artérielle est calculée dans la plage de fréquence [0.6, 4.5] Hz. (e) La valeur de tension artérielle moyenne du signal de référence étant perdue lors du calcul de la CWT, nous l'ajoutons à la représentation en ondelettes en l'additionnant à tous les coefficients. (f) Les représentations en ondelettes du signal PPG et du signal de tension artérielle sont utilisées pour entrainer les réseaux neuronaux présentés en section III-D.

Le signal PPG extrait de chaque vidéo a été découpé en petits extraits de 2.56 secondes à une fréquence d'échantillonnage de 100 Hz (256 valeurs par extrait). Un jeu de 4123 extraits a ainsi été constitué. 70 % du jeu (2887 extraits de 2.56 secondes) sont réservés pour l'entrainement du modèle d'IA, 15 % (618 extraits) pour la phase de validation et 15 % (618 extraits) pour la phase de test. Les tensions artérielles systolique (Systolic Blood Pressure, SBP), diastolique (Diastolic Blood Pressure, DBP) et moyenne (Mean Arterial Pressure, MAP) ont été calculées à partir du signal de référence fourni dans la base de données.

B. Constitution du signal PPG

La chaine de traitement est similaire à une méthode récemment proposée par notre groupe de recherche [15]. Nous avons dans un premier temps utilisé une technique récente de segmentation du visage reposant sur un modèle neuronal convolutif [21] permettant de segmenter la peau du visage. Ce modèle a déjà été utilisé dans le contexte de l'extraction de signal PPG à partir de vidéos [22]. Le signal PPG est calculé grâce à une moyenne spatiale des intensités des pixels de peau sur le canal vert. la figure 3a présente un signal PPG brut calculé à partir d'une des vidéos de la base BP4D+. Les signaux sont ensuite rééchantillonnés sur 100 Hz. Un algorithme spécifique de suppression de tendances [17] permettant d'atténuer les basses fréquences du signal est appliqué. la figure 3b montre l'impact de cette opération sur le signal PPG. Les extraits de 2.56 secondes sont ensuite calculés sur le signal PPG estimé à partir des vidéos ainsi que sur les signaux de tension de référence (voir figure 3c pour un exemple). Les signaux PPG ont été standardisés via la formule du z-score ($\mu = 0$ et $\sigma = 1$). Les jeux d'entrainement, de validation et de test ont été constitués à partir de tous ces extraits (voir section III-A).

C. Transformée en ondelettes continue

La transformée en ondelettes continue (Continuous Wavelet Transform, CWT) du signal PPG et de tension artérielle (Blood Pressure, BP) est utilisée pour entrainer le modèle neuronal présenté en section III-D. Une illustration générale de l'approche est présentée en figure 3. La CWT d'un signal correspond à une représentation temps-fréquence calculé à partir d'une fonction prototype appelée aussi ondelette mère. Contrairement à la transformée de Fourier, la CWT permet de détecter des changements abrupts de fréquence à l'aide d'une famille d'ondelettes calculée à partir de l'ondelette mère [15].

La CWT des signaux PPG et de tension artérielle est calculée dans la plage de fréquence [0.6, 4.5] Hz (plage des fréquences cardiaques chez l'être humain). La valeur de tension artérielle moyenne du signal de référence étant perdue

lors du calcul de la CWT, nous l'ajoutons à la représentation en ondelettes en l'additionnant à tous les coefficients (voir figure 3e) :

$$CWT_{BP} = CWT_{BP} + \mu_{BP} \quad (1)$$

Ici, μ_{BP} correspond à la valeur moyenne du signal de tension artérielle et CWT_{BP} aux coefficients de la transformée en ondelettes de ce même signal. Les CWT ont une dimension de $256 \times 256 \times 2$ pixels. Elles sont utilisées pour entrainer le modèle d'IA présenté dans la prochaine section.

D. Modèle d'IA

L'architecture neuronale développée dans ce travail a d'ores et déjà été proposée et testée dans un travail de recherche précédent [15]. Il s'agit d'une version modifiée du réseau U-Net initialement proposé par Ronneberger et al. [23] et soutenue par une ossature pré-entrainée (backbone). Ce type de réseau est très utilisé dans le milieu médical pour des tâches de segmentation sur des scanners [24]. L'architecture est composée d'une partie descendante (encodeur) complétée par une partie ascendante (décodeur), donnant ainsi une forme en U au réseau. La branche descendante contient un ensemble de couches convolutive et de sous-échantillonnage (pooling). La partie ascendante intègre des couches de déconvolution connectées aux convolutions de la partie descendante. Une vue schématique du réseau a été présentée en figure 1. Chaque couche convolutive intègre un noyau de taille $(3, 3)$ couplé à une fonction d'activation Rectified Linear Unit (ReLU).

Une ossature correspond à un réseau (e.g. VGG16 ou ResNet) pré-entrainé sur des bases de données d'images très volumineuse telle qu'ImageNet [25]. Cette ossature est intégrée dans la partie descendante du réseau U-Net et l'entrainement consistera à optimiser les paramètres internes de la partie ascendante. Nous avons, dans ce travail, initialisé le réseau U-Net avec une ossature ResNeXt101 [26]. Le nombre de variables modifiables pendant l'entrainement (poids et biais) est de 53 millions. Le choix de cette ossature particulière est motivé par les conclusions de notre précédent travail où les différentes ossatures les plus couramment utilisées ont été comparées [15].

