Video-based Cardiac Physiological Measurements Using Joint Blind Source Separation Approaches

by

Huan Qi

B. Eng., Zhejiang University, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Electrical and Computer Engineering)

The University of British Columbia
(Vancouver)

July 2015

Abstract

Non-contact measurements of human cardiopulmonary physiological parameters based on photoplethysmography (PPG) can lead to efficient and comfortable medical assessment. It was shown that human facial blood volume variation during cardiac cycle can be indirectly captured by regular Red-Green-Blue (RGB) cameras. However, few attempts have been made to incorporate data from different facial sub-regions to improve remote measurement performance. In this thesis, we propose a novel framework for non-contact video-based human heart rate (HR) measurement by exploring correlations among facial sub-regions via joint blind source separation (J-BSS). In an experiment involving video data collected from 16 subjects, we compare the non-contact HR measurement results obtained from a commercial digital camera to results from a Health Canada and Food and Drug Administration (FDA) licensed contact blood volume pulse (BVP) sensor. We further test our framework on a large public database, which provides subjects' left-thumb plethysmograph signal as ground truth. Experimental results show that the proposed framework outperforms the state-of-the-art independent component analysis (ICA)-based methodologies.

Driver physiological monitoring in vehicle is of great importance to provide a comfortable driving environment and prevent road accidents. Contact sensors can be placed on the driver's body to measure various physiological parameters. However such sensors may cause discomfort or distraction. The development of non-contact techniques can provide a promising solution. In this thesis, we employ our proposed non-contact video-based HR measurement framework to monitor the drivers heart rate and do heart rate variability analysis using a simple consumer-level webcam. Experiments of real-world road driving demonstrate that the pro-

posed non-contact framework is promising even with the presence of unstable illumination variation and head movement.

Preface

This thesis is based on the following works:

- Huan Qi, Z. Jane Wang and Chunyan Miao, "Non-contact Driver Cardiac Physiological Monitoring Using Video Data", accepted for the Third IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), 2015.

The research was jointly initiated by Dr. Z. Jane Wang and the thesis author, and the majority of the research, including literature survey, model design, algorithm implementation, experimental data collection, data analysis and paper writing, was conducted by the author of this thesis, with valuable suggestions from Dr. Z Jane Wang. Dr. Zhenyu Guo and Dr. Xun Chen also helped on the methods part in Chapter 2. Dr. Xun Chen, Mr. Liang Zou and Mr. Yiming Zhang helped greatly on data collection of the road driving experiment in Chapter 3.

Table of Contents

List of Tables

List of Figures

Glossary

ECG Electrocardiogram

HRM Heart Rate Monitoring

PPG Photoplethysmography

HR Heart Rate

HRV Heart Rate Variability

BVP Blood Volume Pulse

IBI Inter-beat Interval

SDNN Standard Deviation of the IBI Series

RMSSD Root Mean Square of Successive Differences of the IBI Series

LF Low Frequency

HF High Frequency

PSD Power Spectral Density

LS Lomb-Scargle

HRVAS HRV Analysis Software

RGB Red-Green-Blue

SCV Source Component Vector

J-BSS Joint Blind Source Separation

CCA Canonical Correlation Analysis

M-CCA Multiset Canonical Correlation Analysis

BSCM Between-set Source Correlation Maximization

ESCM Eigenvalue-maximization of Source Correlation Matrix

RGCCA Regularized Generalized Canonical Correlation Analysis

IVA Independent Vector Analysis

FFT Fast Fourier Transform

C-MCCA Connectivity Multiset Canonical Correlation Analysis

CDM Connectivity Design Matrix

SSQCOR Sum of Squared Correlation

M3L Max-margin Multi-label Classification

EEG Electroencephalography

EOG Electrooculography

Acknowledgments

I want to express my great appreciation to my supervisor, Dr. Z. Jane Wang for her persistent support, constant encouragement and profound insight in the research area throughout my master study. I would like to thank Dr. Chunyan Miao from Nanyang Technological University for financial support and research guidance during my visit in Singapore. Many thanks go to Dr. Zhenyu Guo and Dr. Xun Chen for their research advice.

I would like to thank all my dear friends and labmates. Thanks a lot for their help and feedback. Special thanks go to Liang Zou and Yiming Zhang for their friendship and lunch companion since the day we began to study together at UBC.

I would like to thank all committee members of my master exam for their valuable time and suggestions.

Last but not least, I own my deepest gratitude to my parents in China, Mr. Xiaodong Qi and Mrs. Juyan Wang, for their endless love and support. They are the spiritual idols in all aspects of my life.

Chapter 1

Introduction

1.1 Motivation

Various human physiological parameters provide direct or indirect evidence of human health state. Measurements of these physiological parameters, which are often interdependent, have always been one of the most fundamental questions in the area of modern medicine. Among numerous parameters, cardiovascular parameters are of great research interest, including heart rate (HR), heart rate variability (HRV), blood pressure, and respiratory rate. Large-scale clinical studies show that surveillance and prevention of certain cardiovascular diseases requires regular medical assessment of HR and HRV [15], which is also known as heart rate monitoring (HRM). The history of HRM partially reflects the development of medical technologies. In traditional Chinese medicine, which dates back to more than 2,000 years ago, therapists can diagnose illness based on patients' wrist pulse patterns. For centuries, HRM was carried out by placing an ear on the patient's chest. The invention of stethoscope by French physician René Laennec nearly 200 years ago was a milestone in HRM [2]. It provides instant and clear heart beat feedback to therapists in a non-invasive fashion. Later in the year of 1903, the Dutch physiologist and Nobel laureate Willem Einthoven invented the first practical electrocardiogram (ECG). With the ECG technique, it is possible to observe and record the entire cardiac electrical activity of heart beat cycle.

Since the invention of ECG, great efforts have been made to develop conve-

Figure 1.1: Popular heart rate monitoring devices.

Table 1.1: Popular heart rate monitoring techniques

Device	Target Signal	Accessory
Electrocardiogram	Heart electrical activity	Adhesive electrodes
Holter monitor	Heart electrical activity	Adhesive electrodes
Chest strap	Heart electrical activity	Transmission module
Doppler fetal monitor	Electronic audio	Ultrasound couplant
Finger pulse oximeter	Blood volume pulse	None
Watch-like heart rate sensor	Blood volume pulse	None
Non-contact video technique	Blood volume pulse	None

nient and comfortable HRM devices, as shown in Fig. 1.1. Some of the most popular HRM techniques are listed in Table 1.1. A Holter monitor is a portable medical device for continuously monitoring various cardiovascular electrical activities for more than 24 hours. Electrodes need to be attached to the human body together with the monitor itself and no intensive exercise is allowed during monitoring. A commercial-level chest strap is designed to measure HR in situations such as racing, hiking, and various sports exercises. A segment of the chest strap is made of multi-layered textile with good conductivity. A small and light processor is at-

tached to measure heart electrical activity and transmit signals to other devices such as smartphones and computers. A Doppler fetal monitor uses the Doppler effect to generate electronic audio simulation of fetal heart beat. To enhance the simulation, ultrasound couplant (usually liquid) is often used to facilitate the transmission of ultrasonic energy from the transducer into the target. A finger pulse oximeter is a non-invasive device designed for convenient HRM, which can provide accurate instant HR within just a few seconds. It is based on the photoplethysmography (PPG) effect generated by cardiac cycle. A PPG sensor is clipped on a thin part of the human body such as a fingertip. It measures the change of tissue optical absorbance, which is known as the blood volume pulse (BVP) signal. No other accessory is required. Another type of PPG-based devices is the popular watch-like heart rate monitor, such as Apple Watch. All aforementioned HRM devices can be classified as contact techniques since they require physical contact between the electrodes or sensors and the human body. Placement and removal of these attachments can cause discomfort, stress and even epidermal stripping [1].

With the advances of imaging sensors and computer vision technologies, many vision algorithms have been successfully adapted and applied to biomedical engineering applications [21]. Video-based physiological measurement was also born in this exciting trend. Without requiring any physical contact, video-based physiological measurement technique allows remote detection of human blood volume pulse signals (thus heart rate measurements) using designated imaging sensors or even low-cost webcams. This technique can potentially bring HRM to the next level of comfort and convenience. Non-contact heart rate measurement benefits from the integration of computer vision and biomedical signal processing. Both computer vision and biomedical signal processing areas have witnessed important advances in recent years. For instance, state-of-the-art face tracking algorithms are more robust to various background, occlusion, illumination change, and intensive head motion. Advanced multi-set analysis of biomedical signals can reveal deeper correlations across multiple datasets. Investigating the interaction between these two areas has been receiving increasing research attention. It is expected that such interaction would achieve more accurate and robust non-contact physiological measurement [40].

Studying the robust of non-contact HRM is also of great importance to facilitate

the utilization of non-contact cardiac physiological measurements in real world applications. Convenient measurement of HR is of great potential for both clinic diagnosis and daily healthcare. In a clinical setting, non-contact methods work without attaching any medical electrode or sensor to the subject. Some of such methods have been clinically tested, such as vital signs monitoring during haemodialysis [40], neonatal intensive care unit [1], quantification of limb movement in epileptic seizure [29], and dynamic tissue phantoms evaluation [43]. Family healthcare can also benefit a lot from non-contact detection techniques, especially with the rapid dissemination of smartphones [37]. For instance, commercial apps such as Cardiio (Cardiio, Inc., San Francisco, CA, USA) and Vital Signs Camera [33] (Philips, Inc., Amsterdam, Netherlands) enable users to measure heart rates using continuous recordings of their faces by front cameras on the phones. Another potential application is the driver medical assistance in the automobile environment, which is considered as one of the most promising ways to effectively prevent accidents and augment intelligence in transportation systems [14, 47]. Such an assistance system should incorporate reliable measurements of the drivers vital signals in order to depict his/her driving condition.

In this thesis, with the intention to enhance the accuracy and robustness of existing non-contact techniques, we plan to develop a non-contact video-based heart rate monitoring framework based on the combination of advanced computer vision and multi-set data analysis methods. Moreover, we attempt to test the proposed framework under different indoor and outdoor environments.

1.2 Background of Cardiac Physiological Measurements

As a prognostic factor and potential therapeutic target, HR has been verified in large epidemiological studies to be an independent predictor of cardiovascular and all-cause mortality for people with or without diagnosed cardiovascular disease [15]. Currently, ECG devices are widely used in HRM due to the high reliability. ECG records the electrical activity of the heart over a certain period of time using multiple electrodes attached on a patient's body. During each heart beat cycle, heart muscle depolarizing would result in the tiny electrical variation on the skin, which can be detected by ECG electrodes. Variations from multiple

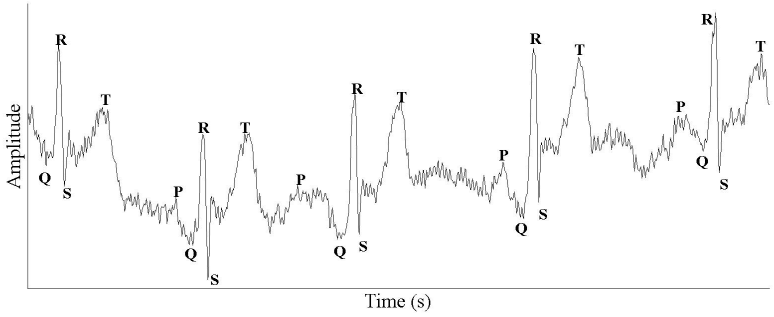


Figure 1.2: A segment of ECG waveforms.

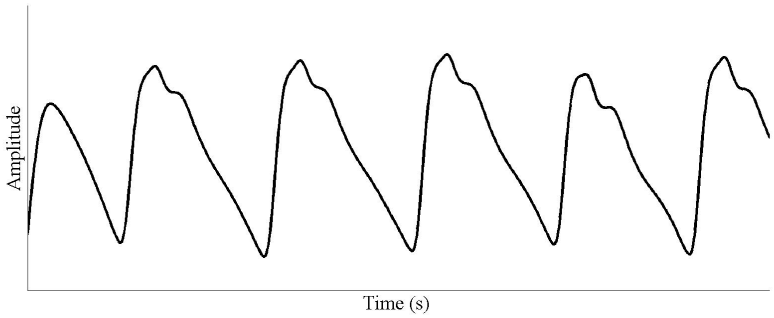


Figure 1.3: A segment of BVP waveforms of a subject from the DEAP database [25].

electrodes are processed to form heart beat waveforms, as shown in Fig. 1.2. Normally, each individual heart beat is represented on the ECG as a PQRST complex. Different parts of the PQRST complex are related to different sub-processes of the cardiac cycle. One common method to estimate HR is to use the mean R-R interval:

$$\mathrm{HR} = \frac{60}{\overline{T_{\mathrm{RR}}}}, \tag{1.1}$$

where $\overline{T_{\mathrm{RR}}}$ denotes the average time interval between adjacent R peaks given a segment of heart beat waveforms. Since physiological interpretation of PQRST complex is beyond the scope of this thesis, interested readers are referred to [20] for more information.

PPG is an indirect but effective technique to measure cardiovascular BVP. During a cardiac cycle, variations of tissue blood volume in certain human body segments modulate the transmission or reflection of visible light at these segments. PPG sensors are used to capture such variations in the dedicated light source, and the heart rate can be estimated correspondingly by measuring time intervals between consecutive peaks of the signal [40]. As shown in Fig. 1.3, the BVP signal contains less information about cardiac cycle than the ECG signal does. However, the BVP signal is sufficient to estimate HR using a similar method as in ECG:

$$\mathrm{HR} = \frac{60}{\overline{IBI}}, \tag{1.2}$$

where $\overline{IBI}$ denotes the average inter-beat interval (IBI) between adjacent heart beat peaks given a segment of BVP waveforms.

Besides the heart rate, another important cardiac physiological parameter is HRV. HRV is the physiological phenomenon of variation in the time interval between heartbeats. Once the heart beat signal is obtained, a sequence of IBIs computed from every pair of adjacent peaks can be extracted for HRV analysis, which is, to some extent, more informative than the heart rate alone. HRV analysis usually focuses on the time and frequency domain measures of the IBI series. Various time domain measures have been proposed, such as the standard deviation of the IBI series (SDNN), the root mean square of successive differences of the IBI series (RMSSD), the number of successive differences that are great than x milliseconds (NNx, often $x = 50$), and the percentage of total intervals that successively differ by more than x milliseconds (pNNx, often $x = 50$).

Clinical pathological studies of HRV reveal that the low frequency (LF) and high frequency (HF) oscillations of the IBI series are of great research interest. It is believed that LF is associated with sympathetic and parasympathetic activity and HF is associated with respiratory sinus arrhythmia. The nominal frequency ranges of LF and HF are 0.04$\sim$0.15 Hz and 0.15$\sim$0.4 Hz, respectively. Powers within certain frequency bands are useful for quantitative description. For example, the LF power measures the amount of power within [0.04 Hz, 0.15 Hz], calculated by integrating the power spectral density (PSD) over the frequency band. The HF power can be calculated in a similar way over the range [0.15 Hz, 0.4 Hz]. The

ratio of the LF power to the HF power (i.e., LF/HF) also provides insight into the sympatho-parasympatho balance. A popular way to estimate PSD is the Lomb-Scargle periodogram method (LS-Periodogram), which does not require the data to be uniformly sampled. If we segment the IBI series temporally and generate the LS-Periodogram with respect to time, the resulting plot is a spectrogram. Currently there are many open source toolboxes such as HRVAS [36] for HRV analysis in time and frequency domains. The physiological interpretation of the time and frequency domain measures in HRV analysis is beyond the scope of this thesis. Interested readers are referred to [7] for more information.

