Master Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Robust visual heart rate estimation

Bc. et Bc. Radim Špetlík

Supervisor: Ing. Jan Čech, PhD.
May 2018

CTU
CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Špetlík Radim** Personal ID number: **421050**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Branch of study: **Computer Vision and Image Processing**

II. Master’s thesis details

Master’s thesis title in English:

Robust Visual Heart Rate Estimation

Master’s thesis title in Czech:

Robustní vizuální odhadování tepu

Guidelines:

1. Visual heart rate estimation is a well-known method . However, it has been demonstrated in near ideal conditions, i.e. in a high quality video with a large dominant frontal face.
2. Review the literature and conduct experiments to establish limitations of the published methods.
3. Propose a robust method applicable in the broadest set of conditions that, besides the pulse, reports the confidence of the estimate
4. Implement and experimentally validate the method.

Bibliography / sources:

[1] J. Allen. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3):R1, 2007.
[2] H. Liu, Y. Wang, and L. Wang. A review of noncontact, low-cost physiological information measurement based on photoplethysmographic imaging. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2012:2088-2091, 2012.
[3] Y. Sun and N. Thakor. Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging. IEEE Transactions on Biomedical Engineering, 63(3):463-477, Mar. 2016.
[4] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, B. Karasfi, Y. S. Ali, and F. Meriaudeau. Heart rate estimation using facial video: A review. Biomedical Signal Processing and Control, 38:346-360, Sept. 2017

Name and workplace of master’s thesis supervisor:

Ing. Jan Čech, Ph.D., Visual Recognition Group, FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **12.01.2018** Deadline for master's thesis submission: **25.05.2018**

Assignment valid until: **30.09.2019**

Ing. Jan Čech, Ph.D. Supervisor’s signature	doc. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Ing. Pavel Ripka, CSc. Dean’s signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student's signature

Acknowledgements

I would like to express my deep gratitude to Assistant Professor Ing. Jan Čech, Ph.D. and Professor Ing. Jiří Matas, Ph.D., my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank Assistant Professor Ing. Vojtěch Franc Ph.D., for his advices and assistance.

I would like to extend my thanks to the system administrators of the Computer Vision group of Center for Machine Perception for their help in keeping the computational resources in working order.

Finally, I wish to thank my parents for their support and encouragement throughout my study and my wife Barborka, the best of wives.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague 10 May 2018

....
signature

Abstract

A novel heart rate estimator, HR-CNN – a two-step convolutional neural network, is presented. The network is trained end-to-end by alternating optimization to be robust to illumination changes and relative movement of the subject and the camera. The network works well with images of the face roughly aligned by an of-the-shelf commercial frontal face detector.

An extensive review of the literature on visual heart rate estimation identifies key factors limiting the performance and reproducibility of the methods as: (i) a lack of publicly available datasets and incomplete description of published experiments, (ii) use of unreliable pulse oximeters for the ground-truth reference, (iii) missing standard experimental protocols.

A new challenging publicly available ECG-Fitness dataset with 205 sixty-second videos of subjects performing physical exercises is introduced. The dataset includes 17 subjects performing 4 activities (talking, rowing, exercising on a stepper and a stationary bike) captured by two RGB cameras, one attached to the currently used fitness machine that significantly vibrates, the other one to a separately standing tripod. With each subject, “rowing” and “talking” activity is repeated with a halogen lamp lighting. In case of 4 subjects, the whole recording session is also lighted by an LED light.

HR-CNN outperforms the published methods on the dataset reducing error by more than a half. Each ECG-Fitness activity contains a different combination of realistic challenges. The HR-CNN method performs the best in case of the “rowing” activity with the mean absolute error 3.94, and the worst in case of the “talking” activity with the mean absolute error 15.57.

Keywords: heart, heart rate, HR, visual, estimation, heart rate estimation, photoplethysmography, reflectance, non-contact, remote, video, video heart rate estimation, robust, motion robust, illumination robust

Supervisor: Ing. Jan Čech, PhD.
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Karlovo namesti 13
121 35 Prague 2
Czech Republic

Abstrakt

Je představena nová metoda odhadu srdeční frekvence, HR-CNN – dvoustupňová konvoluční neuronová síť. Síť je trénována *end-to-end* alternující optimalizací a je robustní vůči změnám osvětlení a relativnímu pohybu snímaného objektu a kamery. Síť funguje dobře s nepřesně registrovaným obličejem z komerčního obličejového detektoru.

Z rozsáhlého rozboru relevantních zdrojů vyplývají klíčové faktory omezující přesnost a reprodukovatelnost metod jako: (i) nedostatek veřejně dostupných datových sad a nedostatečně popsané experimenty v publikovaných článcích, (ii) použití nespolehlivého pulzního oximetru pro referenční ground-truth, (iii) chybějící standardní experimentální protokoly.

Je představena nová veřejně dostupná datová sada ECG-Fitness, která obsahuje 205 minutových videí, v nichž 17 dobrovolníků cvičí na posilovacích strojích. Dobrovolníci provádí celkem 4 aktivity (rozhovor, veslování, cvičení na stepperu a na rotopedu). Každá aktivita je zachycena dvěma RGB kamerami, z nichž jedna je připevněna k právě používanému posilovacímu stroji, který výrazně vibruje, a druhá je uchycena na samostatně stojícím stativu. Aktivity “veslování” a “rozhovor” opakují dobrovolníci dvakrát. Při druhém opakování jsou osvětleni halogenovou lampou. 4 dobrovolníci jsou osvětleni LED světlem ve všech šesti videích.

HR-CNN má o více jak polovinu lepší výsledky než dosud publikované metody. Každá aktivita v ECG-Fitness datasetu představuje jinou kombinaci realistických výzev. HR-CNN má nejlepší výsledky v případě aktivity “veslování” s průměrnou absolutní chybou 3.94 a nejhorší v případě aktivity “rozhovor” s průměrnou absolutní chybou 15.57.

Klíčová slova: srdce, srdeční tep, tep, puls, tepová frekvence, srdeční puls, vizuální, odhad, odhad srdečního pulsu, photoplethysmografie, reflektivní, bezkontaktní, video, odhad srdečního pulsu z videa, robustní, robustní vůči pohybu, robustní vůči změně osvětlení

Překlad názvu: Robustní vizuální odhadování tepu

Contents

Figures

Tables

Chapter 1
Introduction

Heart rate (HR) is a basic parameter of cardiovascular activity [1]. The measurement of HR is broadly used – from monitoring of exercise activities to prediction of acute coronary events. The HR measurement is commonly performed simply by palpating the pulse or by dedicated devices, e.g. pulse oximeters or electrocardiographs. The more expensive the device, the more precise and reliable the measurement. These methods of measurement require physical contact.

Visual HR estimation, i.e. HR estimation from a stored video sequence or a direct feed from a camera, has recently received significant attention [2, 3]. In suitable conditions [4], the accuracy of visual HR estimation methods is comparable to the accuracy of contact methods and by not requiring physical contact, the subject's comfort is improved. Moreover, the measurement can be done at a distance. Also, the recorded material need not to be primarily designed for HR estimation allowing *ex post* analysis.

The accuracy of the visual HR estimation depends on acquisition conditions. Best known visual HR methods are highly sensitive to motion and lighting and thus require subject's cooperation. Commonly, the ground truth is provided by the pulse oximeter, which is sensitive to the quality of contact with the skin and the lighting setup. The datasets used for evaluation of the visual HR methods reflect their assumptions – subjects do not move and they are illuminated by daylight or a professional studio light source. As a rule, an engineered, complicated signal processing pipelines consisting of several consecutive steps have been used (e.g. [5] or [6]) and thus it is non-trivial to robustify such approaches.

The thesis presents HR-CNN, a novel heart rate estimator which is a two-step convolutional neural network. The network is trained end-to-end by alternating optimization and is robust to illumination changes and relative movement of the subject and the camera. The network works well with an image of the face roughly aligned by an of-the-shelf commercial frontal face detector.

HR-CNN is evaluated on a newly collected dataset. The dataset, called ECG-Fitness, contains 205 videos of 17 subjects. The subjects perform rapid movements in unconstrained directions. Lighting conditions include daylight and interfering lighting. The videos are uncompressed and the ground-truth

HR is given by an electrocardiograph.

An experimental protocol evaluating robustness to motion and illumination conditions is introduced alongside the dataset. The protocol adopts the error statistics of Heusch et al. [7] and follows the proposed methodology. We note that there is no common methodology and experimental protocol used in the visual HR estimation research.

In a summary, we consider multiple aspects of the visual HR estimation problem. We develop a motion and lighting robust HR estimation method and evaluate on a newly collected challenging dataset with a standardized protocol.

The rest of the thesis is organized as follows. Chap. 2 presents the related work and discusses methodology and terminology of the visual HR estimation research in detail. Chap. 3 introduces the developed method – the two-step convolutional neural network “HR-CNN”. In Chap. 4, an extensive experimental evaluation including a large-scale comparative study, two experiments on the network's interpretation, and five experiments inspecting the network's robustness are given.

Chapter 2
Related Work

Visual HR estimation methods compute heart rate (HR) by analyzing subtle changes of skin-color. It is believed that these changes are caused by peripheral circulation of blood – the analyzed signal is a “blood volume signal”. Historically, analysis of peripheral circulation was a domain of plethysmography. Plethysmography, from Greek πληϑος (fullness) and γραφός (to write) [8], measures changes in volume inside a living body. In 1936, Molitor and Kniazuk [9] introduced photoplethysmography (PPG) that performs the measurements remotely with a photosensitive device. Today, PPG is in fact a synonym for non-contact monitoring of cardiovascular activity [10]. PPG based devices monitor the human heart rate and estimate the level of oxygen in blood. PPG may be performed in two basic modes. Transmittance PPG (tPPG) and reflectance PPG (rPPG). In tPPG, the photodetector captures light transmitted through the body tissue, in rPPG, the reflected light is recorded. Both forms exist in contact and non-contact versions.

We are interested in HR estimation performed remotely by monitoring peripheral circulation of blood, i.e. in non-contact reflective photoplethysmographic (NrPPG) HR estimation or simply “visual HR estimation”. To the best of our knowledge, there are four recent studies that review the NrPPG research.

The earliest is the work of Allen [10] from 2007. Allen focuses on the clinical application of PPG approaches and includes references to the rPPG as well. The rPPG is represented by several papers. The topics range from plastic surgery post-operative monitoring to oculoplethysmography, a non-contact method of detecting carotid occlusive arterial disease.

A more recent work of Liu et al. [11] tracks the rapid development of the rPPG approaches between years 2007 and 2012. Authors interpret the cause of the development as the introduction of cheap and relatively precise measuring devices, e.g. web cameras and alike. Liu et al. conclude that although the NrPPG is comparable with the traditional contact tPPG systems it needs further improvement for the clinical use in terms of signal-to-noise ratio.

A paper by Sun and Thakor [3], published in September 2015, provides a survey of a large body of the literature focused on contact and NrPPG methods, there referred to as imaging PPG. The differences between the discussed methods are shown on the different choices taken during the procedure of obtaining the blood volume signal. The authors conclude that the NrPPG "will dramatically change our lifestyle in the near future".

The most recent work of Hassan et al. [2] from September 2017 provides a comparison of the heart rate estimation methods that employ a video recording. Authors provide a review of the methods based on illumination variance and subtle head motion induced by ballistic forces of the heart and they conclude that "non-invasive nature [of the NrPPG] opens possibilities for health monitoring towards various fields such as health care, telemedicine, rehabilitation, sports, ergonomics and crowd analytics".

The reviews present over 60 studies on HR estimation using NrPPG. Majority of them is performed on private datasets with *ad hoc* evaluation procedures. Only one of them [5] is validated on a publicly available dataset.

Recently, Heusch et al. [7] reimplemented two baseline HR estimation approaches [12, 13] and a method of Li et al. [5]. Also, an experimental protocol was introduced in [7] enabling a comparison of the rPPG methods. The authors tested the three works on the MAHNOB HCI-Tagging dataset [14]. Interestingly, they were not able to fully reproduce the results reported by Li et al. They argue that it might be caused by an unknown parameter setting of the blood volume signal extraction pipeline. Heusch et al. provided all reimplemented codes and also collected a publicly available dataset COHFACE[1]. Since the three reported studies are the only ones tested on public datasets, we will discuss only these three.

An approach of Haan et al. [12] (referred to as CHROM) is based on combining color difference, i.e. chrominance, signals. First, skin-color pixels are found in each frame of input sequence. Then, an average color of skin pixels is computed in each frame and projected on a proposed chrominance subspace. The projected signals are bandpass filtered separately in the XY chrominance colorspace and projected into a one-dimensional signal. The algorithm is shown to outperform blind source separation methods on a private dataset of 117 static subjects.

Li et al. [5] (referred to as LiCVPR) is the only approach validated on a publicly available dataset. Bottom part of a face is found in the first frame of a sequence and tracked with Lucas-Kanade tracker [15]. An average intensity of the green channel over the area of measurement is computed in each frame and corrected for illumination changes. Background is segmented and its average green intensity is used to mitigate illumination variations with a Normalized Least Mean Squares filter. Then, subject's non-rigid motions are eliminated simply by discarding the motion-contaminated segments of the signal. Finally, temporal filters are applied and Welch's power spectral density estimation method is used to estimate the HR frequency. Experiments are performed on two datasets, a private one and the MAHNOB HCI-Tagging dataset. Pearson's correlation coefficient of 0.81 is reported in the experiments on the MAHNOB dataset.

The last considered approach is Spatial Subspace Rotation (referred to as 2SR) by Wang et al. [16]. First, skin pixels are found in each frame.

[1] https://idiap.ch/dataset/cohface

Then, subspace of skin pixels in the RGB space is built for each frame in the spatial domain. The rotation angle between the spatial subspaces is computed and analyzed between consecutive frames. Authors claim that no bandpass filtering is required to obtain the blood volume signal. The method is validated on a private dataset consisting of 54 videos. Performance of the algorithm under various conditions such as skin tone, subject's motion and recovery after a physical exercise is examined resulting in correlation coefficient of 0.94.

The rest of the chapter is organized as follows. In Sec. 2.1, a taxonomy of the NrPPG approaches is given. Sec. 2.2 comments on the methodology in the NrPPG research. Sec. 2.3 discusses the gold standards of NrPPG. Sec. 2.4 and Sec. 2.5 analyzes the difficulties in the blood volume signal reconstruction. In Sec. 2.6, subtle movements of human body caused by cardiovascular activity are discussed. Sec. 2.7 concludes the whole chapter.

2.1 Terminology and Taxonomy

Sun and Thakor [3] were the first to provide a detailed survey on the NrPPG methods, there referred to as imaging PPG. We consider this naming convention potentially misleading. It suggests that there is something unique to the NrPPG approaches employing CMOS and CCD cameras. We find the only difference in the number of photodetectors performing the readings. It comes natural that any two methods that process the same type of signal coming from the same type of sensor should be in the same category. The fact that in one case only a single sensor and in the other millions of them are used should not play a role. Furthermore, the term is very similar to the "PPG imaging" (e.g. [17]). PPG imaging refers to a process of mapping spatial blood-volume variations in living tissue with a video camera [18]. Not surprisingly, there is a line of research by Kamshilin et al. [19, 20] in which the term imaging PPG is used when the PPG imaging is actually thought. Therefore we propose a taxonomy based on a clear distinction between the approaches.

We recognize **tPPG** and **rPPG** methods as described in the beginning of Chap. 1. These may be performed in either **contact** or **non-contact** manner.