La transformée en ondelettes (passée en entrée et prédite en sortie du réseau) contient une partie réelle et une partie imaginaire. Elle est donc définie sur deux canaux. Le carré moyen des erreurs (Mean Square Error, MSE) a été retenu en tant que fonction de cout pour l'entrainement du modèle :

$$MSE = \frac{1}{n} \sum_{i,j} \left(CWT_{i,j} - \widehat{CWT}_{i,j} \right)^2 \quad (2)$$

CWT correspond à la transformée en ondelettes du signal de tension artérielle de référence (voir figure 3d). $\widehat{CWT}$ est la représentation en ondelettes prédite par le réseau de neurones (à partir de la CWT du signal PPG).

L'implémentation de l'IA a été effectuée sous Python via l'API Keras et la bibliothèque Tensorflow. La librairie Segmentation Models [27] proposée par P. Yakubovskiy a été utilisée pour développer l'architecture neuronale. L'entrainement a été lancé sur 500 époques via des lots de 16 images.

IV. Résultats et discussion

Le réseau U-Net transforme un signal PPG, estimé à partir d'une analyse sur la vidéo de la personne, en un signal de tension artérielle par le biais de leurs représentations en ondelettes. La figure 4 illustre des exemples typiques d'estimation. Nous observons une similitude entre les signaux prédits et de référence. Les amplitudes et les formes d'onde prédites sont généralement bien retranscrites, en particulier au niveau des

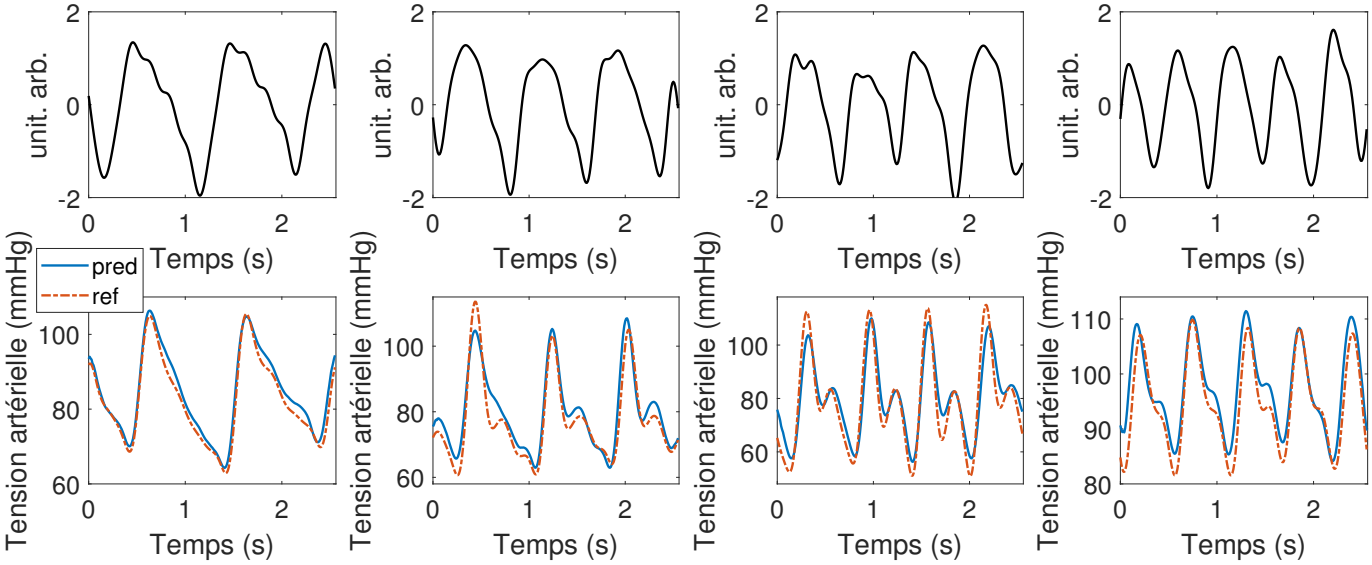


FIGURE 4: Exemples typiques de tensions artérielles continues prédites par le réseau U-Net pour différentes fréquences cardiaques. En haut : signaux PPG mesurés par caméra. En bas : tensions artérielles prédites et de référence.

pics maximums (qui serviront à calculer la tension systolique) et des pics minimums (tension diastolique). Nous avons évalué les performances globales à l'aide des standards internationaux [28], [29] définis par l'Association for the Advancement of Medical Instrumentation (AAMI) et par la British Hypertension Society (BHS). Nous soulignons toutefois que la base BP4D+ contient des vidéos et données physiologiques qui n'ont pas été enregistrées dans un contexte clinique. Aussi, le sous-ensemble constitué pour l'étude (voir section III-A) intègre 57 participants là où l'AAMI recommande d'évaluer les techniques d'estimation de la tension artérielle sur un minimum de 85 sujets.