1.3 Background of Non-contact Physiological Measurements

Many efforts have been made to provide non-contact HR measurements. Some works used dedicated sensors such as Doppler wave sensors [5, 17, 44] and thermal imaging sensors [16]. The study in [45] showed, for the first time, that BVP signals can be remotely acquired from the human face using consumer-level digital cameras in *ambient light*. Poh *et al.* [34] presented an independent component analysis (ICA) framework to measure HR using a low-cost webcam in ambient light. Later, the authors extended their previous work with measurements of the respiratory rate and low & high frequency components of HRV [35]. To overcome the frequency resolution limitation of traditional red-green-blue (RGB) sensors, Mcduff *et al.* [30] presented a modified five band digital camera with the cyan and orange frequency bands being added to the original red, green and blue color channels. Experimental results showed that such modifications improve the performances of physiological measurements of HR and HRV. Real-time measurements using continuous wavelet transform can be achieved despite the existence of light and motion artifacts [8]. In [6], an interesting motion-based approach was presented to recover the heart beat signal. It showed that the influx of blood during a cardiac cycle causes detectable head motion according to Newton's third law of motion. The frequency component of such a motion can be extracted by using facial feature tracking and principal component analysis. A recent publication [26] reported the results under more challenging conditions, where the subject's

motions and illumination variations are involved. The proposed normalized least mean square adaptive filtering method was tested on a difficult public database and achieved the state-of-the-art performance when compared to other methods.

Based on the aforementioned literature of recent years, it is concluded that the accuracy of non-contact measurements is highly susceptible to video recording environment. For example, in [30], facial video recording was conducted under well-controlled laboratory environment with stable indoor illumination and little head motion artifact. In that case, the performance of non-contact measurements is almost as good as the performance of the contact PPG sensor (ground truth). In [26], a much more challenging experimental environment impaired the performance of non-contact methods to a large extent. Therefore it suggests that there is still much space for improvement in video-based non-contact physiological measurements.

1.4 Research Objectives and Methodology

The technical objective of this thesis is to investigate possible solutions to enhancing the performance of video-based non-contact HR measurements under challenging experimental environments in order to facilitate the utilization of non-contact physiological measurements in real-world applications. It is worth noting that almost all previous non-contact techniques extract the face color channel data by averaging over the entire facial region without considering potential variations among different facial sub-regions [1]. Therefore we plan to investigate how such variations among different facial sub-regions might contribute to the non-contact HR measurement. We also plan to conduct three types of experiments to evaluate the performances of non-contact HR measurements:

i Experiment under the well-controlled laboratory environment using self-collected video and physiological data.

ii Experiment under a more challenging laboratory environment using video and physiological data from public affective computing database.

iii Experiment under the difficult *road-driving* setting using self-collected video and physiological data.

[1] A facial 'sub-region' is a region containing only one part of a human face.

In all three experiments, we attempt to compare the proposed non-contact HR measurement framework with traditional HRM devices. In the road-driving experiment, we also plan to compare HRV time and frequency measures.

In order to achieve the above objective, we propose a non-contact video-based human heart rate measurement framework by exploiting data correlations among specified facial sub-regions via advanced facial landmark localization and joint blind source separation approaches. The two main components of the proposed framework, facial landmark localization and joint blind source separation, are summarized as follows:

- **Facial Landmark Localization**
  Collect facial color channel data is the first step of most non-contact methods. It is desirable that this data collection procedure is robust to potential head movement and illumination variation. In this thesis, we employ an advanced real-time facial landmark localization algorithm to collect facial color channel data. Specifically, based on detected facial landmark coordinates, we design a division pattern to divide facial region into four subregions and collect data respectively for subsequent data analysis. Experimental results verify the robustness of this algorithm under different settings including road-driving condition.

- **Joint Blind Source Separation**
  To extract BVP from color channel data of different facial sub-regions, we propose to use joint blind source separation methods including multi-set canonical correlation analysis (M-CCA) and independent vector analysis (IVA). A BVP extraction pipeline is designed to ensure accurate heart rate measurement by using techniques such as frequency analysis, signal detrending, correlation clustering via normalized cut. We also develop a BVP peak detection method using adaptive sliding window size. We observe that by altering connectivity among different sub-regions, HR measurement performance actually varies. Based on this observation, we propose to use a max-margin multi-label classification method to learn interaction among data from different sub-regions in order to further enhance the performance of our framework. A learning-based connectivity multi-set canonical cor-

relation analysis (C-MCCA) algorithm is proposed. Experimental results demonstrate that the proposed non-contact framework outperforms current state-of-the-art method.

The organization of the remainder thesis is as follows. In Chapter 2, we elaborate the employment of facial landmark localization and joint blind source separation approaches. The proposed non-contact measurement framework is described in detail, including its pipeline and several proposed methods. Chapter 3 contains the setting descriptions and results of three types of experiments to evaluate the performance of the proposed framework. The conclusions are given in Chapter 4 along with discussion of future work.

Chapter 2

Method

2.1 Introduction

Two major concerns that determine the performance of a non-contact HR measurement method are (i) data collection and (ii) BVP signal recovery, especially under challenging environment where illumination variation and motion artifacts are involved. To tackle the first concern, we employ a real-time facial landmark localization algorithm to track subjects' facial regions. Based on coordinates of certain facial landmarks such as eyes, nose, and mouth, we divide facial regions into four non-overlapping sub-regions and extract four sets of color channel signals respectively. Compared with traditional methods which mainly focus on the entire facial region, this sub-region data collection method allows the investigation of interaction among different sub-regions and whether it would contribute to the improvement of HR measurement accuracy. J-BSS methods can deal with this multi-set analysis problem by using correlation maximization. In this thesis, we introduce a framework for non-contact HR measurement by using landmark localization to designate facial sub-regions and J-BSS to recover BVP signals. We elaborate the details in the following sections.

2.2 Facial Landmark Localization

In this section, we first introduce the importance of facial data collection using landmark localization methods in brief, and then focus on an advanced vision algorithm that is recently proposed and how to divide facial regions using designated division pattern.

For each frame of video signals, we divide the face into $M = 4$ sub-regions and extract color channel data from each of them. There are two important issues about dividing facial sub-regions in video signals. Firstly, a face should be accurately detected in each frame of the recorded video. Secondly, it should be guaranteed that the locations of corresponding facial sub-regions remain approximately the same over all frames. The second one is a fundamental requirement since we intend to study correlations among different sub-regions. It only makes sense when color channel data in each sub-region are extracted from the same physical part no matter what absolute coordinates of the face might change from one frame to another due to possible head movements.

Facial landmark localization is a natural approach to address these requirements. Specifically, we employ the landmark localization method proposed in [4] to divide sub-regions based on coordinates of detected facial landmarks (e.g. eyebrows, pupil, nose bridge, and lips). In [4], an efficient facial landmark localization algorithm is introduced to achieve real-time 2D face and eyes landmark detection and tracking. By incrementally updating a discriminative facial deformable model, the algorithm achieves state-of-the-art performance for face alignment in static images and face tracking in videos. In this thesis, we use a pre-trained model provided by the authors of [4], which allows detection and tracking of 49 facial landmarks. As shown in Fig. 2.1, the landmark localization method works in different viewpoints and returns (x, y) coordinates of detected facial landmarks (marked as green dots) in each frame of the video for each subject. Based on the coordinates of these landmarks, we specify a division pattern by using certain landmarks and their geometric connection lines as vertexes and edges to constitute $M = 4$ polygons or regions of interest as facial sub-regions. In Fig. 2.2(a), filled dots denote facial landmarks while black squares denote division vertexes. $\{r_i\}$ represent coordinates for selected landmarks and $\{V_i\}$ for division vertexes. $\{V_1, V_2, V_3, V_4, V_6\}$ are

Figure 2.1: Robust facial landmark localization in different viewpoints, generated by a pre-trained model in [4]

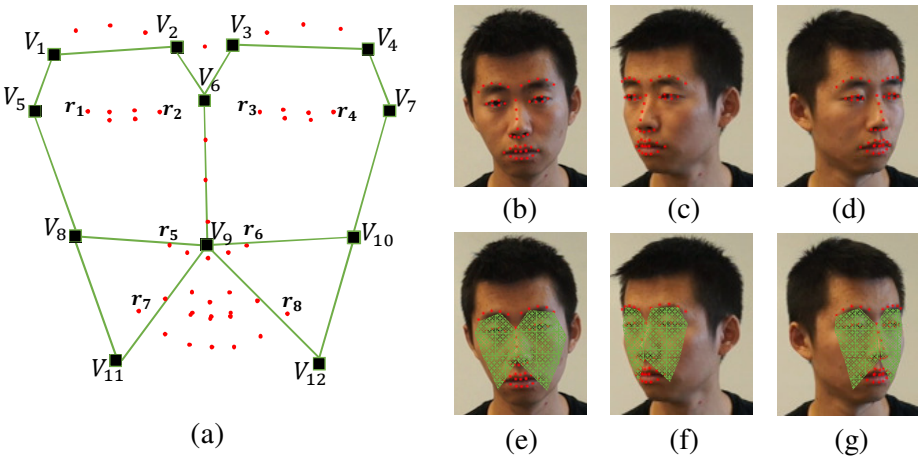


Figure 2.2: Facial sub-region division. (a) Division vertexes distribution. (b)-(d) Facial landmark localization in different viewpoints. (e)-(g) Areas covered by four facial sub-regions.

corresponding facial landmarks. The remaining vertexes are specified as follows:

$$V_5 = r_1 + \frac{3}{4}(r_1 - r_2) \quad V_7 = r_4 + \frac{3}{4}(r_4 - r_3)$$
$$V_{11} = r_7 + \frac{3}{4}(r_7 - r_5) \quad V_{12} = r_8 + \frac{3}{4}(r_8 - r_6)$$
$$V_9 = \frac{1}{2}(r_5 + r_6) \quad V_8 = \frac{1}{2}(V_5 + V_{11}) \quad V_{10} = \frac{1}{2}(V_7 + V_{12})$$

Such division pattern is tested in different experimental environments. In Fig. 2.3, we show that this division pattern guarantees a robust data collection in different facial sub-regions under three different settings: a well-controlled laboratory environment (the first row in Fig. 2.3), a challenging recording environment of DEAP affective computing public database [25] (the second to fourth rows in Fig. 2.3), a real-world road driving setting where illumination and motion artifacts are constantly involved (the last row in Fig. 2.3).

It is shown that most facial sub-regions are robustly captured, as shown in Fig. 2.2(e)-(g). Among eight division vertexes, two are facial landmarks, the rest are geometric coordinates constituted based on facial landmarks. For each sub-region, color channel data is computed by averaging over all pixels within the sub-region. By aligning data temporally, we can acquire four facial sub-region datasets, each containing a multi-dimensional color channel signal for the corresponding sub-region during the recording.

2.3 Joint Blind Source Separation

After acquiring multiple datasets from facial sub-regions, we perform joint blind source separation (J-BSS) methods based on the assumption that these datasets share common underlying sources, which come from blood volume variations during cardiac cycles.

2.3.1 Problem Formulation

First we describe the mathematical formulation and notations of J-BSS problems in general. Given M datasets, let $X^{[m]} \in \mathbb{R}^{V \times N}$ denote the m-th dataset, where V is the number of variables and N is the number of samples (in our case $V = 3$ for

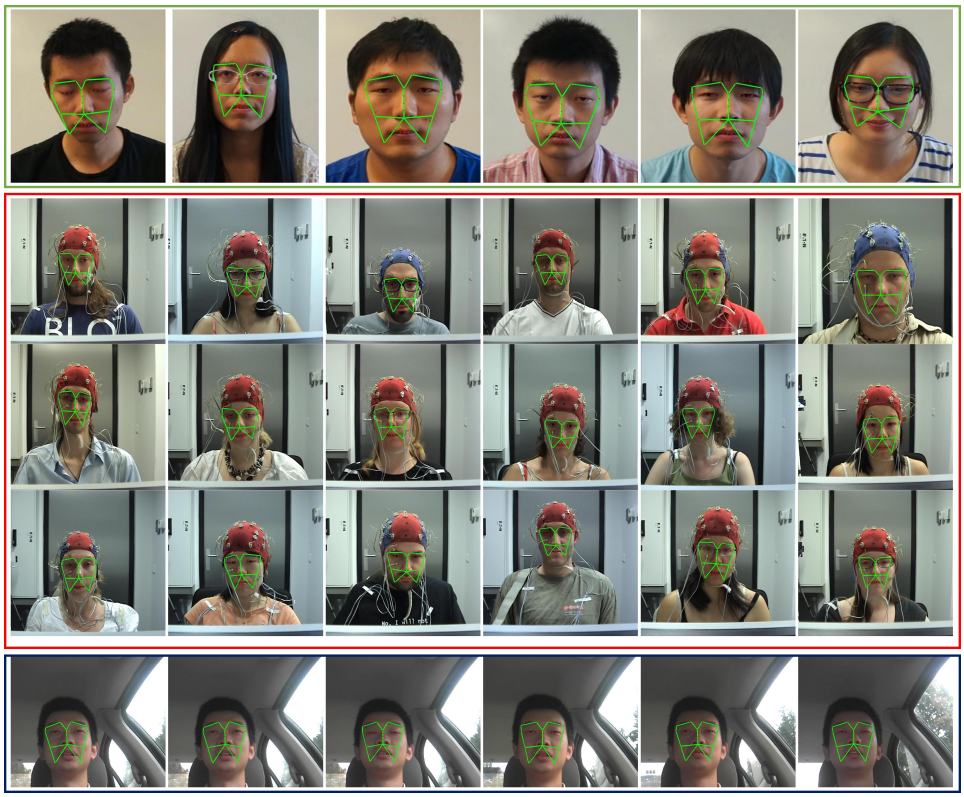


Figure 2.3: Facial landmark localization and sub-region division pattern under three different experimental settings. From top to bottom: the setting of our self-collected laboratory experiment, the setting of DEAP affective computing database [25], the setting of our self-collected road-driving experiment.

the RGB color channels and N is the number of video frames). $X^{[m]}$ can be further denoted with respect to column vectors $x_n^{[m]}$ $(1 \leq n \leq N)$ as:

$$X^{[m]} = \left[x_1^{[m]}, \cdots x_N^{[m]}\right] \qquad \text{for } 1 \leq m \leq M \tag{2.1}$$

where $x_n^{[m]} \in \mathbb{R}^{V \times 1}$ is the n-th observation. It is assumed that each dataset is a linear mixture of L underlying independent sources:

$$x^{[m]} = A^{[m]} s^{[m]} \qquad \text{for } 1 \leq m \leq M \tag{2.2}$$

where $A^{[m]} \in \mathbb{R}^{V \times L}$ means the mixing matrix to be determined and $s^{[m]} \in \mathbb{R}^{L \times 1}$ is the random source vector that can be further expressed as $s^{[m]} = [s_1^{[m]}, \cdots, s_L^{[m]}]^{\mathrm{T}}$ for $1 \leq m \leq M$, where the superscript T means the transpose operation.

The concept of source component vector (SCV) in J-BSS is based on the assumption that common underlying sources are shared among multiple datasets [23]. Formally, let $s_l = [s_l^{[1]}, \cdots, s_l^{[M]}]^{\mathrm{T}}$ denote the l-th SCV, for $1 \leq l \leq L$. s_l is a random vector uncorrelated with all other SCVs but the components within itself are mutually correlated. The general goal of J-BSS is to find SCVs by estimating the mixing matrices $A^{[m]}$'s or their inverse matrices $W^{[m]}$'s and the corresponding source vector estimations $y^{[m]} = W^{[m]} x^{[m]}$. The estimation of $s^{[l]}$ is hereby expressed as $y_l = [y_l^{[1]}, \cdots, y_l^{[M]}]^{\mathrm{T}}$, where $y_l^{[m]}$ is the estimation of the l-th component in dataset m:

$$y_l^{[m]} = (w_l^{[m]})^{\mathrm{T}} x^{[m]} \tag{2.3}$$

where $w_l^{[m]}$ is a demixing vector. It is the l-th column vector of $W^{[m]}$.

