Biomedical Signal Processing and Control 74 (2022) 103457

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A deep learning approach for remote heart rate estimation

Jaromir Przybyło

AGH University of Science and Technology, 30 Mickiewicza Ave., 30-059 Krakow, Poland

ARTICLE INFO

Keywords:
Deep learning
Imaging photoplethysmography
Videoplethysmography
Telemedicine
Image processing
Heart rate estimation

ABSTRACT

Remote monitoring of elderly people or patients in home isolation is an essential part of modern telemedicine. Videoplethysmography (VPG) is a method of noncontact assessment of heart rate and other cardiovascular parameters. Many algorithms have been developed to extract and improve the quality of the VPG signal. The main objective of this study is to design a method that replaces existing multistage algorithms and provides continuous monitoring of the user's pulse. The article presents a method of heart rate measurement based on the Long Short Term Memory (LSTM) Deep Neural Network. The proposed method outperforms the algorithm based on the analysis of the green component (G) and provides comparable results to the state-of-the-art methods such as Independent Component Analysis (ICA) and Plane Orthogonal to the Skin (POS). The best result for G was 6.49 bpm (beats per minute), ICA = 3.02 bpm, POS = 2.61 bpm, and for the proposed method was 3.26 bpm. While maintaining the accuracy comparable to ICA and POS algorithms, the LSTM network works well also beyond the visible spectrum, e.g., with infrared lighting when the color signal is not available and is easily adaptable to telemedicine applications.

1. Introduction

Photoplethysmography (PPG) is a noninvasive, low-cost optical technique used to detect volumetric changes in blood in the peripheral circulation. It has many medical applications including clinical physiological monitoring (i.e., blood oxygen saturation and heart rate). The PPG sensor has to be applied directly to the skin, which limits its practicality in situations such when free movement is required.

Videoplethysmography (VPG) has recently become popular as a method of noncontact measurement of biosignals, including cardiovascular parameters [1]. It can also be used in remote vital signs monitoring of COVID-19 patients in home isolation. The advantage of such an approach, compared to a standard PPG technique, is that it does not require uncomfortable wearable accessories and enables easy adaptation to different requirements in a variety of applications.

In short, the VPG technique makes it possible to remotely measure a user's heart rate (HR) or breathing rate (BR) using a consumer grade camera that observes a person's face or other skin areas. Compared to standard photoplethysmography (PPG) techniques, the advantages of this approach are that it does not require cumbersome wearable accessories and allows easy adaptation to different requirements in various applications, such as: optimization of training in sports [2], feedback control for fitness exercise[3], emotional communication in the field of human-machine interaction [4], and monitoring the driver's vital signs in the automotive industry [5]. Monitoring of other cardiovascular parameters, such as heart rate variability (HRV) [6] and blood pressure [7], are also described in the literature.

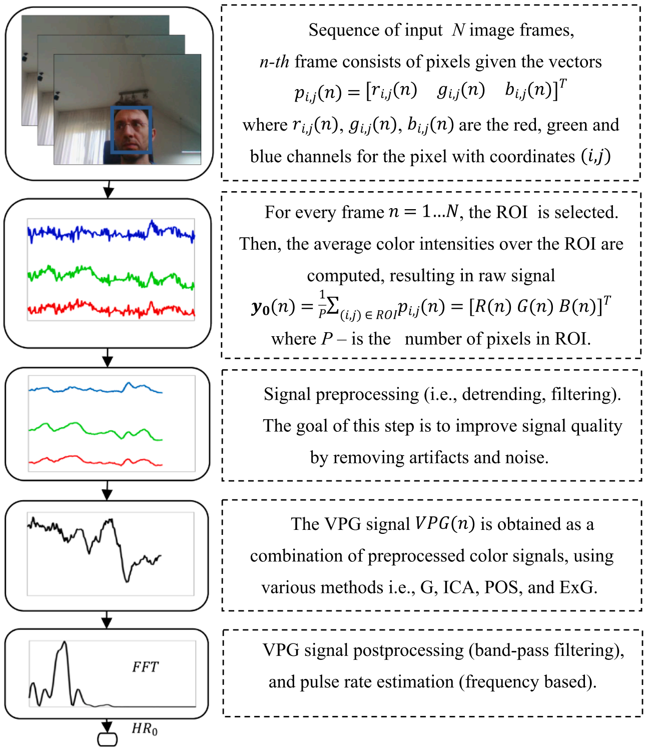
In real world applications, reliable estimation of heart rate (HR) from video is a complicated task as many factors can contaminate the VPG signal with artifacts. The VPG signal is formed by averaging pixel values inside a selected region of interest (ROI) for consecutive video frames (Fig. 1). Pixel values around the face region can be affected by a combination of rigid (head tilt, change of position) and nonrigid movements (facial actions, eye blinking) of the subject, resulting in signal artifacts. Changes in the environment such as fluctuations of lighting (indoor lights, blinks of computer screens, a flash of reflected light), and internal noise of a digital camera can also affect pixel values.

Therefore, many signal preprocessing algorithms have been developed to improve the quality of the signal. For example, detrending [8] has been used to remove the trend from the signal without affecting the heart rate bandwidth. Moving average (MA) filtering smooths the signal and suppresses high-frequency noise.

At the next stage of the analysis, the calculation of the VPG signal is performed based on the previously processed and refined color signals. Because the commonly used green signal component may contain artefacts, other techniques have been used as a robust alternative to G: green–red difference (GRD) [9], blind source separation by independent component analysis (ICA) [10], a plane orthogonal to the skin (POS)

E-mail address: przybylo@agh.edu.pl.

https://doi.org/10.1016/j.bspc.2021.103457
Received 29 July 2021; Received in revised form 30 November 2021; Accepted 16 December 2021
Available online 6 January 2022



Sequence of input N image frames, *n-th* frame consists of pixels given the vectors $p_{i,j}(n) = [r_{i,j}(n) \quad g_{i,j}(n) \quad b_{i,j}(n)]^T$ where $r_{i,j}(n)$, $g_{i,j}(n)$, $b_{i,j}(n)$ are the red, green and blue channels for the pixel with coordinates (i,j)

For every frame $n = 1 \ldots N$, the ROI is selected. Then, the average color intensities over the ROI are computed, resulting in raw signal $\boldsymbol{y}_{\mathbf{0}}(n) = \frac{1}{P}\sum_{(i,j) \in ROI} p_{i,j}(n) = [R(n)\ G(n)\ B(n)]^T$ where P – is the number of pixels in ROI.

Signal preprocessing (i.e., detrending, filtering). The goal of this step is to improve signal quality by removing artifacts and noise.

The VPG signal $VPG(n)$ is obtained as a combination of preprocessed color signals, using various methods i.e., G, ICA, POS, and ExG.

VPG signal postprocessing (band-pass filtering), and pulse rate estimation (frequency based).

Fig. 1. Steps of pulse rate estimation from facial video using videoplethysmography.

[11], excess green (ExG) [12]). All of them require RGB image representation (two or three signal components) to work.