A. Métriques générales

L'erreur absolue moyenne (Mean Absolute Error, MAE, equation 3) et la racine carrée de l'erreur quadratique moyenne (Root Mean Square Error, $RMSE$, equation 4) sont utilisées pour quantifier la concordance entre le signal de tension prédit par le modèle d'IA ($\widehat{BP}$) et la référence (BP). Nous avons calculé ces métriques pour la DBP, MAP et SBP sur tous les extraits du jeu de test. Les résultats sont reportés dans le tableau I. Une analyse comparative présentant les résultats des études proches est proposée.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|BP_i - \widehat{BP_i}| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(BP_i - \widehat{BP_i}\right)^2} \quad (4)$$

		MAE (mmHg)	RMSE (mmHg)
Rong et Li [30]	DBP	7.59	–
	SBP	9.97	–
Schrumpf et al. [31]	DBP	10.3	–
	SBP	13.6	–
iPPG2BP (résultats de cette étude)	DBP	5.1	6.85
	MAP	4.47	6.01
	SBP	6.73	9.34

TABLE I: Erreurs sur les estimations de la tension artérielle. Les métriques ont été calculées entre le signal de tension prédit par le modèle d'IA et le signal de référence. les résultats des études proches sont présentés à titre de comparaison.

B. Métrique de la BHS

La BHS évalue les techniques estimant la tension artérielle par le pourcentage cumulatif des erreurs [28]. Trois niveaux de performance (A, B et C) ont été établis (voir tableau II). Le niveau est calculé en fonction des prédictions sur le jeu de test dont la valeur doit être inférieure à trois seuils définis empiriquement : 5, 10 et 15 mmHg. Le tableau II présente en plus les résultats de l'étude de Rong et Li [30], seule étude à notre connaissance à avoir calculé les métriques du BHS sur des estimations basées sur des signaux PPG mesurés à partir d'une analyse vidéo.

Les résultats de la méthode que nous proposons sont tout à fait intéressants : plus de 60 %, 87 % et 95 % des échantillons de la base de test présentent des erreurs respectivement inférieures à 5, 10 et 15 mmHg pour la DBP et la MAP (niveau A). Plus de 50 % and 79 % des prédictions de la SBP sont situées sous les 5 et 10 mmHg (grade B) tandis que 89.6 % des prédictions sont sous la barre des 15 mmHg, ce qui est très proche du seuil de 90 %.

		Pourcentage d'erreur		
		$\leq$ 5 mmHg	$\leq$ 10 mmHg	$\leq$ 15 mmHg
Rong et Li [30]	DBP	55.4%	85.7%	98.2%
	SBP	48.2%	78.6%	94.6%
iPPG2BP (résultats de cette étude)	DBP	60.2%	87.1%	95.8%
	MAP	66.8%	90.9%	96.4%
	SBP	50.2%	79.0%	89.6%
BHS	niveau A	60%	85%	95%
	niveau B	50%	75%	90%
	niveau C	40%	65%	85%

TABLE II: Métriques du BHS sur les prédictions de la DBP, MAP et SBP.

C. Métriques de l'AAMI

L'AAMI propose d'évaluer les techniques d'estimation de la tension en analysant l'erreur moyenne (Mean Error, ME) et l'écart-type des erreurs (Standard Deviation of Errors, SDE) sur le jeu de test [29]. La technique doit présenter une ME inférieure à 5 mmHg et un SDE inférieur à 8 mmHg pour pouvoir respecter le standard international.

Le tableau III présente les résultats de l'évaluation selon les critères présentés précédemment. Nous avons reporté les valeurs présentées par Luo et al. [32] et Rong et Li [30]. Nos résultats se situent globalement sous les seuils pour la DBP et la MAP. La ME est faible et le SDE est inférieur à 8 mmHg. Nous pouvons remarquer que les prédictions de la SBP présentent une ME faible mais un SDE légèrement supérieur au seuil de 8 mmHg. Nous notons que les techniques dédiées à la conversion du signal de tension artérielle à partir du signal PPG en contact [33] produisent aussi des estimations de la SBP plus erronées que les estimations de la DBP.

		ME (mmHg)	SDE (mmHg)
Luo et al. [32]	DBP	-0.20	6.00
	SBP	0.39	7.30
Rong et Li [30]	DBP	0.79	2.58
	SBP	2.1	3.35
iPPG2BP (résultats de cette étude)	DBP	-1.001	6.781
	MAP	-0.205	6.007
	SBP	1.51	9.221
Standard AAMI		$\leq$ 5	$\leq$ 8

TABLE III: Métriques de l'AAMI sur les prédictions de la DBP, MAP et SBP. ME : Mean Error (erreur moyenne). SDE : Standard Deviation of Errors (écart-type des erreurs).

V. CONCLUSION

Nous avons proposé, dans cet article, une solution basée sur l'IA permettant d'estimer la tension artérielle à partir d'une vidéo délivrée par une caméra. Cette estimation est effectuée par le biais d'un réseau de neurones en forme de U et de la représentation en ondelettes du signal PPG sans contact, ce

signal ayant été calculé par analyse vidéo. Il s'agit à notre connaissance de la première étude proposant une estimation d'un signal de tension artérielle continu à partir d'une vidéo.

Cette recherche permet d'envisager une détection précoce de l'hypertension ainsi que d'autres pathologies cardiovasculaires avec un moyen bas-coût et d'ores et déjà accessible. Dans le domaine du handicap, ces résultats sont tout autant pertinents, certaines pathologies cardiovasculaires apparaissant en moyenne plus fréquemment que chez les sujets sains [2].

RÉFÉRENCES

[1] Organisation Mondiale de la Santé, "Plan d'action 2013-2020 pour la Stratégie mondiale de lutte contre les maladies non transmissibles."