2.3.2 IVA and M-CCA

Several approaches to J-BSS have been developed in recent years based on different statistical assumptions. One direction focuses on extensions of the ICA idea. Methods such as group ICA [9], parallel ICA [28], IC-PLS [11], and independent vector analysis (IVA) [3, 23] were proposed. Among these methods, IVA is a natural extension of ICA from one to multiple datasets by ensuring that the extracted sources are independent within each dataset and meanwhile well correlated across multiple datasets. IVA is designed to minimize the mutual information $\mathscr{I}_{\mathrm{IVA}}$

among the estimated SCVs [3]:

$$\begin{aligned}
\mathscr{I}_{\mathrm{IVA}} &\triangleq \mathscr{I}[y_1;\cdots;y_L] \\
&= \sum_{l=1}^{L} \mathscr{H}[y_l] - \mathscr{H}[y_1;\cdots;y_L] \\
&= \sum_{l=1}^{L} \mathscr{H}[y_l] - \mathscr{H}\left[W^{[1]}x^{[1]},\cdots,W^{[M]}x^{[M]}\right] \\
&= \sum_{l=1}^{L}\left(\sum_{m=1}^{M} \mathscr{H}[y_l^{[m]}] - \mathscr{I}[y_l]\right) \\
&\qquad - \sum_{m=1}^{M} \log|\det(W^{[m]})| - C_1
\end{aligned} \tag{2.4}$$

where $\mathscr{H}(\cdot)$ denotes the entropy of certain random variables (or vectors) and C_1 is the constant term $\mathscr{H}[x^{[1]},\cdots,x^{[M]}]$. The derivation shows that minimizing $\mathscr{I}_{\mathrm{IVA}}$ is equivalent to minimizing the entropy of all components $y_l^{[m]}$ and maximizing the mutual information within each estimated SCV y_l for $l = 1,\cdots,L$. We use IVA-G [3] for the implementation of IVA, which exploits second-order statistical information across multiple datasets by assuming that each SCV follows a multivariate Gaussian distribution.

Another popular method, named multiset canonical correlation analysis (M-CCA), is an extension based on the concept of canonical correlation analysis (CCA). The goal of CCA is to find two linear transformation vectors a, b for two random vectors x_1, x_2 such that the correlation between $y_1 = a^{\mathrm{T}}x_1$ and $y_2 = b^{\mathrm{T}}x_2$ is maximized. The random variables y_1 and y_2 are called the first pair of canonical variates (CVs). The following pairs of CVs can be iteratively obtained by achieving maximum correlation under the constraint that they are statistically uncorrelated to the previous ones. M-CCA extends the idea of correlation maximization by optimizing a certain objective function of the correlation matrix of CVs from multiple random vectors to achieve maximum overall correlation [22]. It is presented in [27] that J-BSS can be achieved by two different yet interrelated approaches: between-set source correlation maximization (BSCM) and eigenvalue maximization of source correlation matrix (ESCM), respectively. In BSCM approach, the group of sources with largest between-set correlation values are first extracted from datasets, by op-

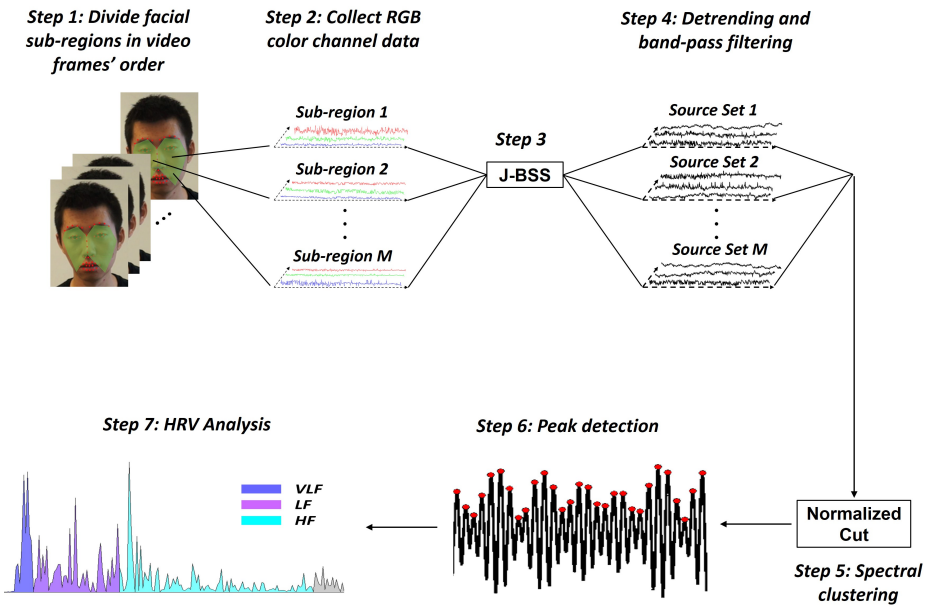


Figure 2.4: Overview of the proposed video-based (non-contact) HR measurement method using facial landmark localization and J-BSS techniques. First, subjects' faces are divided into several sub-regions according to coordinates of facial landmarks. Then color channel data from each sub-region are collected into temporal signals and fed to J-BSS algorithms. The obtained source sets are clustered after certain detrending and filtering operations. Finally we could recover the BVP signal and conduct HR estimation and HRV analysis.

timizing objective functions with respect to correlation magnitudes, i.e. measure of overall correlation. The ESCM approach, on the other hand, focused on the maximum eigenvalue $\lambda_{\max}(R^{(l)})$ of $M \times M$ source correlation matrix $R^{(l)}$ for the l-th SCV.

2.4 Identify BVP Signal

The major steps of the proposed framework are shown in Fig. 2.4. In step 1, uncompressed videos of each subject are analyzed frame by frame using facial landmark localization to obtain sub-region division. Details have been discussed in Section 2.2. We calculate average RGB channel values of pixels within each sub-

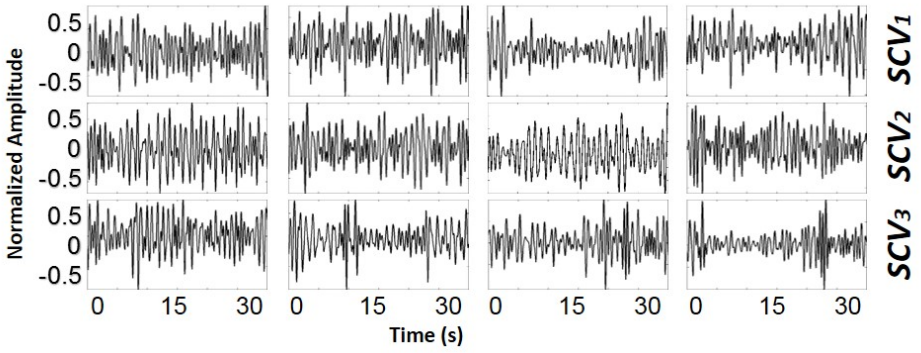


Figure 2.5: One example of recovered SCVs using the M-CCA method. They are computed from datasets of four facial sub-regions, and each has three color channels and three underlying sources to recover.

region to generate a feature $\mathscr{X}(k)$ for frame k in a recorded video:

$$\mathscr{X}(k) = [x_k^{[1]}, \cdots, x_k^{[M]}]^{\mathrm{T}} \tag{2.5}$$

where $x_k^{[m]} = [x_k^{m_R}, x_k^{m_G}, x_k^{m_B}]$ contains average RGB color channel values of sub-region m in frame k. M is the number of sub-regions (datasets).

In step 2, by combining features from all frames, we can obtain a feature sequence of one subject $[\mathscr{X}(1), \mathscr{X}(2), \cdots, \mathscr{X}(N)]$, where N is the number of frames in the video. The feature sequence of sub-region m is referred to as dataset $X^{[m]}$, where $X^{[m]} \in \mathbb{R}^{3\times N}$. For $m = 1, \cdots, M$, we have:

$$[\mathscr{X}(1), \cdots, \mathscr{X}(N)] = \begin{bmatrix} X^{[1]} \\ \vdots \\ X^{[M]} \end{bmatrix} \tag{2.6}$$

All datasets are detrended using a technique that has been widely used in HRV analysis to remove slow linear or more complex trends in signals [41]. Similar performance is achieved when we set the smoothness parameter between 1500 and 2000. The detrended signals are further normalized to have zero mean and unit variance. They are then fed into J-BSS algorithms to recover source signals under

the assumption that the number of sources is equal to the number of observations for each dataset in step 3 of Fig. 2.4. One example of recovered SCVs are shown in Fig. 2.5. The resulting source signals are band-pass filtered between [0.5 Hz, 2.5 Hz], corresponding to 30 bpm and 150 bpm as the lower and upper bounds of human HR measurements in step 4 of Fig. 2.4.

In [34], Poh *et al.* empirically selected the second source to be the BVP signal. This can be practically problematic since there is no guarantee on the order of recovered sources using ICA. Source selection method in [30, 35] is simple yet effective. All sources were band-pass filtered and then calculated by the normalized Fast Fourier Transform (FFT). The one with the largest peak in frequency domain was selected to be the BVP signal. In J-BSS framework, the number of recovered sources can be large and differences among them are sometimes difficult to tell even in the frequency domain. We instead consider the basic assumption of J-BSS that SCVs are uncorrelated with all other SCVs but the components within each SCV are correlated. It is reasonable to assume that those mutually correlated sources in source sets are likely to be the estimates of BVP signals since human heart beat is an underlying source for all facial sub-region datasets.

Therefore, in step 5 of Fig. 2.4, we propose performing spectral clustering on the similarity matrix of all source variables, which is calculated based on the normalized cross correlation among all recovered source signals. The cluster number is carefully appointed to be slightly smaller than the total number of recovered sources such that most resulting clusters have only one component. The cluster with largest number of components is selected, which contained several BVP signal candidates according to our assumption. In our case, 12 source signals were assigned to 8 clusters. The largest cluster usually contained 3 or 4 BVP signal candidates. In step 6, we run the same source selection approach discussed above in frequency domain to determine the best BVP signal. Here we employ the Normalized Cut algorithm [38] to achieve spectral clustering. One example of the source selection process is shown in Fig. 2.6. Most clusters have only one or two components while the largest cluster contains four. The resulting BVP signal reflects the subject's heart beat process during the recording and can be used to estimate HR. The signal is first interpolated to increase its sampling frequency to $F_s = 256$ Hz. A peak detection algorithm is then applied to find peaks that are at least separated

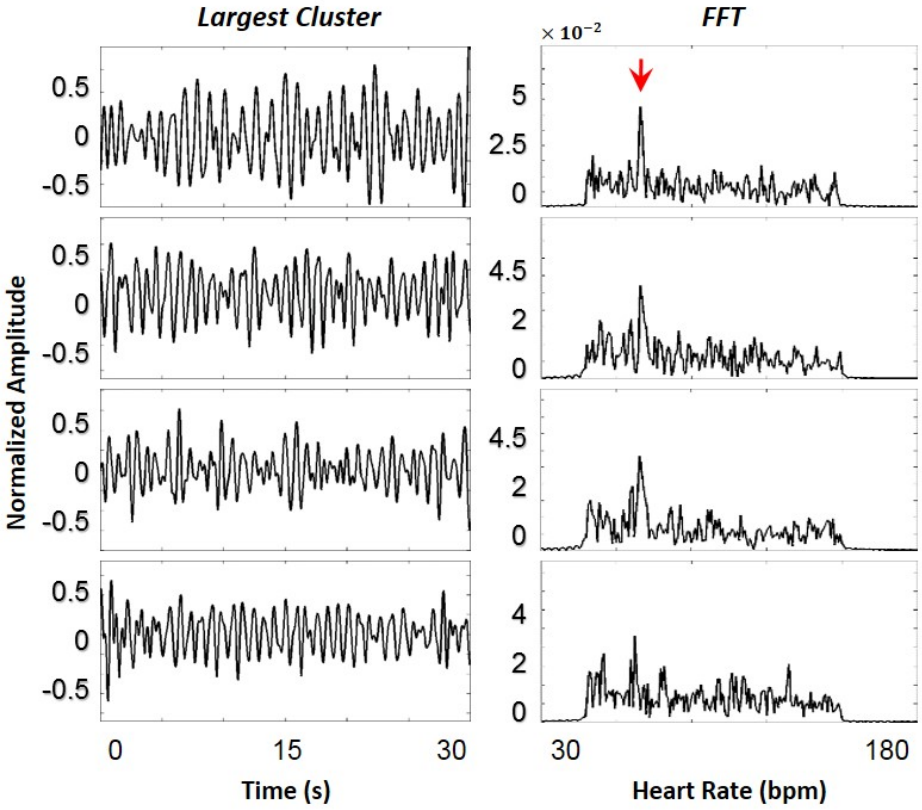


Figure 2.6: Results in Fig. 2.5 were clustered by Normalized Cut [38]. The largest cluster has four elements and their frequency spectra all contain peaks near 1Hz, which is close to human resting HR's. The arrow indicates the largest peak among all spectra, which belongs to the BVP signal estimates. Here the cluster number is set to 8.

by δ sampling points, which is called δ-correction. Here δ is a positive integer designed to ignore smaller peaks that might occur in close proximity to a large one, which can be quite common cases in BVP signals. For instance, if there is a large local peak at index x, then all smaller peaks in the range $(x-\delta, x+\delta)$ are ignored.

In [30], different values of δ are tested and the one giving the best peak detection performance are chosen by visual verification. However, it is reasonable to select the value of δ in an adaptive way because the IBI of BVP signals varies from time to time. Peak detection algorithm should adjust its step, in this case δ,

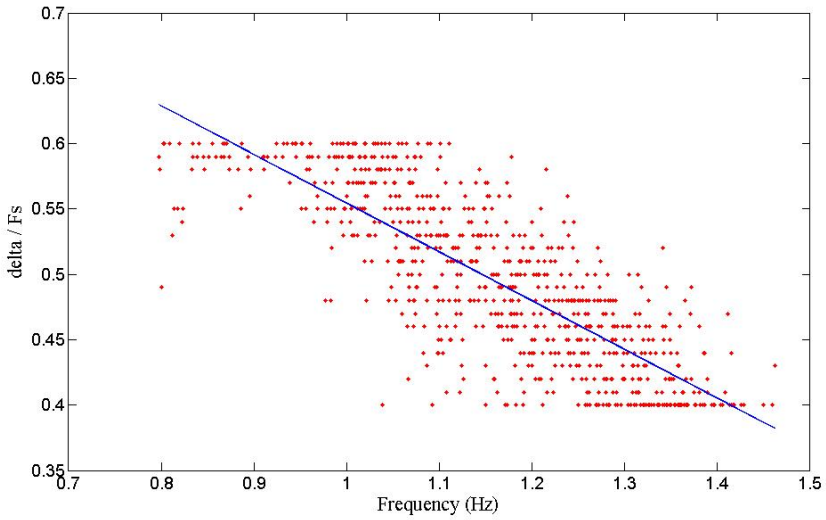


Figure 2.7: Scatter plot of δ test on DEAP affective computing database. The line shows the linear regression model that is fit using test data. Here F_s denotes the sampling rate of the BVP signal after interpolation.

according to the frequency of BVP signals. Intuitively, a segment of BVP signal with a frequency of 1.5 Hz (i.e. 90 beats per minutes) should have more peaks than the one with a frequency of 0.9 Hz. Accordingly, a smaller value of δ should be used to detect peaks of a signal with larger frequency. To verify this point, we run a test on the DEAP affective computing database [25]. Given BVP signals of different frequency, we manually choose different values of δ to estimate HR and select the δ that would yield the most accurate HR measurement. A scatter plot of the experimental results are shown in Fig. 2.7. It is clear that a linear correlation between the frequency of input signal f and the optimal value of δ^* exists. We fit a linear regression model based on these pairs of (f, δ^*) and use it to predict the optimal δ value given an input BVP signal with frequency f. It is called the adaptive δ-correction. Fig. 2.8(c) shows the performance of our adaptive δ-correction peak detection method.

