Stress and Emotion Recognition based on Remote Photoplethysmography

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Elektrotechnik und Informationstechnik
des Karlsruher Instituts für Technologie (KIT)

angenommene

Dissertation

von

M.Sc.

Kai Zhou

aus Wuhan

Tag der mündlichen Prüfung:	22. 04. 2024
Hauptreferent:	Prof. Dr. rer. nat. Wilhelm Stork
Korreferent:	Prof. Dr. rer. nat. Werner Nahm

Abstract

In this dissertation, the measurement principle of remote Photoplethysmography (rPPG) is discussed in detail and on this basis, a system for rPPG-based stress and emotion recognition is designed. The focus of this work is on the design of algorithmic approaches.

Within this scope, the work introduces a deep learning-based core algorithm for rPPG. The design of the algorithm considers signal processing from both temporal and spatial dimensions. The algorithm is evaluated on multiple benchmark datasets in various measurement setups with different evaluation metrics, demonstrating the state-of-the-art performance of the algorithm.

Furthermore, this work delves into the exploration of methods for stress and emotion recognition based on the camera-derived measurements. For stress recognition, the investigation prioritizes the feasibility of end-to-end methods, while the focus for emotion recognition is on the automatic assessment using a dimensional model for emotion description.

Kurzfassung

In dieser Dissertation wird das Messprinzip der remote Photoplethysmography (rPPG) detailliert diskutiert und auf dieser Grundlage wird ein System zur Stress- und Emotionserkennung auf Basis von rPPG entworfen. Der Fokus der Dissertation liegt auf dem Entwurf algorithmischer Ansätze.

In diesem Rahmen wird ein auf Deep Learning basierender Kernalgorithmus für rPPG vorgestellt. Das Design des Algorithmus berücksichtigt die Signalverarbeitung sowohl aus zeitlicher als auch aus räumlicher Dimension. Der Algorithmus wird anhand mehrerer Benchmark-Datensätze in verschiedenen Messaufbauten mit unterschiedlichen Metriken bewertet, was seine Performance von Stand der Technik unter Beweis stellt.

Darüber hinaus vertieft diese Arbeit die Erforschung von Methoden zur Stress- und Emotionserkennung, die auf kamerabasierten Messungen basieren. Der Teil der Stresserkennung priorisiert die Untersuchung der Machbarkeit von End-to-End-Methoden, während der Teil der Emotionserkennung die Präzision der Erkennung unter Verwendung eines dimensionalen Modells zur Emotionsbeschreibung betont.

Danksagung

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am FZI Forschungszentrum Informatik im Forschungsbereich Embedded Systems and Sensors Engineering (ESS).

Ich bedanke mich besonders bei Herrn Prof. Dr. rer. nat. Wilhelm Stork vom Institut für Technik der Informationsverarbeitung (ITIV) für die Möglichkeit der Promotion und der Betreuung dieser Arbeit. Es war eine große Ehre für mich, als Doktorand bei Prof. Stork zu promovieren. Seine Hilfe und Ratschläge in akademischen und beruflichen Angelegenheiten werden mich nachhaltig prägen. Prof. Dr. rer. nat. Werner Nahm vom Institut für Biomedizinische Technik (IBT) danke ich für die Übernahme des Korreferats und das damit verbundene Engagement.

Für die Zusammenarbeiten im Projekt PulsCam danke ich Timon Blöcher und Simon Krause. Es war definitiv eine der besten Zeiten in FZI für mich. Ein großer Dank gilt auch den Kollegen und Freunden Christoph Zimmermann, Johannes Schneider, Friedrich Gauger, Lukas Kohout, Markus Schinle, Jennifer Zeilfelder, Marc Schroth, Matthias Diehl, Christina Erler, Oliver Werthwein und Lara Schweickart. Ich danke auch Gergely Biri und Julia Hofmann für das Korrekturlesen.

Zuletzt danke ich meinen Eltern und meiner Schwester für ihre Liebe und Zuwendung aus der Ferne über die Ozeane hinweg in den letzten zehn Jahren.

Contents

Notation

General notation

Scalars	Italic Roman and Greek letters	I, α
Vectors	Bold Roman lowercase letters	$\mathbf{t}$
Matrices	Bold Roman uppercase letters	$\mathbf{R}$
Sets	Blackboard bold uppercase letters	$\mathbb{V}$

In multidimensional sets of elements related to time series, the first index denotes time.