[2] A. Stevens, E. Courtney-Long *et al.*, "Hypertension Among US Adults by Disability Status and Type, National Health and Nutrition Examination Survey, 2001–2010," *Preventing chronic disease*, vol. 11, 2014.

[3] A. Al-Naji, K. Gibson *et al.*, "Monitoring of Cardiorespiratory Signal : Principles of Remote Measurements and Review of Methods," *IEEE Access*, 2017.

[4] S. Zaunseder, A. Trumpp *et al.*, "Cardiovascular assessment by imaging photoplethysmography–a review," *Biomedical Engineering/Biomedizinische Technik*, 2018.

[5] D. McDuff, "Camera measurement of physiological vital signs," *arXiv preprint arXiv :2111.11547*, 2021.

[6] F. Bousefsaf, C. Maaoui, and A. Pruski, "Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 568–574, 2013.

[7] ——, "Remote detection of mental workload changes using cardiac parameters assessed with a low-cost webcam," *Computers in biology and medicine*, vol. 53, pp. 154–163, 2014.

[8] C. Maaoui, F. Bousefsaf, and A. Pruski, "Automatic human stress detection based on webcam photoplethysmographic signals," *Journal of Mechanics in Medicine and Biology*, vol. 16, no. 04, p. 1650039, 2016.

[9] F. Bousefsaf, C. Maaoui, and A. Pruski, "Peripheral vasomotor activity assessment using a continuous wavelet analysis on webcam photoplethysmographic signals," *Bio-medical materials and engineering*, vol. 27, no. 5, pp. 527–538, 2016.

[10] ——, "Automatic Selection of Webcam Photoplethysmographic Pixels Based on Lightness Criteria," *Journal of Medical and Biological Engineering*, vol. 37, no. 3, pp. 374–385, 2017.

[11] A. Ni, A. Azarang, and N. Kehtarnavaz, "A Review of Deep Learning-Based Contactless Heart Rate Measurement Methods," *Sensors*, vol. 21, no. 11, p. 3719, May 2021. [Online]. Available : https://www.mdpi.com/1424-8220/21/11/3719

[12] C.-H. Cheng, K.-L. Wong *et al.*, "Deep Learning Methods for Remote Heart Rate Measurement : A Review and Future Research Agenda," *Sensors*, vol. 21, no. 18, p. 6296, Sep. 2021. [Online]. Available : https://www.mdpi.com/1424-8220/21/18/6296

[13] F. Bousefsaf, A. Pruski, and C. Maaoui, "3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video," *Applied Sciences*, vol. 9, no. 20, p. 4364, Oct. 2019. [Online]. Available : https://www.mdpi.com/2076-3417/9/20/4364

[14] D. Djeldjli, F. Bousefsaf *et al.*, "Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera," *Biomedical Signal Processing and Control*, vol. 64, p. 102242, Feb. 2021.

[15] F. Bousefsaf, D. Djeldjli *et al.*, "iPPG 2 cPPG : reconstructing contact from imaging photoplethysmographic signals using U-Net architectures," *Computers in Biology and Medicine*, vol. 138, p. 104860, Sep. 2021. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S0010482521006545

[16] W. Wang, "Robust and automatic remote photoplethysmography," PhD Thesis, Technische Universiteit Eindhoven, 2017.

[17] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE transactions on biomedical engineering*, vol. 49, no. 2, pp. 172–175, 2002, publisher : IEEE.

[18] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[19] M. Hassan, A. Malik *et al.*, "Heart rate estimation using facial video : A review," *Biomedical Signal Processing and Control*, vol. 38, pp. 346–360, 2017.

[20] Z. Zhang, J. M. Girard *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[21] Y. Nirkin, I. Masi *et al.*, "On face segmentation, face swapping, and face perception," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 98–105.

[22] Y. Ouzar, D. Djeldjli *et al.*, "LCOMS Lab's Approach to the Vision for Vitals (V4V) Challenge," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2750–2754.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net : Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[24] S. Leclerc, E. Smistad *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, 2019.

[25] E. C. Too, L. Yujian *et al.*, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019, publisher : Elsevier.

[26] S. Xie, R. Girshick *et al.*, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[27] P. Yakubovskiy, *Segmentation Models*. GitHub, 2019, publication Title : GitHub repository. [Online]. Available : https://github.com/qubvel/segmentation_models

[28] E. O'Brien, J. Petrie *et al.*, "The british hypertension society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems." *Journal of hypertension*, vol. 8, no. 7, pp. 607–619, 1990.

[29] G. S. Stergiou, B. Alpert *et al.*, "A universal standard for the validation of blood pressure measuring devices : Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement," *Hypertension*, vol. 71, no. 3, pp. 368–374, 2018, publisher : Am Heart Assoc.

[30] M. Rong and K. Li, "A Blood Pressure Prediction Method Based on Imaging Photoplethysmography in combination with Machine Learning," *Biomedical Signal Processing and Control*, vol. 64, p. 102328, Feb. 2021. [Online]. Available : https://linkinghub.elsevier.com/retrieve/pii/S1746809420304444

[31] F. Schrumpf, P. Frenzel *et al.*, "Assessment of Non-Invasive Blood Pressure Prediction from PPG and rPPG Signals Using Deep Learning," *Sensors*, vol. 21, no. 18, p. 6022, Sep. 2021. [Online]. Available : https://www.mdpi.com/1424-8220/21/18/6022

[32] H. Luo, D. Yang *et al.*, "Smartphone-based blood pressure measurement using transdermal optical imaging technology," *Circulation : Cardiovascular Imaging*, vol. 12, no. 8, p. e008857, 2019.