After localizing peaks in a BVP signal, we can compute the average IBI using coordinates of detected peaks and estimate HR using 1.2.

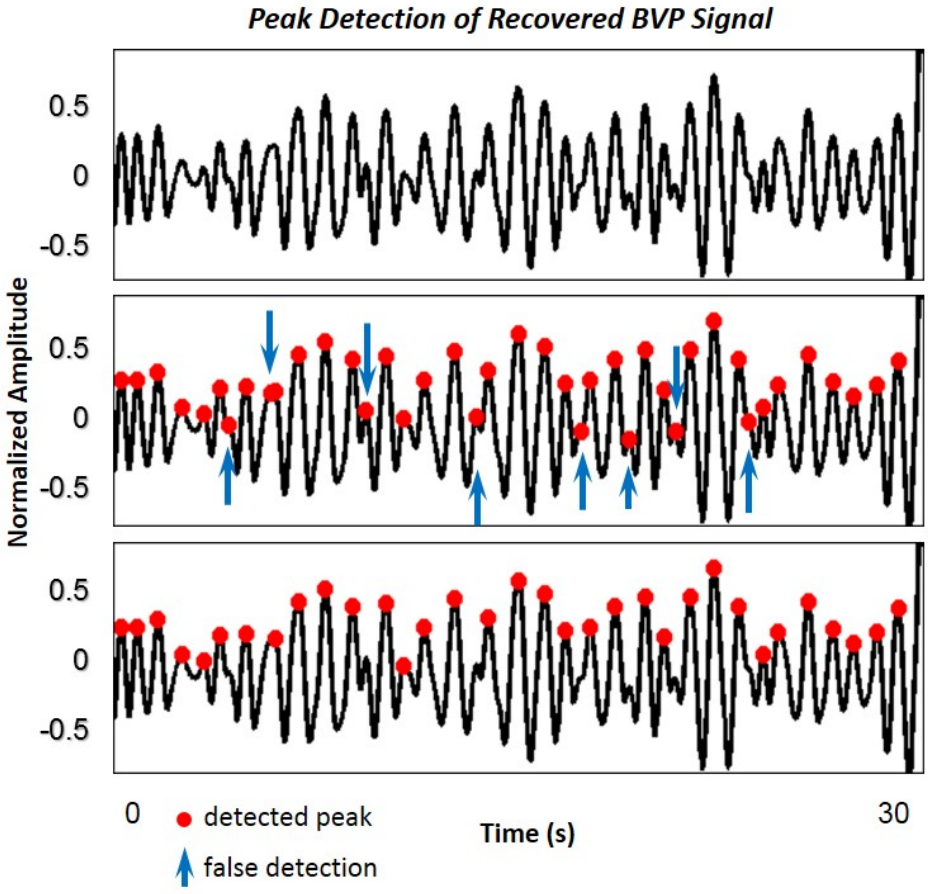


Figure 2.8: The top row is an interpolated BVP signal before peak detection. The remaining two figures show different detection performances. With a fixed δ, several small local peaks are also incorporated as labeled by blue arrows in the middle row. Using adaptive δ-correction by incorporating frequency knowledge of input BVP signal, false detections are removed and almost all large local peaks corresponding to heart beats are successfully detected.

2.5 Connectivity Multiset Canonical Correlation Analysis

In the M-CCA setting, all datasets are symmetrically incorporated to calculate the overall correlation. However, asymmetry might occur in certain applications. For instance, sampling areas, illumination angles or even reflectivity are likely different among facial sub-regions, which may cause inaccuracy if being treated equally. Is there an *optimal correlation combination* among datasets that would give the best measurement performance? Previously discussed J-BSS methods such as IVA and M-CCA are not capable of exploring this question. Hence, we propose a learning-based J-BSS approach for non-contact HR measurement based on M-CCA, named the connectivity multiset canonical correlation analysis (C-MCCA), to explore the existence of such an optimal combination.

Inspired by [27, 42], we plan to combine the correlation maximization methodology with the flexibility of connectivity designation and modify the method to achieve J-BSS. The proposed method is termed as C-MCCA since we introduce connectivity among datasets. We notice that doing J-BSS via M-CCA would involve all datasets unbiasedly when optimizing correlation objective functions, even though some datasets might not be as useful as others. However, since no prior knowledge about connectivity among different facial sub-regions can be directly used, we plan to design a data-driven method using a training set to *learn* a potentially non-linear mapping from multiset facial color channel signals (input) to an optimal connectivity pattern (output).

In [42], a method named regularized generalized canonical correlation analysis (RGCCA) was proposed for multiset data analysis. Unlike J-BSS via M-CCA, which involves a multi-stage deflationary correlation maximization scheme, RGCCA tries to find a group of linear mixtures that achieves overall correlation maximization for multiple datasets using the knowledge of partial least square path modeling algorithms [12]. In their framework, a design matrix $C = \{c_{m,n}\}$ is introduced based on prior knowledge about between-set correlation of multiple datasets: $c_{m,n} = 1$ if dataset m and n are connected and 0 otherwise.

Similarly, we introduce a binary connectivity design matrix (CDM) $C = \{c_{m,n}\}$ in our non-contact HR measurement method. For any two different facial sub-

regions m and n, where $m,n = 1,\cdots,M$, we can choose to either incorporate data correlation between m and n by setting $c_{m,n} = 1$ or simply discard the correlation by setting $c_{m,n} = 0$. Generally a CDM is used to describe whether color channel data between any two facial sub-regions are jointly analyzed (connected) to maximize overall correlation. The major difference between our method and RGCCA is that ours is a multi-stage method without any prior knowledge on the connectivity of different datasets. In order to explore whether there exists an optimal CDM for non-contact HR measurement, we have to consider all possible CDMs. Theoretically, there are $2^{M(M-1)/2}$ combinations given the number of datasets M, which yields 64 possible CDMs in our case ($M = 4$).

Similar to M-CCA, C-MCCA contains L stages, where L is the number of sources variables to be extracted. In the l-th stage, the following optimization problem is solved:

$$\begin{aligned} \max_{w_l^{[m]}, w_l^{[n]}} \quad & \sum_{\substack{m,n=1 \\ m\neq n}}^{M} c_{m,n} \left| r_l^{[m,n]} \right|^2 \\ \text{s.t.} \quad & w_l^{[m]} \perp \{w_1^{[m]}, \cdots, w_{l-1}^{[m]}\} \quad \text{except for } l = 1. \end{aligned} \tag{2.7}$$

where $r_l^{[m,n]} \triangleq \text{corr}[(w_l^{[m]})^{\mathrm{T}} x^{[m]}, (w_l^{[n]})^{\mathrm{T}} x^{[n]}]$ denotes the correlation for $m,n = 1,\cdots,M$. The orthogonality constraint indicates that the newly obtained demixing vector should be uncorrelated to the previous ones. Here we use the sum of squared correlation (SSQCOR), which is presented as one of the five objective functions in [22]. SSQCOR is shown to be the most robust one in [27]. After identifying demixing vectors, we could recover SCVs using (2.3). It is obvious that M-CCA is a special case of C-MCCA where its CDM is $\mathbf{1}^{M\times M}$, i.e. a $M \times M$ all-ones matrix.

C-MCCA works similarly as aforementioned J-BSS methods, so it can easily fit into our non-contact HR measurement framework. The only difference is that we have to designate a CDM *a priori* before multi-set signals are fed into C-MCCA. Intuitively, it is expected that given multi-set signals (input), an optimal CDM (output) can be determined to best recover the latent BVP signal. To this end, we propose to use a max-margin multi-label (M3L) classification method [18] to learn non-linear mapping between the input and output.

Formulated in [18], the objective of multi-label classification is to predict *a set*

of relevant binary labels for a given input, which is in contrast to the multi-class classification problem where one has to predict the single, most probable label. Formally, it is to learn a mapping f from a point $\mathbf{x}$ to a set of labels $\mathbf{y} \in \mathscr{Y}$. Here $\mathscr{Y}$ denotes the set of all possible binary labels with $|\mathscr{Y}| = L$. We assume that N training samples are in the form of $(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^D \times \{\pm 1\}^L$ with ${y_i}^l$ being $+1$ if label l has been assigned to sample i and -1 otherwise. One way to formulate this problem as a max-margin one would be to define a loss function Δ between ground truth label and predicted label and minimize it over the training set subject to regularization, which can be formulated as the following primal:

$$\begin{aligned} \min_{f} \quad & \frac{1}{2}\|f\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & f(\mathbf{x}_i, \mathbf{y}_i) \geq f(\mathbf{x}_i, \mathbf{y}) + \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \quad \forall i, \mathbf{y} \in \{\pm 1\}^L \setminus \{\mathbf{y}_i\} \end{aligned} \tag{2.8}$$

with a new point $\mathbf{x}$ being assigned $\mathbf{y}^* = \arg\max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. In [18], an efficient M3L approach is proposed to reduce computational complexity even when dense pairwise label correlation is incorporated. The implementation details are beyond the scope of this thesis, please refer to [18] for more information.

To demonstrate how the selection of CDM influences non-contact HR measurement, we use trials in DEAP affective computing database as examples. In the database, each trial contains one subject's facial video recording and blood volume pulse ground truth from a contact PPG sensor. First we use facial landmark localization algorithm to extract 4 sets of facial color channel signals from 4 facial sub-regions. These four datasets serve as the input to our C-MCCA method. Since we have no prior knowledge on how to select the optimal CDM, we shall try all of them one by one by altering $\{y_i\}$ in the following CDM:

$$\begin{pmatrix} 1 & c_{1,2} & c_{1,3} & c_{1,4} \\ c_{2,1} & 1 & c_{2,3} & c_{2,4} \\ c_{3,1} & c_{3,2} & 1 & c_{3,4} \\ c_{4,1} & c_{4,2} & c_{4,3} & 1 \end{pmatrix} \tag{2.9}$$

where $c_{m,n} = c_{n,m} \in \{0, 1\}$, for $m, n = 1, \cdots, 4$. There are 64 possible CDMs in

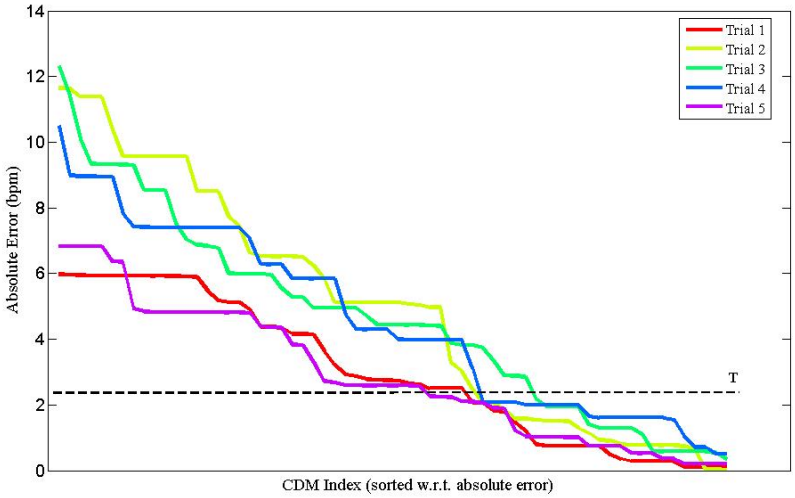


Figure 2.9: Absolute error of non-contact HR measurement from 5 independent trials by altering CDM pattern.

total. Performance varies among different CDMs as they alter the connectivity pattern among different facial sub-regions. Fig. 2.9 shows how different CDMs influence the accuracy of non-contact HR measurement on five independent trials. We measure the performance in terms of absolute error $\varepsilon = |\mathrm{HR}_{gt} - \mathrm{HR}_{nc}|$. Here 'gt' refers to ground truth acquired by contact PPG sensor while 'nc' refers to non-contact measurement using C-MCCA. Results in Fig. 2.9 have been sorted. It is clear that C-MCCA would extract BVP signals with better performance for non-contact HR measurement given certain CDMs.

To collect training samples, we first define a label set $\mathscr{Y}$. As shown in Equation 2.9, it takes six-dimension vector $\mathbf{c}$ to determine a CDM:

$$\mathbf{c} = \{c_{2,1}, c_{3,1}, c_{4,1}, c_{3,2}, c_{4,2}, c_{4,3}\} \tag{2.10}$$

Each element in $\mathbf{c}$ represents a connectivity status between two facial regions. For instance, if $c_{2,1} = 1$, sub-region #1 and #2 are connected. Their correlation would be included to solve Equation 2.7. Therefore we define a label set with size of six: $\mathscr{Y} = \{y_1, y_2, y_3, y_4, y_5, y_6\}$. Each label $y_i \in \{+1, -1\}$ corresponds to one element

of $\mathbf{c}$ with the exact order in Equation 2.10 for $i = 1, \cdots, 6$. The corresponding pair of $(y_i, c_{m,n})$:

$$y_i = \begin{cases} +1 & \text{if } c_{m,n} = 1 \\ -1 & \text{if } c_{m,n} = 0 \end{cases} \tag{2.11}$$

Given an input $\mathbf{x}$, the learnt model would predict its label set $\mathbf{y} \subset \mathscr{Y} = \{y_1, y_2, y_3, y_4, y_5, y_6\}$. In our non-contact HR measurement framework, we define the training pair $(\mathbf{x}, \mathbf{y})$ of a single trial as follows:

$$\mathbf{x} = [\text{Corr}_{2,1}, \text{Corr}_{3,1}, \text{Corr}_{4,1}, \text{Corr}_{3,2}, \text{Corr}_{4,2}, \text{Corr}_{4,3}]^T \tag{2.12}$$

where $\text{Corr}_{m,n}$ represent the Pearson's correlation coefficient between $\mathbf{g}_m$ and $\mathbf{g}_n$. Here $\mathbf{g}_m$ denotes the extracted *green* channel signal of facial sub-region m in the trial. Green channel signal has been widely used in non-contact research because green channel provides the strongest PPG signal according to optical property. To determine label set $\mathbf{y}$, we define a threshold parameter T. As the black dash line shown in Fig. 2.9, T controls the upper bound of the absolute error of one trial using C-MCCA. It is assumed that those CDMs that yield absolute error lower than T contains representative features to determine mapping from $\mathbf{x}$ to the optimal CDM. We can use $\mathbf{c}$ in Equation 2.10 to represent a CDM. Let $\{\mathbf{c}\}_\mathbf{T}$ denote the set of CDMs which yield absolute error lower than T. For a single trial, we compute the expectation of this set, denoted as $\{\bar{\mathbf{c}}\}_\mathbf{T}$, by averaging over all elements in $\{\mathbf{c}\}_\mathbf{T}$. We define label set $\mathbf{y}$ as:

$$\mathbf{y} = \{\bar{\mathbf{c}}\}_\mathbf{T} \succeq \lambda \tag{2.13}$$

where $\lambda \in [0,1]$ is another threshold parameter and $\succeq$ denotes an element-wise operator. For vector $\mathbf{a}$ and scalar A, $\mathbf{a} \succeq A$ returns a binary vector of the same size as $\mathbf{a}$. For each element a_i in $\mathbf{a}$:

$$(\mathbf{a} \succeq A)_i = \begin{cases} 1 & \text{if } a_i > A \\ 0 & \text{if } a_i \leq A \end{cases} \tag{2.14}$$