Inside the NrPPG branch, another important partitioning may be made. One group of approaches preforms the NrPPG capture with an **ambient lighting**, the second uses a **supplementary lighting**. This classification follows a line of research showing the importance of the light source spectral composition [21, 22, 23, 24, 25, 26].

A PPG method may perform a **blood volume imaging** or a **blood volume signal (BVS) reconstruction**. Both the blood volume imaging and BVS refer to the measured quality – the volume of blood passing through the tissue. PPG, on the other hand, refers to the measurement setup – measuring is performed with a specific illumination and photosensitive device setup.

HR may be estimated from the BVS, e.g. by counting the number of peaks in a given time interval of the signal. HR estimated from the BVS is sometimes called the "blood volume pulse". Visual HR estimation is a NrPPG method performing blood volume pulse estimation.

Note that this taxonomy is not used by the researchers. In the discussed literature, the terminology is vague and inconsistent. The only common denominator is that the research is performed with a PPG technology.

2.2 Experimental Methodology

In 2007, Allen [10] complains that there are no internationally recognized standards for a clinical PPG measurement, and that the published research tends to be using "quite differing measurement technology and protocols," thereby limiting the reproducibility of the outcomes.

Schafer and Vagedes [27] review existing PPG studies in 2013 and they conclude that generally speaking, "quantitative conclusions are impeded by the fact that results of different studies are mostly incommensurable due to diverse experimental settings and/or methods of analysis".

Works published during the rapid development period of the NrPPG field in the last decade did not follow the recommendations given by Allen nor Schafer and Vagedes. In 2018, there are still no PPG measurement standards, the researchers in the NrPPG field use different experimental settings, the studies fail to report fundamental details about the setup of the experiments. However, an attempt has been made to improve the reproducibility of the NrPPG research by Heusch et al. [7] as discussed earlier.

2.3 Gold Standard

It was reported by numerous works (e.g. [28, 29, 30, 31, 32, 18, 33, 34, 35, 36, 6, 37]) that a signal obtained from a transmittance-mode pulse oximeter may serve as a gold standard in the evaluation of a NrPPG approach. However, the results of the research discussed bellow suggest that the reliability of the device is limited.

Mardirossian and Schneider showed that various physiological factors, heavy skin pigmentation including, are a source of the erroneous measurement of the device [38]. Trivedi et al. [39] examined five commercially available pulse oximeters during hypoperfusion[2], probe motion, and exposure to ambient light interference. None of the inspected devices performed the best under all conditions with failure rates varying from approximately 5% to 50%. Teng and Zhang [40] showed that the BVS obtained from a pulse oximeter is affected by a "contacting force between the sensor and the measurement site". Moreover, Palve in [41] concludes that a reflection-mode pulse oximeter gives more accurate readings under less than ideal conditions, which is agreed also

[2] Hypoperfusion is the inadequate perfusion of body tissues, resulting in inadequate supply of oxygen and nutrients to the body tissues.

by Wax in [42] and Nesseler in [43]. We consider these findings as a very good reason for abandoning the transmittance-mode pulse oximeter as a gold standard.

Due to the results of Buchs et al. [44], who showed that the BVS measured in the two index fingers and the two second toes differs for diabetic and non-diabetic subjects, and Nitzan et al. [45], who found that the pulse transit time is a function of a subject's age, we also consider the reflectance-mode pulse oximeter as compromised.

Based on the outcomes of the presented works and our own readings (see Fig. 4.5 on how a BVS differs for two devices), we conclude that an electrocardiograph instead of the pulse oximeters should be used as the gold standard for evaluation of a particular NrPPG approach.

2.4 Blood Volume Signal Reconstruction Difficulty

Advanced signal processing methods "needed to recover the [heart rate] information" are presented in [29]. Independent component analysis (ICA), principal component analysis, auto- and cross-correlation are compared and it is concluded that the most suitable method for the purpose of heart rate estimation is the ICA. In the following paragraphs, we discuss the conditions strongly affecting the accuracy of the heart rate estimation methods. As the heart rate in NrPPG approaches is obtained by processing a blood volume signal, we will focus on the BVS reconstruction.

We identify four major causes that can make the BVS reconstruction task difficult: (i) a video compression, (ii) a lighting setup, (iii) subject's movement and (iv) a skin type. The compression is discussed in Sec. 4.2.3.

Subject's movement may be mitigated by precise tracking and weighted spatial averaging [31]. Also a multi-imager array was proposed to improve the motion robustness of the NrPPG reconstruction [46, 47]. When the NrPPG imaging or deeper analysis of the BVS mechanisms is pursued, also the ballistocardiographic movement (BCG), i.e. the movement induced by the ballistic forces of the heart, must be accounted for [13] (see Sec. 2.6).

By the lighting setup the light source position and intensity, both in space and over time, and its spectral composition are meant. Stationary, uniform and orthogonal lighting was shown to minimize artifacts in the BVS that are induced by the BCG movement – the variations in the light flux "amplify the modulation caused by subtle BCG motions" [18]. Effects of the light source spectral range were studied intensively [25, 23, 24, 26, 22], and a model predicting the relative NrPPG-amplitude was proposed [21] and verified [12]. Given the spectral composition of light, absorption spectrum of the oxygenated blood and dermis, and assuming 3% concentration of melanin, the authors were able to determine the spectral response of the BVS in the red, green, and blue channels of a camera.

The blood volume signal-to-noise ratio is typically unfavorable. However, if properly captured, the BVS may be recovered by simple spatial averaging [28, 48, 31, 4, 35, 13, 49, 50] as confirmed in our experiments.

Furthermore, we discourage from use of blind source separation methods (BSS) in the BVS reconstruction. When the BSS methods are employed, an assumption has to be made that a blood volume signal is the only periodic component in the video [51]. This assumption is generally not true [12]. When the cameras have the sampling rate close to the AC current frequency and a common light source illuminates the subject, aliasing effect might occur resulting in a corruption of the signal. Moreover, use of the BSS for the clinical application is limited by the fact that, as to the order of the decomposed components, BSS methods are ambiguous [37], and a heuristic-driven selection must be performed. Hence, instead of trying to recover the signal that might not even be present one should focus on avoiding video compression, improving the lighting setup, and accounting for the subject's movement.

2.5 Motion Corruption

The biggest limitation of the existing NrPPG methods seems to be the inability to cope with the motion-induced artifacts in the captured signal. Recent approaches try to resolve this issue by increasing the dimensionality of the signal, thus improving separability of the BVS from distortions caused by motion. We recognize two major research directions in this matter. The first is represented by [52]. Here the dimensionality of the measurement is increased algorithmically by decomposing the RGB-signals into different frequency bands. In the second, the dimensionality of measurement is increased physically, e.g.. by using a 5-band camera as in [53]. Currently, the approaches based on the algorithmic principles are receiving more attention, probably because the price of the specialized hardware is high and its availability is limited.

2.6 Ballistocardiographic Movements

HR estimation may be performed by analysis of subtle movements of human body caused by cardiovascular activity. The movements are known as ballistocardiographic movements. Ballistocardiography (BCG) studies ballistic forces of the heart, i.e. the inertial forces induced by the blood pulsation. Balakrishnan et al. was the first to recognize the BCG movement of a human face in a video and demonstrated reconstruction of the BVS with a blind source separation based approach [54].

One of the recent works [55] uses a combination of the BCG movement and color information to reconstruct the BVS related measures. A key BCG study was performed by Moço et al. who inspected an extent to which the BCG artifacts, i.e. motion artifacts inflicted by cardiac activity, influence the PPG imaging techniques [18]. Contamination of blood volume imaging maps was showed to be severe implicating that the BCG artifacts must be accounted for in any research in the NrPPG imaging field. Otherwise a misinterpretation of the results is at hand. In this matter a recently proposed new physiological model of the remote PPG introduced by Kamshilin et al. [56] was inspected and needs to be reexamined.

As to the BCG movements, we are not making use of them in the HR estimation but we consider them to be a source of a corruption of the BVS.

2.7 Lessons Learned from Literature

By reviewing the visual HR estimation literature we identified the key factors limiting the research as: (i) vague terminology, (ii) heterogeneous methodology, (iii) incomplete description of the datasets and experimental setups, and (iv) absence of publicly available datasets.

Chapter 3
Method

We develop a two-step convolutional neural network to estimate a heart rate of a subject in a video sequence. Overview of the method is shown in Fig. 3.1. The input of the network is a sequence of images of the subject's face in time. The output is a single scalar – the predicted HR.

The network composes of two parts, each performs a single step. In the first step, the *Extractor* network takes an image and produces a single number. By running the *Extractor* over a sequence of images, a sequence of scalar outputs is produced. In the second step, the *Extractor*-produced sequence is fed to the *HR Estimator* network that outputs the HR. The two networks are trained separately. First, given the true heart rate, the *Extractor* is trained to maximize the signal-to-noise ratio (SNR). Then, the *HR Estimator* is trained to minimize the mean absolute error (MAE) between the estimated and the true HR.

Let $\mathcal{T} = \{(\mathbf{x}_j^1, \dots, \mathbf{x}_j^N, f_j^*) \in \mathcal{X}^N \times \mathcal{F} \mid j = 1, \dots, l\}$ be the training set that contains l sequences of N facial RGB image frames $\mathbf{x} \in \mathcal{X}$ and their corresponding HR labels $f^* \in \mathcal{F}$. Symbol $\mathcal{X}$ denotes a set of all input images and $\mathcal{F}$ is a set of all sequence labels, i.e. the true HR frequencies measured in hertz. We presume the HR to be constant within a given sequence. If the HR changes rapidly, we use a piece-wise constant approximation by a sliding window.

3.1 Extractor

Let $h(\mathbf{x}^n; \mathbf{\Phi})$ be the output of the *Extractor* CNN for the n-th image and $\mathbf{\Phi}$ a concatenation of all convolutional filter parameters. The quality of the extracted signal is measured by the SNR using a power spectral density (PSD). Given frequency f

$$\mathrm{PSD}(f, \mathbf{X}; \mathbf{\Phi}) = \left(\sum_{n=0}^{N-1} h(\mathbf{x}^n; \mathbf{\Phi}) \cdot \cos\left(2\pi f \frac{n}{f_s}\right)\right)^2 + \left(\sum_{n=0}^{N-1} h(\mathbf{x}^n; \mathbf{\Phi}) \cdot \sin\left(2\pi f \frac{n}{f_s}\right)\right)^2 \tag{3.1}$$

where $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^N)$ is a sequence of N facial images, and f_s is a sampling frequency.

Intuitively, given a true HR, amplitude of its frequency should be high

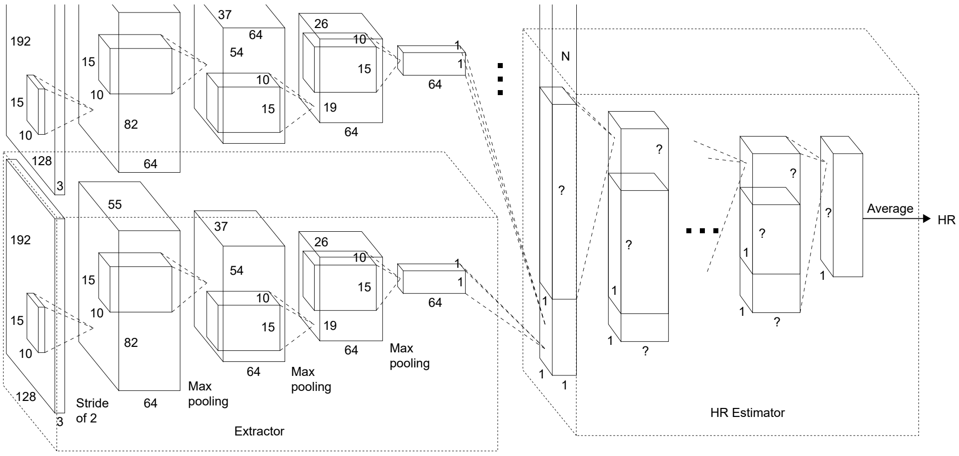


Figure 3.1: HR-CNN – architecture of the heart rate convolutional neural network. The *Extractor* takes an image and produces a single number. By running the extractor over a sequence of images, a sequence of scalar outputs is produced. The sequence is fed to the *HR Estimator* and a heart rate is predicted. The question marks illustrate that the architecture of the HR Estimator differs between datasets.

while amplitudes of background frequencies low. To measure the quality of the *Extractor*, the SNR introduced in [12]

$$\mathrm{SNR}(f^*, \mathbf{X}; \mathbf{\Phi}) = 10{\cdot}\log_{10}\left(\sum_{f\in\mathcal{F}^+} \mathrm{PSD}(f, \mathbf{X}; \mathbf{\Phi}) \Big/ \sum_{f\in\mathcal{F}\setminus\mathcal{F}^+} \mathrm{PSD}(f, \mathbf{X}; \mathbf{\Phi})\right) \tag{3.2}$$

is used where f^* is the true HR, $\mathcal{F}^+ = (f^* - \Delta, f^* + \Delta)$, and a tolerance interval Δ accounts for the true HR uncertainty, e.g. due to the HR non-stationarity within the sequence. The nominator captures the strength of the true HR signal frequency. The denominator represents the energy of the background noise, the tolerance interval excluding.

The structure of the CNN used in our experiments is shown in Table 3.1. The parameter $\mathbf{\Phi}$ is found by minimizing the loss function

$$\ell(\mathcal{T}; \mathbf{\Phi}) = -\frac{1}{l}\sum_{j=1}^{l} \mathrm{SNR}(f_j^*, \mathbf{X}_j; \mathbf{\Phi}). \tag{3.3}$$

3.2 HR Estimator

The *HR Estimator* is another CNN taking 1D signal – output of the *Extractor* CNN – and producing the HR. The training minimizes the average L$_1$ loss between the predicted and the true HR f_j^*

$$\ell(\mathcal{T}; \boldsymbol{\theta}) = \frac{1}{l}\sum_{j=1}^{l}\left| g\left(\left[h(\mathbf{x}^1; \mathbf{\Phi}), \cdots, h(\mathbf{x}^N; \mathbf{\Phi})\right]; \boldsymbol{\theta}\right) - f_j^*\right| \tag{3.4}$$

Layer type	Configuration
Convolution	filt: 1, k: 1×1, s: 1, p: 0
ELU	
MaxPool	15×10, s: $(1,1)$, p: 0
Convolution	filt: 64, k: 12×10, s: 1, p: 0
ELU	
MaxPool	15×10, s: $(1,1)$, p: 0
Convolution	filt: 64, k: 15×10, s: 1, p: 0
ELU	
MaxPool	15×10, s: $(1,1)$, p: 0
Convolution	filt: 64, k: 15×10, s: 1, p: 0
ELU	
MaxPool	15×10, s: $(2,2)$, p: 0
Convolution	filt: 64, k: 15×10, s: 1, p: 0
Input	192×128 RGB image

Table 3.1: Structure of the *Extractor* network. The second column describes the number of filters 'filt', the filter size 'k', stride 's' and padding 'p'.

where $g\left(\left[h(\mathbf{x}^1;\mathbf{\Phi}),\cdots,h(\mathbf{x}^N;\mathbf{\Phi})\right];\boldsymbol{\theta}\right)$ is the output of the CNN for a sequence of N outputs of the *Extractor*, and $\boldsymbol{\theta}$ is a concatenation of all convolutional filter parameters of the *HR Estimator* CNN.