The article presents a method of heart rate measurement based on the LSTM Deep Neural Network, which allows to replace a combination of several existing algorithms. The contribution of this paper is following:

- the method based on a trained LSTM network that efficiently processes the raw signal and extracts VPG data, replacing several existing algorithms and consequently reducing the number of parameters and improving the accuracy of HR estimation,
- unlike other methods, the proposed algorithm works with both color and grayscale images, that allows its use in situations where RGB camera is not an option (i.e., infrared person monitoring during nighttime),
- the framework is applicable to continuously monitor a user's heart rate in various real-world applications (telemedicine, heart rate monitoring in fitness clubs or in the psychologist's office, diagnose neuropathy).

The paper is structured as follows. Section 2 provides a literature review including a description of the general algorithmic framework of VPG-based heart rate estimation. Section 3 presents the description of the proposed algorithm as well as details of the LTSM training. Results, discussion and experimental setup are presented in Section 4. The paper is summarized in Section 5.

2. Related works

One of the well-known VPG approaches has been presented by Verkruysse et al. [13]. Authors showed that plethysmographic signals can be measured remotely by observing the human face with a consumer-grade digital camera. There has been a rapid development in this field since then, as reflected in the extensive literature on VPG techniques. A summary of 69 studies related to PPG and VPG can be found in [14]. However, most of them present the results of experiments carried out under controlled conditions (i.e., limited person movements, constant lighting, or short-term monitoring).

The general algorithmic framework of VPG-based heart rate estimation consists of the following steps, depicted on Fig. 1 (described in more detail in [15]):

- ROI selection: for each video frame, the region of interest (ROI) containing heart rate related information is selected, refined, and tracked,
- raw color signal extraction (by averaging pixel values inside ROI) and preprocessing (i.e., detrending, filtering),
- VPG signal extraction: raw VPG signal as a combination of refined color signals: i.e., G (green component), ICA, POS, and ExG.
- VPG signal postprocessing: i.e. band-pass filtering,
- pulse rate estimation: time-based or frequency-based.

In real world applications, reliably estimating heart rate (HR) from video is not an easy task. Various improvements of one or several stages can be found in literature.

To solve the problem of rigid head movements and to reduce the interference from illumination changes, a framework that uses face tracking and the green background value as a reference has been proposed in [16]. To reduce the impact of sudden nonrigid facial movements, noisy signal segments are excluded from the analysis. Additionally, the proposed temporal filters reduce the slow and nonstationary trend of the HR signal.

As a robust alternative to the analysis of the green (G) image component, several new signal extraction methods have been introduced. In [9], the authors proposed green–red-difference (GRD). Poh et al. [10] improved the algorithm by introducing heart rate measurements from video images based on blind source separation. An automatic face detection algorithm was used to detect faces in video frames and locate the region of interest (ROI), thus compensating for the movement of the subject. A basic webcam built into the laptop was used to record videos for analysis.

A study aimed at identifying color combinations for VPG-based HR measurements across various skin tones have been presented in [17]. Authors showed that an optimum RGB color channel combination for camera-based HR measurements exists.

In [12], we proposed a VPG heart rate measurement system designed to continuously monitor a user's heart rate during typical human–computer interaction scenarios (i.e., working at the computer). We evaluated a new signal extraction method based on an excess green (ExG) image representation. The results show that the ExG method provides acceptable accuracy while being much faster to compute than other state-of-the-art methods (i.e., ICA).

Recent studies have shown that deep learning methods could be used for VPG based HR estimation. In [18], the authors proposed an end-to-end approach for estimating HR from facial videos using a deep learning framework. The only requirement is the face detector to capture the facial region of interest (ROI). The presented network consists of 2D convolutional (Conv2D) and LSTM layers. The Conv2D operation extracts spatial features, while LSTM captures temporal information.

Kopeliovich et al. [19] present a study of architectural improvements for a convolutional neural network performing the estimation of heart rate values from color signals used as inputs. The signal values were obtained by averaging the intensity of RGB components over the ROIs extracted from the face. HR estimation problem is addressed as a classification task.

In [20], authors introduced an automated method for measuring pulse rate from video, using a representation learning approach and 3D convolutional neural networks. The video, consisting of a series of frames was used directly as input to the CNN. No prior image processing (e.g., automatic face detection and tracking) is required.

All three approaches are aimed at the analysis of color video and use

Convolutional Neural networks (CNN) to extract pulse information. In contrast, the solution proposed in this paper is intended to replace only the VPG signal pre- and post-processing stages, consequently reducing the number of algorithm parameters. Moreover, the VPG signal can be formed using only one image component (i.e., infrared) in situations where RGB data is not available. The result of its operation is a VPG signal that can then be further analyzed by various methods (time or frequency domain) or other neural network layers.

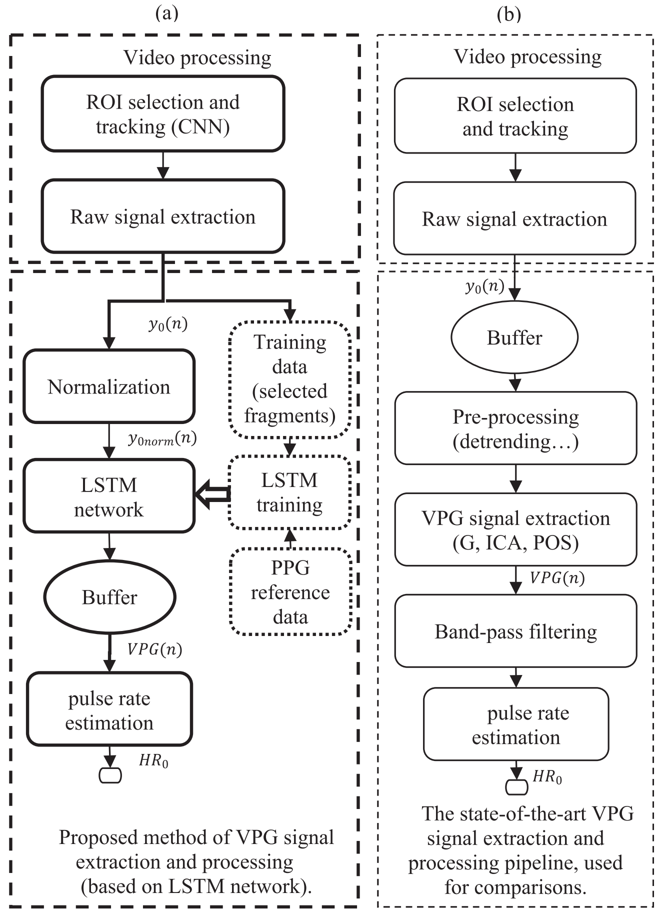
3. Materials and methods

3.1. Overall architecture of the algorithm

The block diagram of the algorithm's operation is shown in Fig. 2a. The entire algorithm has been adapted to continuous operation in real-time framework (MATLAB function, that can be automatically converted to C++ code using MathWorks code generation tools), however the implementation details fall beyond this paper. The proposed framework also allows extracting data for LSTM training (synchronization of video and ground truth HR timestamps) and comparing the results with other state-of-the art methods.