[33] N. Ibtehaz and M. S. Rahman, "PPG2ABP : Translating Photoplethysmogram (PPG) Signals to Arterial Blood Pressure (ABP) Waveforms using Fully Convolutional Neural Networks," *arXiv preprint arXiv :2005.01669*, 2020.

Mesure sans contact de la fréquence par caméra basée sur l'apprentissage profond

Yassine Ouzar, Frédéric Bousefsaf, Choubeila Maaoui

Mesure sans contact de la fréquence par caméra basée sur l'apprentissage profond

Yassine Ouzar, Frédéric Bousefsaf et Choubeila Maaoui
Laboratoire de Conception, Optimisation et Modélisation des Systèmes (LCOMS),
Université de Lorraine, LCOMS, F-57000 Metz, France.
{nom.prénom}@univ-lorraine.fr

Résumé

Nous présentons dans ce papier une nouvelle architecture de bout en bout basée sur un réseau de neurones spatio-temporel profond pour l'estimation de la fréquence cardiaque à partir des trames vidéo issues d'une webcam bas coût. Contrairement aux méthodes existantes, nous estimons la valeur de la fréquence cardiaque directement sans passer par l'extraction du signal iPPG et sans incorporer de connaissances préalables ou d'étapes de traitement supplémentaires. Nous avons construit notre réseau en utilisant des couches de convolution séparable en profondeur 3D avec des connexions résiduelles pour extraire simultanément des caractéristiques spatiales et temporelles. Ceci est très approprié pour la mesure en temps réel car le modèle nécessite un nombre réduit de paramètres et un court fragment vidéo. Les résultats obtenus semblent très satisfaisants et prometteurs, d'autant plus que les expériences ont été menées sur des ensembles de données collectés dans des conditions non contrôlées. La mesure de paramètres physiologiques sans contact est à la fois prometteuse et pertinente dans le contexte du suivi de l'évolution de maladies invalidantes. Les avancées récentes sont aujourd'hui intégrées dans des systèmes d'assistance à la personne et sont utilisées durant les séances de thérapie par réalité virtuelle.

Mots-clés : fréquence cardiaque; sans contact; webcam; iPPG; réseaux de neurones convolutifs.

1 Introduction

Selon les dernières statistiques de l'Organisation Mondiale de la Santé, les maladies cardiovasculaires sont la première cause de décès dans le monde (World Health Organisation, 2018). Elles augmentent avec l'augmentation de la population, l'obésité et la sédentarité. Le contrôle non optimal de ces maladies est responsable de 70% des accidents vasculaires cérébraux, de 50% des crises cardiaques et de plusieurs cas d'insuffisance rénale. Ce type de maladies est souvent asymptomatique ce qui nécessite un contrôle périodique et à long terme via des mesures fréquentes de l'activité cardiaque afin de les prévenir et de mieux les prendre en charge.

L'électrocardiographie (ECG) et la photopléthysmographie (PPG) sont les principaux moyens pour la mesure de l'activité cardiaque. Les deux techniques utilisent des capteurs en contact qui doivent être attachés aux parties du corps et nécessitent le respect de certaines conditions pour obtenir de bonnes mesures. Malgré la grande précision et la robustesse fournies par ces dispositifs intrusifs, le contact avec la peau peut être gênant voire infaisable en raison de certains cas critiques citons par exemple les brûlures, les ulcères cutanés, les maladies contagieuses (Sun, 2016). Par conséquent, ces différentes limites, ainsi que la forte demande pour une technologie fiable, confortable, simple, portable, non

stressante et peu coûteuse, ont incité les chercheurs à développer de nouvelles techniques de mesure sans contact des signaux physiologiques.

Au cours de la dernière décennie, de grands progrès ont été réalisés pour l'estimation sans contact des paramètres vitaux tels que la fréquence cardiaque à l'aide de la photopléthysmographie par imagerie (iPPG) pour surmonter les faiblesses des dispositifs invasifs. La iPPG est une technique optique permettant une évaluation à distance de l'activité cardiaque en observant les variations du volume sanguin sur le visage d'une personne à l'aide d'une simple caméra bas coût. Cette technique est très prometteuse en santé publique, en particulier dans le contexte du vieillissement et des maladies invalidantes. Elle est désormais intégrée aux technologies d'assistance (Tagnithammou, 2021).

Les algorithmes d'iPPG classiques sont basés sur des approches conventionnelles qui impliquent généralement des pipelines à plusieurs étages et nécessitent plusieurs étapes de traitement d'image et de signal (Bousefsaf, 2013; de Haan, 2013; Poh, 2010). Ces méthodes ont été mises en œuvre dans des scénarios contraints et reposent sur certaines hypothèses concernant l'interaction lumière-peau et les mouvements de la tête. Par conséquent, la plupart des méthodes proposées fonctionnent raisonnablement bien sur des ensembles de données collectées dans des environnements contrôlés, mais les performances se dégradent considérablement dans des scénarios réels.