Numbers and indexing

t	Discrete time point
c	Index of color channel

Variables, functions

l	Short signal length
T	Video sequence length
$\mathbf{s}$	Signal trace obtained by color projection
$\mathbf{h}$	Short pulse signal output by core rPPG function
$\mathbf{H}$	Long pulse signal obtained by stacking the short signals via overlap-adding
$\mathbf{I}$	Input face image
$\mathbf{T}$	Triangulated face region
$\mathbf{C}$	Color signal
$\tilde{\mathbf{C}}$	Temporally normalized color signal
$\mathbf{L}$	Temporal sampling matrix
$\mathbf{p}$	Vector in the color space for the linear signal projection
$\boldsymbol{\alpha}_E, \boldsymbol{\alpha}_D$	Parameters of encoder, decoder
x, y, z	Positions in the face model coordinate system
u, v	Positions in the image coordinate system
H, W	Height and width of the normalized image
$\mathbf{N}$	Normalized face image
$\tilde{\mathbf{M}}$	Normalized face image sequence
$\mathbf{PPG}$	Reference PPG signal
ρ	Pearson correlation coefficient

Operators and Functions

$\mu(\cdot)$	Mean operator
$\sigma(\cdot)$	Standard deviation operator
$F(\cdot)$	Core rPPG function to map color signals to a short pulse signal
$L(\cdot)$	Loss function for network training
$FA(\cdot)$	Face alignment

1

Introduction

Over the past few decades, significant advancements in microelectronics and sensing technologies have led to the development of more compact and sophisticated measurement devices for physiological data. These advancements not only greatly enhance the accuracy and convenience of the measurement, but have also prompted the transition of these devices from clinical environments into everyday settings. Alongside these advancements, physiological parameter measurements have been integrated into wearable devices such as smartwatches or smart patches. These devices enable the continuous measurement of physiological data in daily life, which can provide more comprehensive information about an individual's physical condition. As a result, the objectives behind measuring these physiological parameters have evolved. These measurements are now not limited to clinical scenarios of disease treatment and monitoring, but also include the management of an individual's physical, mental and emotional health in everyday life.

Furthermore, non-contact measurement for physiological parameters are actively being researched and developed, utilizing technologies such as WiFi [Wan17a, Kha21, Gu17], radars [Car71, Zha18], ultrasound [Nan15, Wan18b, Wan21], thermal cameras [Sha12], and remote Photoplethysmography. Non-contact measurements for physiological parameters allow for the information extraction without requiring close proximity to the subject or constraining their activity. Among these methods, remote Photoplethysmography (rPPG) stands out as one of the most extensively studied techniques. It utilizes a camera to measure physiological parameters such as Pulse Rate (PR), Pulse Rate Variability (PRV), Respiratory Rate (RR), blood oxygen saturation (SpO2) or even Blood Pressure (BP). This versatile technique finds its applications in the physical monitoring across various scenarios. It has been utilized in neonatal care [Kev17]. In daily life, it can be used not only for fitness monitoring, as described in [Wan16b], but also for assessing the physical condition of drivers, as studied in [Blö17] and [Wu17]. In the context of the COVID-19 pandemic, the increased utilization of telehealth has also underscored the significance of remote physiological measurement techniques. By providing healthcare professionals with flexible and efficient access to patients' physiological information, rPPG technology could play a vital role in supporting the effectiveness of telehealth systems, thereby contributing to the improvement of healthcare services.

Besides monitoring of physical status, the application of rPPG technology has ushered in new opportunities in the realm of affective computing. It offers more adaptive and user-friendly methods for assessing stress and emotions. This potential of rPPG in affective computing, specifically in stress and emotion recognition, will be the central focus of this dissertation.

1.1 Motivation

Recognition of stress and emotion status holds potential implications across various domains. Decades of research have demonstrated the consistent and compelling relationship of stress and negative emotions with disease risk and mortality. It has been proven that stress and negative emotions are significantly associated with depression, hypertension, cardiovascular events and metabolic disorders [Cro20, Alb16].

Mental illnesses add significant burdens on the healthcare system. Based on data from the German Federal Statistical Office [Bun23], Germany allocated up to 9.45 billion Euros for addressing depression-related illnesses in 2020. Mental illness and behavioral disorders constitute a major portion, accounting for 13% of overall healthcare spending, equivalent to cardiovascular diseases, as shown in Figure 1.1. These staggering figures highlight the urgent need for reliable and efficient tools for stress and emotional state assessment. The development of such tools should be able to enhance diagnostic accuracy of mental health issues, improve the efficacy of therapeutic interventions, and, ultimately, elevate the quality of care provided in mental health settings.