The algorithm consists of the following steps:

- The sequence of input video frames $p_{i,j}(n) = \left[r_{i,j}(n) \quad g_{i,j}(n) \quad b_{i,j}(n) \right]^T$ is passed to the video processing part of the algorithm. The purpose of this step is to extract raw color (or grayscale) signal $y_0(n) = [R(n)G(n)B(n)]^T$, by averaging the pixel values inside the region of interest (ROI). ROI selection is based on the face detection results using CNN. Section 3.1.1 describes the video processing step in detail.
- The signal $y_0(n)$ from previous step is then passed to the proposed method of VPG signal extraction and processing. The LSTM deep network architecture has been chosen as the basis of the proposed solution. Section 3.1.2 describes LSTM network in detail.
- To estimate the heart rate from the signal $VPG(n)$, the algorithm based on the calculation of the power spectral density (PSD) has been used. Section 3.1.3 describes the pulse rate estimation in detail.



(a)
(b)
Video processing
ROI selection and tracking (CNN)
Raw signal extraction
$y_0(n)$
Normalization
Training data (selected fragments)
$y_{0norm}(n)$
LSTM network
LSTM training
PPG reference data
Buffer
$VPG(n)$
pulse rate estimation
HR_0
Proposed method of VPG signal extraction and processing (based on LSTM network).
Video processing
ROI selection and tracking
Raw signal extraction
$y_0(n)$
Buffer
Pre-processing (detrending…)
VPG signal extraction (G, ICA, POS)
$VPG(n)$
Band-pass filtering
pulse rate estimation
HR_0
The state-of-the-art VPG signal extraction and processing pipeline, used for comparisons.

Fig. 2. Algorithm outline. (a) the proposed solution based on the LSTM network. (b) the state-of-the-art HR estimation pipeline. Results from both approaches are compared and presented in the Section 4 of this paper. The goal of the video processing part of the algorithm is to form raw color signal $y_0(n)$ from sequence of video frames. The $y_0(n)$ is then processed by LSTM network, forming VPG signal. Based on VPG signal the heart rate is estimated using frequency methods. LSTM training and data preparation is presented in Section 3.2.

The signal $y_0(n)$, together with the ground truth PPG data (obtained from PolarOH1 device), was also used for LSTM training. The methodology of data preparation and details of the LSTM training are provided in Section 3.2.

An additional processing pipeline was used to compare the proposed method with other state-of-the-art algorithms (Fig. 2b). It consists of the following steps:

- Video processing (same as in main pipeline).
- Signal buffering in a cyclic buffer of length M, and then pre-processing using a detrending algorithm based on mean centering with scaling (2), described in [2].
- VPG signal extraction using three methods: analysis of the green (G) image component, blind source separation by independent component analysis (ICA) [10], and a plane orthogonal to the skin (POS) [11].
- Pulse rate estimation (same as in main pipeline).

3.1.1. Region of interest (ROI) selection and tracking

For each video frame *n*, the region of interest (ROI) containing VPG-related information is selected and tracked using the face detector. In our previous work [12], we used the face detection algorithm based on Dlib's frontal face detector combined with a Kanade-Lucas-Tomasi (KLT) tracking algorithm. As a result, it was possible to track faces successfully in the video sequence. In this work, an alternative framework based on Intel® Distribution of OpenVINO™ Toolkit [21] has been adapted for face tracking. This toolkit is intended for quick development of applications and solutions that emulate human vision. It is based on CNNs and designed to maximize the performance of the target implementation of computer vision algorithms. For face detection and tracking, the pretrained Deep Learning model face-detection-0105 from Intel's Model Zoo was selected. This model [22] uses MobileNetV2 architecture as a backbone with an anchor-free head detector for indoor and outdoor scenes recorded with a front-facing camera.

To form a raw color signal $y_0(n)$ the most popular approach of ROI choice, that is, full rectangular region from face detector, was used. This might introduce additional noise because of the included nonskin pixels (i.e., eye, eyebrows). However, non-skin pixel elimination (by removing outliers from ROI) or free-form ROI selection are also possible in our framework.

3.1.2. VPG signal extraction and processing

The selected region of interest was used to calculate the average color intensities $p_{i,j}(n) = \left[r_{i,j}(n) \quad g_{i,j}(n) \quad b_{i,j}(n) \right]^T$ over the ROI for each subsequent image frame *n*, forming a signal $y_0(n) = \frac{1}{P}\sum_{(i,j)\in ROI} p_{i,j}(n)$. In the case of grayscale or infrared videos, this signal consists of only one image component.

After normalization to the range $< 0–1 >$, the signal $y_{0norm}(n)$ is sent to the input of the LSTM network, and the output of the network is then buffered in a cyclic buffer of length M, forming signal VPG($n - i$), where $i = 1..M$. To simplify notation, this signal is marked as VPG(n) in this article. Unlike other methods, the proposed LSTM processing does not require band-pass filtering.

3.1.3. LSTM network

A general algorithmic framework of VPG-based heart rate estimation consists of several steps (i.e., detrending, moving-average filtering, VPG signal extraction) that are intended to form VPG signal and eliminate noise. However, in some cases such preprocessing does not improve or even degrades the signal quality. Moreover, proper selection of preprocessing parameters is difficult. The LSTM network is a type of recurrent neural network (RNN) architecture that is able to learn a long-term dependency between time steps in time series and sequence data. Therefore, this type of network has the potential to replace other VPG signal formation and preprocessing algorithms and it was chosen as the basis of the proposed solution. The architecture of the LSTM network used is shown in Fig. 3.

The dimension of the sequence input layer is equal to the number of video components (three for RGB, it is also possible to use one for grayscale/IR video). The LSTM layer is the core of the network and consists of the hidden and cell states that can learn signal properties and time dependencies. The length of the LSTM layer was selected experimentally. To improve network generalization, the dropout layer was used during training. The goal of the fully connected layer is to form one-dimensional output VPG signal, based on activation from the LSTM layer, a weight matrix, and a bias vector. The regression layer follows the final fully connected layer and computes the mean-squared-error loss for the regression problem. The details and parameters of the LSTM network are provided in Section 3.2.2.

3.1.4. Pulse rate estimation

The algorithm based on the calculation of the periodogram power spectral density (PSD) has been used to estimate the heart rate from the signal $VPG(n)$. The following parameters was set: FFT length = 1024 samples, and a rectangular window with the same length as the input signal. To find the pulse frequency HR_0, the highest frequency peak was located in the PSD. The frequency resolution f_{res} depends on the length of the FFT (which is in our case equal to the length of signal buffer used), equation (1):

$$f_{res} = \frac{F_s}{M} \tag{1}$$

where: M is the length of the buffer and F_s is the sampling frequency (frame rate of the video).

For $F_s = 60$ and buffer length $M = 1024$, the $f_{res} = 3.52$ bpm.

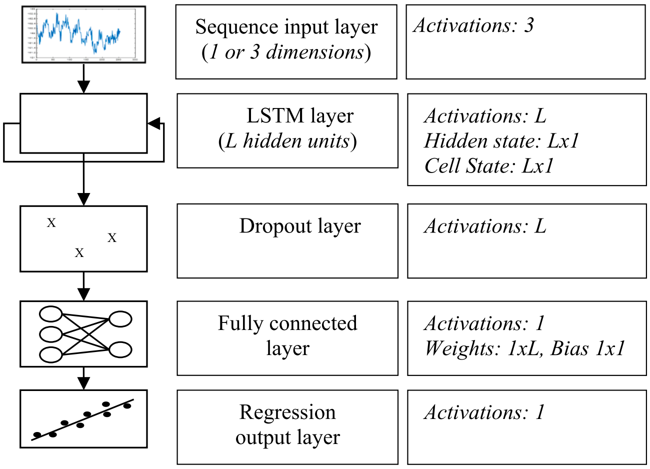


Fig. 3. The architecture of LSTM network.