Given a single trial, we can now determine a training sample pair $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^6 \times \mathbb{B}^6$ by choosing certain threshold parameters $T \in \mathbb{R}^+$ and $\lambda \in [0,1]$. It is clear that T controls the accuracy and λ controls the sparsity of $\mathbf{y}$. After collecting

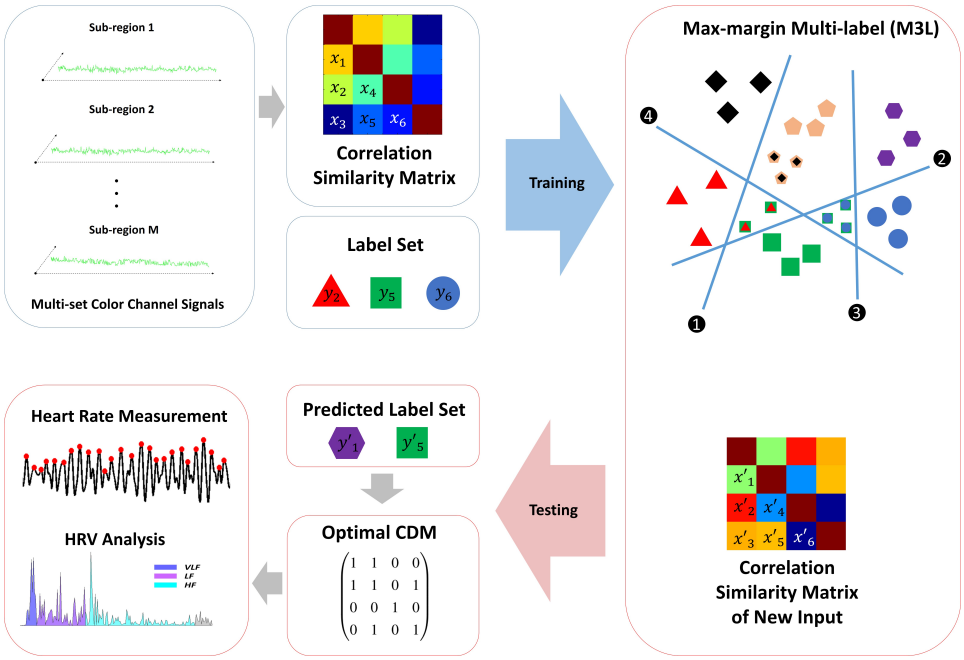


Figure 2.10: Proposed learning-based C-MCCA based on an M3L model [18]. Given multi-set color channel signals, we train the model using extracted feature $\mathbf{x}$ and label set $\mathbf{y}$. The trained model can predict the optimal label set $\mathbf{y}'$ (i.e. CDM) given any input feature $\mathbf{x}'$ extracted from new multi-set color channel signals. The predicted CDM is then used for subsequent heart rate measurement and HRV analysis.

certain amount of training samples, we can train a M3L model in Equation 2.8 and predict the optimal CDM given a new input. Fig. 2.10 depicts the general idea of our M3L training and testing procedure. Using the predicted CDM, we then perform C-MCCA to recover BVP signal for HR estimation. More discussion and experimental results of the proposed learning-based C-MCCA will be presented in Chapter 3. For easy reading, we summarize the proposed non-contact framework as pseudo-codes in Algorithm 1, using C-MCCA as an example in the J-BSS step.

Algorithm 1 Video-based HR Measurement via C-MCCA

Input: Video frame sequences
Output: IBI sequence, HR estimation

procedure TRAINING M3L MODEL
    Collect N training samples.
    Train a M3L model $\mathscr{M}$ by solving Equation 2.8.
end procedure
procedure LANDMARK LOCALIZATION
    for $m = 1 \rightarrow M$ **do**
        for $k = 1 \rightarrow K$ **do**
            if *a face is detected in frame k* **then**
                Divide it into M sub-regions using [4].
                $x_k^{[m]} \leftarrow \overline{rgb}(m)$.
            end if
        end for
        $X^{[m]} \leftarrow [x_1^{[m]}, \cdots, x_K^{[m]}]$
    end for
end procedure

procedure JOINT BLIND SOURCE SEPARATION
    Compute $\mathbf{x}$ *for the current trial using Equation 2.12*
    Predict optimal CDM: $\mathbf{C}^* \leftarrow f(\mathscr{M}, \mathbf{x})$
    Solve $\{W^{[m]}\text{'s}\} \leftarrow \text{argmin}_W(SSQCOR)$ *in Equation 2.7 given* $\mathbf{C}^*$.
    for $m = 1 \rightarrow M$ **do**
        $Y^{[m]} \leftarrow W^{[m]} X^{[m]}$
    end for
end procedure

procedure HEART RATE ESTIMATION
    Compute similarity matrix $S \leftarrow \{Y^{[m]}\} \times \{Y^{[m]}\}$
    Obtain $\{y_L\text{'s}\}$ *as the largest cluster via* $NormCut(S)$.
    $y^* \leftarrow \text{argmax}_{y,f} |FFT(\{y_L\text{'s}\})|$.
    HR, IBI $\leftarrow AdaptivePeakDetection(y^*)$
end procedure

2.6 Summary

We propose a non-contact HR measurement framework to accurately and robustly recover human BVP signals from multiple color channel signals of different facial sub-regions by exploiting their data correlation interactions. An advanced real-time facial landmark localization algorithm is used to track facial regions. A facial division pattern is designed using coordinates of certain facial landmarks to form

four facial sub-regions. In each sub-region, color channel data are collected as the form of temporal signals. Facial video data recorded under different experimental settings are used to test its performance. The results demonstrate that the combination of facial landmark localization and facial region division is suitable for data collection even when intensive illumination variation and motion artifacts are involved. It is much more reliable than previously used face tracking methods for non-contact HR measurement.

We use joint blind source separation methods (M-CCA and IVA) to extract latent BVP signals from multiset color channel signals. An adaptive δ-correlation peak detection method is proposed by fitting a linear regression model between BVP signal frequency and δ-correlation. It is expected that such an adaptive method would yield a better peak detection performance. Experimental results will be presented in next chapter. Finally, a learning-based connectivity multiset canonical correlation analysis (C-MCCA) algorithm is proposed to investigate how color channel data from different facial sub-regions would influence the performance of non-contact HR measurement. A max-margin multi-label classification model is used to map from multiset signals (input) to optimal connectivity design matrix (CDM) based on training samples collected from DEAP affective computing database. One limitation of the proposed C-MCCA is the requirement of training data. More discussion will be made in next chapter regarding training and testing stage of C-MCCA.

Chapter 3

Experiments

3.1 Introduction

We evaluate our non-contact HR measurement framework using three different experiment settings: (i) well-controlled laboratory environment with stable illumination and little motion artifacts (EXP1); (ii) challenging laboratory environment where illumination variations and head motion are involved (EXP2); (iii) real-world road driving situation with strong illumination variation and head movement (EXP3). A summary of experimental data source, and HRM device involved for evaluation purpose is presented in Table 3.1. In all three settings, traditional HRM devices are used to provide heart rate ground truth for evaluation purpose. In the following sections, we give detailed introductions on these experiments including their settings, subjects involved, number of trials. Then we report experimental results of our proposed non-contact HR measurement framework on these experiments respectively. All data processing procedures are implemented in a desktop (Intel Core i7 @3.20 GHz) using MATLAB.

i Experiment under well-controlled laboratory environment using self-collected video and physiological data.

ii Experiment under challenging laboratory environment using video and physiological data from public affective computing database.

Table 3.1: A summary of experiments

	Non-contact data source	HRM device involved
EXP1	Self-collected facial video recording	Commercial finger oximeter
EXP2	DEAP affective computing database [25]	Clinical PPG sensor
EXP3	Self-collected facial video recording	Commercial chest strap

iii Experiment under difficult road-driving setting using self-collected video and physiological data.

3.2 EXP1: Laboratory Experiment

We first carried out a self-collected non-contact HR measurement experiment to study the performance of the proposed method. All recordings were conducted in an indoor laboratory environment with stable ambient light, which was the mixture of sunlight from nearby windows and fluorescent lamps from the ceiling. Subjects were asked to sit at a chair, wearing a pulse oximeter on their left index fingers. Meanwhile a consumer-level digital camera (Sony, Corp., NEX-5R, Tokyo, Japan) was used to take the video recordings at a distance of 1.5 meter. Video data were stored in a laptop for offline processing. An illustration of the system setup is shown in Fig. 3.1 with annotation. 16 subjects of both genders (5 females), various ages (22-40, avg. 27.9 y.o.), and multiple skin colors (East Asian, Semu, Caucasian) participated in the experiments. Six subjects were wearing glasses. All subjects were in good health status, and one subject was pregnant. We recorded a 60-second video for each subject in 1920×1080 resolution (raw data) and 50 frames per second (fps). Before recording, subjects were asked to manually count their carotid pulse for exactly one minute and report the result. Then they were asked to wear a pulse oximeter and sit *without intentional head movement for 60 seconds*. Video recording as well as pulse oximeter measurement began simultaneously after that. During the 60-second recording, subjects were asked to keep the body still and face the camera. Any mild facial expression was acceptable. One subject smiled occasionally and it turned out to have no salient impact on the performance of our method. Since finger motion may impact the accuracy of

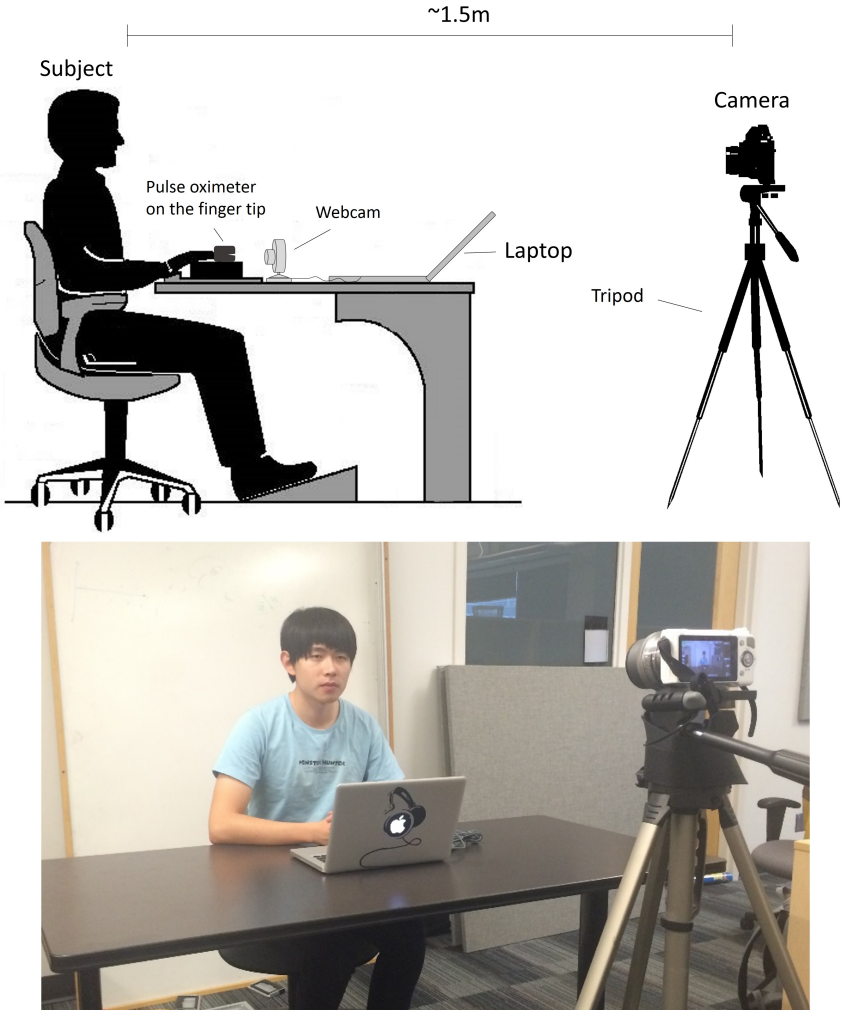


Figure 3.1: Illustration of the system setup. The pulse oximeter was slightly clamped on subject's finger tip. A webcam was programmed to take pictures of pulse oximeter's OLED screen every one second. A consumer-level digital camera recorded the subject with the support of a tripod. All drawing materials in the upper figure are from the Internet. The lower figure shows a subject is being recorded in one trial.

pulse oximeter measurement, subjects were asked to keep their finger as steady as possible during video recording. When recording was over, subjects were asked to manually count their carotid pulse again for exactly one minute and report the result. All data acquired were then analyzed by the proposed method to recover BVP signals and hereby estimate HR.

To evaluate the performance of our method, we compare the non-contact HR measurement to the readings of a pulse oximeter (Choice Electronic Technology Co., MD300C2, Beijing, China), which is officially licensed by Health Canada and FDA. Such a device is widely used in family healthcare due to its low price and portability. A probe with dedicated light source provided by LEDs is slightly clamped on human finger tip to measure pulsatile variations in the light transmitted through tissue [40]. The variations are extracted as BVP signals to estimate HR in term of beats per minute (bpm).

The pulse oximeter updates its reading every second and display it on an OLED screen. By collecting its readings during the 60-second recording, we could visualize HR variations of each subject. Since the oximeter has no data transmission module, we designed a data collection approach using an external webcam (Microsoft, Corp., Xbox Live Vision, Redmond, WA, USA) connected to the laptop, as shown in Fig. 3.2(a). Subjects were asked to place their fingers properly so that the webcam would focus on the OLED screen of the pulse oximeter, as shown in Fig. 3.2(b). We programmed the webcam to take a picture every one second, synchronizing with the oximeter. Once video recording was taken, the webcam was automatically and simultaneously activated so that the video signals were temporally aligned with the contact measurements. Fig. 3.2(c) shows five example curves of the pulse oximeter's readings. We collected video signals from 16 subjects and recovered BVP signals using our J-BSS methods. These signals were then used to estimate HR. All 16 videos were recorded in 1920×1080 resolution and 50 fps, with 60-second length for testing. For comparison use, we reimplemented the ICA-based non-contact HR measurement algorithm in [34, 35]. The experimental results of all methods on EXP1 are shown in Table 3.2. For our framework, we tested on both IVA and M-CCA algorithm in the step of J-BSS and report the results here. Several statistical measures are used for evaluation purpose, which have been used in previous research work. First we have ground truth

(a) (b) (c)

Figure 3.2: (a) Webcam is focused on the OLED screen of pulse oximeter and programmed to take pictures each second. (b) An example picture taken by the webcam in (a). (c) Smoothed 60-second samples of five subjects' oximeter readings. The average HR is shown in parenthesis.

HR_{gt} acquired from contact finger pulse oximeter by averaging over the HR reading during video recording period. For a single trial, the trial error is computed as $\mathrm{HR}_{err} = \mathrm{HR}_{nc} - \mathrm{HR}_{gt}$, where HR_{nc} denotes non-contact HR measurement result. The first two are the mean error and corresponding standard deviation of the trial error sequence with $N = 16$ elements, denoted as M_e and SD_e respectively. The root mean squared error of trial error sequence is also used, denoted as $RMSE_e$:

$$RMSE_e = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |\mathrm{HR}_{err}(n)|^2} \tag{3.1}$$

The mean error rate of trial sequence, M_r, is defined as:

$$M_r = \frac{1}{N}\sum_{n=1}^{N} \frac{|\mathrm{HR}_{err}(n)|}{\mathrm{HR}_{gt}(n)} \times 100\% \tag{3.2}$$

The Pearson's correlation coefficient between ground truth sequence and non-contact measurement sequence is also computed, denoted as r. The symbol $*$ in Table 3.2 means correlation coefficient satisfies $p < 0.01$, indicating that there are statistically significant correlations between ground truth and the non-contact method. The scatter plot of all results are shown in Fig. 3.3.