3.2.1 Discussion

Our first experiments were conducted on a non-challenging dataset. A simple argument maximum in the PSD of the *Extractor*'s output

$$\hat{f} = \arg\max_{f} \mathrm{PSD}(f, \mathbf{X}, \boldsymbol{\theta}) \tag{3.5}$$

gave MAE less than 3 (see Sec. 4.4.1). However, this simple HR estimation was not robust to challenges in videos from other datasets where a video compression was used, the subject's HR was not stationary, or subject's motion was present. Therefore, we introduced the *HR Estimator* CNN.

3.3 Implementation Details

In all experiments, the *Extractor* was trained on the training set of the PURE dataset (see Sec. 4.1) and fixed. Data augmentation including random translation (up to 42 pixels in y-axis and 28 pixels in x-axis), cropping and rotation (±5 degrees) was applied to each frame of the training sequence. Brightness was randomly adjusted for a whole sequence. The HR Estimator was trained for each dataset separately. During the training of the ECG-Fitness dataset, the sequences were split to 10 seconds clips to account for the rapid HR changes.

Block	Layer type	Configuration
	Convolution	filt: 1, k: 1, s: 1, p: 0
12	ELU	
	MaxPool	p: 0
	Convolution	s: 1, p: 0
11	ELU	
	Convolution	s: 1, p: 0
10	ELU	
	Convolution	s: 1, p: 0
⋮	⋮	⋮
3	ELU	
	MaxPool	p: 0
	Convolution	s: 1, p: 0
2	ELU	
	Convolution	s: 1, p: 0
1	ELU	
	Convolution	s: 1, p: 0
Input	192 × 128 RGB image	

Table 3.2: Modular structure of the *HR Estimator*. The final structure is configured on-the-fly by selecting active blocks. The block number 1 is always activated. The first column denotes the number of the block, the second the type of a layer and the third describes the number of filters 'filt', the filter size 'k', stride 's' and padding 'p'. The number of convolutional filters, filter sizes and MaxPool kernel sizes is different for each dataset.

Both networks use a standard chain of convolution, MaxPool and activation functions and share the following settings. Before the first convolution layer and after every MaxPool layer, a batch normalization was inserted. Exponential Linear Units [57] were used as the activation functions. Dropout was used. Batch normalization was initialized with weights randomly sampled from a Gaussian distribution with $\mu = 0$ and $\sigma = 0.1$, convolution layers were initialized according to the method described in [58]. Both networks were trained using PyTorch library, Adam optimizer was used with learning rate set to 0.0001 in case of the Extractor and to 0.1 in case of the HR Estimator.

For both training setups, a set of all input facial RGB images $\mathcal{X} = \mathbb{R}^{192\times 128}$. Faces were found by a face detector, the bounding boxes were adjusted to the aspect ratio 3 : 2 to cover the whole face, cropped out and resized to 192×128 pixels. The set of true HR $\mathcal{F} = \{\frac{40}{60}, \frac{41}{60}, \ldots, \frac{240}{60}\}$ in case of extractor and $\mathcal{F} = \mathbb{R}^{0+}$ in case of estimator.

The *Extractor*'s configuration (shown in Tab. 3.1) was the same for all experiments. The structure of the *HR Estimator* was "task specific" – it was configured for a particular dataset or its subset.

Algorithm 1 Metropolis-Hastings Monte-Carlo Random Walk

```
Given $X_t$,
Generate $Y_t \sim X_t + \varepsilon_t$
Take
    $X_{t+1} = \begin{cases} Y_t & \text{with probability} \quad \min\left\{1, \frac{f(Y_t)}{f(X_t)}\right\} \\ X_t & \text{otherwise.} \end{cases} \qquad (3.7)$
```

Configuring Structure of HR Estimator with Metropolis-Hastings Random Walk

In the first experiments, the structure of the *Extractor* was selected *ad hoc*. Challenging videos motivated us to perform the HR estimation by another CNN – the *HR Estimator*. To configure the structure of the HR Estimator, namely the depth of the network, the number of filters, and the conv and MaxPool sizes, the Metropolis-Hastings Monte-Carlo Random Walk was used.

In the Metropolis-Hastings Monte-Carlo Random Walk, a Markov chain denoted by (X_t) is used to sample from some target probability distribution $p(x) = f(x)/C$, where C is an unknown constant, for which a direct sampling is difficult [59]. To do so, a proposal distribution q is defined. Candidate Y_t is sampled depending on the current state X_t as

$$Y_t = X_t + \varepsilon_t \tag{3.6}$$

where ε_t is a random perturbation with distribution q, independent of X_t. To perform the sampling, a heuristic is implemented: if $f(Y_t) \geq f(X_t)$, keep the proposed state Y_t and set it as next state in the chain, otherwise accept the proposed state with a probability $f(Y_t)/f(X_t)$. Note that any constant C cancels out. The whole algorithm is depicted in Alg. 1. For a Markov chain to settle in a stationary distribution, probability of the transition $X_t \to X_{t+1}$ must be equal to the probability of the reverse transition $X_{t+1} \to X_t$. This constraint is fulfilled when the proposal distribution q is symmetric. Symmetric proposal distribution is the Normal, Cauchy, Student's-t, and Uniform distribution. In our case, we used the uniform proposal distributions in all cases.

In our setting, we presumed the target distribution f to be a multivariate with four dimensions corresponding to "layer activation/deactivation", "number of convolution filters", "size of convolutional filter" and "size of MaxPool kernel". The scaled probability density function is computed as

$$f(Y_t) = \frac{1}{\min\limits_{e=\{1,\ldots,500\}} (\ell(\mathcal{T};\boldsymbol{\theta}^e))} \tag{3.8}$$

where $\ell(\mathcal{T};\boldsymbol{\theta}^e)$ is the Mean Average Error from (3.4) for an epoch e. We applied a component-wise approach – in every iteration of the algorithm, we performed four samples. In case of the first component Y_t^1 (at the step t), an identification number of the block (see Tab. 3.2) was drawn uniformly from $\{2, 3, 4, \ldots, 12\}$ and the corresponding block was activated, if it was deactivated before, or deactivated, in the other case. The HR Estimator was trained for 500 epochs and the $f(Y_t)$ was computed using the minimum value of $\ell(\mathcal{T}; \boldsymbol{\theta})$ on the validation set. In the same manner, the procedure was repeated for the size of the MaxPool kernel $= \{3, 5, 10, 15\}$, the size of the convolutional filter $= \{3, 8, 16, 32, 64, 90\}$, and the number of convolutional filters $= \{4, 8, 16, 32, 64, 128, 256\}$. Non-valid combinations of parameters were skipped.

Chapter 4
Experiments

This chapter is organized as follows. In Sec. 4.1, four datasets used in the experiments are introduced. Sec. 4.2 presents the results of the two preliminary experiments showing the impact of the precise registration and compression on the quality of the reconstructed blood volume signal (BVS). An interpretation of the *Extractor* and the *HR Estimator* is given in Sec. 4.3 in two qualitative studies. The last Sec. 4.4 presents a thorough comparative study and five introspective experiments on the HR-CNN method.

4.1 Datasets

The visual HR datasets are small, usually 1 to 20 subjects, and private. To the best of our knowledge, there are three publicly available datasets for evaluation of HR estimation methods. In MAHNOB dataset [14], the ground truth is derived from an electrocardiograph. The PURE dataset [60] and COHFACE dataset [7] contain the ground truth from pulse oximeters. Devices performing contact PPG differ in both software and hardware implementation. Also, they are prone to inaccuracy due to various conditions (subject's health status, motion, external lighting) [38, 40, 39] and produce errors in the ground truth. HR error statistics, especially when using a pulse oximeter as a gold standard, might not be the best choice – here the HR may be obtained by various approaches, e.g. by computing number of peaks detected in one minute of a BVS for every consecutive sample, or by calculating the HR from distances between a couple of peaks. In both cases, averaging of HR over a certain time window may be applied. The peak detection algorithm and the averaging window length are not known for a particular device and its different setting was shown to have negative effects on the derived measures [61]. As discussed in Sec. 2.3, an electrocardiograph synchronized with the capturing device should be preferred as a gold standard reference. The issues of the available visual HR datasets inspired us to create a novel challenging dataset. We collected the ECG-Fitness dataset described in Sec. 4.1.4 where the ground-truth HR is given by an electrocardiograph.

The experiments are performed on the three publicly available datasets and on the newly collected ECG-Fitness dataset. For the purpose of evaluation, the following factors affecting the datasets must be taken into account:

	MAHNOB	COHFACE	PURE	ECG-Fitness
lighting	studio	daylight, studio	daylight	daylight, halogen lamp, LED
subject's head movement	none	none	none talking translation rotation	talking translation rotation scale
number of subjects	30	40	10	17
number of videos	3490	160	60	102[1]
sequence storage	compressed video	compressed video	lossless PNG	raw video
sequence compression	H.264	MPEG-4 Visual	none	none
sequence bits per pixel	$\approx 1.5 \times 10^{-4}$	$\approx 5 \times 10^{-5}$	24	24
sequence frame rate	61	20	30	30
frame resolution	780 × 580	640 × 480	640 × 480	1920 × 1080

1 51 videos of the same action from two cameras.

Table 4.1: Datasets used for visual heart rate estimation experiments in Chap. 4.

(i) the lighting conditions, (ii) the amount of subject's movement during the recording, and (iii) the data compression level. Tab. 4.1 contains the details about the datasets including the evaluation-relevant facts. The MAHNOB dataset (see Sec. 4.1.1) contains videos from 30 subjects. Majority of them sits still and watches a screen positioned in front of them lighted uniformly with a studio lighting. The COHFACE dataset (see Sec. 4.1.2) contains 40 subjects starring at a camera, studio and natural lighting setups are used. The PURE dataset consists of 10 subjects performing 6 different tasks, including head rotation and translation, in daylight. The ECG-Fitness dataset contains 17 subjects practicing on fitness machines in daylight, halogen lamp light and LED lamp light.

Note that a five-subject dataset used in the preliminary experiments is not covered in this section (see Sec. 4.2 for the description of both the dataset and experiment).

The rest of the section contains a more detailed description of the datasets including technical details.

4.1.1 MAHNOB HCI-Tagging

3739 videos of 30 young healthy adult participants are available. However, only 3490 videos are used in the experimental protocol. The corpus contains one color and five monochrome videos for each recording session. The videos were recorded in a controlled studio setup (see Fig. 4.1 for a sample video frame). For full details, please see the dataset manual[1] [14]. The lengths of the videos vary from 1 to 259 seconds. Subjects in the videos watch emotion-eliciting clips. Every session is accompanied by rich physiological data that include readings from electroencephalograph, electrocardiograph, temperature sensor and respiration belt. The videos are compressed in H.264/MPEG-4 AVC compression, bit rate ≈ 4200 kb/s, 61 fps, 780×580 pixels, which gets $\approx 1.5 \times 10^{-4}$ bits per pixel – the videos are heavily compressed.

4.1.2 COHFACE

The COHFACE dataset[2] consists of 40 subjects (12 females and 28 males) sitting still in front of a camera (see Fig. 4.1 for a sample video frame). With each subject, two 60 seconds long videos for two different lighting conditions are recorded. This gives a total of 160 one-minute long RGB video sequences).

The video sequences have been recorded with a Logitech HD C525. Physiological recordings, namely blood volume signal (BVS) and breathing rate have also been recorded. Physiological signals have been acquired using devices from Tought Technologies and using the provided BioGraph Infiniti software suite, version 5.

The videos are compressed in MPEG-4 Visual, i.e. MPEG-4 Part 2, bit rate ≈ 250 kb/s, resolution 640×480 pixels, 20 frames per second, which gets $\approx 5 \times 10^{-5}$ bits per pixel. In other words, the videos were heavily compressed and in the light of recent findings of McDuff et al. [62] the BVS is almost certainly corrupted.

4.1.3 PURE

The PURE dataset[3] [60] consists of 10 persons performing 6 different, controlled head motions in front of a camera (see Fig. 4.1 for a sample video frame). The video is captured by a professional grade camera with frame rate of 30 Hz and resolution 640×480 pixels. There are 8 male and 2 female subjects, each recording lasts 60 seconds. During the camera recordings, the BVS is recorded from a clip pulse oximeter. The oximiter delivers blood volume signal, heart rate and SpO2 readings.

The test subjects were placed 1.1 meters from the camera. The only source of light was a daylight coming from a large window frontal to the subject. The illumination conditions vary slightly for different videos due to weather conditions.

The subjects were asked to perform the following tasks: (i) *sit still* and look directly into the camera, (ii) *talk* while trying to avoid head motion, (iii) *move head slowly* parallel to the camera plane, (iv) *move head quickly*, (v) *rotate head a little*, (vi) *rotate head a lot*.

[1] https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf

[2] Available at https://www.idiap.ch/dataset/cohface.

[3] Available at http://www.tu-ilmenau.de/neurob/data-sets/pulse.

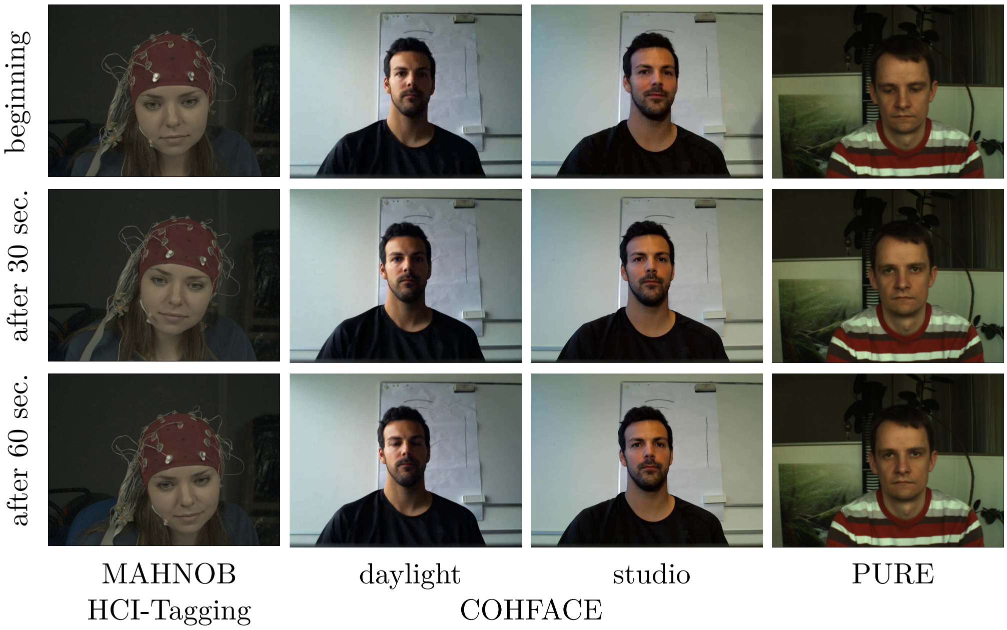


Figure 4.1: MAHNOB, COHFACE, and PURE datasets – frames from selected videos at the beginning of the sequence, after 30 and after 60 seconds.

Video frames in the PURE dataset are stored separately in PNG image files. Unfortunately, this dataset only contains ground truth in form of the BVS and SpO2 readings captured from the clip pulse oximeter.

Unlike the two previously described datasets, here the signal from the camera was stored in lossless PNG files. A frame 640×480 pixels is ≈ 390kB, which gets ≈ 10 bits per pixel. That is 2×10^5 times more than COHFACE and $\approx 6 \times 10^4$ times more than MAHNOB.

4.1.4 ECG-Fitness

We collected a realistic corpus of subjects performing physical activities on fitness machines. The dataset includes 17 subjects (14 male, 3 female) performing 4 different tasks (speaking, rowing, exercising on a stationary bike and on an elliptical trainer, see Fig. 4.3 and Fig. 4.4) captured by two RGB Logitech C920 web cameras and a FLIR thermal camera (see Fig. 4.2 for the capture setup). The FLIR camera was not used in the current study. The subjects were informed about the purpose of the research and signed an informed consent.