Moreover, as a vital task in the field of affective computing, stress and emotion recognition facilitates the creation of user-friendly human-machine interface (HMI), which can seamlessly integrate into diverse application environments such as ambient assisted living [Pud19], education [Con18] and robotics [San17].

Stress and emotion status manifest through various avenues:

- Affective perception
- Affective behavior
- Physiological reactions

While affective perception focuses on the emotional state an individual consciously acknowledges and is often gauged through self-reported surveys or questionnaires, objective evaluations of a person's affective state are derived

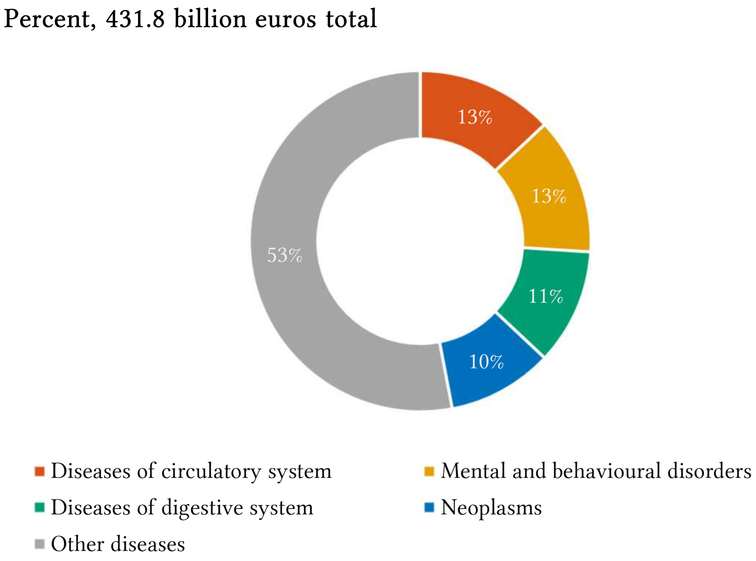


Figure 1.1: Cost of illness 2020 in Germany. Source: [Bun23].

from either observing the outward expressions characterized as affective behavior or by analyzing the inherent physiological reactions in response to emotional stimuli.

One of the traditional methods for objectively assessing affective states is through the measurement of cortisol level [Gri06], typically conducted via sample tests (blood, saliva, urine, hair, etc). Since real-time analysis of these samples is currently difficult, this method is not suitable for applications where low latency in the feedback process is required.

Presently, research in automatic stress and emotion recognition using physiological reactions predominantly relies on contact sensors such as Electroencephalogram (EEG) [Liu17, Meh17, Lu15, Qin19, Sub21, Zha20b], Electrocardiography (ECG) [Bug17, Suz21, Nar15], or Galvanic Skin Response (GSR)

[Mar13, Shu19, Udo17]. These sensors measure parameters such as brain activity, heart rates, and sweat levels, respectively. Given that they require direct physical contact with the skin, their deployment for daily use or large-scale monitoring poses significant challenges. In recent years, the use of cameras for stress or emotion measurement has emerged as a vibrant area of research, with behavior analysis such as pose recognition or facial expression recognition standing as one of the most investigated methodologies. However, affective behaviors can be influenced by both conscious and subconscious manipulations, potentially leading to a discrepancy between displayed emotions and the true affective state of an individual.

Against this backdrop, remote Photoplethysmography technology presents a novel possibility for assessing individuals' affective states by measuring physiological parameters via cameras, a process that is not under conscious control. One notable advantage of rPPG lies in its straightforward measurement setup, requiring only a camera, a light source, and a processing unit. This stands in contrast to other techniques like EEG and ECG, which often necessitate complex sensor placement procedures and expert assistance from medical professionals. Additionally, the contactless nature of rPPG measurements eliminates any discomfort associated with sensor attachment, making it particularly beneficial for long-term continuous measurement.

This dissertation focuses on investigating the potential of the rPPG technology in the capture and evaluation of affective states of individuals. Leveraging the inherent advantages of rPPG, this research seeks to contribute to the development of accurate, flexible and reliable methods for automatic stress and emotion recognition.