Fig. 3.3(a) used ICA based on the joint approximate diagonalization of eigenmatrices (JADE) algorithm [10], the same one used in [35]. The facial data collect

Table 3.2: Performance on EXP1 using different non-contact HR measurement methods.

Method	M_e (SD_e) (bpm)	$RMSE_e$ (bpm)	M_r (%)	r
Poh [35]	1.6195 (1.0644)	1.9196	2.4853	0.9575*
Ours (IVA)	1.3922 (1.1741)	1.7974	2.1449	0.9510*
Ours (M-CCA)	0.9514 (0.7865)	1.2186	1.4519	0.9797*

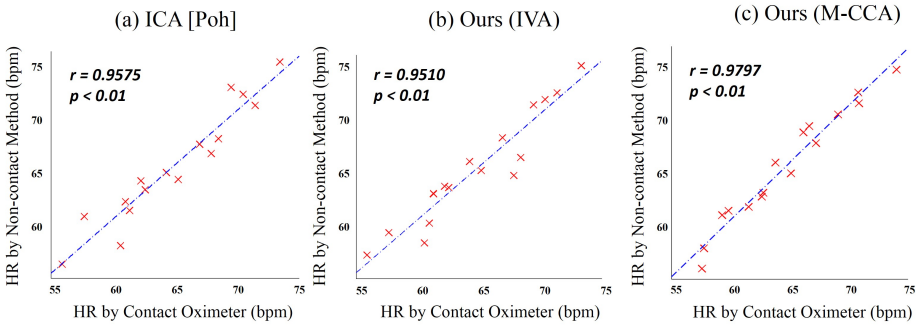


Figure 3.3: Scatter plots of three non-contact methods (a) ICA by [35] (b) ours using IVA (c) Ours using M-CCA.

method is based on Viola-Jones algorithm [46]. In Fig. 3.3(b)-(c) the same facial sub-region datasets were computed using different J-BSS methods: (b) IVA, (c) M-CCA. The proposed learning-based C-MCCA algorithm is not used in this experiment for two reasons: (i) C-MCCA requires certain amount of training samples, usually at least 30% of the trial numbers. In EXP1, only 16 trials were conducted, one for each subject. The training samples is too small in this case to learn significantly meaningful model that would actually improve the performance. (ii) As mentioned above, EXP1 has a well-controlled experiment setting, where indoor illumination is much more stable and head movement artifact is relatively small comparing to EXP2 and EXP3. In this case, all non-contact methods listed in Table 3.2 yield good performance with respect to all statistical measures. With the mean error less than 2 bpm and mean error rate less than 3%, it is safe to say that the performance of a non-contact measurement method is *similar* to that of a contact PPG sensor. It is also noted that J-BSS-based methods outperform ICA-based method with respect to mean error, RMSE, and mean error rate. In Fig 3.4, a bar

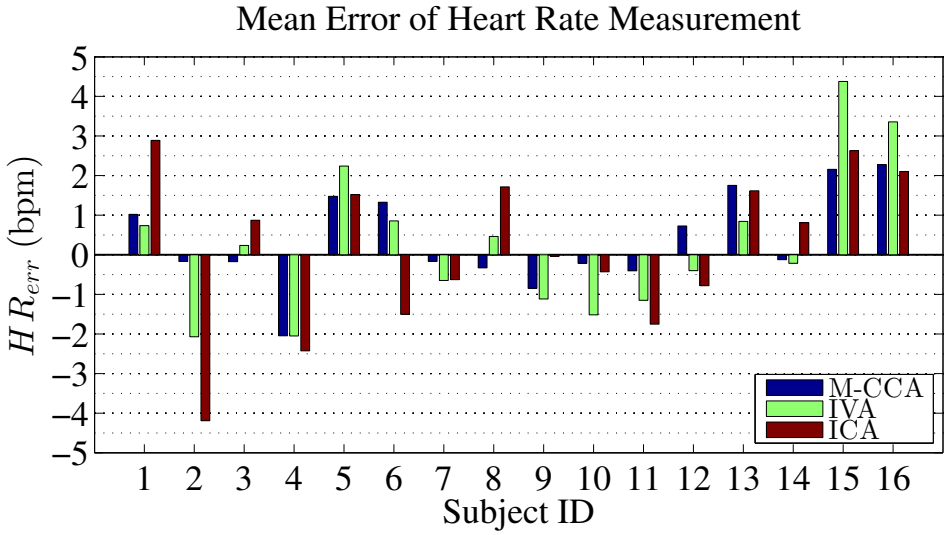


Figure 3.4: Bar plot of experimental results in EXP1.

plot of the EXP1 results are illustrated with respect to trial errors. Here we use subject IDs to index trials. It can be seen that M-CCA manages to keep all trial errors within 2 bpm, while ICA is not. Discussion on why J-BSS-based methods show more stable performance comparing to ICA-based method will be made at the end of this chapter.

3.3 EXP2: Public Database Experiment

In EXP2, we test the proposed framework using IVA, M-CCA, and C-MCCA methods on the DEAP affective computing database [25]. DEAP is a public multi-modal database for analysis of human affective states in terms of the levels in arousal, valence, like/dislike, dominance, and familiarity. It provides electroencephalography (EEG) and other peripheral physiological signals recordings of 32 participants under designated multimedia emotional stimuli (music video). Basic information of DEAP database is listed in Table 3.3. For more details, please refer to [25]. For 22 of the 32 participants, who consent to publish audio-visual recordings, their frontal face video recordings are publicly available. We test our proposed framework on these videos. Among all kinds of available peripheral physiological signal, we use

Table 3.3: Basic information of DEAP database such as subject statistics, physiological parameters

Participant Information	
Number of participants	32
Number of participants with visual consent	22 (11 females)
Number of trials	$\sim$ 40/participant
Trial duration	60 seconds/trial
Physiological parameters (downsampled to 128 Hz)	
Electroencephalography	Channel 1-32
Electrooculography	Channel 33-34
Electromyography	Channel 35-36
Galvanic skin response	Channel 37
Respiration belt feedback	Channel 38
Blood volume pulse	Channel 39

BVP (channel-39) as ground truth for evaluation purpose. In DEAP, frontal videos were recorded in DV PAL format using a SONY DCR-HC27E camcorder. They were then segmented and transcoded to 50 fps de-interlaced videos using h264 codec.

22 participants are of both genders (11 females) and various ages (19-37, avg. 26.5 y.o.). Ten of them wore glasses during the recording. Prior to the experiment, EEG and peripheral physiological sensors were placed and signals checked. The plethysmograph sensor was attached on left thumb. All participants' frontal faces were occluded by electrooculography (EOG) sensors to varying degrees, as shown in Fig. 3.5. During the experiment, they were asked to sit on a chair and watch different music videos. For each trial, a 60-second frontal face video was recorded in 720$\times$576 resolution at 50 fps, along with other physiological signals. Each participant had 40 trials. The total amount of trials (test videos) in this experiment is 874 [1]. We use BVP signals captured by the plethysmograph sensor as ground truth HR_{gt}. In DEAP's pre-processed data release, the raw BVP signal (channel 39) was downsampled to 128 Hz and segmented appropriately to temporally align with the face video recording.

[1] Due to technical issues (i.e. tape ran out), participate # 3,5,14 have only 39 face videos for each of them, and participant # 11 has only 37 face videos.

Figure 3.5: A participant's frontal face video during the experiment. Electrodes, wires, and tapes occlude parts of the facial regions.

In Section 3.1, we describe EXP2 as an experiment with *challenging* laboratory environment. The reasons are (i) Part of facial regions are occluded by electrodes, wires, and tapes, which lowers the performance of almost all facial tracking algorithms regardless whether the algorithm is tracking the entire facial region or designated facial landmarks. This may increase the rate of false detection or frame drop. (ii) As illustrated in Fig. 2.3 and Fig. 3.5, frontal facial videos of DEAP were recorded in a relatively dark environment comparing to EXP1. Participants sat in front of a color monitor, which was playing music videos. The illumination variations on participants' facial regions caused by the color monitor were large enough even to be visible by naked eyes. Such variation acts as artifacts in color channel data collection step and potentially interferes the accurate extraction of facial BVP signals [26]. (iii) In EXP2, participants can move their heads naturally without being asked to keep still. This results in motion artifacts in the collected color channel signals.

To evaluate our non-contact HR measurement framework, we first perform facial landmark localization algorithm to all EXP2 trials. Due to aforementioned

reasons, false detection and frame drop appeared in some trials. Here we pick 782 video trials[2], which do not have any frame drops, out of the total number of 874 trials. For each video trial with duration of 60 seconds, we extract the last 40 seconds and measure the non-contact HR using Equation 1.2. Their corresponding BVP signals are used as ground truth. Experimental results of EXP2 are shown in Table 3.5. The statistical measures are computed in the same way as we do in EXP1.

Table 3.4: Performance on EXP2 using different non-contact HR measurement methods (with $\delta = 0.5$)

Method	M_e (SD_e) (bpm)	$RMSE_e$ (bpm)	M_r (%)	r	Divergence
Poh [35]	7.9177 (6.3074)	10.1203	10.6647	0.4299*	12
IVA	6.7792 (5.1651)	8.5206	9.3684	0.4413*	0
M-CCA	6.7434 (5.2486)	8.5432	9.3505	0.4084*	0

Table 3.5: Performance on EXP2 using different non-contact HR measurement methods (with adaptive δ-correlation)

Method	M_e (SD_e) (bpm)	$RMSE_e$ (bpm)	M_r (%)	r	Divergence
Poh [35]	5.7089 (5.1334)	7.6752	8.5844	0.5530*	12
IVA	4.9227 (4.7683)	6.8513	7.0197	0.5964*	0
M-CCA	4.7001 (4.6633)	6.6189	6.8171	0.6358*	0
C-MCCA	**3.6554 (3.4169)**	**5.0017**	**5.1712**	**0.7423***	0

For C-MCCA, we randomly select 200 (out of 782) trials to compose training set using methods mentioned in Section 2.5. Two threshold parameters are set to the following values: $T = 3$ and $\lambda = 0.53$. Both are determined by cross validation. Comparing to the results of EXP1, performance of all non-contact methods drops significantly. This verifies that EXP2 has a much more challenging experimental setting than EXP1. Our proposed non-contact framework using IVA, M-CCA, and learning-based C-MCCA methods all outperform the ICA-based method with respect to all statistical measures. In Fig. 3.6, the error distributions of the pro-

[2]Participate #20 has frame drop problems in his every trial (Average 20 frames per trial). Although it is not a big problem given the frame per second (fps) being 50, we remove all of his trials anyway.

posed C-MCCA and the ICA-based method are compared. In general, it is clear that C-MCCA outperforms ICA. It's worth noting that ICA-based method does not converge in 12 (out of 782) trials, while our proposed methods converge in all trials. In Section 2.4, we propose an adaptive peak detection method, namely adaptive δ-correction. Unlike previous non-contact methods [30, 35] which fix δ, the proposed method uses a simple linear regression model to predict δ value from the frequency information of the recovered BVP signal. To verify its performance, we test EXP2 with two δ-correlation strategies: (i) fix $\delta = 0.5$ (ii) use adaptive δ-correlation. For the first strategy, $\delta = 0.5$ yields the best performance comparing to other fixed values. The top three rows of Table 3.4 and Table 3.5 show the experimental results for (i) and (ii). Using δ-correlation gives more than 27% performance improvement by all methods, which proves its effectiveness. We further observe that such an adaptive method helps decrease measurement bias and generally moves the error mean back to zero. In Fig. 3.7, adaptive and non-adaptive δ-correlation are compared in three different non-contact methods. With adaptive δ-correlation, the mean of error distribution tends to move towards zero. The variance of error distribution is also smaller.

Correlation coefficient r in Table 3.4 and Table 3.5 also indicates that the proposed non-contact framework has better performance. In Fig. 3.8(a), it is shown that with adaptive δ-correlation, non-contact method such as [35] is better correlated with the ground truth than the one with fixed δ. In Fig. 3.8(b), we further compare our learning-based C-MCCA with ICA-based method. In the wide range from 50 bpm to 90 bpm, C-MCCA gives good HR estimations in most cases. Outliers which fall far from the perfect correlation line (ground truth is perfectly fitted.) exist in all non-contact methods. Most of them result from participants' intensive head motion during video recordings, such as unintentionally rotating heads from left side to right side in a fast speed.

In [26, 34], the authors claim that for application scenarios such as detecting vital signs of an emergency situation, HR measurement with error less than $Theta = 5$ bpm is likely to be acceptable. Here we show how aforementioned non-contact methods perform under different Θ values. The measure we use is called

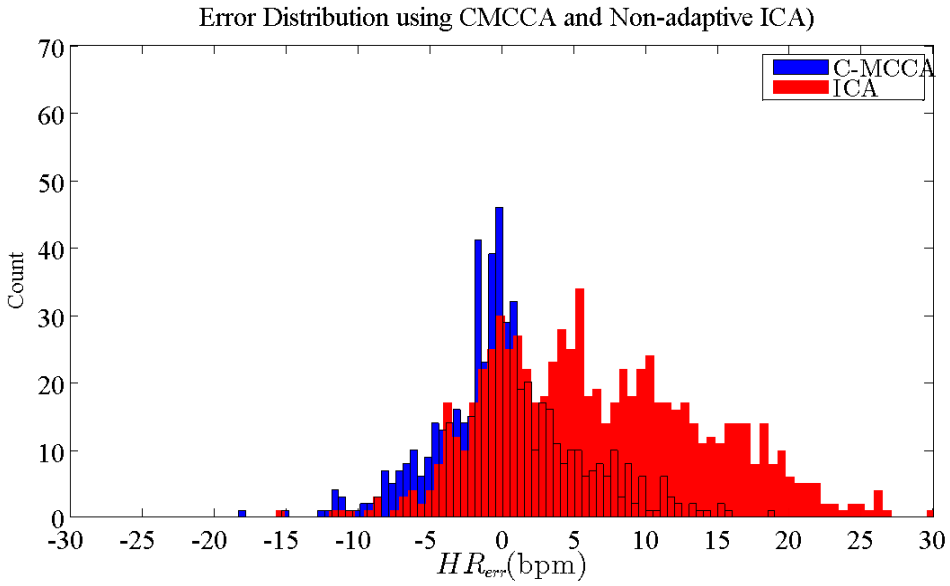


Figure 3.6: Error distribution of the proposed C-MCCA and ICA-based method [35] without adaptive δ-correlation.

acceptance rate Γ:

$$\Gamma = \frac{\text{Number of Trials s.t. abs}(\text{HR}_{nc}) < \Theta}{\text{Number of All Trials}} \times 100\% \tag{3.3}$$

where abs$(\cdot)$ denotes the absolute value. In Table 3.6, we compare the acceptance rate of different non-contact methods by changing Θ. C-MCCA manages to control the HR error under 5 bpm for more than 70% of trials.