One Logitech camera was attached to the currently used fitness machine, the other was positioned on a tripod as close to the first camera as possible. Three lighting setups were used: (i) natural light coming from a nearby window, (ii) a standard 400W halogen light and (iii) a 50W led light source composed of 20W and 30W light (COB CN LED-FT-20W, COB CN LED-FT-30W). The artificial light sources were positioned to bounce off the walls and illuminate the subject indirectly.

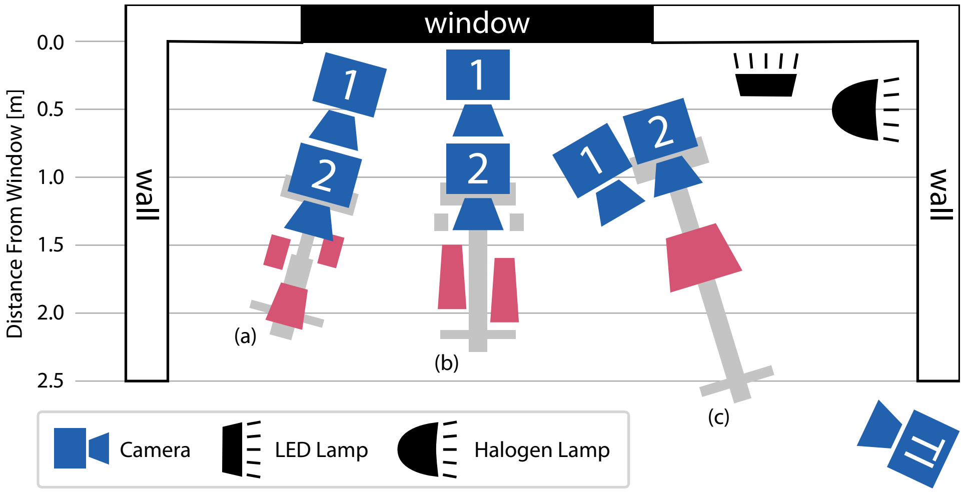


Figure 4.2: ECG-Fitness dataset – camera and illumination setup: (a) stationary bike, (b) elliptical trainer, (c) rowing machine. RGB camera 1 and thermal imaging camera TI were placed on a tripod, RGB camera 2 was attached to the fitness machine. A standard 400W halogen lamp and a 50W led light source composed of 20W and 30W light were used.

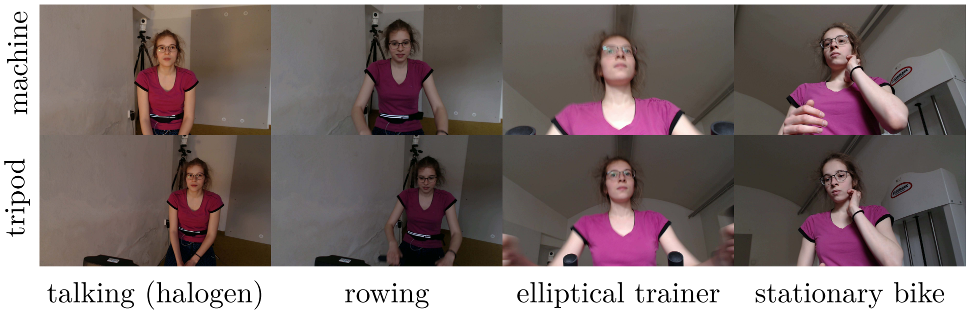


Figure 4.3: ECG-Fitness dataset. First row: camera attached to the currently used fitness machine (camera 2 in Fig. 4.2), second row: camera placed separately on a tripod.

Two activities (speaking and rowing) were performed twice – once with the halogen lighting resulting in a strong 50 Hz temporal interference, once without. In case of 4 subjects, an LED light was used during the recording of all activities. In total 204 videos from the web cameras, 1 minute each, were recorded with 30 fps, 1920 × 1080 pixels and stored in an uncompressed YUV planar pixel format. The age range of subjects is 20 to 53 years. During the video capture, an electrocardiogram was recorded with two-lead Viatom CheckMe$^{\text{TM}}$Pro device with the CC$_5$ lead. The ground-truth HR was computed with a Python implementation of Pan-Tomkins algorithm [63]. The lowest measured HR is 56, the highest 159 – a 10 second sequence was used for the computations. The mean HR is 108.96, the standard deviation 23.33 beats per minute.

The dataset covers the following challenges: (i) large subject's motion (possibly periodic) in all three axes, (ii) rapid motions inducing motion blur,

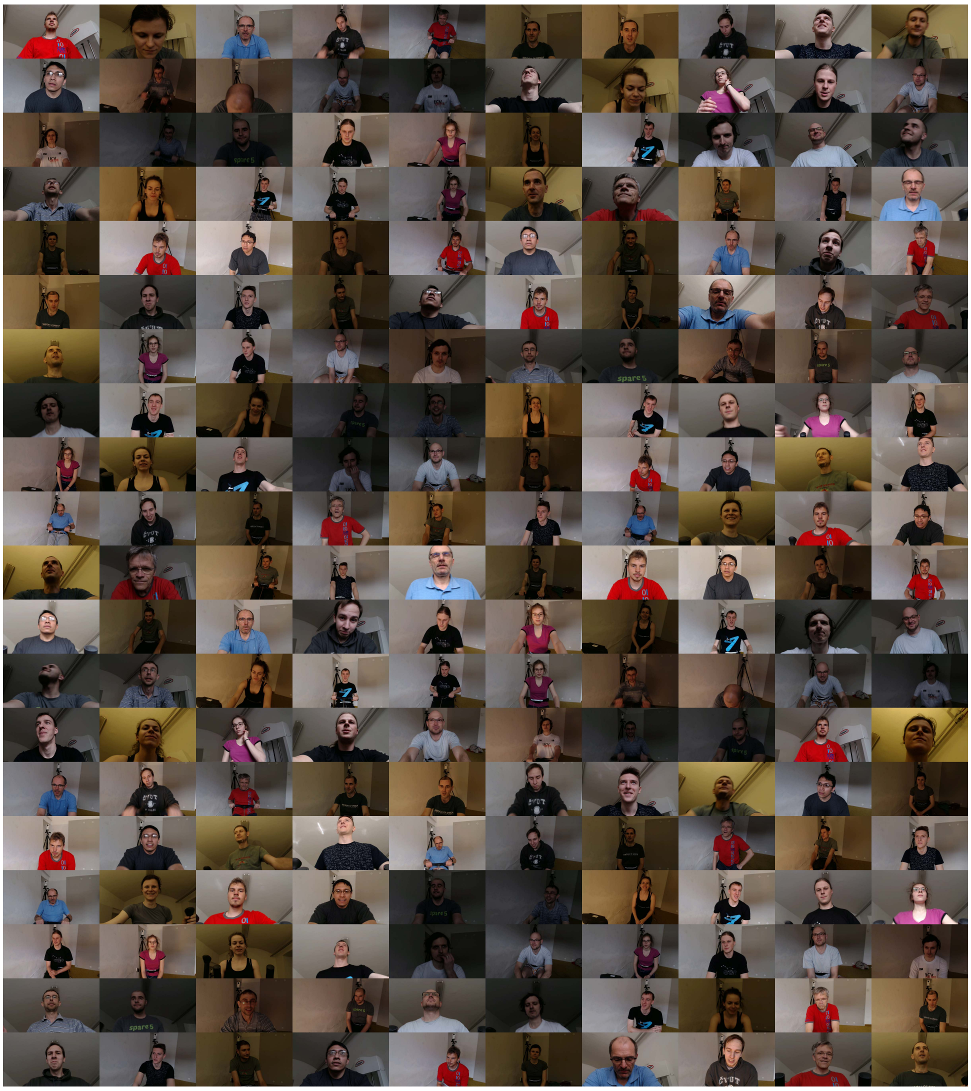


Figure 4.4: ECG-Fitness dataset. Overview of the pose and illumination variability present in the ECG-Fitness dataset.

(iii) strong facial expressions, (iv) wearing glasses, (v) non-uniform lighting, (vi) light interference, (vii) atypical non-frontal camera angles.

For the purpose of the dataset creation, a custom capture program was developed in the C++ programming language. The two C920 web-cameras were controlled by the OpenCV library[4]. Before the capture, the exposition settings (shutter speed, ISO and aperture) were set manually and were frozen during the capture. The cameras were focused manually. The FLIR thermal camera uses an analogue PAL color encoding system, therefore the Blackmagic Design Intensity Shuttle frame grabber was used to capture the analogue thermal images. The grabber was controlled through a provided software development kit.

[4] Available at https://opencv.org/.

4.2 Preliminary Experiments

Before the HR-CNN method was developed, effects of video compression and precise face registration were examined. In the experiments, a simplistic BVS extraction method was used – the signal was computed by spatial averaging over the green channel of regions shown in Fig. 4.6 (a). The experiments show that the quality of the BVS is adversely affected by the compression and improved by a precise face registration.

4.2.1 Preliminary Experiments Dataset

Preliminary experiments were performed with 5 volunteers (4 male, 1 female) aged 22 to 30 years, all with Fitzpatrick skin type III. The subjects were informed about the purpose of the research and signed an informed consent. 60 seconds long videos were captured in 1920 × 1080 @ 29.97 fps to an uncompressed YUV420 format, AVI container, by Logitech C920 web camera with a hardware chromatic subsampling 4:2:0. A single video size was approx. 7 GB. A BVS tPPG signal from the right index finger and an electrocardiograph signal was recorded by a clinically certified two-electrode Viatom CheckMe$^{\mathrm{TM}}$Pro. Clinically certified pulse oximeter Beurer PO 80 was used to record a BVS tPPG signal from the left index finger. Both devices were synchronized with the camera. In case of four subjects, the light source was an overcast light coming from a nearby window. In case of one subject, the light source was an indirect light coming from a standard 500W halogen light.

Two videos were recorded for each subject. In the first, four photogrammetric markers were attached to the subject's forehead and the subject was asked to sit calmly, see Fig. 4.6 (a). In the latter, the subject's head was stabilized in a custom made frame, see Fig. 4.5 (a), and the subject was asked to turn the palms to the camera.

To quantitatively asses the strength of the reconstructed signal, we employ a signal-to-noise ratio (3.2). Here, $\mathcal{F} = \{\frac{40}{60}, \frac{41}{60}, \ldots \frac{240}{60}\}$, $\mathcal{F}^+ = (f^* - \frac{9}{60}, f^* + \frac{9}{60})$ and f^* is the median of heart rates (measured in hertz) computed from the peak-to-peak distances from a pulse oximeter signal. Before the SNR is computed, the signal is weighted by the Hann window over the entire sequence to mitigate boundary effects.

4.2.2 Precise Face Registration

In this experiment, we examine the extent to which a precise registration affects the SNR of the BVS.

An influence of the precise tracking and registration on the quality of a BVS is inspected. A video stabilized by pixel-to-pixel registration is compared to a non-stabilized case. Videos with subjects having four photogrammetric markers attached to their foreheads were used (see Fig. 4.6 (a)). The stickers were manually set as interest points in the first frame, the reference frame,

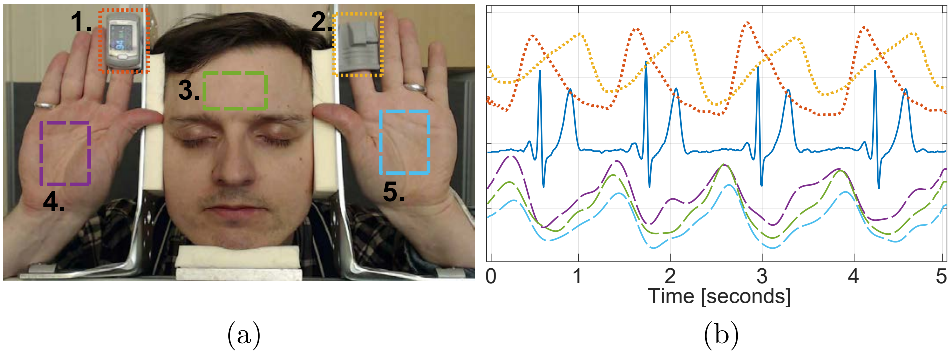


Figure 4.5: Precise face registration experiment: (a) experimental setup with a head fixed in a custom made stabilization frame, (b) 5 seconds of reference electrocardiogram – navy blue, distinguishable by the QRS-complex, blood volume signal measured by contact transmittance PPG – pulse oximeters on the left and right index fingers (color-coded like areas 1. and 2.), and by non-contact reflectance PPG – average over areas 3., 4., and 5. captured by a video camera. The blood volume signal is high and low pass filtered, and amplified 500 times. Arbitrary units.

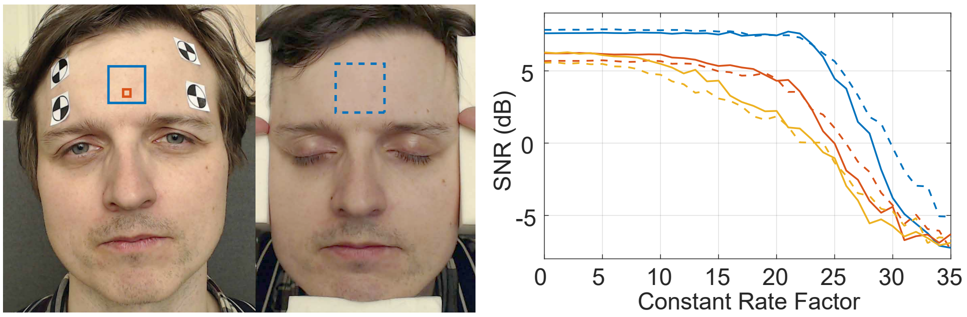
and were tracked with a MATLAB implementation of Lukas-Kanade tracker. A homography in each of the remaining frames was found between the reference and the tracked points. The homographies were then used to register the pixels of the forehead over frames. A linear interpolation was used. Two rectangular areas of measurement (ROI) were examined: 15×15 and 75×75 pixels. Both were positioned at the first frame of the video and the BVS was calculated by spatial averaging over a ROI in a green channel of every video frame.

A power spectral density of the BVS for the subject number five is shown in Fig. 4.7 (a). In both cases the heart rate frequency is clearly visible. Without the registration, we observe false frequencies with significant energy, while in the registered case the energy of these frequencies is reduced.

The results for all subjects are presented in Fig. 4.7 (b). After the registration, the SNR improves in all cases. The experiment suggests that a slight movement does not corrupt only NrPPG imaging as discussed in Sec. 2.6 but that it also corrupts the BVS. The corruption is probably caused by a combination of small ROI size and uneven texture of a skin. The smaller the ROI, the stronger the influence of the imperfections present on the skin surface. If an average over a small ROI is computed, the fluctuations of the image intensity, caused by the moving texture, produce a false signal. In a larger ROI, the fluctuations are averaged out. Note that the low SNR in case of subject #2 and #5 is caused by the low power of the heart rate frequency.