Avec le grand succès de l'apprentissage profond pour les tâches d'imagerie médicale et de vision par ordinateur, les travaux récents ont incorporé des architectures d'apprentissage profond à différentes étapes du pipeline de photopléthysmographie conventionnelle (Chen, 2018; Niu, 2020; Yu, 2019). Bien que les méthodes proposées permettent d'extraire avec précision le signal iPPG, mais plusieurs limites restent à surmonter. Tout d'abord, ces systèmes ne sont pas de bout en bout, ce qui nécessite encore des étapes de pré-traitement ou de post-traitement supplémentaire. De plus, la fréquence cardiaque doit être mesurée même dans des scénarios non contrôlés. De nombreuses situations peuvent impacter la mesure : la personne peut bouger la tête ou exprimer des émotions, son visage peut être partiellement occlus ou les conditions d'éclairage peuvent changer en permanence. Cela peut affecter la qualité du signal iPPG extrait et donc dégrader la précision des résultats.

Pour remédier à ces faiblesses, nous avons développé une méthode d'apprentissage profond de bout en bout pour l'estimation instantanée de la fréquence cardiaque directement à partir des séquences vidéo faciales. Notre architecture est entièrement automatique et ne nécessite aucune connaissance préalable ni aucun pré-traitement ou post-traitement particulier. Elle se concentre automatiquement sur les zones les plus vascularisées du visage, analyse les subtiles variations de couleur sur ces régions pour enfin estimer la fréquence cardiaque correspondante.

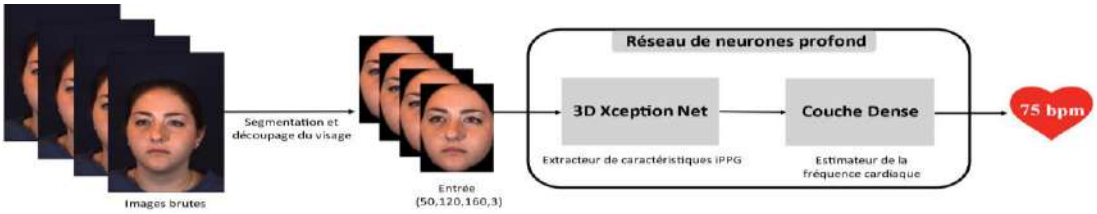


Figure 1 : Aperçu de notre solution proposée pour l'estimation de la fréquence cardiaque instantanée.

2 Matériel et Méthodes

Le framework général de notre méthode est illustré dans la figure 1. Nous considérons la tâche d'estimation de la fréquence cardiaque à partir de vidéos faciales comme une tâche de régression en une étape. Une segmentation du visage est effectuée en premier lieu pour éliminer le fond et les zones non cutanées (Nirkin, 2017). Ensuite, sans aucune étape de prétraitement ou de post-traitement supplémentaire, des lots de 50 images (correspondant à 2 secondes) sont introduits dans un réseau 3D entièrement convolutif pour estimer la fréquence cardiaque correspondante.

2.1 Base de données

Pour fonctionner avec précision dans des scénarios bien contrôlés ainsi que dans des scénarios difficiles, nous avons entraîné notre modèle sur une base de données publique à grande échelle (nommée BP4D+). Cette base de données est dédiée principalement à la reconnaissance multimodale des émotions spontanées à l'aide d'expressions faciales et de paramètres physiologiques tels que la fréquence cardiaque (Zhang, 2016). Par rapport aux bases de données de fréquence cardiaque existantes, BP4D+ est considérablement plus importante en termes de quantité de données et de diversité ethnique (noir, blanc, asiatique, hispanique/latino). Cette base de données peut ainsi fournir un apprentissage plus robuste car elle contient de nombreuses scénarios difficiles tels que des mouvements significatifs de la tête, des expressions faciales et des variations de fréquence cardiaque importantes, ainsi qu'une importante diversité en termes de teint de peau qui n'est pas disponible dans les autres bases de données

2.2 Segmentation du visage

L'extraction des régions d'intérêt (ROI) est la première étape de tous les systèmes de mesure de la fréquence cardiaque par caméra (Niu, 2020; Poh, 2010; Yu, 2019). Elle vise à maximiser le rapport signal/bruit (SNR) en éliminant les régions non cutanées qui ne contiennent aucun changement de couleur associé au rythme cardiaque. À notre connaissance, la plupart des systèmes iPPG existants basés sur l'apprentissage profond ont utilisé soit le visage entier, soit une région du visage sélectionnée grâce à des connaissances empiriques. Plusieurs détecteurs de visages et de repères faciaux ont été utilisés pour localiser la ROI (King, 2009; Viola and Jones, 2004; Zhang, 2016). Cependant, ils échouent souvent lorsque les visages présentent des mouvements de tête importants, des variations de pose, des occlusions ou des expressions faciales. De nombreux autres défis affectent également la capacité d'extraction de la ROI, tels que la couleur de peau, l'éclairage et l'arrière-plan.

Pour surmonter les limitations des algorithmes de détection de visage, nous effectuons une segmentation de visage en utilisant un algorithmes de l'état de l'art proposé initialement pour l'échange de visage (Nirkin, 2017). Cette méthode fonctionne idéalement dans toutes les conditions mentionnées ci-dessus sans manquer aucune image. Les visages sont correctement segmentés des arrière-plans et des occlusions avec une grande précision.