1.2 Research objectives

This dissertation aims to make a contribution to the development of a measurement system that enables stress and emotion state monitoring using remote Photoplethysmography. The system captures Blood Volume Pulse (BVP) signals and vital parameters from designated facial regions using a camera.

Then, based on the extracted information, it assesses the affective condition of the subjects being measured. Within the scope of this research, a new algorithm for rPPG will be proposed. Given the remarkable performance of deep learning in various computer vision and signal processing tasks, this dissertation will try to utilize deep learning for camera-based pulse signal extraction. Thus, the following question should be discussed:

- What advantages do neural networks offer compared to traditional methods for rPPG signal extraction?
- How can the strengths of neural networks be combined with those of traditional methods, and what signal quality can be achieved compared to state-of-the-art rPPG algorithms?

To ensure scalability and cost-effectiveness, a low-cost webcam was selected as the primary measurement sensor. The focus of the algorithm development is to answer the following question:

- What measurement accuracy can be achieved for vital parameters such as pulse rate and pulse rate variability using a low-cost camera? Is this level of accuracy sufficient for stress and emotion recognition?

Based on these considerations, the evaluation will investigate the following questions within two application scenarios:

- What level of accuracy can be attained for stress recognition using rPPG from a low-cost camera?
- Is it feasible to recognize dimensional emotions using rPPG? How does the quality of the measurement compare to that of facial expression analysis?

2

Basics

This chapter presents the fundamental principles necessary for a comprehensive understanding of this dissertation. It starts with an exploration of the descriptive models of stress and emotion. This is followed by an in-depth look at the cardiovascular system, covering the anatomical description of the heart and vascular system, the physiology underlying cardiac conduction, and the derivation of blood volume pulse along with crucial parameters like heart rate and heart rate variability. Following this, properties of the skin tissue are introduced, including its structure, the vascular circulation of the face skin, the blood flow regulation and properties of skin colors. Finally, the chapter introduces the underlying principle of Photoplethysmography (PPG), coupled with an introduction to the remote Photoplethysmography measurement model.

2.1 Descriptive models for stress and emotion recognition

Recognition of stress and emotion is a subset of the larger field of affective computing. For the recognition task, a descriptive model is required to define the target affective characteristics that the system is intended to identify and categorize. The following sections will discuss the descriptive models for emotion and stress recognition.

2.1.1 Emotion model

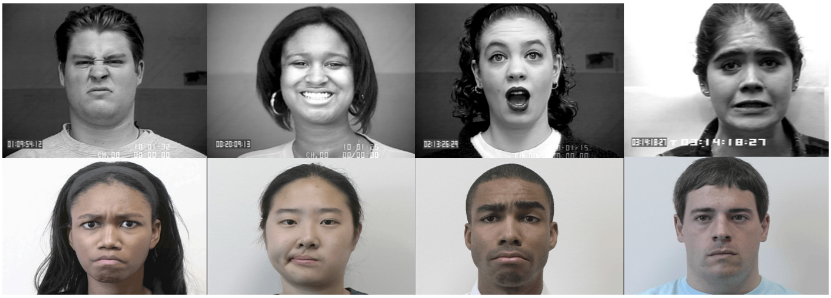


Figure 2.1: Emotion categories as described by Ekmann [Ekm73]: disgust, joy, surprise, fear, anger, contempt, sadness, and the neutral emotion.

The study on emotions has a long history, including antique descriptions such as the seven emotions (joy 喜, anger 怒, sorrow 哀, fear 惧, love 爱, hatred 憎, and desire 欲) by Confucian philosophy, and Cicero's four basic emotions (fear *metus*, pain *aegritudo*, lust *libido*, and pleasure *laetitia*). In the 19th century, Darwin and Prodger [Dar98] proposed the theory that emotions have evolved via natural selection, and were universally shared across cultures. In the 1970s, Ekmann [Ekm73] expanded upon this theory, proposing the categorized model that has a profound influence on the subsequent emotion recognition research. The categorized model groups the emotions into seven classes: joy, sadness, fear, anger, disgust, surprise and contempt. These seven emotions are illustrated in Figure 2.1. However, not all researchers accepted the theory of universal emotions. Averill [Ave74] challenged Darwin's theory, arguing that the social aspect and the relationship between language and emotion should be considered in the description for emotions as well. Moreover, categorical models can not fully capture the complexity of emotional states, as they do not represent certain emotions such as calmness or serenity that are part of a broader linguistic range of emotions.