3.2. LSTM training methodology

Photoplethysmography (PPG) is based on the principle that blood absorbs light more than surrounding tissue, so variations in blood volume affect the amount of light measured by the photodetector (changes in transmission or reflectance). Videoplethysmography (VPG), relies on the measurement of changes in the intensity of light reflected by the skin, and requires only ambient light and a digital camera. Thus, VPG is sometimes referred as iPPG (imaging photoplethysmography) or rPPG (remote photoplethysmography).

The basic principle of both methods is the same, but the sources of errors are different. Assuming the PPG signal is less susceptible to noise, it can be used as a reference for training the LSTM network. Thus, the purpose of LSTM is to eliminate artifacts resulting from, for example, rigid or nonrigid subject movements or fluctuations of scene lighting.

3.2.1. Training data preparation

Employing the same algorithmic framework as depicted in Fig. 2, the signal $y_0(n)$ was logged to disk and used for LSTM training. Training of the LSTM network was carried out on a data set prepared as follows. First, several fragments of input signal $y_0(\mathrm{n})$ were extracted from the selected video sequences. The length of those samples was approximately 17 s (1024 samples). Because they may contain a trend resulting from, e.g., head movements, facial expressions, or changes in lighting, only fragments with low amplitude variance were selected as training and validation data. Samples with significant trend were used later as a base for data augmentation.

Next, all fragments were normalized to have zero mean and range $<-0.5, 0.5>$, and finally were divided into two sets – training and validation. Many combinations of input signals from different videos and various numbers of fragments with different heart rates were used for LSTM training. Ultimately, the best network performance has been achieved for the dataset consisting of only three training samples with heart rate: 70, 88, and 96 bpm. Four fragments have been selected for validation. Fig. 4 shows the signals and their parts chosen as training and validation sets.

As a network output reference signal, the ground truth PPG was used (corresponding to the selected input signal fragments). Because the trend can be present in PPG signal, to remove it – the detrending algorithm based on mean centering with scaling (2) have been used [8], followed by normalization (zero mean and range < -0.5, 0.5 >).

$$y_1(n) = \frac{y_0(n) - \mu(n, L)}{\mu(n, L)} \tag{2}$$

where $\mu(n, L)$ is an L-point running mean vector of the signal $y_0(n)$.

The example input and reference signal is depicted in Fig. 5.

To increase the size of the training set, data augmentation was applied to all input fragments. Parts of the input signal, previously excluded from training due to the presence of significant trend, have been selected as a base for augmentation (Fig. 6). This trend is a result from typical noise sources, like head movements, lighting changes, etc.

To obtain a clean trend data (i.e., without higher frequencies), the zero-phase digital bandpass filter (Hamming) with the following parameters has been used: range 0.167–0.67 Hz (1–40 bpm), length 128. To generate more augmented data, each of the smoothed trend fragments was shifted and scaled (resulting in 50 more fragments). Then, this trend dataset was added to each of the training input signal fragments, resulting in 603 training input sequences total. The validation set was not augmented and consists of 4 fragments.

3.2.2. LSTM training

The LSTM training was carried out on an Nvidia GeForce RTX 2070 SUPER GPU and Intel Core i9 CPU with 32 GB RAM, and MATLAB R2020b (with Deep Learning Toolbox).

The network was initialized with random weights and bias using

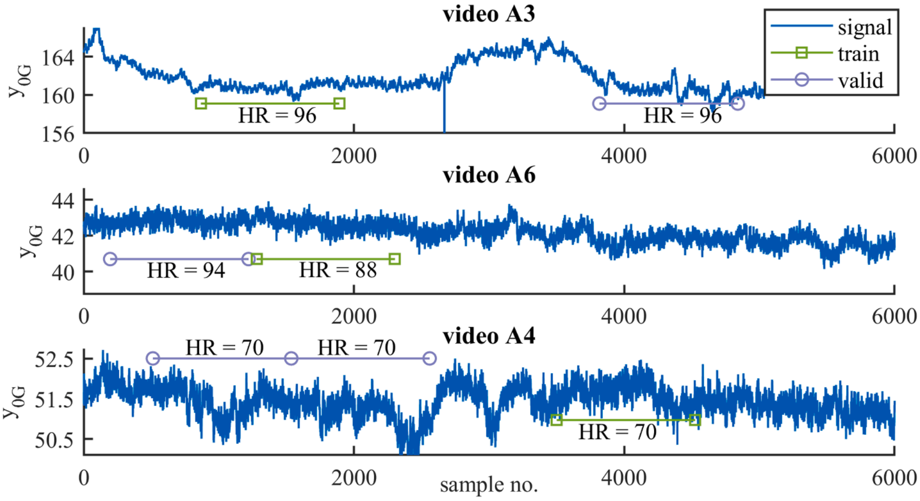


Fig. 4. Video sequences used for training and validation of the LSTM network. Fragments selected for training/validation are marked by horizontal lines.

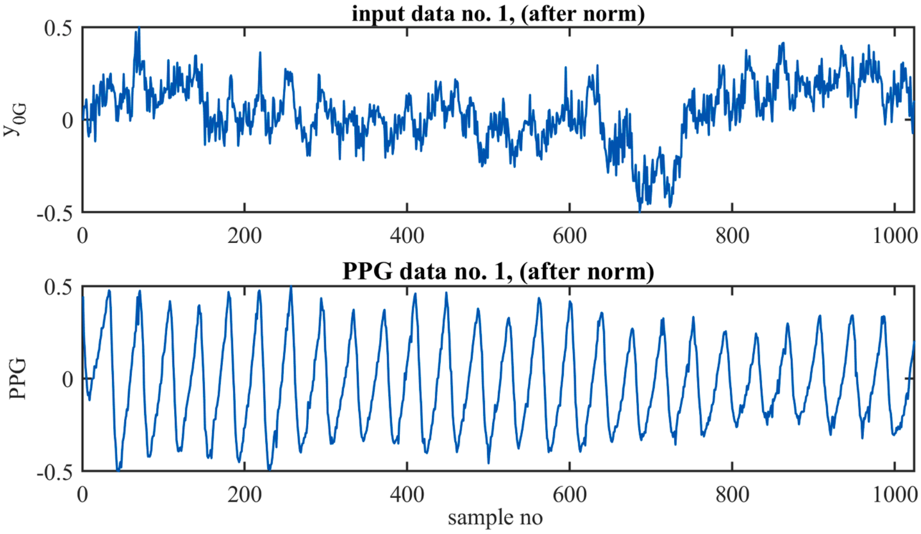


Fig. 5. The example input and reference (PPG) signal fragments used for training.