2.3 Architecture

Le réseau proposé est inspiré de l'architecture Xception (Chollet, 2017) qui utilise la convolution séparable en profondeur (CSP) au lieu de la convolution classique. Cette dernière est couteuse en termes de temps de calcul et de besoins en mémoire. L'architecture globale de notre modèle est composée de 36 couches convolutives structurées en 14 modules, tous liés par des raccourcis comme dans les réseaux ResNet à l'exception du premier et du dernier module (figure 2). Le réseau étant très profond, ces connexions résiduelles permettent d'éviter le problème de disparition du gradient. Chaque CSP est suivie d'une normalisation par lots pour stabiliser le processus d'apprentissage et accélérer la convergence, et également une fonction d'activation ReLU pour effectuer une cartographie non linéaire. La sortie de l'extraction des caractéristiques est aplatie et introduite à deux couches denses avec respectivement 1024 et 1 neurones, pour estimer la valeur de la fréquence cardiaque.

2.4 Implémentation

Le modèle proposé est mis en œuvre à l'aide du framework Keras et tensorflow, et exécuté sur NVIDIA Quadro P400. Pour toutes les expériences, l'entrée est fixée à 50 images. Inspiré par la procédure d'optimisation SWATS (Keskar, 2017), nous commençons l'apprentissage avec l'optimiseur Adam rectifié (RAdam) (Liu, 2020), et nous passons à la descente de gradient stochastiques (SGD)

lorsque la précision de l'ensemble de validation cesse de s'améliorer. Nous entrainons le réseau pendant 25 époques avec une taille de lot de 64. Le taux d'apprentissage a été fixé à 10^{-4}. En plus d'une couche dropout d'un ratio de 0,4 appliqué avant la couche dense finale du réseau, des stratégies de régularisation L1 et L2 sont utilisées, ce qui permet de surmonter le problème de surapprentissage et d'améliorer la capacité de généralisation du modèle à de nouvelles données.

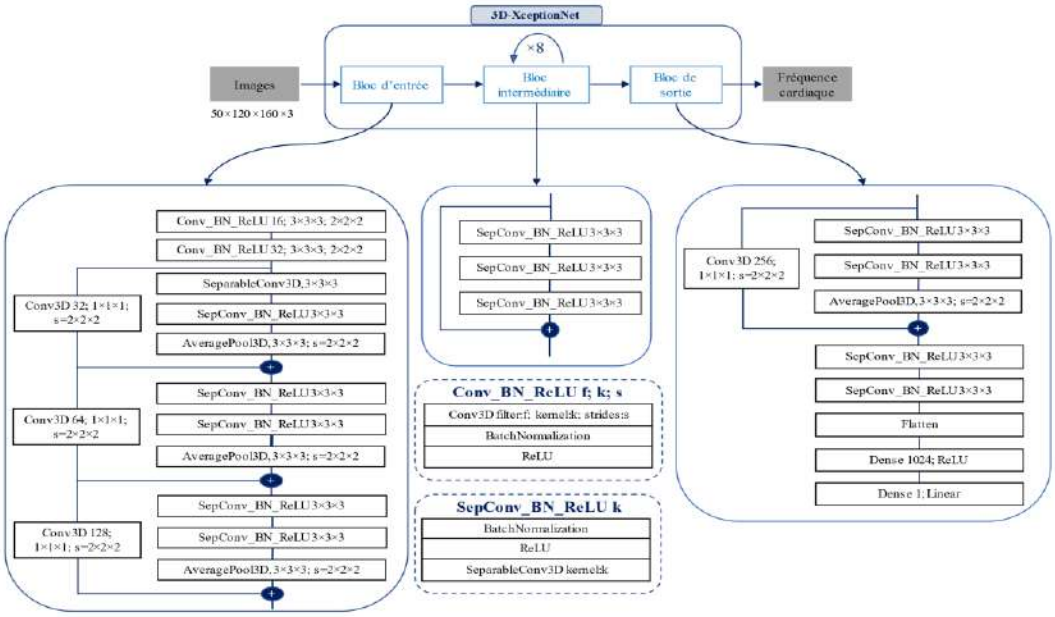


Figure 2 : L'architecture proposée : Elle correspond à une version modifiée du réseau Xception. L'entrée passe d'abord par le flux d'entrée, puis par le flux intermédiaire qui se répète huit fois, et enfin par le flux de sortie qui régresse la valeur de la fréquence cardiaque.

3 Résultats

Afin d'étudier la capacité de généralisation et l'efficacité du modèle proposé présenté, trois bases de données ont été utilisées, à savoir MMSE-HR (Zhang, 2016), MAHNOB-HCI (Soleymani, 2012) et UBFC-RPPG (Bobbia, 2017). MMSE-HR est directement utilisée pour les tests sans aucun traitement supplémentaire car elle a été collectée dans les mêmes conditions que la base d'apprentissage. Alors que UBFC-RPPG et MAHNOB-HCI sont sous-échantillonnés de 30 fps et 61 fps respectivement à 25 fps. Nous évaluons les performances de notre approche avec d'autres techniques de l'état de l'art en utilisant différentes métriques. Les résultats de comparaisons présentés dans les tableaux suivants montrent la grande précision de notre méthode qui surpasse tous les algorithmes de l'état de l'art.

Méthode	**MAE (bpm)**	**RMSE (bpm)**	**r**
PhysNet	12.76	13.25	0.44
SAMC	12.24	11.37	0.71
RhythmNet	6.98	7.33	0.78
AutoHR	5.71	5.87	0.89
Méthode proposée	**4.13**	**5.34**	**0.89**

Tableau 1: Cross-dataset sur MMSE-HR.