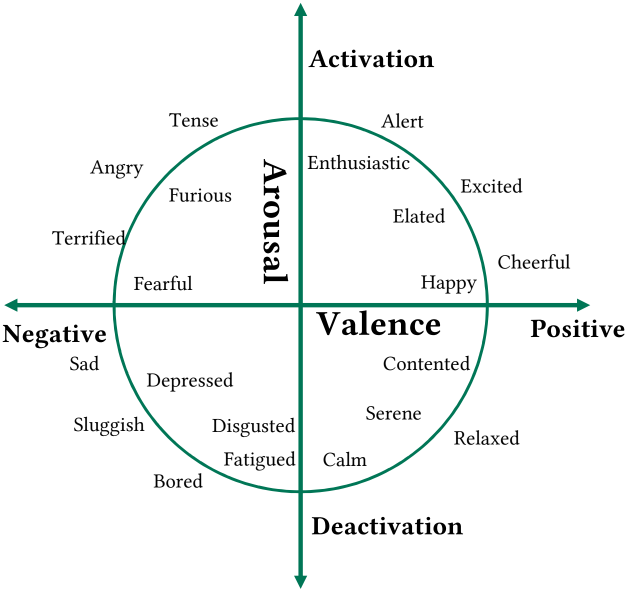


Figure 2.2: Circumplex model by Russell [Rus79].

Dimensional models represent emotions along continuous dimensions. One of the most widely adopted dimensional models for emotion recognition is the circumplex model, proposed by Russell [Rus79] in the late 1970s. The circumplex model describes emotions along two dimensions: valence and arousal. Valence represents the pleasantness or unpleasantness of the affective status, while arousal describes the degree of activation or energy associated with the emotion. These two dimensions divide the emotion space into four quadrants: Low-Arousal Low-Valence (LALV), exemplified by emotions such as sadness; Low-Arousal High-Valence (LAHV), with calmness being an example; High-Arousal Low-Valence (HALV); and High-Arousal High-Valence (HAHV), as shown in Figure 2.2. Occasionally, a third dimension is considered, such as Dominance [Aba15] or Liking [Koe11]. Compared to categorical models, dimensional models are able to represent more complex emotions and are more adaptable when dealing with emotions that are non-discrete or that have a high degree of variability [Por17].

2.1.2 Stress model

In 1936, Hungarian endocrinologist Hans Selye conducted an experiment in which he exposed rats to various stressors (cold, hunger, shock, or excessive muscular exercise). He found that regardless of the stressor, the rats always showed the same physiological responses, such as adrenal hyperactivity, lymphatic atrophy, and peptic ulcers [Sel36]. He used the term *stress* to describe this phenomenon, conceptualizing it as *the nonspecific response of the body to any demand that either causes or results in pleasant or unpleasant conditions* [Sel76]. Selye proposed the General Adaptation Syndrome (GAS) model to explain the total body response to stressors. In the first alarm stage of the GAS, the body mobilizes all resources in preparation to either fight off or escape from immediate stressors. During this stage, the body releases hormones such as epinephrine and norepinephrine to increase heart rate, respiratory rate and blood pressure; the blood glucose level is increased by elevated cortisol to provide more energy to the body. This process is also known as the “fight or flight” response, which was first described by Cannon [Can53] at the beginning of the 20th century. In the second stage, the body begins to repair itself and return to its status prior to the stress, accompanied by a decreased release of alarm hormones. If stress continues, the body could enter the third exhaustion phase, where the body is no longer able to cope with the stress and chronic health conditions can develop. To prevent the risks present in the exhaustion stage, an effective detection of the acute stress in the first stage is particularly critical, which is also one of the focal points of research in this dissertation.

To describe stress and emotion in a unified model, several works expanded an extra dimension in the circumplex emotion model for stress description. For example, Thayer [Tha90] along with Schimmack and Rainer [Sch02] split the arousal dimension into tense-arousal and energetic arousal, where tense-arousal stands for the dimension related to stress. While it has been demonstrated that there are connections between stress and negative emotions [San10], stress is typically not considered as an emotion in the affect computing research [Sch19]. In this dissertation, emotion and stress will be considered as two separate aspects of affective status, and therefore, the tasks of stress and emotion recognition will be investigated in two independent Chapters 7 and 8.