4.2.3 Video Compression

In this section, we first discuss specifics of works that use videos as a container for the captured data. Then an experiment showing how a video compression



(a) (b)

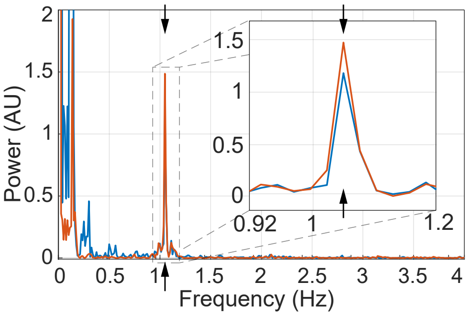
Figure 4.6: Precise face registration and compression experiment: (a) regions used in the experiments; solid blue – 75 × 75 px, solid orange – 15 × 15 px, dashed blue – 100 × 100 px, (b) blood volume signal-to-noise ratio as a function of video compression level defined by the Constant Rate Factor; average for 5 subjects. Results for 60 second videos with resolution 1920 × 1080 pixels (blue), downscaled to 878 × 494 pixels (orange) and to 434 × 234 pixels (yellow). Dashed lines – results with tracking stickers, full lines – with stabilized head, see Fig. 4.6 (a) left and right respectively.

affects the quality of the reconstructed BVS is presented.

Discussion

Surprisingly, many published studies fail to describe the dataset used to perform the experiments. We believe that this failure comes from the fact that the researches are not completely familiar with details of storing the captured data in a video file.

A common denominator of the NrPPG studies is that they report the captured data as being stored with 3 × 8 bits in some kind of video format. Without specifying that the video signal was not compressed, this information is useless. Let us explain why. In [64] we read that the videos "were recorded in 24-bit RGB (with 8 bits per channel)", 25 frames per second. Also, a capturing device is introduced – a Handycam Camcorder (Sony HDR-PJ580V) with resolution of 1440 × 1080 pixels. However, this particular camcoder records the videos (at the best) in the MPEG-4 AVC/H.264 format with a bitrate up to 24 Mbps. MPEG-4 AVC/H.264 is a block-oriented motion-compensation-based video compression standard. This standard permits to employ several kinds of compression principles including inter frame compression. This particular compression method stores the frames as expressed in terms of one or more neighboring frames. In other words, there is an image at the beginning and at the end of some sequence. The images in between are reconstructed from the two images. In between, only data needed for the reconstruction are stored, not the whole images. Now, how much is 3 × 8 bits? In case of [64], we can record up to 25 Mbps information per second. With 25 frames per second, we have 1 Mb per frame, and inside a frame, we have 1440x1080 pixels. $1000000/(1440 \times 1080) \approx 0.64$, i.e. we ended up with 3×8 bits ≈ 0.64. In [65], the camera used is Sony XDR-XR500 recording in H.264, 1920 × 1080



(a)

subject #	1	2	3	4	5
15 × 15 px ROI					
not registered	1.70	-6.17	-2.74	1.88	-7.08
registered	5.47	-6.03	2.60	2.99	-6.42
75 × 75 px ROI					
not registered	8.88	-5.42	6.77	6.80	-6.59
registered	9.15	-5.34	7.39	7.52	-5.68

(b)

Figure 4.7: Precise face registration experiment: (a) power spectral density of the blood volume signal before (blue) and after (orange) registration for subject #1; the signal computed by spatial averaging over a ROI of size 15 × 15 pixels in a green channel of a video; the true heart rate is marked by an arrow, (b) signal-to-noise ratio in decibels of the blood volume signal for 5 subjects. The signal is computed by spatial averaging over the green channel of regions shown in Fig. 4.6 (a). Results before and after registration of the regions.

pixels, 29.97 frames per second, and a bitrate up to 16 Mbps, i.e. the situation is even worse.

In [37] we read that the videos “were further compressed in mp4 format”. A clear distinction between a compression and a format must be made. When one speaks about a video format, a video container is actually thought. A *container* or wrapper format is a metafile format specifying how different elements of data coexist in a particular file. A container does not describe how the data are encoded. So, there is no such thing as “mp4 format compression”.

The influence of the compression on the quality of the BVS reconstruction was examined by McDuff et al. in [62]. The experiments were performed on combinations of different compression algorithms with different motion tasks. The two tested compression standards were H.264 and H.265. The videos were compressed with a different constant rate factor (CRF), a setting for which we cannot find a more precise description other than that it “control[s] the adaptive quantization parameter to provide constant video quality across frames” [62]. It is not a surprise we can't find a better description. The crucial information here is that both H.264 and H.265 are *standards*. In other words, H.264 is a sum of instructions on how to encode a video so an arbitrary H.264 decoder can process it. The standard only specifies a structure of the compressed stream and does not tell anything about its content, i.e. quality. That is a domain of an encoder and since encoders are implementation specific, we have no guarantees of quality at all. McDuff et al. solve this by using a particular publicly available freeware implementation of the H.264 and H.265 standards, x264 and x265. But the message is clear – if a BVS reconstruction is pursued, only none or lossless compression is safe.

McDuff et al. also mentioned the chroma subsampling, i.e. a method of reducing the number of samples used to represent the chromacity. Although the chroma subsampling may be used to represent an amount of color information loss in a standard compression scheme, we would like to emphasize its role in the design of the capturing devices. The most common way of capturing an uncompressed signal is with use of a web camera. However, even if stored in a raw format, still the "quality" of the signal might vary for different capturing devices. The web cameras typically perform the chroma subsampling already on the hardware level, before the captured data is sent to the USB port. Therefore, we also find important to always include an information whether there was any kind of "hardware" chroma subsampling present for a particular capturing device.

Experiment

In the experiment, effects of a video compression on the strength of the recovered BVS are inspected. Every video file in the dataset was compressed with a constant rate factor (CRF) setting varied from 0 to 35. Usually, the CRF is explained as a setting that induces "constant video quality", as opposed to the constant bit rate. CRF set to 0 means that a lossless compression is performed. FFmpeg program (version 2.8.11) was used to compress the videos with an x264 encoder, a publicly available implementation of H.264 standard. The default CRF setting in x264 is 23. The BVS was obtained by spatial averaging over a ROI of size 100 × 100 pixels (see Fig. 4.6 (a)) in a green channel of a video. The videos with tracking markers were stabilized first (as described in Sec. 4.2.2).

Results are shown in Fig. 4.6 (b). Originally, only experiments with the full resolution videos, i.e. 1920 × 1080 pixels (Full HD), were used. However, we did not experience the gradual decrease of the SNR reported by McDuff et al. who used videos with resolution 658 × 492 pixels. Therefore we performed the experiment also with videos downscaled to 878 × 494 and 434 × 234 pixels. Bi-cubic interpolation was used. The ROIs were scaled proportionally. Here the gradual SNR decrease is visible (see Fig. 4.6 (b)). Note that downscaling the video also lowered the SNR, and in case of the Full HD videos, the SNR remained high until CRF 23. We conclude that reducing the video resolution negatively affects the SNR of the recovered BVS. Furthermore, steeper SNR loss is experienced when the H.264 compression is applied to the videos with a reduced resolution.

Next, we discuss results of Blackford and Estepp [46] who performed a similar experiment – they reduced the resolution of videos from 658 × 492 to 329 × 246 pixels and concluded that there was "little observable difference in mean absolute error" between the two reconstructed BVS. We identify four reasons why their conclusion differs from the results reported by us. First, independent component analysis, a powerful blind source separation (BSS) method, was used to obtain the BVS. We argue, that use of BSS methods in clinical application is not desirable (see Sec. 2.4). Second, the ICA was computed with signals from five industry grade cameras that were part of a 9 camera array, each camera capturing images with resolution 658 × 492 pixels. An array of high quality cameras loses the benefits of NrPPG approaches built on cheap capturing devices. Third, a whole image, not a ROI, was

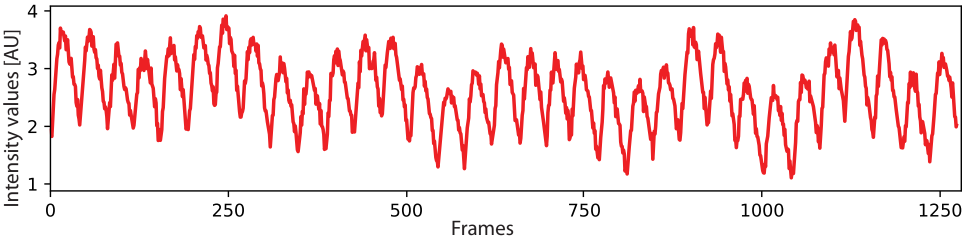


Figure 4.8: The *Extractor* output for 1270 facial images of a calmly sitting subject #8 from the PURE dataset. Intensity values in arbitrary units.

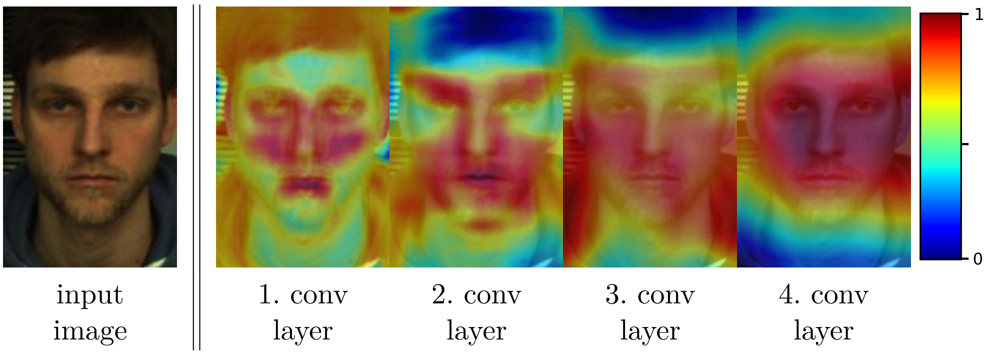


Figure 4.9: The *Extractor* introspection. Sequence of Grad-CAM heatmaps of convolutional layers through the Extractor network (from the earliest 1. convolutional layer on the left to the latest 4. on the right). The activations in cheek and lips areas contribute to the output the most.

used in their approach. Fourth, the blood volume signals recovered after the downscaling were evaluated against the full sized videos with a mean absolute error, not with a SNR.

An approx. 16 dB difference in the SNR reported by us and by McDuff et al. remains to be explained. First, McDuff et al. use the same experimental setup and approach as Blackford and Estepp [46]. Second, they compute the SNR with a different, unspecified formula. Third, we compute the BVS by spatial averaging over a ROI from the green channel of a single camera, they compute by applying ICA on spatial averages of the whole images for red, green and blue channels from 5 cameras.

4.3 Interpretation of HR-CNN

To provide an interpretation of what the CNNs have actually learned, we present two insights. First, we give a "visual explanation" of the *Extractor* network based on the Grad-CAM method [66] (Fig. 4.9) adapted to our settings. Then a plot of the true and an estimated HR for a sequence with a rapid HR change is presented.

4.3.1 Extractor

Given a convolutional layer with a filter k, we compute activations A_{ij}^k of each neuron ij and derivatives $\frac{\partial y}{\partial A_{ij}^k}$ of the output y with respect to the activations. Importance weights read

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y}{\partial A_{ij}^k}, \tag{4.1}$$

where Z is the number of neurons in a given feature map. Weight α_k captures the "importance" of the feature map k for the output y. The result is a coarse heat-map that is computed as a linear combination

$$L_{Grad-CAM} = \sum_k \alpha_k A^k. \tag{4.2}$$

The heatmap is resized and laid over an input image (see Fig. 4.9).

Sequence of heatmaps L provides a clue about the *Extractor*'s function. In our case, the first layer (left) "focuses" on cheeks and lips, the next one increases the importance of cheeks and reduces importance of hair, and this trend follows in the next two layers. The fact that the extractor "focuses" on the lips was surprising. We inspected lips as a possible source of the BVS during the preliminary experiments but we were not able to obtain stable results. However, we did not track the lips and in this case it is the segmenting ability of the *Extractor* network that makes the difference.

4.3.2 HR Estimator

To inspect the behavior of the *HR Estimator*, a plot of the ground truth HR and the estimated HR for a sequence with a significant change of HR (Fig. 4.10) is presented. The plot shows the true HR and an estimated one for a "rowing" activity of the subject #0 from the ECG-Fitness dataset. The camera attached to the rowing machine was used – strong vibrations of the machine are clearly visible in the video. Both the true and estimated HR were computed from 10 second windows at 1 second intervals. The predicted HR follows the ascending trend of the true HR. Around the frame number 1500, the estimated HR deviates from the true HR for several tens of frames. Visual inspection of the video revealed that the subject shows strong facial expressions reflecting the difficulty of the rowing activity. However, the subject shows facial expressions in the whole video to some extent so the nature of the deviation might be of a different kind.

4.4 Evaluation

In this section, the introspection of the HR-CNN method is given along with a comparative study.

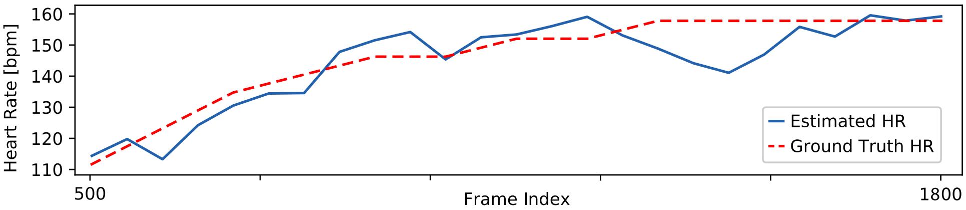


Figure 4.10: The *HR Estimator* introspection. Output of the *HR Estimator* for a video with a significant increase of the subject's HR. The estimated (blue solid) and the true HR (red dashed) computed from a 10 second window at a 1 sec. time interval.

An open source Python package `bob.rppg.base`[5] provided by Heusch et al. [7] was used for the computations. The same error metrics reflecting discrepancy between the true and the predicted HR were used. In particular, the root mean square error (RMSE) and the Pearson's correlation coefficient, were used as in [7]. In addition, mean absolute error (MAE) was computed. We test the developed HR-CNN method on four datasets (standard: COHFACE, MAHNOB, PURE, newly collected: ECG-Fitness) against three baseline methods (LiCVPR [5], CHROM [12], 2SR [16]). In addition, we compare the methods against a naive "baseline". The baseline always outputs a constant HR – the average HR of the training set.

For the purpose of the face bounding boxes detection in case of the PURE and ECG-Fitness datasets, a commercial implementation of *WaldBoost* [67] based detector[6] was used. Bounding boxes for the MAHNOB and COHFACE datasets were provided by the `bob.rppg.base` package.

Experimental protocol. Inspired by Heusch et al. [7], we define an experimental protocol for evaluating visual HR estimation methods. The protocol prescribes that the visual HR estimation method: (i) receives a sequence of facial images and outputs a single number (an estimated HR), any other output is considered to be invalid, (ii) is permitted to learn its parameters on the training set, (iii) is evaluated on the test set with the Pearson's correlation coefficient, the mean absolute error, the root mean squared error, and the percentage of videos with a successful HR estimation.

We adopt the training and test split for COHFACE and MAHNOB datasets defined in the "all" experiment performed by Heusch et al. [7] in the *bob.rppg.base* Python package. On the following pages, several experiments are presented. The splits for the PURE and ECG-Fitness datasets were performed randomly. The splits were made "subject-wise" – all videos of a particular subject were either in the training set or in the test set. In case of the ECG-Fitness dataset, "activity" subsets of the original set were created containing videos of a particular activity. Again, the splits were made "subject-wise" and also "protocol-wise" – once a video was assigned to the training set in one experiment, it never appears in the test set of any other

[5] https://gitlab.idiap.ch/bob/bob.rppg.base

[6] Eyedea Recognition Ltd. `http://www.eyedea.cz/`.

experiment. In all presented experiments, the parameters of each method were trained on the training set of the particular dataset. The testing was done on previously unseen data – it used to be a common practice in the community to tune parameters of the methods directly on the test sets.