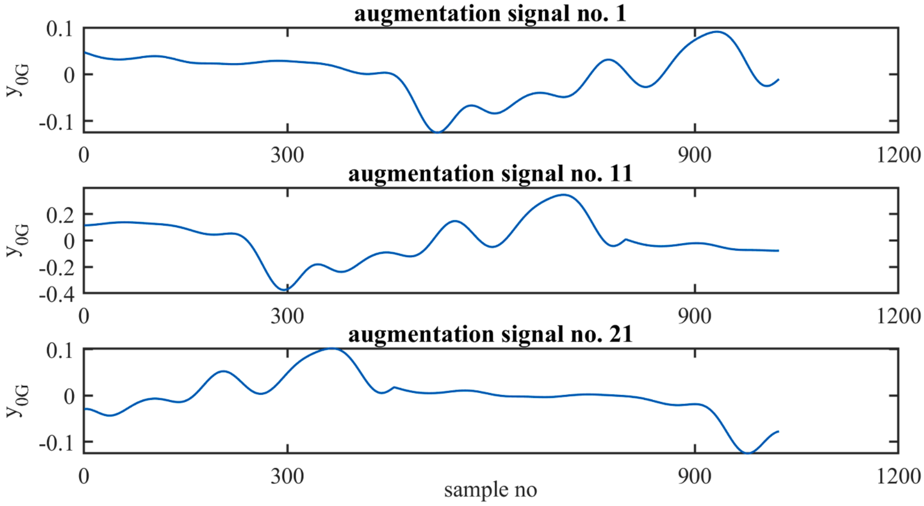


Fig. 6. The example trend signals used for augmentation. Signals no.11 and 21 were made from no.1 by shifting and/or scaling.

Xavier initializer. The other training parameters were: minibatch size = 4, number of epochs = 16, initial learning rate = 0.0002 (updated every 4 epochs by multiplying with a factor of 0.2). To improve network generalization, the dropout layer was used during training (with a probability of 0.5). The length of the LSTM layer was selected experimentally to L = 450 (input weights 1800x3, recurrent weights 1800x450, bias 1800x1). Fully connected layer parameters, corresponding to the assumed LSTM length, were as follows: weights 450x1, bias 1x1. The Adam solver was used as the optimizer and the loss was computed as the mean squared error between the ground truth (PPG) and output from the network (VPG(n)).

Because the training procedure involves randomness (initialization and dropout layers), LSTM has been trained with the same set of parameters (i.e., the length of LSTM) several times, and the best network was selected for further testing.

4. Results

4.1. Dataset description

Three data sets, characterized by great diversity, were used to evaluate the proposed algorithm. Two different cameras, eight participants with age ranging from 22 to 70 years, seven locations with various lighting conditions (both natural and artificial), and various activities performed by the participants. Ethical review and approval are not applicable for this study, because the article presents a noncontact and non-invasive method of measuring pulse rate. This is only a preliminary research and the results have not been used to assess human health. All devices used for collecting the ground truth pulse rate are battery powered and are commercially available products for personal use. Participants were not exposed to any stress - they performed only daily activities, as they did every day (i.e., working with a computer, sitting). Informed consent was received from all human subjects.

The first set of data was recorded using the following configuration. RGB and infrared video sequences were captured using the Intel® RealSense™ camera (model D425). The video acquisition parameters were the following: a resolution of 640 × 480 pixels and a frame rate of 60 FPS. The camera was located 0.5 to 0.6 m from the volunteers (depending on the experiment). Video duration ranges approx. from 2 to 5 min. Details are provided in Table 1. Three different locations with various illumination levels was selected. Additional signals were also recorded (ambient light level) using a SimpleLink™ SensorTag CC2650 (Texas Instruments, Dallas, TX, USA). It is a low energy Bluetooth device that includes 10 low-power MEMS sensors (i.e. ambient light sensor). The SensorTag was placed on the chest of the subject near the neck and face. To measure the ground truth HR and PPG signal, two devices connected via Bluetooth were used. The ECG-based $H10$ Heart Rate Sensor (Polar, Kempele, Finland) measured the reference HR. The optical heart rate sensor OH1 (Polar, Kempele, Finland) captured the PPG signal. Recorded data was used for both: training LSTM network and testing. The ground truth heart rate (HR) varies from 48 bpm to 128 bpm.

To assess the generalization potential of the LSTM neural network, videos (collected for several years) from our previous work were used as test data. In most of those videos, participants were asked to perform different tasks reflecting typical user-computer interaction scenarios (details can be found in [11]). Those video sequences were captured using a different camera model – an Intel® RealSense™ SR300 Depth Camera (Creative BlasterX Senz3D, Intel, Santa Clara, CA, USA) with a resolution of 640 × 480 pixels and a frame rate of 60 FPS. The ground truth HR were measured using the ECG-based H7 Heart Rate Sensor (Polar, Kempele, Finland) connected via Bluetooth. The PPG data was not collected. Video duration ranges approx. from 2 to 6 min. Participants were performing HCI tasks on all videos except B10 and B11 where they were standing 1 m from the camera and performed minor exercises. Furthermore, for B10 and B11, the 2nd simultaneous video stream was recorded using an infrared camera. Details are summarized in Table 2.

All devices used for collecting ground truth pulse rate (polar H7, polarH10, polar OH1) are battery powered and are commercially available products for personal use.

In addition, several video sequences from the MR-NIRP video dataset (indoor) [23] were used to further assess the network generalization capability. However, VPG signals obtained from those videos (30 FPS), need to be upsampled to match the sampling frequency of the training database (60 FPS).

4.2. Evaluation methodology

The testing procedure follows the algorithm outline depicted in Fig. 2a. After normalization, the signal $y_{0\text{norm}}(n)$ is sent (without any preprocessing) to the input of the LSTM network, and the output of the network is then buffered in a cyclic buffer of length N and used to estimate heart rate. To estimate the heart rate, the algorithm based on the

Table 1
First set of recorded video sequences (with ground truth data: PPG, HR, light level).

Video no.	Description
A1	Participant 1: male, ~45 years old, ground truth HR = 70–74 bpm, participant is sitting still, facial actions present; daylight (average illumination ~ 400 lx)
A2	Participant 1: male, ~45 years old, HR = 68–73 bpm, the participant is sitting still; daylight (average illumination ~ 390 lx)
A3	Participant 1: male, ~45 years old, HR = 75–96 bpm, the participant performs light exercises; daylight (average illumination ~ 430 lx)
A4	Participant 2: female, ~68 years old, HR = 70 bpm, pacemaker, heart arrhythmia, participant is sitting still; daylight (average illumination ~ 83 lx)
A5	Participant 3: male, ~70 years old, HR = 52–55 bpm, after taking beta blockers, participant is sitting still; daylight (average illumination ~ 83 lx)
A6	Participant 1: male, ~45 years old, HR = 73–122 bpm, participant is sitting still, but after exercise; daylight (average illumination ~ 50 lx)
A7	Participant 1: male, ~45 years old, HR = 79–98 bpm, participant performs light exercises, head movements, and facial actions present; daylight (average illumination ~ 180 lx)
A8, A8ir	Participant 1: male, ~45 years old, HR = 64–79 bpm, Participant is sitting; daylight (illumination ~ 180 lx); second simultaneous video stream recorded – infrared camera

Table 2
Second set of recorded video sequences (with ground truth data: HR, light level).