Méthode	MAE (bpm)	RMSE (bpm)	r
rPPGNet	5.51	7.82	0.78
SAMC	4.96	6.23	0.83
AutoHR	3.78	5.10	0.86
RhythmNet	-	3.99	0.87
Méthode proposée	**3.17**	**3.93**	**0.88**

Tableau 2 : Résultats sur MAHNOB-HCI.

Méthode	MAE (bpm)	RMSE (bpm)	std
Green	10.2	20.6	20.2
POS	5.12	10.5	10.4
3DCNN	5.45	8.64	8.55
PRNet	5.29	7.24	6.45
Méthode proposée	**4.99**	**6.26**	**6.25**

Tableau 3 : Résultats sur UBFC-RPPG.

4 Conclusion et Perspectives

Dans cet article, nous proposons une nouvelle architecture de bout en bout basée sur un réseau spatio-temporel profond qui prédit la fréquence cardiaque sans passer par l'extraction du signal iPPG et sans utiliser des connaissances préalables. Le réseau proposé s'inspire d'un modèle Xception qui s'est avéré efficace pour les bases de données d'images 2D à usage général en termes de précision, de vitesse de convergence rapide et de faibles coûts de calcul. Nos expériences approfondies ont montré l'efficacité de notre approche qui atteint une plus grande précision et surpasse les méthodes existantes sur trois ensembles de données de référence populaires tels que MMSE-HR, UBFC-RPPG et MAHNOB-HCI. Cependant, nous avons identifié plusieurs problèmes qui peuvent encore être améliorés dans des études futures. Premièrement, les mauvaises performances des techniques d'apprentissage profond pour les échantillons minoritaires dans le cas d'ensembles de données déséquilibrés qui sont fortement biaisés vers une peau plus claire et des fréquences cardiaques moyennes. L'application de stratégies avancées d'augmentation des données ou l'utilisation de données synthétiques pourrait améliorer encore les performances en augmentant le nombre d'échantillons pour les peaux foncées ou pour les fréquences cardiaques faibles et élevées. De plus, nous avons remarqué un taux élevé de valeurs aberrantes et de signaux PPG de mauvaise qualité dans les bases de données que nous avons utilisées. La préparation et le nettoyage des données avant la formation sont essentiels pour entraîner correctement le réseau et éviter les problèmes de surapprentissage. Enfin, les réseaux existants sont souvent constitués d'un grand nombre de paramètres et nécessitent des coûts de calcul élevés, entravant largement son application sur des appareils à faible consommation d'énergie tels que les téléphones portables. Par conséquent, l'étude de modèles de réseau légers peut considérablement améliorer la vitesse et la précision tout en maintenant des performances similaires ou meilleures.

Nos travaux futurs aborderont les problèmes mentionnés ci-dessus pour construire une architecture sophistiquée qui fonctionne avec précision dans des situations réalistes.

References

World Health, O. (2018). global health estimates 2016: death by Cause, Age, Sex, by country and Region, 2000-2016. Geneva.

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J., 2017. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters.

Bousefsaf, F., Maaoui, C., Pruski, A., 2013. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. Biomedical Signal Processing and Control 8, 568–574.

Chen, W., McDuff, D., 2018. Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–365.

Chollet, F., n.d. Xception: Deep Learning with Depthwise Separable Convolutions 8.

de Haan, G., Jeanne, V., 2013. Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering 60, 2878–2886.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., n.d. ImageNet: A Large-Scale Hierarchical Image Database 2.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

Keskar, N.S., Socher, R., 2017. Improving Generalization Performance by Switching from Adam to SGD. arXiv:1712.07628 [cs, math].

King, D.E., n.d. Dlib-ml: A Machine Learning Toolkit 4.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2020. On the Variance of the Adaptive Learning Rate and Beyond. arXiv:1908.03265 [cs, stat].

Nirkin, Y., Masi, I., Tran, A.T., Hassner, T., Medioni, G., 2017. On Face Segmentation, Face Swapping, and Face Perception. arXiv:1704.06729 [cs].

Niu, X., Shan, S., Han, H., Chen, X., 2020. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. IEEE Trans. on Image Process. 29, 2409–2423.

Poh, M.-Z., McDuff, D.J., Picard, R.W., 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express 18, 10762.

Ruder, S., 2017. An overview of gradient descent optimization algorithms. arXiv:1609.04747 [cs].

Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. IEEE Trans. Affective Comput. 3, 42–55.

Sun, Y., Thakor, N., 2016. Photoplethysmography revisited: from contact to noncontact, from point to imaging. IEEE Transactions on Biomedical Engineering 63, 463–477.

Tagnithammou, T., Monacelli, É., Ferszterowski, A., Trénoras, L., 2021. Emotional state detection on mobility vehicle using camera: Feasibility and evaluation study. Biomedical Signal Processing and Control 66, 102419.

Viola, P., Jones, M., n.d. Rapid Object Detection using a Boosted Cascade of Simple Features 9.

Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G., 2020. AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. IEEE Signal Process.

Yu, Z., Li, X., Zhao, G., 2019. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, in: BMVC.

Zhang, K., Zhang, Z., Li, Z., n.d. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.

Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L., 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).