Parameter tuning. Parameters of the LiCVPR, CHROM and 2SR methods needed to be tuned for each training set (of each dataset or its "experiment subset"). LiCVPR has 12 parameters, CHROM 6 and 2SR 4. The range of the parameter space is unknown and no learning procedures were given by the authors. Tuning then becomes an unpleasant an difficult task. Exhaustive evaluation of even a sparsely sampled parameter space is virtually impossible. Fortunately, in case of the COHFACE and MAHNOB datasets, Heusch et al. [7] provide the best-result-yielding parameters for all three methods. We tried our best to find the right parametrization in case of the ECG-Fitness and PURE datasets. We followed a strategy applied by Heusch et al. – we first optimized the parameters of the first step of the signal processing pipeline, fixed them, and kept on with the optimization of the parameters of the second step, and so on. Obviously, failure to find the right parametrization would lead to an unfair comparison. However, such failure is an inherent part of the processing pipeline of the three methods.

Note that in case of the ECG-Fitness dataset and the HR-CNN method, the *Extractor* and *HR Estimator* were trained on the dataset by alternating optimization – in every iteration, the parameters of one network were fixed, the other network was minimizing the MAE on the training set. In the next iteration, the roles of the networks switched. The network tuple yielding the lowest validation MAE was selected. A limited time schedule did not allow us to apply the alternating optimization in case of other datasets.

4.4.1 HR Estimator Variants

As pointed out earlier (see Sec. 3.2.1), our first experiments were conducted on a non-challenging PURE dataset. The video sequences in this dataset are uncompressed and the subjects perform a little to none movement in a controlled fashion (see Fig. 4.1). In this case, a simple argument maximum in the PSD (3.1) of the Extractor's output (3.5) gives MAE less than 3.

Tab. 4.2 shows results of five different HR estimation methods. The input of these methods is the signal coming from the *Extractor* network. First four lines for each of the measures (Pearson's corr. coeff., MAE and RMSE) show results of different types of estimation with the *HR Estimator* network. The first line shows the situation where the *HR Estimator* is fed by 10 second windows evaluated at 10 second intervals. The estimated HR is compared to the true HR computed at the corresponding 10 second window. Note, that this only applies for the ECG-Fitness dataset. In case of the other datasets, the results from the 10 second windows are compared to the true HR of the whole sequence. Results in the following two lines represent mean and median of the *HR Estimator*'s results for non-overlapping 10 second windows evaluated at 10 second intervals. The estimated HR is compared to the true

		SIZE OF HR estimation / ground truth HR WINDOW	COHFACE	ECG-Fitness	MAHNOB	PURE	PURE MPEG-4 Visual
Pearson's corr. coeff.	HR Estimator						
	+ ———	10 s. / 10 s.	0.15	0.80	0.44	0.89	0.44
	+ average	10 s. / whole	0.26	0.86 ①	0.52 ①	0.98 ②	0.70 ①
	+ median	10 s. / whole	0.30 ①	0.84 ②	0.52 ①	0.98 ②	0.65
	+ ———	whole / whole	0.29 ②	0.82	0.51	0.98 ②	0.70 ①
	$\arg\max_f \mathrm{PSD}(f)$	whole / whole	−0.21	0.10	−0.04	0.99 ①	0.43
MAE	HR Estimator						
	+ ———	10 s. / 10 s.	11.17	8.65	7.32 ②	4.55	14.01
	+ average	10 s. / whole	8.39	8.28 ①	7.38	1.97 ②	9.01
	+ median	10 s. / whole	8.24 ②	8.63 ②	7.40	2.39	9.97
	+ ———	whole / whole	8.10 ①	9.46	7.26 ①	1.84 ①	8.72 ②
	$\arg\max_f \mathrm{PSD}(f)$	whole / whole	21.38	46.33	23.32	2.00	6.15 ①
RMSE	HR Estimator						
	+ ———	10 s. / 10 s.	14.27	11.80	9.51	6.74	17.43
	+ average	10 s. / whole	10.98 ②	10.59 ①	9.37 ②	2.79	11.08 ②
	+ median	10 s. / whole	11.08	11.02 ②	9.39	3.43	11.62
	+ ———	whole / whole	10.78 ①	11.77	9.24 ①	2.37 ①	11.00 ①
	$\arg\max_f \mathrm{PSD}(f)$	whole / whole	26.80	55.49	28.67	2.50 ②	12.85

Table 4.2: HR-CNN evaluation – HR estimation variants. HR estimation is performed with the *HR Estimator* and by the argument maximum in the power spectral density (PSD) of the blood volume signal (3.1). The estimation is made either from 10 second windows at 10 second intervals or on a whole sequence. The same applies for the ground truth HR. When evaluating estimation from 10 second windows against the ground truth HR of a whole sequence, results for median and average of the estimated heart rates are presented. Pearson's correlation coefficient, mean average error and root-mean-square error is computed on the test sets of the datasets.

HR of the whole sequence. Next, the input to the *HR Estimator* is the whole output of the *Extractor* and the estimated HR is compared to the true HR of the whole video sequence. The last presented HR estimation approach is the argument maximum in the PSD (3.1) of the *Extractor*'s output.

The results show: (i) Pearson's correlation coefficient is generally high in all cases when the *HR Estimator* network is used, no matter which estimation procedure is used. This also holds for the case of the uncompressed PURE dataset. As mentioned before, this dataset is not challenging and output of the *Extractor* (see Fig. 4.8) strongly resembles a sine wave, therefore the 0.99 Pearson's correlation coefficient for the argument maximum. (ii) MAE is the best in case of the *HR Estimator* approaches, but the argument maximum yields the best results in case of the compressed PURE dataset. This is hard to interpret since the effects of video compression on the quality of the extracted signal are severe. (iii) RMSE is the best in case of the *HR Estimator*. (iv) The argument maximum yields low MAE, RMSE and high Pearson's correlation coefficient when the dataset is not challenging. Video compression

		COHFACE	ECG-Fitness	MAHNOB	PURE	PURE MPEG-4 Visual
Pearson's corr. coeff.	baseline	—	—	—	—	—
	2SR	−0.32	0.06	0.06	0.98 ②	0.43
	CHROM	0.26 ②	0.33 ②	0.21	0.99 ①	0.55 ②
	LiCVPR	−0.44	−0.58	0.45 ②	−0.38	−0.42
	HR-CNN	0.29 ①	0.82 ①	0.51 ①	0.98	0.70 ①
MAE	baseline	8.98	17.35 ②	9.19	9.29	9.29
	2SR	20.98	43.66	17.37	2.44	5.78 ①
	CHROM	7.80 ①	21.37	13.49	2.07 ②	6.29 ②
	LiCVPR	19.98	31.90	7.41 ②	28.22	28.39
	HR-CNN	8.10 ②	9.46 ①	7.26 ①	1.84 ①	8.72
RMSE	baseline	10.19 ①	21.60 ②	11.39	11.67	11.67
	2SR	25.84	52.86	26.81	3.06	12.81
	CHROM	12.45	33.47	22.36	2.50 ②	11.36 ②
	LiCVPR	25.59	45.30	10.21 ②	30.96	31.10
	HR-CNN	10.78 ②	11.77 ①	9.24 ①	2.37 ①	11.00 ①

Table 4.3: Evaluation of visual HR methods – Experiment "All" (see Sec. 4.4.2). Pearson's correlation coefficient, mean average error and root-mean-square error on test sets of the datasets for four baseline methods and the developed HR-CNN.

significantly decreases the performance of the argument maximum even when the subject's HR is stationary as it is the case with the COHFACE and MAHNOB datasets. In cases when the HR changes rapidly during the video recording, the argument maximum is not suitable for predicting the average frequency – there is usually no dominant peak in the PSD spectrum of the signal.

4.4.2 Experiment "All"

In this section, a large-scale comparative study is presented. Tab. 4.3 contains the central result of the evaluation.

Since this is the first time the results of all compared methods on all available datasets are presented, we make the discussion "dataset-wise" and report the sizes of the training and test sets for this experiment.

MAHNOB HCI-Tagging

The training set consists of 2302 sequences with an average length of 1812 frames. The test set contains 1188 sequences with an average length of 1745 frames.

The results are presented in Tab. 4.3. The HR-CNN clearly dominates over the other methods. This is true even for LiCVPR that was developed directly on the MAHNOB dataset. Interestingly, Li et al. [5] reports Pearson's correlation coefficient of 0.81, but neither we nor Heusch et al. were able to reproduce the result. The reason is probably the unknown parameter setting of the signal extraction pipeline. In the dataset, the most informative area for

HR estimation is the lower part of a face. The subjects in the dataset wear an electroencephalographic caps that either cover the forehead completely or force hair in the forehead's direction. Also, the cap's color is very similar to the skin's tone. With these limitations, the selection of a measuring area is less or more given – LiCVPR estimates HR only from the lower part of the face. Also, subjects in the dataset rarely move. If a subject moves, LiCVPR removes such sub-sequence as not suitable for the estimation since it contains a "non-rigid motion".

COHFACE

The COHFACE training set contains 24 subjects, the test set 32 subjects.

The dataset contains the most compressed videos. The results are presented in Tab. 4.3. CHROM method yields the best MAE for the test set and HR-CNN performs the best in all other cases. 2SR and LiCVPR perform significantly worse. CHROM and HR-CNN methods use the whole input sequence to reconstruct the BVS and estimate the HR, while the other aggregate local estimates. The first approach seems to best account for the heavy compressed COHFACE videos.

PURE

The PURE training set contains 36 videos of 6 subjects, the test set 24 videos of 4 subjects.

The results are depicted in Tab. 4.3. Surprisingly, MAE on the test set is less than 3 in case of three methods out of four. Poor results of LiCVPR are probably caused by the fact that unlike in the MAHNOB dataset, the subjects in the PURE dataset were asked to perform various head movements in two tasks and to talk in one task. Also, a different video compression method was used. We believe that the main reason behind the good prediction accuracy of the methods is the fact that the PURE dataset is *not compressed*. To confirm our hypothesis, we decided to perform another experiment.

The PURE dataset was compressed with the same compression method and to the same average bit rate as videos from the COHFACE dataset. The results shown in Tab. 4.3 confirm our hypothesis. A drop of the accuracy is visible in the table in case of three methods.

ECG-Fitness

There is 72 videos of 12 subjects in the training set and 24 videos of 4 subjects in the test set of the ECG-Fitness dataset. Videos from both cameras (one positioned on a tripod and the other attached to the currently used fitness machine) were used.

The results presented in Tab. 4.3 show that our method is the most robust one when a strong motion and heavy light interference is present in the videos (see Fig. 4.11 for an example of facial images from the ECG-Fitness dataset used in the experiments). Due to the rapid movement of subjects,

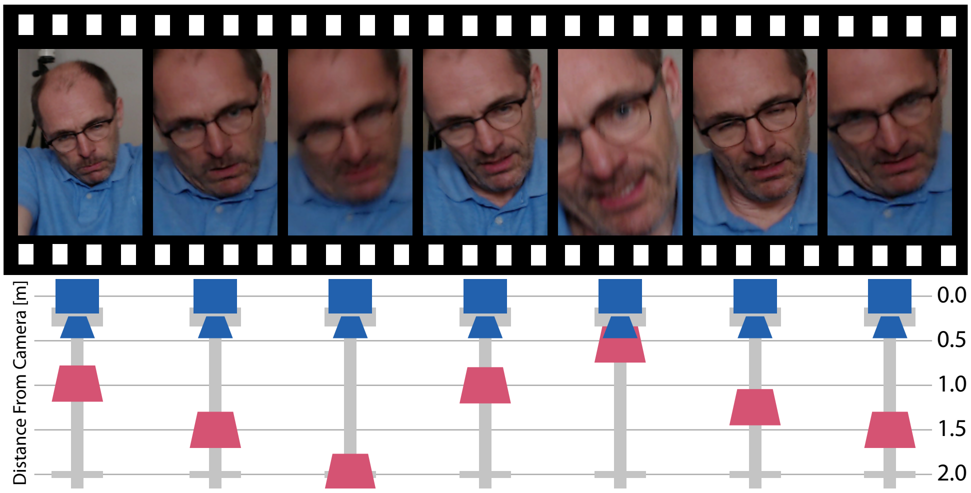


Figure 4.11: Facial images extracted from a rowing session video of male subject from the ECG-Fitness dataset. Bellow, the subject's position (pink) with respect to the camera (blue) is depicted.

the face bounding boxes were not found in videos in all frames. In that cases, the last found bounding box was used. Visual inspection of the extracted faces revealed a strong clutter. The clutter and motion blur are the reason why the LiCVPR and 2SR methods do not perform well. CHROM performs better, because it averages skin-colored pixels in each frame and then performs computations on the sequence as a whole.

4.4.3 Experiment "Single Activity"

The "single activity" experiment inspects robustness of the HR estimation methods to different amounts of motion present in the videos. The methods were trained on the training set of the experiment "all" and evaluated on the testing sets of each activity separately. Videos from both cameras, i.e. the one placed on a tripod and the one attached to the fitness machine, were used.

The results in Tab. 4.4 imply high robustness of the HR-CNN method to strong subject's motion (represented by videos from the "Rowing" activity). Considering MAE and RMSE, in all but one case HR-CNN yields the best or the second best results. Continuing with the discussion of HR-CNN, compared to the other methods, if a halogen light source was used in the "Talking" activity, the results improved. This has two explanations: (i) the method is robust to 50Hz perturbation, and (ii) the method performs better in good illumination conditions. By comparing the MAE and RMSE results of the "Rowing" activity with and without the halogen lamp light, it seems that we can't easily explain the effects of the halogen lamp light on the accuracy of the HR-CNN method. On the other hand, Pearson's corr. coeff. is better if the halogen lamp light was used. Still, more experiments would be needed to give a conclusion.

		Talking	Talking (Halogen)	Rowing	Rowing (Halogen)	Elliptical Trainer	Stationary Bike
% of success	2SR	100.0	100.0	100.0	100.0	80.0	80.0
	CHROM	100.0	100.0	100.0	100.0	80.0	80.0
	LiCVPR	80.0	80.0	12.0	10.0	20.0	30.0
Pearson's corr. coeff.	baseline	—	—	—	—	—	—
	2SR	0.47	0.28 ②	0.30 ②	−0.40	0.27	0.61 ②
	CHROM	0.98 ①	−0.90	0.22	0.67 ②	−0.19	0.45
	LiCVPR	−0.57	−0.66	—	—	1.00 ①	0.23
	HR-CNN	0.66 ②	0.94 ①	0.93 ①	0.95 ①	0.61 ②	0.62 ①
MAE	baseline	19.72	14.37 ②	16.13 ②	22.83 ②	17.78 ②	13.03 ②
	2SR	17.04	38.60	47.58	67.85	47.24	45.57
	CHROM	4.23 ①	19.77	17.59	30.45	36.66	21.94
	LiCVPR	28.76	23.16	68.81	116.37	53.71	19.56
	HR-CNN	15.57 ②	7.78 ①	3.94 ①	9.31 ①	8.61 ①	10.44 ①
RMSE	baseline	22.72	18.63 ②	19.15 ②	28.33 ②	19.50 ②	19.12 ②
	2SR	23.87	47.78	50.36	75.45	57.37	48.49
	CHROM	4.73 ①	38.21	25.21	38.83	49.43	27.34
	LiCVPR	38.06	39.30	68.81	116.37	58.08	28.08
	HR-CNN	17.29 ②	9.16 ①	5.02 ①	10.53 ①	10.25 ①	13.59 ①

Table 4.4: Evaluation of visual HR methods – Experiment "Single Activity". Percentage of videos with successful HR estimation, Pearson's correlation coefficient, mean average error and root-mean-square error on test sets of the ECG-Fitness dataset for four baseline methods and the developed HR-CNN.