2.2 Physiological basics of the cardiovascular system

2.2.1 Anatomical structure and physiology of the heart

The cardiovascular system is composed of the heart and blood vessels. Acting as the system's core, the heart provides the essential mechanical force, through contractions, needed to circulate blood throughout the body. It contains four chambers: two atria and two ventricles, which are divided into left and right sections by the interatrial and interventricular septa, respectively, as illustrated in Figure 2.3. Each atrium and ventricle pair is interconnected by a leaflet valve —the mitral valve on the left and the tricuspid valve on the right —governing the direction of blood flow within the heart. The exit of blood from the heart is regulated by the semilunar valves, specifically the aortic valve at the aorta and the pulmonary valve at the pulmonary artery.

The cardiovascular system is responsible for the delivery of vital oxygen and nutrients to the organs and tissues of the body. This system encompasses two primary sectors:

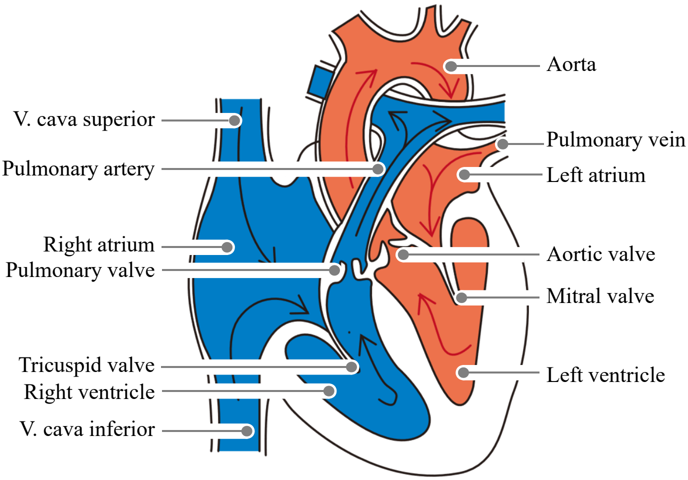


Figure 2.3: Anatomical structure of the human heart.

- *Systemic circulation*, pertaining to the delivery of nutrients and oxygen to the body's tissues
- *Pulmonary circulation*, concerned with the oxygenation of blood

Deoxygenated blood from the body's periphery is returned to the right side of the heart via the venae cavae. During the heart's relaxation phase, or diastole, this blood enters the right ventricle through the tricuspid valve and subsequently departs the heart via the pulmonary valve, transitioning into the pulmonary arteries. Here, a gas exchange occurs in the lungs, with the blood relinquishing carbon dioxide and acquiring oxygen. This oxygen-enriched blood is then transported to the left atrium via the pulmonary veins before progressing into the left ventricle. From this point, during the subsequent heart contraction, or systole, the oxygenated blood is propelled into the systemic circulation via the aortic valve, reaching the arterial vessels, or arterioles and capillaries, within the body's periphery.

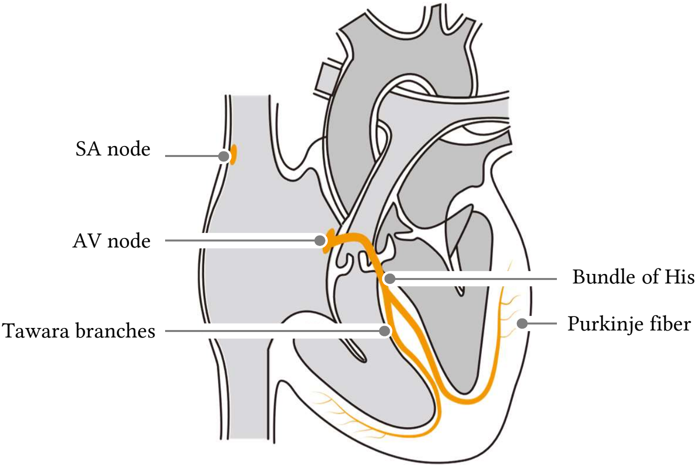


Figure 2.4: Cardiac conduction system.

2.2.2 Cardiac conduction

The heart contraction is induced by electrical impulses that are initiated and propagated by the cardiac conduction system. The heart is composed of two types of cells: (i) contractile muscle cells, responsible for executing the mechanical contraction of the heart, and (ii) pacemaker cells, which generate electrical impulses and regulate the heart rate. While contractile muscle cells constitute approximately 99% of the heart wall's cellular makeup, the remaining 1% comprises pacemaker cells [Bet13]. These pacemaker cells have the capacity to not only receive signals from the brain, but also to spontaneously generate impulses, a phenomenon referred to as autorhythmicity.