Video no.	Description
B1	Participant 4: male, ~23 years old, HR = 48–70 bpm; natural light (avg. illumination ~ 106 lx)
B2	Participant 1: male, ~44 years old, HR = 62–79 bpm; daylight + artificial light (avg. illumination 54 lx)
B3	Participant 1: male, ~44 years old, HR = 66–81 bpm; daylight + artificial light (avg. illumination 40 lx)
B4	Participant 1: male, ~44 years old, HR = 69–83 bpm; daylight (avg. illumination ~ 460 lx)
B5	Participant 1: male, ~44 years old, HR = 71–89 bpm; artificial light (avg. illumination ~ 27 lx)
B6	Participant 5: female, ~42 years old, HR = 60–68 bpm; daylight (avg. illumination ~ 150 lx)
B7	Participant 6: male, ~34 years old, HR = 79–88 bpm; daylight + artificial light (avg. illumination ~ 72 lx)
B8	Participant 7: male, ~22 years old, HR = 88 bpm; daylight + artificial light (avg. illumination ~ 86 lx)
B0	Participant 1: male, ~44 years old, HR = 70–84 bpm; daylight + artificial light (avg. illumination ~ 50 lx)
B10, B10ir	Participant 6: male, ~33 years old, HR = 115–128 bpm; avg. illumination ~ 388 lx
B11, B11ir	Participant 8: female, ~22 years old, HR = 72–94 bpm; avg. illumination ~ 293 lx

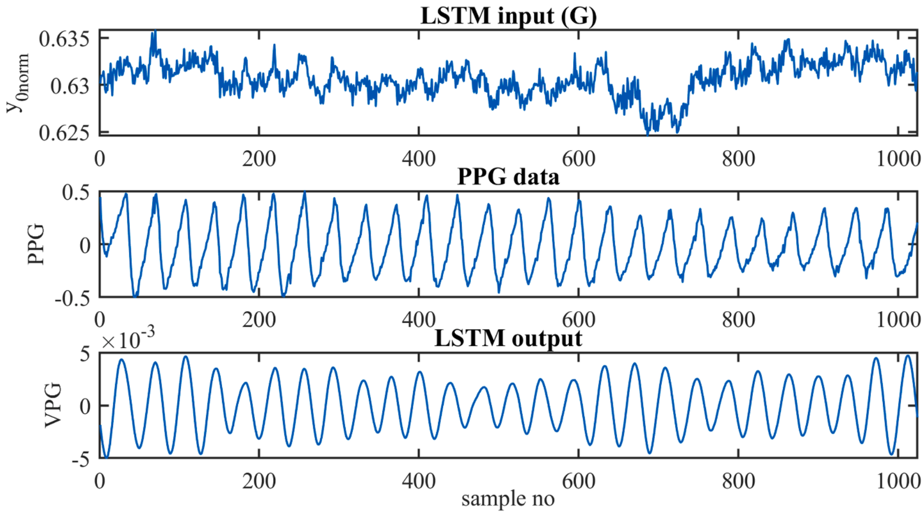


Fig. 7. The example LSTM input and output signal, together with PPG ground truth data.

calculation of the power spectral density (PSD) has been used (Section 3.1.3). Fig. 7 shows an example of the LSTM output signal.

It is worth noting that the input signal is sent sample-by-sample to the LSTM network, and only normalization by scaling to the range $< 0,1 >$ has been used. This is different than the normalization used during training, where the signal is buffered and then mean-centered. It is possible to buffer the input signal first, however, the computation time increases greatly in such a case, and results indicate that this lowers the accuracy of the HR estimation.

To assess the accuracy of signal processing by the LSTM network, the VPG signal was additionally extracted and processed by other state-of-the-art methods (as depicted on the Fig. 2b): analysis of the green (G) channel, independent component analysis (ICA), and a plane orthogonal to the skin (POS). For all those algorithms, the raw VPG signal was additionally processed by the following methods:

- detrending using mean centering with scaling (2),
- smoothing and suppressing high-frequency noise of the signal by using moving average (MA):

$$y_2(n) = \frac{1}{L}\sum_{k=0}^{L-1} y_1(n-k) \tag{3}$$

- band-limited (range 0.67–2.5 Hz, which corresponds to 40–150 beats per minutes) by a zero-phase digital filter (Hamming) of the length 320, forming the output signal denoted as $VPG(n)$.

Results (heart rate) from our algorithm based on LSTM network, are compared with those obtained from other state-of-the-art algorithms (G, ICA,POS). Different kinds of metrics were presented in other works for evaluating the accuracy of HR measurement methods. The most common is a root mean square error denoted as RMSE (4), used in this paper:

$$\mathrm{RMSE} = \sqrt{\frac{1}{k}\sum\nolimits_{i=1}^{K}\left[HR_{error}(i)\right]^2} \tag{4}$$

$$HR_{error} = HR_{video} - HR_{gt} \tag{5}$$

where: HR_{video}– the HR estimated from video, HR_{gt} – the ground truth HR values, K – length of the signal.

4.3. Results of the experiments

All videos (Tables 1 and 2 and Fig. 8) were used as test sequences. To assess whether the LSTM network can replace VPG signal extraction and preprocessing, first the results of HR estimation based on analysis of the green component were compared with the results obtained from LSTM network.

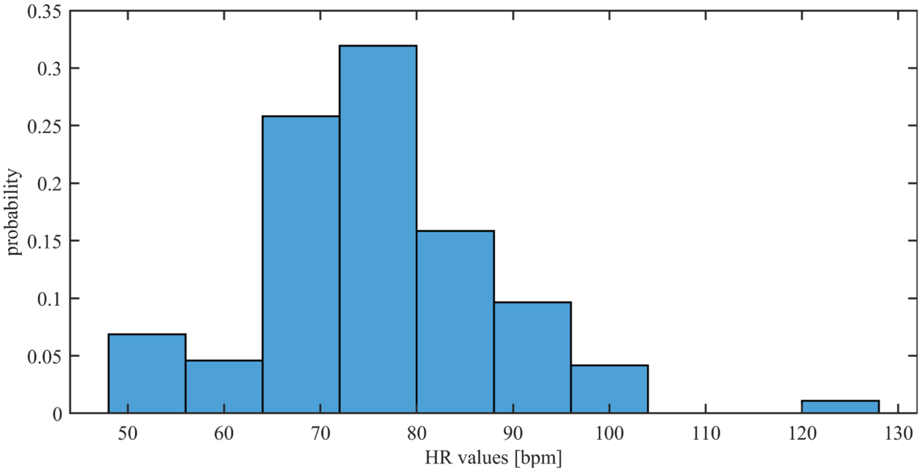


Fig. 8. Histogram of HR ground truth values of all test video sequences.

Because the goal of this comparison was to confirm that the LSTM network performs better than the algorithm based on the analysis of G signal, three sets of parameters were used:

- G_{raw} – raw G signal, without preprocessing and filtering,
- G_{filt} – raw G signal, without preprocessing but with bandpass filtering,
- G_{pre} – VPG signal formed from G using preprocessing (detrending and smoothing) and bandpass filtering.

The LSTM network worked on the raw signal, without any preprocessing and filtering (experiments showed that bandpass filtering of the LSTM results in slightly higher errors). Table 3 presents the results for HR estimation based on G component and LSTM.