CHROM method is the most accurate in case of the "Talking" activity. This might be accounted to the fact that in the talking videos, the least amount of motion was present. The only significant movement present was that of the lips. In case of other activities, CHROM performs significantly worse.

The results of the baseline estimator, i.e. predicting the HR by returning the average HR of video sequences from the training set, gives an important insight about the practicality of the HR estimation methods – the only activity in which the methods are significantly better is "Talking", in case of other activities the only method beating the average from the training set is the HR-CNN method.

4.4.4 Experiment "Single Activity – Retrained"

The "Single Activity – Retrained" experiment uses the same settings as the "Single Activity" experiment with one difference – the methods were retrained for each activity separately on its training set. That being said, the primary focus of this experiment is the ability of a particular method to adapt to a new environment with a limited amount of training samples. Due to the requirement of the CHROM, LiCVPR and 2SR methods to manually

		Talking	Talking (Halogen)	Rowing	Rowing (Halogen)	Elliptical Trainer	Stationary Bike
Pearson's corr. coeff.	RE-TRAINED						
	baseline	—	—	—	—	—	—
	HR-CNN	−0.46	−0.04	0.02	0.20	−0.22	0.52
	ALL-TRAINED						
	baseline	—	—	—	—	—	—
	HR-CNN	0.66	0.94	0.93	0.95	0.61	0.62
MAE	RE-TRAINED						
	baseline	20.27	20.63	11.38 ②	20.60	12.57 ②	14.60
	HR-CNN	21.48	17.59	20.53	20.30 ②	20.31	17.80
	ALL-TRAINED						
	baseline	19.72 ②	14.37 ②	16.13	22.83	17.78	13.03 ②
	HR-CNN	15.57 ①	7.78 ①	3.94 ①	9.31 ①	8.61 ①	10.44 ①
RMSE	RE-TRAINED						
	baseline	24.18	27.80	15.09 ②	23.63	14.91 ②	17.29 ②
	HR-CNN	25.24	26.27	25.10	23.02 ②	22.89	21.97
	ALL-TRAINED						
	baseline	22.72 ②	18.63 ②	19.15	28.33	19.50	19.12
	HR-CNN	17.29 ①	9.16 ①	5.02 ①	10.53 ①	10.25 ①	13.59 ①

Table 4.5: Evaluation of visual HR methods – Experiment "Single Activity – Retrained". Pearson's correlation coefficient, mean average error and root-mean-square error on the test sets of the ECG-Fitness dataset for the HR-CNN method and a baseline method.

tune their parameters for a given dataset, we did not include them in the experiment. Efforts to do so would be much higher than the possible profits. Hence, this experiment only compares the baseline method, i.e. returning the average HR of the training set for all testing sequences, and HR-CNN.

If we compare the results of the "all-trained" and "re-trained" methods in Tab. 4.5, we clearly see that HR-CNN is very sensitive to the size of the training set. This result follows our observations on the performance of the CNNs – when there was not enough training examples, we were not able to train the network to minimize the error on the validation set. There is not a single case in the experiment where the HR-CNN method would yield better results when retrained on a smaller training set.

Taking look at the baseline method, one would expect to see better results after retraining on a particular set but that is not the case for the "Talking" activity. This might be caused by the fact that we randomly permuted the sequence in which the subjects performed the activities. We did so to record a more diverse dataset. Hence, the subject's HR differs greatly in this activity.

	trained evaluated	COHFACE	ECG-Fitness	MAHNOB	PURE	PURE MPEG-4 Visual
Pearson's corr. coeff.	COHFACE	0.29 ①	0.32 ②	0.03	−0.01	0.32
	ECG-Fitness	0.09	0.50 ①	0.07 ②	0.06	−0.18
	MAHNOB	0.06	−0.13	0.51 ①	−0.21	−0.15
	PURE	0.13 ②	0.20	0.00	0.98 ①	0.59 ②
	PURE MPEG-4 Visual	0.07	0.29	0.00	0.88 ②	0.70 ①
MAE	COHFACE	8.10 ①	48.14	26.35	18.44	12.08
	ECG-Fitness	26.38	14.48 ①	11.14 ②	22.11	29.59
	MAHNOB	14.80	35.81 ②	7.26 ①	18.17	16.56
	PURE	9.82 ②	40.54	33.22	1.84 ①	8.33 ①
	PURE MPEG-4 Visual	12.04	44.24	35.63	9.72 ②	8.72 ②
RMSE	COHFACE	10.78 ①	51.78	28.37	20.66	14.28
	ECG-Fitness	28.63	19.15 ①	12.82 ②	25.71	32.54
	MAHNOB	17.40	41.96 ②	9.24 ①	20.74	19.12
	PURE	11.93 ②	46.14	35.53	2.37 ①	9.42 ①
	PURE MPEG-4 Visual	15.12	49.22	37.34	11.38 ②	11.00 ②

Table 4.6: HR-CNN evaluation – experiment "HR Estimator Cross-dataset". Pearson's correlation coefficient, mean average error and root-mean-square error on the test sets of the datasets evaluated in a cross-dataset setting by the developed HR-CNN. The *Extractor* network trained on the PURE dataset was used in all cases. The *HR Estimator* was evaluated in a cross-dataset setting.

4.4.5 Experiment "HR Estimator Cross-dataset"

The *HR Estimator* network was introduced to account for specifics of the recording setup of a particular dataset, i.e. various compression methods and different amounts of relative subject's and camera movement. It is thus interesting to inspect the *HR Estimators* trained for a particular recording setup on a different recording setup – in a cross-dataset setting. Note that in all cases the *Extrator* network trained on the PURE dataset was used. The estimators were trained on the training sets of the datasets from the experiment "all".

The results are presented in Tab. 4.6. Each column represents the *HR Estimator* trained on a particular dataset, e.g. the first column represents the estimator trained on the COHFACE dataset and each row contains its result on a particular dataset. All combinations of the "trained-on-dataset×evaluated-on-dataset" pairs were computed.

In the table, the expected pattern is visible. The best results for a particular dataset are received when the *HR Estimator* trained particularly for that dataset is used. However, there is an exception. The *HR Estimator* trained on the uncompressed PURE dataset gives better MAE and RMSE than the one trained directly on the compressed dataset. Even if the difference is small, still we would expect this not to be the case. On the other hand, this result implicates that it is better to train the model on a non-compressed dataset and then use it in a compressed setting and not the other way.

		Talking MACHINE	Talking TRIPOD	Talking (Halogen) MACHINE	Talking (Halogen) TRIPOD	Rowing MACHINE	Rowing TRIPOD
% of success	2SR	100.0	100.0	100.0	100.0	100.0	100.0
	CHROM	100.0	100.0	100.0	100.0	100.0	100.0
	LiCVPR	60.0	80.0	80.0	60.0	0.0	0.0
Pearson's corr. coeff.	baseline	—	—	—	—	—	—
	2SR	0.55	0.67	−0.40	0.80 ②	0.30	0.30 ②
	CHROM	0.98 ①	0.98 ①	−0.86	−0.95	0.78 ②	−0.36
	LiCVPR	0.48	0.93 ②	−0.29 ②	−0.26	—	—
	HR-CNN	0.63 ②	0.69	0.93 ①	0.95 ①	0.97 ①	0.96 ①
MAE	baseline	19.72	19.72	14.37 ②	14.37	16.13	16.13 ②
	2SR	22.80	11.27 ②	39.87	37.32	47.93	47.22
	CHROM	4.10 ①	4.36 ①	18.18	21.37	12.97 ②	22.21
	LiCVPR	23.41	19.01	14.42	8.92 ②	—	—
	HR-CNN	16.24 ②	14.90	8.38 ①	7.17 ①	2.88 ①	5.00 ①
RMSE	baseline	22.72	22.72	18.63 ②	18.63	19.15	19.15 ②
	2SR	29.98	15.52 ②	52.86	42.09	50.66	50.06
	CHROM	4.61 ①	4.84 ①	37.48	38.92	15.76 ②	31.98
	LiCVPR	28.75	22.82	22.06	9.04 ②	—	—
	HR-CNN	17.98 ②	16.58	9.56 ①	8.74 ①	3.95 ①	5.91 ①

Table 4.7: Evaluation of visual HR methods – experiment “Camera Vibration” Part 1. Percentage of videos with successful HR estimation, Pearson's correlation coefficient, mean average error and root-mean-square error on the test sets of the ECG-Fitness dataset for four baseline methods and the developed HR-CNN. The videos from the camera attached to the currently used fitness MACHINE and the camera attached to the TRIPOD were evaluated separately for each activity.

The next interesting part of this experiment is the second best *HR Estimator* and the amount of the difference between the first and the second best. Here we only discuss the MAE. We see three interesting findings. The first one was discussed in the previous paragraph. The second one is the second best estimator for the COHFACE dataset – the PURE trained estimator. How come that an estimator trained on an uncompressed dataset, such as the PURE dataset, also works for heavily compressed videos? We argue that in this case, the reason is that the subjects in the COHFACE dataset do not move which results in two things: (i) since there is no movement in the sequences, the video compression algorithm has much easier job reducing the output bitrate and the BVS measured by a camera might not be corrupted so heavily as when there is a significant movement in the video, (ii) HR estimation performed on still subjects is easier due to various reasons (as discussed in Sec. 4.2.2). We believe that the third surprise, i.e. the reasonable result of the ECG-Fitness-trained estimator on the MAHNOB dataset, is of the same kind as the second one – subjects in the MAHNOB dataset move a little or none at all.

		Rowing (Halogen) MACHINE	Rowing (Halogen) TRIPOD	Elliptical Trainer MACHINE	Elliptical Trainer TRIPOD	Stationary Bike MACHINE	Stationary Bike TRIPOD
% of success	2SR	100.0	100.0	80.0	80.0	80.0	80.0
	CHROM	100.0	100.0	80.0	80.0	80.0	80.0
	LiCVPR	0.0	0.0	0	20.0	20.0	40.0
Pearson's corr. coeff.	baseline	—	—	—	—	—	—
	2SR	−0.34	−0.69	−0.02	0.47 ②	0.91 ①	0.11
	CHROM	0.63 ②	0.71 ②	0.30 ②	−0.78	0.86 ②	−0.43
	LiCVPR	—	—	—	—	—	1.00 ①
	HR-CNN	0.95 ①	0.96 ①	0.59 ①	0.66 ①	0.59	0.78 ②
MAE	baseline	22.83 ②	22.83 ②	17.78 ②	17.78 ②	13.03 ②	13.03
	2SR	62.42	73.27	55.29	39.19	49.68	41.45
	CHROM	28.70	32.19	19.14	54.19	30.77	13.11
	LiCVPR	—	—	—	31.61	47.56	5.56 ①
	HR-CNN	9.79 ①	8.83 ①	9.46 ①	7.76 ①	12.49 ①	8.38 ②
RMSE	baseline	28.33 ②	28.33 ②	19.50 ②	19.50 ②	19.12 ②	19.12
	2SR	71.42	79.27	60.36	54.23	50.37	46.55
	CHROM	37.81	39.82	32.02	62.13	31.80	21.99
	LiCVPR	—	—	—	31.61	47.56	7.19 ①
	HR-CNN	11.00 ①	10.04 ①	9.92 ①	10.58 ①	13.95 ①	13.22 ②

Table 4.8: Evaluation of visual HR methods – experiment "Camera Vibration" Part 2. Percentage of videos with successful HR estimation, Pearson's correlation coefficient, mean average error and root-mean-square error on the test sets of the ECG-Fitness dataset for four baseline methods and the developed HR-CNN. The videos from the camera attached to the currently used fitness MACHINE and the camera attached to the TRIPOD were evaluated separately for each activity.

4.4.6 Experiment "Camera Vibration"

The "Camera Vibration" protocol inspects the effect of the camera vibration on the performance of the methods. The methods were trained on the training set of the experiment "all" and evaluated "activity-" and "vibration-"wise – the vibrations were either present (the camera was attached to the currently used fitness MACHINE) or not present (the camera was firmly attached to a TRIPOD).

The results are presented in Tab. 4.7 and Tab. 4.8. In all activities but "Talking", HR-CNN dominates the results.

HR-CNN performs the best in the "Rowing" activity. The "Rowing" activity is the one where the strongest camera vibration is present. Interestingly, the method yields better results if attached to the vibrating rowing machine and the halogen lamp light is not present. The reason for this behavior might come from the positioning of the cameras. The MACHINE camera attached to the rowing machine sees the subject *en face* all the time. That is not the case with the TRIPOD camera. Although the best efforts were made to position the TRIPOD camera as close as possible to the MACHINE camera, in case of the rowing machine the TRIPOD camera needed to be positioned to the side of the machine (see Fig. 4.3) to make sure that no vibrations are present at the TRIPOD camera. If the halogen lamp light is present, the difference between the two cameras is less extreme. Just to remind, the *Extractor* was trained on the PURE dataset first, and then by an alternating optimization (together with the *HR Estimator*). The distance between the TRIPOD camera and the subject was ≈ 10 cm greater than in case of the MACHINE camera.

The results from the "Stationary Bike" activity are somewhat surprising. The LiCVPR method is better than the HR-CNN method by more than a half in case of the camera placed on a TRIPOD. This is however not true in case of the MACHINE camera. The reason is that the MACHINE camera captures the subject from a low angle (see Fig. 4.3). On contrary, the TRIPOD camera has a nice view of the subject's face and during the "Stationary Bike" activity, there is only a little movement present. LiCVPR seems incapable of handling the low angle of view, but when a little movement is present and the face is clearly visible, the method works well.

HR-CNN yields the most stable results in case of the "Elliptical Trainer" activity – the difference between the recording angles was not so dramatic as in the previous cases. Here, the most important factor was ≈ 10 cm greater distance between the TRIPOD camera and the subject than the distance between the MACHINE camera and the subject (see Fig. 4.3).

To briefly comment the results of other methods, CHROM and 2SR preform the best in the "Talking" activity, probably because of lack of subject's movement, and LiCVPR yields very bad results for the cameras with extreme angles of view since it requires to track the subject's face, which is very difficult given the challenging recording setup. LiCVPR fails completely for the "Rowing" activity with the most rapid movement.

4.4.7 Summary of Experiments

The developed method performs significantly the best on the ECG-Fitness dataset that contains realistic challenges. In contrast to the commonly used COHFACE and MAHNOB datasets, the videos are not compressed. In terms of practical impact, there is a little point in validating the heart rate estimation methods on datasets where the only challenge is the compression. If one is interested in visual HR estimation from compressed videos, raw material may be always compressed with the desired compression standard and the required compression level.

Chapter 5

Conclusion

A novel two-step convolutional neural network for heart rate estimation, called HR-CNN, was introduced. The HR-CNN network comprises of the *Extractor* and the *HR Estimator* network. Both networks use a standard chain of convolution, MaxPool and activation blocks. The *Extractor* is trained on the PURE dataset. Structure of the *HR Estimator* is configured for each target dataset – it was introduced to cope with the compression and motion artifacts. The structure, namely the depth of the network, the number of filters, and the conv and MaxPool sizes, is found by the Metropolis-Hastings Monte Carlo Random Walk algorithm.