Table 3.6: Acceptance rate using different non-contact methods

Method		$\Theta = 5$ (%)	$\Theta = 3$ (%)	$\Theta = 2$ (%)	$\Theta = 1$ (%)
Poh [35]	$\delta = 0.5$	42.08	27.27	20.78	11.69
	Adaptive δ	56.49	38.57	28.80	16.62
IVA	$\delta = 0.5$	44.76	29.41	21.61	10.87
	Adaptive δ	63.04	44.25	34.40	19.44
M-CCA	$\delta = 0.5$	47.57	30.31	21.36	11.51
	Adaptive δ	64.32	48.34	38.87	21.87
C-MCCA	Adaptive δ	**72.79**	**53.53**	**43.64**	**25.80**

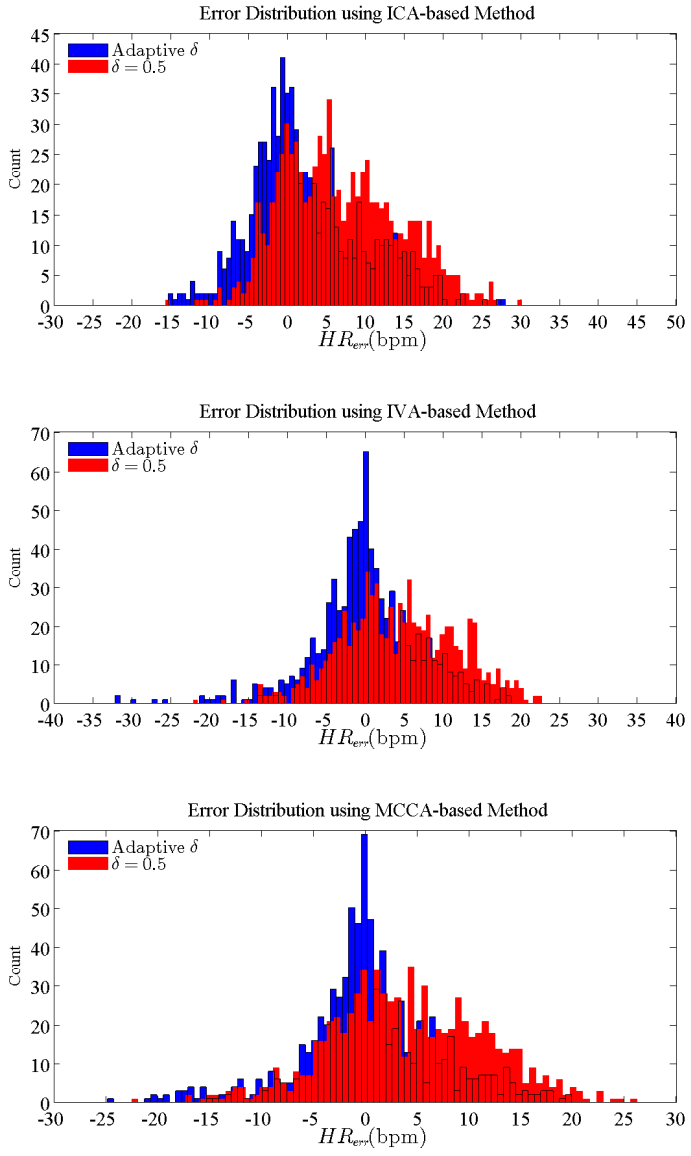
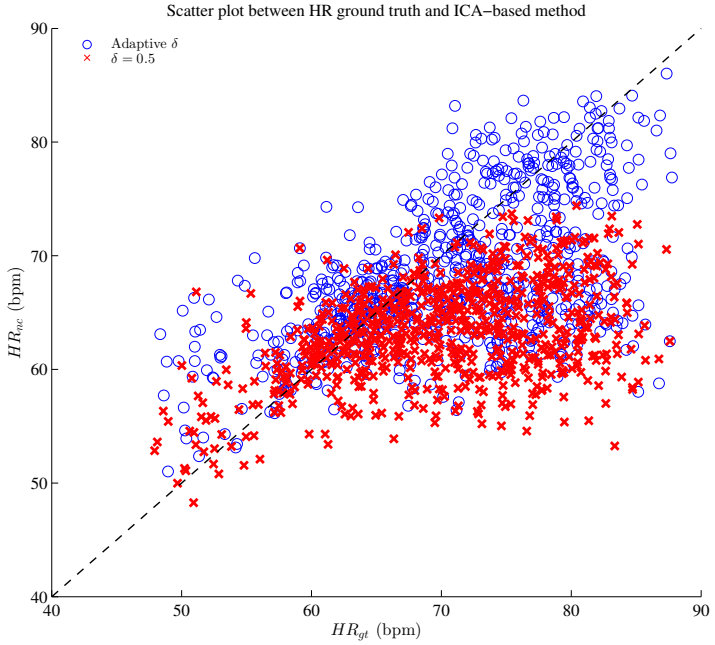
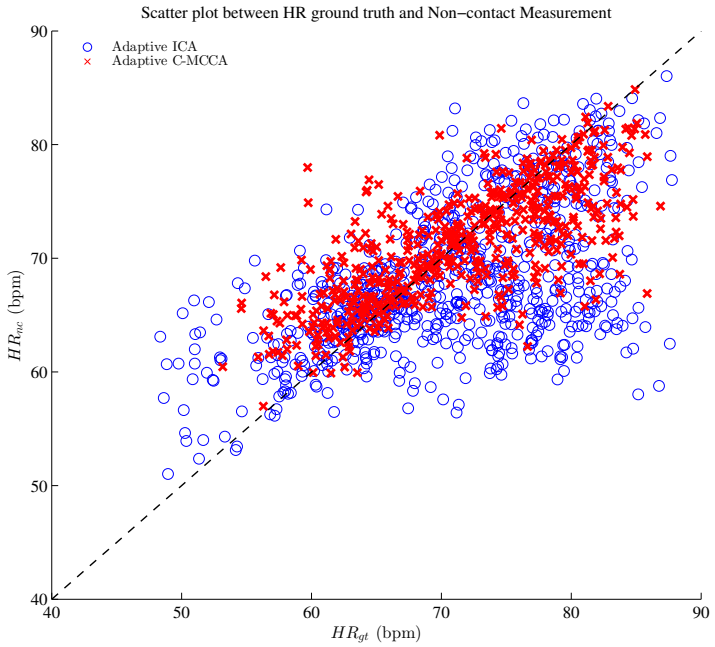


Figure 3.7: Error distribution of three non-contact methods with and without adaptive δ-correlation.



(a)



(b)

Figure 3.8: The scatter plot comparing HR_{gt} with HR_{nc} between (a) Adaptive δ-correlation and fixed δ-correlation (b) Adaptive ICA and Adaptive C-MCCA.

3.4 EXP3: Road-Driving Experiment

Official statistics have shown that driver fatigue, distraction and paroxysm account for a considerable proportion of road fatal accidents every year. Driver medical assistance in automobile environment has been considered as one of the most promising ways to effectively prevent accidents and augment intelligence in transportation systems [19, 47]. Such assistance should incorporate reliable measurements of driver's vital signals in order to depict his/her driving condition. In [19], physiological sensors were distributed in a testing vehicle as well as on driver's body. Driving experiments indicated that physiological signals such as skin conductivity can provide a metric of driver stress level and a measure of how different road and traffic conditions affect drivers. In [14], driver fatigue and distraction were combined into one superclass named driver inattention. The paper reviews that physiological signals such as EEG, ECG, EOG, surface electromyogram and PPG have been jointly used to measure driver inattention level and detect abnormal driving conditions such as arrhythmia, hypovigilance and drowsiness.

Among these physiological parameters, cardio-related parameters are of great interest. The most fundamental ones are HR and HRV. Paroxysm of cardiovascular disease is usually accompanied by early heart failure, which means heart would function decreasingly to pump blood all over the body. As a result, detectable parameter variations in autonomic nervous system occur, leading to reduced vagal-cardiac activity and enlarged sympathetic activity [13]. Studies show that these vital signs are highly associated with HR and HRV [24]. In [32], experiments under laboratory conditions were carried out to detect early onset of driver fatigue using HRV frequency-domain measures. An artificial neural network system was built up to classify between fatigue and alert condition. In [31], highway driving experiments revealed the high reliability of HR and HRV measures in distinguishing between single task driving and low/high cognitive workload driving. Therefore, it is reasonable to believe that driver medical assistance system with constant monitoring of driver's vital signals, especially cardiac parameters, can provide physiological evidence to reflect his/her driving conditions.

In EXP3, we test our proposed non-contact HR measurement framework on a real-world road driving experimental setting. As shown in Fig. 3.9, a commercial-

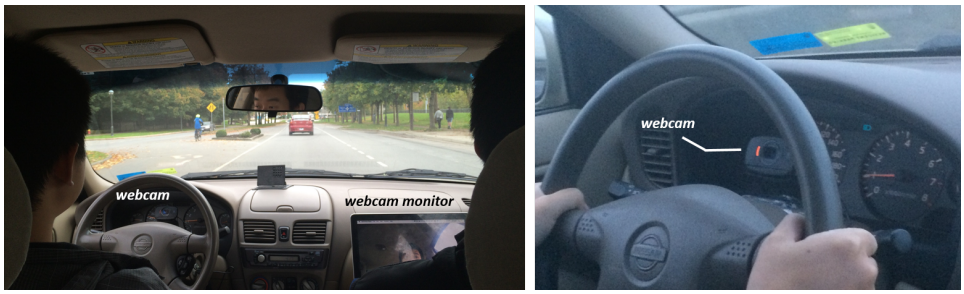


Figure 3.9: Road driving experiment (EXP3) setting-up. In the left figure, we show that the webcam is placed behind the wheel and a laptop is used to monitor the video recording. In the right figure, a zoom-in picture is provided.

level webcam is installed behind the wheel and a laptop is used to monitor the video recording. A zoom-in picture is also provided. Two examples of HRV analysis are shown in Fig. 3.10. The top row show results from the laboratory experiment and the bottom are from real road driving experiment. First, subjects' facial BVP signals were extracted with the real-time facial landmark detector and C-MCCA method. The locations of heart beat peaks were then determined in order to acquire IBI series for further analysis. Finally, we calculated different measures to analyze IBI series. We compute both time and frequency domain measures and draw corresponding LS-Periodogram as well as spectrogram based on Lomb-Scargle method (LS-Spectrogram). Time-domain measures include mean HR during the recording, SDNN and RMSSD for IBI series. In the frequency domain, we use LS-Periodogram to estimate PSD and compute LF power, HF power and their ratio LF/HF. Our non-contact framework provides a promising solution to driver physiological monitoring. More advanced method that considers illumination variation may further improve the performance, which is beyond the scope of this thesis.

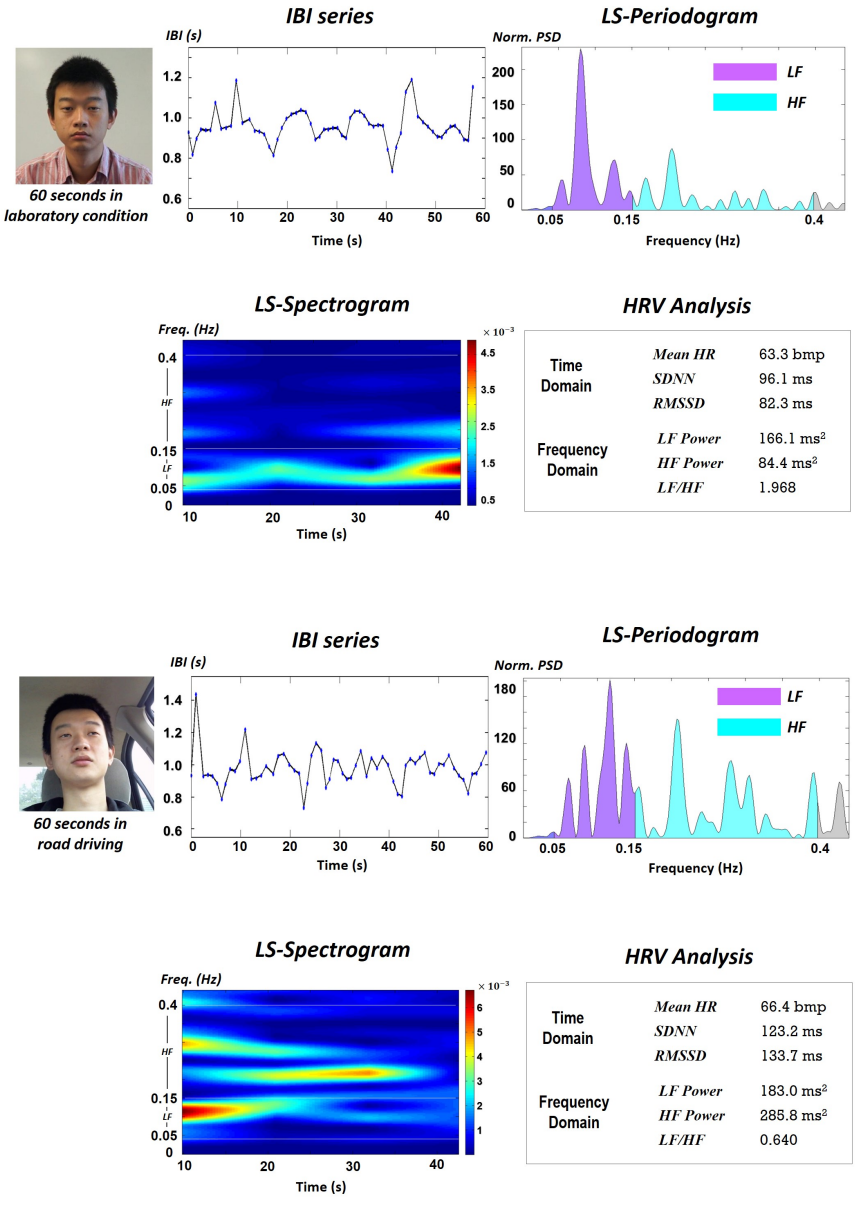


Figure 3.10: HRV analysis examples. The top row is from the laboratory setting. The bottom row is from the real road driving experiment. Six measures in time and frequency domains are computed based on IBI series. LS-Periodogram and LS-Spectrogram are also given.

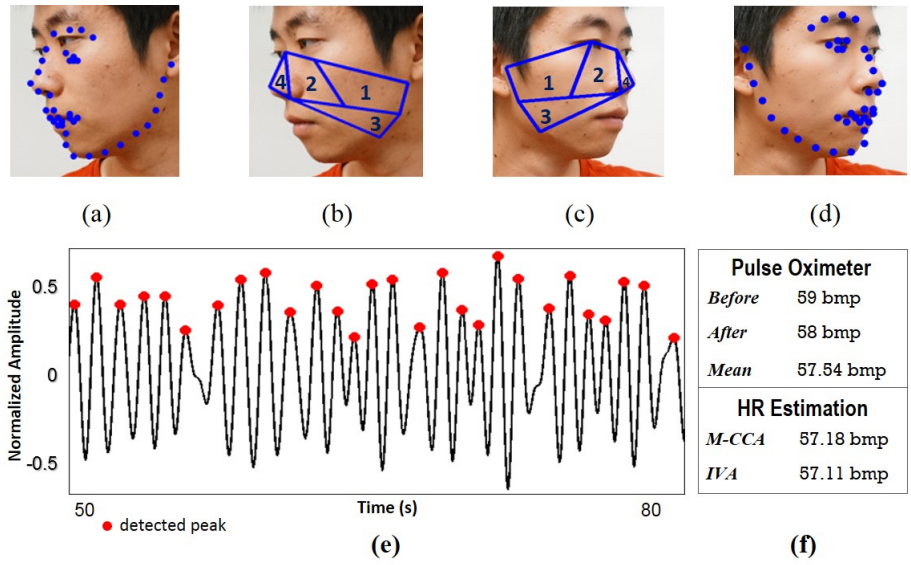


Figure 3.11: (a)-(d) Division pattern for profiles. (e) Part of recovered BVP signal with detected peaks. (f) Readings from pulse oximeter and HR estimates using M-CCA and IVA.