For 12 out of 19 videos, RMSE for LSTM network is lower. For 2 videos, RMSE is higher. For 7 test sequences, the results are similar, assuming that a similar result is when the RMSE value is within the accuracy limits computed as $-\frac{1}{2}\bullet f_{res} < \mathrm{RMSE} < \frac{1}{2}\bullet f_{res}$ (where f_{res} for 60FPS is equal to 3.52 bpm).

Analysis of the results shows that the two videos for which LSTM performs worse, have a mean HR equal to about 53–55 bpm. The LSTM network was trained on fragments with mean heart rate: 70, 88, and 96 bpm. Because the training data did not contain fragments with low HR values (54 bpm), another fragment was added to the training data (from video A5 with HR = 54 bpm). However, this did not improve the results, what may mean that the heart rate value is not the only factor of the variability in the VPG signal.

It is interesting to note that the results for videos recorded by the infrared camera are also improved by using LSTM network (except for one sequence). Without any modifications of the LSTM structure and training procedure (the signal from one infrared channel was duplicated on all RGB components). Similarly, using the same procedure (duplicating G component) for RGB sequences, obtained results outperforms HR estimation based on G component (Table 3, column $LSTM_G$).

The comparison of the LSTM network and other methods (ICA and POS) are summarized in Table 4.

For ICA and POS methods, the signal has been preprocessed (detrending and smoothing) and bandpass filtering has been used. Comparing the ICA method with LSTM shows that LSTM performs better for 4 videos and it is comparable for 11 sequences. The ICA is in turn better for the other 4 videos. Two of them are the same as for G comparison (the HR was about 54 bpm), for one the mean HR was about 122 bpm (which was not included in the training dataset). One video (A6) for which ICA is slightly better than LSTM, has a mean HR about 84 bpm, however for part of this sequence the HR was about 120 bpm. This result confirms that the LSTM performs worse for the sequences with heart rate significantly different from the training set. Regarding the comparison, POS and LSTM – deep network performs better on only one video and for most of the sequences LSTM results are comparable to POS method (slightly higher RMSE for LSTM).

Table 3
Comparison of signal extraction methods: G and LSTM (RMSE values).

Video no.	G_{raw}	G_{filt}	G_{pre}	LSTM	$LSTM_G$
A1	4,93	4,58	3,68	**1,54**	2,78
A2	4,87	2,57	1,25	1,39	1,84
A3	24,86	20,39	17,41	**11,30**	**11,76**
A4	0,75	0,49	0,31	0,66	0,31
A5	6,47	2,97	3,33	19,70	18,55
A6	15,07	7,21	5,80	4,26	8,63
A7	11,64	9,60	5,50	**2,28**	5,67
A8	5,26	2,90	1,94	3,01	3,71
B1	9,08	9,45	10,71	21,03	17,95
B2	9,98	7,97	6,10	**3,01**	**3,84**
B3	13,68	10,37	6,88	**3,25**	**4,28**
B4	15,06	11,66	7,20	**3,26**	**3,38**
B5	14,25	10,87	4,61	4,34	5,11
B6	6,84	5,14	3,90	2,83	3,38
B7	22,03	19,01	15,27	**2,90**	**7,51**
B8	23,08	19,34	15,08	**9,47**	**12,45**
B9	17,06	14,94	11,43	**3,74**	**5,06**
B10	67,15	52,81	50,45	**43,78**	**47,25**
B11	23,49	19,72	15,50	**5,51**	**7,25**
A8ir	10,92	6,91	2,53	1,86	–
B10ir	56,64	49,52	43,68	48,50	–
B11ir	27,08	24,67	19,22	**8,84**	–

bolded – better results for LSTM, underline – worse results for LSTM.
$LSTM_G$ – results for LSTM network with input signal G duplicated on other components.

Table 4
Comparison of signal extraction methods: ICA, POS, and LSTM (RMSE values).

Video no.	POS	ICA	$LSTM_{RGB}$
A1	1,14	1,68	1,54
A2	1,13	1,09	1,39
A3	11,56	11,16	11,30
A4	0,31	0,33	0,66
A5	1,62	2,03	19,70
A6	2,04	1,55	4,26
A7	1,88	1,90	2,28
A8	1,87	1,67	3,01
B1	17,53	14,30	21,03
B2	2,61	2,71	3,01
B3	3,61	3,96	3,25
B4	3,08	3,02	3,26
B5	7,53	6,79	4,34
B6	1,88	1,73	2,83
B7	3,72	10,17	2,90
B8	9,51	11,10	9,47
B9	4,96	8,83	3,74
B10	2,60	30,64	43,78
B11	5,72	7,94	5,51

$LSTM_{RGB}$ – RMSE values provided for all videos excluding IR.

The summary of all signal extraction methods is presented in Fig. 9 and Table 5. In the box plot, it can be seen that more outliers are present for the case of LSTM network compared to other methods. This may suggest that the generalization potential of the LSTM network is limited or that there is a need to consider more diverse training data.

The results of the tests (presented in Table 6) carried out on several videos from the MR-NIRP database (both RGB and IR) confirm the network's ability to generalize. Obtained results are consistent with the results for videos from our own database, despite the need to adjust the sampling frequency for the NIRP database.

5. Conclusion

A reliable noncontact monitoring of cardiovascular parameters with the use of videoplethysmography can be difficult, because many factors can introduce noise to the signal, e.g., subject movement and illumination changes. Thus, many signal preprocessing algorithms and VPG extraction methods have been developed to form the VPG signal and improve its quality. They usually consist of several steps, which require careful parameter selection. In contrast, in this paper, a new method of VPG signal processing based on the LSTM network is introduced. It is intended to replace the VPG signal pre- and post-processing stages, consequently reducing the number of algorithm parameters that affect the accuracy of the HR estimation.

Testing the LSTM network on videos recorded using different cameras, with various recording conditions (distance, lightning) and monitored persons, shows that LSTM generalizes well. The LSTM network was trained on selected parts of the signal using data augmentation technique. Accuracy of heart rate estimation is good for most of the test sequences and outperforms the basic methods based on G channel. It is worth noting that the LSTM network does not differ much from other methods in terms of accuracy and allows for pulse estimation on the basis of one component (either green or IR), which is not possible for

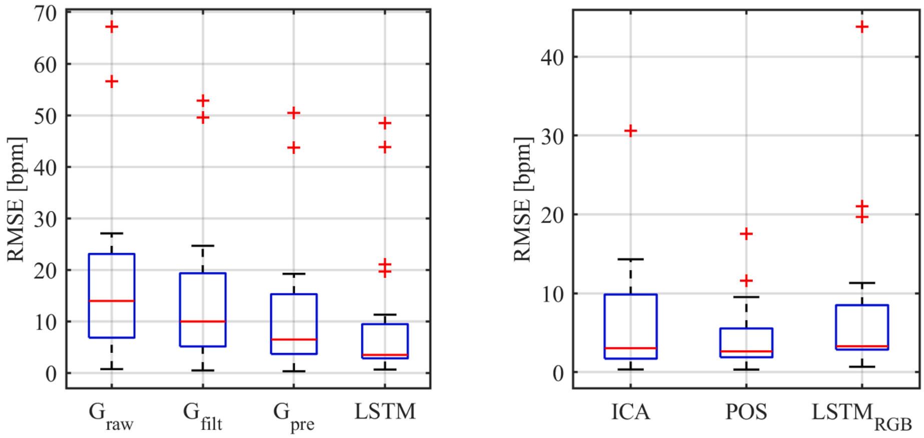


Fig. 9. A comparison of signal extraction methods – box plots for RMSE. Blue lines – IQR range, red line – median value.