The HR-CNN network yields the state-of-the-art results outperforming three published methods [12, 5, 16] and a baseline according to a new experimental protocol. The protocol prescribes that the visual HR estimation method: (i) receives a sequence of facial images and outputs a single number (an estimated HR), any other output is considered invalid, (ii) is permitted to learn its parameters on the training set, (iii) is evaluated on the test set with the Pearson's correlation coefficient, mean absolute error, root mean squared error, and a percentage of videos with a successful HR estimation.

A new challenging publicly available ECG-Fitness dataset with 205 sixty-second videos of subjects performing physical exercises has been introduced. The dataset includes 17 subjects performing 4 activities (talking, rowing, exercising on a stepper and a rowing machine) captured by two RGB cameras, one attached to the currently used fitness machine that significantly vibrates, the other one to a separately standing tripod. With each subject, the "rowing" and "talking" activity is repeated with a halogen lamp lighting. In case of 4 subjects, the whole recording session is also lighted by an LED light.

The performance of the methods differs the most on the ECG-Fitness dataset. In contrast to the other datasets, the ECG-Fitness dataset contains realistic challenges. HR-CNN outperforms the published methods on the dataset reducing error by more than a half.

The structure of the *HR Estimator* requires to be configured for each target dataset. We believe that a single structure should be sufficient for a group of videos stored with the same compression method. Due to time constraints, we leave this research for a future work.

Bibliography

[1] Y.-H. Chen, H.-H. Chen, T.-C. Chen, and L.-G. Chen, “Robust heart rate measurement with phonocardiogram by on-line template extraction and matching,” *Engineering in Medicine and Biology Society, 2011 Annual International Conference*, vol. 2011, pp. 1957–1960, 2011.

[2] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, B. Karasfi, Y. S. Ali, and F. Meriaudeau, “Heart rate estimation using facial video: A review,” *Biomedical Signal Processing and Control*, vol. 38, pp. 346–360, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809417301362

[3] Y. Sun and N. Thakor, “Photoplethysmography Revisited: From Contact to Noncontact, From Point to Imaging,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 463–477, Mar. 2016.

[4] R. Spetlik, J. Cech, and J. Matas, “Non-Contact Reflectance Photoplethysmography: Progress, Limitations, and Myths.” Xi’An, China: IEEE Computer Society, May 2018.

[5] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, “Remote Heart Rate Measurement from Face Videos under Realistic Situations,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 4264–4271.

[6] W. Wang, S. Stuijk, and G. d. Haan, “Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 415–425, Feb. 2015.

[7] G. Heusch, A. Anjos, and S. Marcel, “A Reproducible Study on Remote Heart Rate Measurement,” *arXiv:1709.00962 [cs]*, Sep. 2017, arXiv: 1709.00962. [Online]. Available: http://arxiv.org/abs/1709.00962

[8] R. W. Brown, *Composition of Scientific Words.* Washington, D.C. London: Smithsonian Books, Jul. 2000.

[9] H. Molitor and M. Kniazuk, “A New Bloodless Method for Continuous Recording of Peripheral Circulatory Changes,” *Journal of Pharmacology*

and Experimental Therapeutics, vol. 57, no. 1, pp. 6–18, May 1936. [Online]. Available: http://jpet.aspetjournals.org/content/57/1/6

[10] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, p. R1, 2007. [Online]. Available: http://stacks.iop.org/0967-3334/28/i=3/a=R01

[11] H. Liu, Y. Wang, and L. Wang, "A review of non-contact, low-cost physiological information measurement based on photoplethysmographic imaging," *Engineering in Medicine and Biology Society, 2012 Annual International Conference*, vol. 2012, pp. 2088–2091, 2012.

[12] G. d. Haan and V. Jeanne, "Robust Pulse Rate From Chrominance-Based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.

[13] 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, Dec. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717852/

[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[15] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623264.1623280

[16] W. Wang, S. Stuijk, and G. de Haan, "A Novel Algorithm for Remote Photoplethysmography: Spatial Subspace Rotation," *IEEE transactions on bio-medical engineering*, vol. 63, no. 9, pp. 1974–1984, Sep. 2016.

[17] U. Rubins, V. Upmalis, O. Rubenis, D. Jakovels, and J. Spigulis, "Real-Time Photoplethysmography Imaging System," in *15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011)*, ser. IFMBE Proceedings. Springer, Berlin, Heidelberg, 2011, pp. 183–186. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-21683-1_46

[18] A. V. Moço, S. Stuijk, and G. d. Haan, "Ballistocardiographic Artifacts in PPG Imaging," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 9, pp. 1804–1811, Sep. 2016.

[19] A. A. Kamshilin, I. S. Sidorov, L. Babayan, M. A. Volynsky, R. Giniatullin, and O. V. Mamontov, "Accurate measurement of the pulse

wave delay with imaging photoplethysmography," *Biomedical Optics Express*, vol. 7, no. 12, pp. 5138–5147, Dec. 2016. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=boe-7-12-5138

[20] A. A. Kamshilin, O. V. Mamontov, V. T. Koval, G. A. Zayats, and R. V. Romashko, "Influence of a skin status on the light interaction with dermis," *Biomedical Optics Express*, vol. 6, no. 11, pp. 4326–4334, 2015. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=boe-6-11-4326

[21] M. Hülsbusch, "An image-based functional method for opto-electronic detection of skin-perfusion," Phd thesis, RWTH Aachen dept. of EE., 2008.

[22] L. F. C. Martinez, G. Paez, and M. Strojnik, "Optimal wavelength selection for noncontact reflection photoplethysmography," vol. 8011. International Society for Optics and Photonics, Nov. 2011, p. 801191. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8011/801191/Optimal-wavelength-selection-for-noncontact-reflection-photoplethysmography/10.1117/12.903190.short

[23] W. Cui, L. E. Ostrander, and B. Y. Lee, "In vivo reflectance of blood and tissue as a function of light wavelength," *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 6, pp. 632–639, Jun. 1990.

[24] B. A. Fallow, T. Tarumi, and H. Tanaka, "Influence of skin type and wavelength on light wave reflectance," *Journal of Clinical Monitoring and Computing*, vol. 27, no. 3, pp. 313–317, Jun. 2013.

[25] J. A. Crowe and D. Damianou, "The wavelength dependence of the photoplethysmogram and its implication to pulse oximetry," in *1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 6, Oct. 1992, pp. 2423–2424.

[26] J. Lee, K. Matsumura, K.-i. Yamakoshi, P. Rolfe, S. Tanaka, and T. Yamakoshi, "Comparison between red, green and blue light reflection photoplethysmography for heart rate monitoring during motion," *Engineering in Medicine and Biology Society, 2013 Annual International Conference*, vol. 2013, pp. 1724–1727, 2013.

[27] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *International Journal of Cardiology*, vol. 166, no. 1, pp. 15–29, Jun. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167527312003269

[28] L. A. M. Aarts, V. Jeanne, J. P. Cleary, C. Lieber, J. S. Nelson, S. Bambang Oetomo, and W. Verkruysse, "Non-contact heart rate

monitoring utilizing camera photoplethysmography in the neonatal intensive care unit — A pilot study," *Early Human Development*, vol. 89, no. 12, pp. 943–948, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378378213002375

[29] B. D. Holton, K. Mannapperuma, P. J. Lesniewski, and J. C. Thomas, "Signal recovery in imaging photoplethysmography," *Physiological Measurement*, vol. 34, no. 11, p. 1499, 2013. [Online]. Available: http://stacks.iop.org/0967-3334/34/i=11/a=1499

[30] M. v. Gastel, S. Stuijk, and G. d. Haan, "Motion Robust Remote-PPG in Infrared," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 5, pp. 1425–1433, May 2015.

[31] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1565–1588, May 2015. [Online]. Available: https://www.osapublishing.org/abstract.cfm?uri=boe-6-5-1565

[32] A. V. Moço, S. Stuijk, and G. de Haan, "Motion robust PPG-imaging through color channel mapping," *Biomedical Optics Express*, vol. 7, no. 5, pp. 1737–1754, Apr. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4871078/

[33] L.-M. Po, L. Feng, Y. Li, X. Xu, T. C.-H. Cheung, and K.-W. Cheung, "Block-based adaptive ROI for remote photoplethysmography," *Multimedia Tools and Applications*, pp. 1–27, Mar. 2017. [Online]. Available: https://link.springer.com/article/10.1007/s11042-017-4563-7

[34] C. Takano and Y. Ohta, "Heart rate measurement based on a time-lapse image," *Medical Engineering & Physics*, vol. 29, no. 8, pp. 853–857, Oct. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350453306001901

[35] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological Measurement*, vol. 35, no. 5, pp. 807–831, May 2014.

[36] M. Villarroel, A. Guazzi, J. Jorge, S. Davis, P. Watkinson, G. Green, A. Shenvi, K. McCormick, and L. Tarassenko, "Continuous non-contact vital sign monitoring in neonatal intensive care unit," *Healthcare Technology Letters*, vol. 1, no. 3, pp. 87–91, Sep. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4612732/

[37] S. Xu, L. Sun, and G. K. Rohde, "Robust efficient estimation of heart rate pulse from video," *Biomedical Optics Express*, vol. 5, no. 4, pp. 1124–1135, Mar. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3985994/

[38] A. Huch, R. Huch, V. König, M. Neuman, D. Parker, J. Yount, and D. Lübbers, "Limitations of pulse oximetry," *The Lancet*, vol. 1, pp. 357–358, 1988.

[39] N. S. Trivedi, A. F. Ghouri, N. K. Shah, E. Lai, and S. J. Barker, "Effects of motion, ambient light, and hypoperfusion on pulse oximeter function," *Journal of Clinical Anesthesia*, vol. 9, no. 3, pp. 179–183, May 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952818097000391

[40] X. F. Teng and Y. T. Zhang, "The effect of contacting force on photoplethysmographic signals," *Physiological Measurement*, vol. 25, no. 5, pp. 1323–1335, Oct. 2004.

[41] H. Pälve, "Reflection and transmission pulse oximetry during compromised peripheral perfusion," *Journal of Clinical Monitoring*, vol. 8, no. 1, pp. 12–15, Jan. 1992. [Online]. Available: https://link.springer.com/article/10.1007/BF01618081

[42] D. B. Wax, P. Rubin, and S. Neustein, "A comparison of transmittance and reflectance pulse oximetry during vascular surgery," *Anesthesia and Analgesia*, vol. 109, no. 6, pp. 1847–1849, Dec. 2009.

[43] N. Nesseler, J.-V. Frénel, Y. Launey, J. Morcet, Y. Mallédant, and P. Seguin, "Pulse oximetry and high-dose vasopressors: a comparison between forehead reflectance and finger transmission sensors," *Intensive Care Medicine*, vol. 38, no. 10, pp. 1718–1722, Oct. 2012.

[44] A. Buchs, Y. Slovik, M. Rapoport, C. Rosenfeld, B. Khanokh, and M. Nitzan, "Right-left correlation of the sympathetically induced fluctuations of photoplethysmographic signal in diabetic and non-diabetic subjects," *Medical and Biological Engineering and Computing*, vol. 43, no. 2, pp. 252–257, Apr. 2005. [Online]. Available: https://link.springer.com/article/10.1007/BF02345963

[45] M. Nitzan, B. Khanokh, and Y. Slovik, "The difference in pulse transit time to the toe and finger measured by photoplethysmography," *Physiological Measurement*, vol. 23, no. 1, pp. 85–93, Feb. 2002.

[46] E. B. Blackford and J. R. Estepp, "Effects of frame rate and image resolution on pulse rate measured using multiple camera imaging photoplethysmography," B. Gimi and R. C. Molthen, Eds., Mar. 2015, p. 94172D. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2083940

[47] J. R. Estepp, E. B. Blackford, and C. M. Meier, "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 1462–1469.

[48] F. Andreotti, A. Trumpp, H. Malberg, and S. Zaunseder, “Improved heart rate detection for camera-based photoplethysmography by means of Kalman filtering,” in *2015 IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO)*, Apr. 2015, pp. 428–433.

[49] T. Wu, V. Blazek, and H. J. Schmitt, “Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes,” vol. 4163. International Society for Optics and Photonics, Nov. 2000, pp. 62–71. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/4163/0000/Photoplethysmography-imaging--a-new-noninvasive-and-noncontact-method-for/10.1117/12.407646.short

[50] H. Zhu, Y. Zhao, and L. Dong, “Non-contact detection of cardiac rate based on visible light imaging device,” vol. 8498. International Society for Optics and Photonics, Oct. 2012, p. 849806. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8498/849806/Non-contact-detection-of-cardiac-rate-based-on-visible-light/10.1117/12.929203.short

[51] G. de Haan and A. van Leest, “Improved motion robustness of remote-PPG by using the blood volume pulse signature,” *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, Aug. 2014.

[52] W. Wang, A. C. d. Brinker, S. Stuijk, and G. d. Haan, “Robust heart rate from fitness videos,” *Physiological Measurement*, vol. 38, no. 6, p. 1023, 2017. [Online]. Available: http://stacks.iop.org/0967-3334/38/i=6/a=1023

[53] D. McDuff, S. Gontarek, and R. W. Picard, “Improvements in Remote Cardiopulmonary Measurement Using a Five Band Digital Camera,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, Oct. 2014.

[54] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting Pulse from Head Motions in Video,” 2013, pp. 3430–3437. [Online]. Available: http://www.cv-foundation.org/openaccess/content_cvpr_2013/html/Balakrishnan_Detecting_Pulse_from_2013_CVPR_paper.html

[55] M. A. Haque, K. Nasrollahi, and T. B. Moeslund, “Estimation of Heartbeat Peak Locations and Heartbeat Rate from Facial Video,” in *Image Analysis*, ser. Lecture Notes in Computer Science. Springer, Cham, Jun. 2017, pp. 269–281. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-59129-2_23

[56] A. A. Kamshilin, E. Nippolainen, I. S. Sidorov, P. V. Vasilev, N. P. Erofeev, N. P. Podolian, and R. V. Romashko, “A new look at the essence of the imaging photoplethysmography,” *Scientific Reports*, vol. 5, May 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4440202/

[57] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," *arXiv:1511.07289 [cs]*, Nov. 2015, arXiv: 1511.07289. [Online]. Available: http://arxiv.org/abs/1511.07289

[58] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Mar. 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html

[59] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 2004. [Online]. Available: //www.springer.com/gp/book/9780387212395

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

[61] T. J. Cross, M. Keller-Ross, A. Issa, R. Wentz, B. Taylor, and B. Johnson, "The Impact of Averaging Window Length on the"Desaturation Indexes during Overnight Pulse Oximetry at High-Altitude"," *Sleep*, vol. 38, no. 8, pp. 1331–1334, Aug. 2015.

[62] D. J. McDuff, E. B. Blackford, and J. R. Estepp, "The Impact of Video Compression on Remote Cardiac Pulse Measurement Using Imaging Photoplethysmography," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 63–70.

[63] M. Sznajder and M. Łukowska, "Python Online and Offline ECG QRS Detector based on the Pan-Tomkins algorithm," Jul. 2017. [Online]. Available: https://zenodo.org/record/826614#.WuBhhohuZPY

[64] Y.-P. Yu, P. Raveendran, and C.-L. Lim, "Dynamic heart rate measurements from video sequences," *Biomedical Optics Express*, vol. 6, no. 7, pp. 2466–2480, Jun. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4505702/

[65] Y. Hsu, Y. L. Lin, and W. Hsu, "Learning-based heart rate detection from remote photoplethysmography features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4433–4437.

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *arXiv:1610.02391 [cs]*, Oct. 2016, arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391

[67] J. Sochman and J. Matas, “WaldBoost - learning for time constrained sequential detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 2, Jun. 2005, pp. 150–156 vol. 2.