3.5 Discussion

3.5.1 HR Estimation Using Side Profile

The idea of facial sub-region division can lead to some interesting applications. With landmark localization techniques such as [48], we can estimate HR using subjects' profile alone. Fig. 3.11(a)-(d) show two new patterns to acquire facial sub-region datasets, corresponding to left and right side profiles. Fig. 3.11(e) displayed the BVP signal acquired from one subject's left profile. The proposed method can achieve desirable HR performance, as shown in Fig. 3.11(f).

3.5.2 Dynamic HR Estimation

In EXP1, we take the entire 60-second video signals into account. In Fig. 3.12, we show one subject's smoothed HR curve according to readings of the pulse oximeter (black dash line). In order to capture HR variation during recordings, we introduce a sliding window with size τ_w and a 95% overlap. HR at time t was estimated

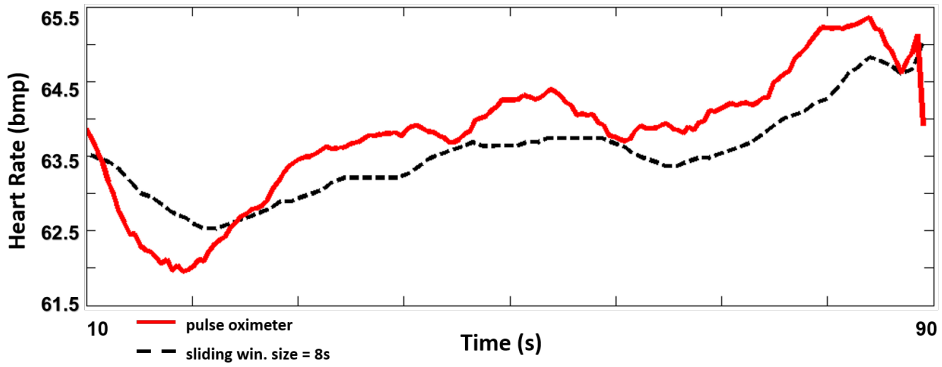


Figure 3.12: Black dash line reflects one subject's HR variation during the recording with sampling rate 1Hz. The red line is the HR estimate based on a slide window of past 10 seconds and a 95% overlap. Both curves were smoothed by moving average method with span 20.

using video signals during the time period $(t-\tau_w, t)$. The red line in Fig. 3.12 was computed with $\tau_w = 10s$. We can see that, when compared with results of pulse oximeter, our approach generally reflects the subject's HR variation with a slight bias towards higher HR.

3.5.3 Performance Analysis

One common step of all non-contact approaches is facial data collection. Traditional ICA-based frameworks [34, 35] used a fast face detector [46] to localize the entire face. It sacrifices accuracy for speed. A considerable portion of non-PPG instances (e.g., hair, recording background) are included and uniformly averaged to get color channel data. Our approach, on the other hand, gets rid of almost all unrelated instances and focuses only on facial regions. Our division patterns, as shown in Fig. 2.2, avoid mouth (useless if opened or half-opened), forehead (useless if there are bangs), and chin (useless if there are beards).

Besides improved facial data collection method, the introduction of J-BSS also makes non-contact HR measurement more robust than ICA-based approaches. Many practical issues, such as ambient light changes or shadow caused by facial expression variation, might cause fluctuation of local color channel values and thus influence the accuracy of BVP signal recovery. We cannot solve this problem by

designing better facial data collection method. However, with J-BSS we can reduce such negative impacts. Given color channel data of different facial sub-regions, J-BSS methods attempt to recover the underlying source set for every sub-region dataset. An important assumption in our proposed method is that the BVP signal is the shared source among all datasets. Results of spectral clustering support this assumption. The largest cluster actually contains candidate BVP signals recovered from different sub-regions. Then the signal with strongest frequency component is selected among all candidate BVP signals. Therefore even if all sub-region datasets are contaminated, to various degrees, by the local fluctuation, we could still recover the one with the minimal impact. In summary, landmark-based facial data collection and J-BSS-based BVP signal extraction together contribute to the better performance of the proposed method, compared to ICA-based approaches.

Chapter 4

Conclusion and Future Work

4.1 Conclusion and Contribution

In this thesis, we proposed a novel framework for non-contact HR measurement by exploiting correlations between different facial sub-regions to enhance the robustness of the measurements when illumination variation and head motions are involved. We tested the proposed framework on three experimental settings: (i) a well-controlled laboratory environment, (ii) a more challenging laboratory environment, and (iii) a real-world road driving environment. Results show that the proposed non-contact framework can be a promising solution to both clinical diagnosis and family healthcare.

In Chapter 2, we presented the proposed non-contact HR measurement framework step by step. Starting from facial data collection, we proposed using an advanced facial landmark localization algorithm that gives real-time tracking even when the video fps is as high as 50. The algorithm returns physical coordinates of 49 facial landmarks. We designed a specific facial division pattern based on these coordinates in order to divide the entire facial region into four sub-regions. Compared with previously used face detectors, this algorithm is robust to various intensive head motions, as demonstrated in subsequent experiments. Based on this facial division pattern, we collected four sets of facial color channel signals and fed them into a joint blind source separation (J-BSS) system to extract latent facial BVP signals. Two recently proposed J-BSS algorithms, IVA and M-CCA, were

used respectively in our framework. Using post-processing methods such as signal detrending, temporal filtering, and spectral clustering, we recovered BVP signals from these sub-regions. We further proposed an adaptive peak detection method that uses frequency information of the BVP signal to better detect heart interbeat. It is named adaptive δ-correlation. Based on M-CCA, we proposed a learning-based J-BSS algorithm, connectivity multi-set canonical correlation (C-MCCA), in order to improve the HR measurement performance. A max-margin multi-label (M3L) classification algorithm is used to learn the optimal connectivity design matrix (CDM) from a certain amount of training samples.

In Chapter 3, we designed three types of experiments to test the proposed non-contact HR measurement framework. In a well-controlled experimental setting (EXP1), 16 subjects were included. Experimental results show that all non-contact methods yield good performance. Our proposed framework works a little better than the ICA-based method. In a much more challenging setting (EXP2), we tested our framework on the DEAP affective computing public database. Random illumination variation and head motions are included in all video recordings, and general performance degradation is observed in our tested non-contact methods. We compared these methods when using a fixed δ correlation and the proposed adaptive δ-correlation respectively. Results showed that the adaptive method can improve the performance by more than 27 % for all tested methods. We randomly collected a training set with 200 (out of 782) trials and tested on the rest. We note that the proposed C-MCCA method yields the best performance with respect to tested statistical measures such as the mean error, RMSE and mean error rate. The scatter plot of correlations between non-contact HR measurements and contact BVP (ground truth) and the acceptance rate analysis in Section 3.3 further demonstrated the effectiveness of the proposed adaptive δ-correlation and C-MCCA method. Generally, experimental results indicate high consistency between traditional contact PPG-sensors and the proposed non-contact methods. In a real-world road driving setting (EXP3), we tested the proposed HR measurement framework by installing a commercial webcam behind the wheel. Experimental results showed that such a non-contact driver monitoring method can be a promising solution to both HR measurement and HRV analysis in both time and frequency domains. We also illustrated that the proposed method can work well when given video signals

of subjects' facial profiles only.

4.2 Future Work

In this thesis, since most previously proposed non-contact HR measurement methods use the entire facial regions to collect signals, we extend the idea by dividing the facial regions into multiple sub-regions which is verified to be more effective to extract facial BVP signals for HR measurement. However, it is just one way to enhance the robustness of non-contact techniques and several limitations are still associated with the non-contact framework. The most important limitation of non-contact measurement is still related to illumination variation and head motion artifacts. In this thesis, attempts have been made to improve the performance of non-contact HR measurement under a challenging laboratory setting. In more severe settings, such as the road driving experiment (EXP3), non-contact HR measurement cannot perform as well as it does in the laboratory due to the intensive illumination variation and head motion artifacts. Our efforts in EXP3 show some promising results and more work will be done in the future to further improve the robustness. Moreover, in the end of Chapter 3, we propose to measure HR using only facial profile. It is natural to extend the current method into a full viewpoint measurement framework.

In this thesis, we focus on video-based non-contact physiological measurement. Specifically, we select heart rate and heart rate variability as research targets. In the future, more physiological parameters will be investigated. Non-contact technique can be a promising solution to both clinical diagnosis and family healthcare. How to achieve non-contact measurement in an both accurate and robust way is a promising research direction that we intend to follow in the future.

Bibliography

[1] L. A. 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 unita pilot study. *Early Hum. Dev.*, 89(12):943–948, 2013. → pages 3, 4

[2] J. Achten and A. E. Jeukendrup. Heart rate monitoring. *Sports medicine*, 33 (7):517–538, 2003. → pages 1

[3] M. Anderson, T. Adali, and X.-L. Li. Joint blind source separation with multivariate gaussian model: algorithms and performance analysis. *IEEE Trans. Signal Process.*, 60(4):1672–1683, 2012. → pages 16, 17

[4] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014. → pages viii, 12, 13, 30

[5] S. Bakhtiari, T. W. Elmer, N. M. Cox, N. Gopalsami, A. C. Raptis, S. Liao, I. Mikhelson, and A. V. Sahakian. Compact millimeter-wave sensor for remote monitoring of vital signs. *IEEE Trans. Instrum. Meas.*, 61(3): 830–841, 2012. → pages 7

[6] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3430–3437, June 2013. → pages 7

[7] G. G. Berntson, J. T. Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648, 1997. → pages 7

[8] F. Bousefsaf, C. Maaoui, and A. Pruski. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous

heart rate. *Biomedical Signal Processing and Control*, 8(6):568–574, 2013. → pages 7

[9] V. D. Calhoun, J. Liu, and T. Adalı. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1): S163–S172, 2009. → pages 16

[10] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157–192, 1999. → pages 36

[11] X. Chen, C. He, Z. J. Wang, and M. J. McKeown. An ic-pls framework for group corticomuscular coupling analysis. *IEEE Trans. Biomed. Eng.*, 60(7): 2022–2033, 2013. → pages 16

[12] W. W. Chin. The partial least squares approach to structural equation modeling. *Modern Method for Business Research*, 295(2):295–336, 1998. → pages 24

[13] M. M. J. De Jong and D. C. Randall. Heart rate variability analysis in the assessment of autonomic function in heart failure. *Journal of Cardiovascular Nursing*, 20(3):186–195, 2005. → pages 46

[14] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *Intelligent Transportation Systems, IEEE Transactions on*, 12(2):596–614, 2011. → pages 4, 46

[15] K. Fox, J. S. Borer, A. J. Camm, N. Danchin, R. Ferrari, J. L. L. Sendon, P. G. Steg, J.-C. Tardif, L. Tavazzi, and M. Tendera. Resting heart rate in cardiovascular disease. *J. Am. Coll. Cardiol.*, 50(9):823–830, 2007. → pages 1, 4

[16] M. Garbey, N. Sun, A. Merla, and I. Pavlidis. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Trans. Biomed. Eng.*, 54(8):1418–1426, 2007. → pages 7

[17] E. Greneker. Radar sensing of heartbeat and respiration at a distance with applications of the technology. *Proc. Conf. RADAR*, 1997. → pages 7

[18] B. Hariharan, L. Zelnik-Manor, M. Varma, and S. Vishwanathan. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 423–430, 2010. → pages ix, 25, 26, 29

[19] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):156–166, 2005. → pages 46

[20] J. W. Hurst. Naming of the waves in the ecg, with a brief account of their genesis. *Circulation*, 98(18):1937–1942, 1998. → pages 5

[21] X. Jiang, M. Dawood, F. Gigengack, B. Risse, S. Schmid, D. Tenbrinck, and K. Schäfers. Biomedical imaging: A computer vision perspective. In *Computer Analysis of Images and Patterns*, pages 1–19. Springer, 2013. → pages 3

[22] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. → pages 17, 25

[23] T. Kim, T. Eltoft, and T.-W. Lee. Independent vector analysis: An extension of ica to multivariate components. In *Proc. Independent Component Analysis and Blind Signal Separation*. Springer, 2006. → pages 16

[24] R. E. Kleiger, J. P. Miller, J. T. Bigger Jr, and A. J. Moss. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American journal of cardiology*, 59(4):256–262, 1987. → pages 46

[25] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on*, 3 (1):18–31, 2012. → pages viii, 5, 14, 15, 22, 33, 38

[26] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4264–4271. IEEE, 2014. → pages 7, 8, 40, 42

[27] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Trans. Signal Process.*, 57(10):3918–3929, 2009. → pages 17, 24, 25

[28] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, and V. Calhoun. Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Hum. Brain Mapp.*, 30(1):241–255, 2009. → pages 16

[29] H. Lu, Y. Pan, B. Mandal, H.-L. Eng, C. Guan, and D. W. Chan. Quantifying limb movements in epileptic seizures through color-based video analysis. *IEEE Trans. Biomed. Eng.*, 60(2):461–469, 2013. → pages 4

[30] D. McDuff, S. Gontarek, and R. W. Picard. Improvements in remote cardio-pulmonary measurement using a five band digital camera. *IEEE Trans. Biomed. Eng.*, 61(10):2593–2601, 2014. → pages 7, 8, 20, 21, 42

[31] B. Mehler, B. Reimer, and Y. Wang. A comparison of heart rate and heart rate variability indices in distinguishing single task driving and driving under secondary cognitive workload. In *Proc Driving Symposium on Human Factors in Driver Assessment, Training & Vehicle Design*, pages 590–597, 2011. → pages 46

[32] M. Patel, S. Lal, D. Kavanagh, and P. Rossiter. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Systems with Applications*, 38(6):7235–7242, 2011. → pages 46

[33] Philips. Vital signs camera, 2011. → pages 4

[34] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, 2010. → pages 7, 20, 35, 42, 50

[35] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. Biomed. Eng.*, 58(1):7–11, 2011. → pages x, 7, 20, 35, 36, 37, 41, 42, 43, 50

[36] J. Rashmur. Design, evaluation, and application of heart rate variability analysis software (hrvas). 2010. → pages 7

[37] C. Scully, J. Lee, J. Meyer, A. M. Gorbach, D. Granquist-Fraser, Y. Mendelson, and K. H. Chon. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Trans. Biomed. Eng.*, 59(2): 303–306, 2012. → pages 4

[38] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. → pages ix, 20, 21

[39] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1):42–55, 2012. → pages

[40] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. Clifton, and C. Pugh. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physiol. Meas.*, 35(5):807–831, 2014. → pages 3, 4, 6, 35

[41] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen. An advanced detrending method with application to hrv analysis. *IEEE Trans. Biomed. Eng.*, 49(2):172–175, 2002. → pages 19

[42] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011. → pages 24

[43] J. E. Thatcher, K. D. Plant, D. R. King, K. L. Block, W. Fan, and J. M. DiMaio. Dynamic tissue phantoms and their use in assessment of a noninvasive optical plethysmography imaging device. In *Proc. SPIE Sensing Technology+ Applications*, 2014. → pages 4

[44] S. S. Ulyanov and V. V. Tuchin. Pulse-wave monitoring by means of focused laser beams scattered by skin surface and membranes. In *OE/LASE'93: Optics, Electro-Optics, & Laser Applications in Science& Engineering*. International Society for Optics and Photonics, 1993. → pages 7

[45] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Opt. Express*, 16(26):21434–21445, 2008. → pages 7

[46] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition (CVPR), 2001 IEEE Conference on*, pages I511–I518, December 2001. → pages 37, 50

[47] T. Wartzek, B. Eilebrecht, J. Lem, H.-J. Lindner, S. Leonhardt, and M. Walter. Ecg on the road: robust and unobtrusive estimation of heart rate. *Biomedical Engineering, IEEE Transactions on*, 58(11):3112–3120, 2011. → pages 4, 46

[48] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886, June 2012. → pages 49