Table 5
Comparison of all signal extraction methods (median RMSE for all videos).

G_{raw}	G_{filt}	G_{pre}	LSTM	POS	ICA	$LSTM_{RGB}$
13.96	9.98	6.49	3.5	2.61	3.02	3.26

$LSTM_{RGB}$ – RMSE values provided for all videos excluding IR.

Table 6
Comparison of signal extraction methods: G, POS, and LSTM (RMSE values) for videos from MR_NIRP database.

Video	G_{pre}	POS	LSTM	$LSTM_{RGB}$
Subject1_still_940 RGB	8.94	2.5	–	3.98
Subject2_still_940 RGB	10.59	4.22	–	7.49
Subject3_still_940 RGB	10.98	5.28	–	4.99
Subject1_still_940 IR	14.53	–	7.96	–
Subject2_still_940 IR	14.73	–	10.96	–

POS and ICA methods.

One important observation is that the high variability of conditions complicates the network training. The impact of factors such as training dataset selection and learning parameters on LSTM learning could not be unequivocally determined. Thus, further research will have to focus on the training dataset selection and preparation. Recording more training videos that include a broader spectrum of HR values and noise from different sources, might be difficult. Thus, a promising research direction will be to generate artificial training signals that make up gaps in the training data. However, the open question is whether artificial data allow for correct network training and ensuring good generalization. The other possible research direction is to use different types of network architectures, for example, encoder-decoder or LSTM networks with attention.

CRediT authorship contribution statement

Jaromir Przybyło: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

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.

Acknowledgements

This research was funded by AGH University of Science and Technology in 2021 as research project No. 16.16.120.773

References

[1] J. Kranjec, S. Beguš, G. Geršak, J. Drnovšek, Rev. Biomed. Signal Process. Control 13 (2014) 102–112, https://doi.org/10.1016/j.bspc.2014.03.004.
[2] W. Wang, A.C. den Brinker, S. Stuijk, G. de Haan, Robust heart rate from fitness videos, Physiol. Meas. 38 (6) (2017) 1023–1044, https://doi.org/10.1088/1361-6579/aa6d02.
[3] C. Zhao, C.-L. Lin, W. Chen, M.-K. Chen, J. Wang, Visual heart rate estimation and negative feedback control for fitness exercise, Biomed. Signal Process. Control 56 (2020), 101680.
[4] D.J. McDuff, J. Hernandez, S. Gontarek, R.W. Picard, 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, ACM, New York, NY, USA, 2016: pp. 4000–4004. https://doi.org/10.1145/2858036.2858247.
[5] Q.i. Zhang, Q. Wu, Y. Zhou, X. Wu, Y. Ou, H. Zhou, Webcam-based, non-contact, real-time measurement for the physiological parameters of drivers, Measurement 100 (2017) 311–321.
[6] R. Favilla, V.C. Zuccala, G. Coppini, Heart rate and heart rate variability from single-channel video and ICA integration of multiple signals, IEEE J. Biomed. Health. Inf. 23 (6) (2019) 2398–2408.
[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, J. Med. Biol. Eng. 39 (1) (2019) 76–85.
[8] G. de Haan, V. Jeanne, Robust pulse rate from chrominance-based rPPG, IEEE Trans. Biomed. Eng. 60 (10) (2013) 2878–2886.
[9] M. Hülsbusch, An image-based functional method for opto-electronic detection of skin-perfusion, Ph.D. Dissertation (in German), Dept. Elect. Eng., RWTH Aachen Univ., Aachen, Germany. (2008).
[10] 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, https://doi.org/10.1109/TBME.2010.2086456.
[11] W. Wang, A.C. den Brinker, S. Stuijk, G. de Haan, Algorithmic principles of remote PPG, IEEE Trans. Biomed. Eng. 64 (7) (2017) 1479–1491.
[12] J. Przybyło, Continuous distant measurement of the user's heart rate in human-computer interaction applications, Sensors 19 (2019) 4205.
[13] W. Verkruysse, L.O. Svaasand, J.S. Nelson, Remote plethysmographic imaging using ambient light, Opt Express. 16 (26) (2008) 21434, https://doi.org/10.1364/OE.16.021434.
[14] Y.u. Sun, N. Thakor, Photoplethysmography revisited: from contact to noncontact, From Point Imag. IEEE Trans. Biomed. Eng. 63 (3) (2016) 463–477, https://doi.org/10.1109/TBME.2015.2476337.
[15] A.M. Unakafov, Pulse rate estimation using imaging photoplethysmography: generic framework and comparison of methods on a publicly available dataset, Biomed. Phys. Eng. Express 4 (4) (2018) 045001, https://doi.org/10.1088/2057-1976/aabd09.
[16] X. Li, J. Chen, G. Zhao, M. Pietikainen, Remote heart rate measurement from face videos under realistic situations, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4264–4271.
[17] H. Ernst, M. Scherpf, H. Malberg, M. Schmidt, Optimal color channel combination across skin tones for remote heart rate measurement in camera-based photoplethysmography, Biomed. Signal Process. Control 68 (2021), 102644.
[18] B. Huang, C.-M. Chang, C.-L. Lin, W. Chen, C.-F. Juang, X. Wu, Visual Heart Rate Estimation from Facial Video Based on CNN, in: in: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020, pp. 1658–1662.

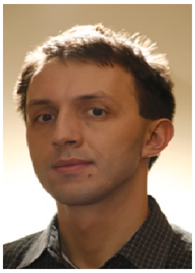
[19] M. Kopeliovich, Y. Mironenko, M. Petrushan, Architectural tricks for deep learning in remote photoplethysmography. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[20] 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.

[21] OpenVINO^{TM} Toolkit, (n.d.). https://docs.openvinotoolkit.org/ (accessed March 1, 2020).

[22] OpenVINO^{TM} Toolkit, version 2020.2, Face detection model, (n.d.). https://docs.openvinotoolkit.org/2020.2/_models_intel_face_detection_0105_description_face_detection_0105.html (accessed March 1, 2020).

[23] E. Magdalena Nowara, T.K. Marks, H. Mansour, A. Veeraraghavan, Sparseppg, Towards driver monitoring using camera-based vital signs estimation in near-infrared, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1272–1281.



Jaromir Przybyło graduated with an outstanding proficiency, with M.Sc. Eng. degree in 2000 from the Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, AGH-UST. Ph.D. degree (with honours) received in 2008 in Computer Science: Biocybernetics and Biomedical Engineering. Since 2000, he has been working at the Automatics and Biomedical Engineering Department, as a research scientist, and (2009) as an assistant professor. Since 2018 he works at the Department of Biocybernetics and Biomedical Engineering AGH-UST. His scientific interests focus on Assistive Technology that use optical observation and image analysis and recognition methods. A potential applications of research results include: systems allowing assessing the emotional state of a person and impact of that state on performed activities, detecting sudden changes in health or enabling disabled people using computers and other devices.