Pacemaker cells are primarily located in the sinoatrial node (SA node), positioned in the right atrium near the coronary sinus on the interatrial septum [Kur10], and the atrioventricular node (AV node), situated in the lower segment of the right atrium near the tricuspid valve, as shown in Figure 2.4. The SA node and AV node serve as the primary and secondary pacemakers, respectively. In a normally functioning heart, electrical impulses originate from the sinoatrial node and are propagated through the muscle cells of the atria to the AV node. There, these impulses experience a delay of approximately 0.09 seconds, ensuring that the atria have completed the ejection of blood prior to ventricular contraction [Cam02]. The impulses then continue along the Bundle of His, the bundle branches (Tawara branches) and the Purkinje fibers, prompting synchronized contractions of the ventricles.

2.2.3 Blood volume pulse

The expulsion of blood from the left ventricle initiates a pulse wave, which disseminates throughout the systemic circuit encompassing the entire body. This propagation of the pulse wave within the arteries can be analogized to fluid dynamics within an elastic pipe. As the ventricles pump blood, the arterial pressure increases, subsequently expanding the arterial cross section. The inherent tension within the artery wall initiates its contraction, pushing the blood towards the subsequent segment of the artery. Three distinctive physical phenomena can be observed during this process: the flow pulse, the pressure pulse, and the volume pulse [Kor09]. The measurement and analysis of the volume pulse form the crux of plethysmography.

Figure 2.5 illustrates the foundational form of the pulse wave, measured from a peripheral artery. The waveform can be broadly categorized into two phases: systole and diastole. The systolic phase commences with the contraction of the left ventricle and concludes with the closure of the aortic valve. The expulsion of blood from the ventricle induces the first steep rise of the wave, referred to as the percussion wave. The tidal wave, observed in the latter portion of systole, may be induced by an echoed pulse returning from the upper body [Sub19]. At the onset of diastole, the third, dicrotic wave can be observed, which is a consequence of the aortic valve's closure and the subsequent alterations in aortic pressure [Nir14].

The morphology of the pulse wave is influenced by both inter- and intra-individual factors. For instance, the waveform observed in the aorta differs from that measured in peripheral arteries. Additionally, factors such as the elasticity of the blood vessels, which in turn depends on variables including

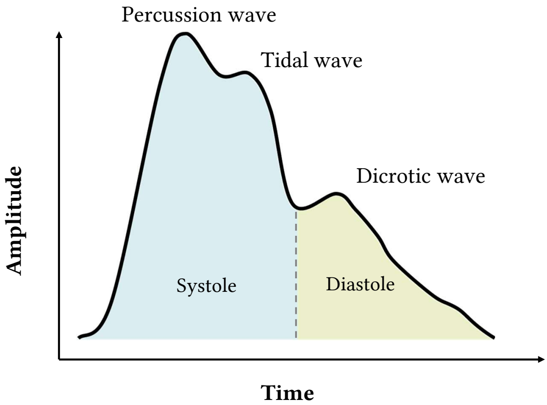


Figure 2.5: Exemplary shape of the pulse wave in peripheral arteries.

age, gender, height, diet, health conditions, and medication, can further influence the waveform [ORo01]. Consequently, these morphological analysis of the pulse wave can provide indicators of cardiovascular conditions such as arterial stiffness or hypertension.

2.2.4 Heart rate and heart rate variability

Vital signs serve as indicators that reflect bodily functions. The most frequently monitored vital signs in clinical settings include heart rate, respiratory rate, temperature, and blood pressure. These parameters are critical indicators used to assess whether an individual is in a non-critical physical condition. Given the relevance of heart rate and heart rate variability to this dissertation, these parameters will be discussed in detail in the ensuing section.

2.2.4.1 Heart rate

Heart rate is defined as the number of heartbeats per minute. The resting heart rate varies depending on factors such as age, weight, and overall cardiovascular health. The range of a typical resting heart rate lies between 60 to 100 beats per minutes (bpm). Neonates can exhibit a resting heart rate higher than 120 bpm, while long-term athletes may have resting heart rates as low as approximately 40 bpm. Heart rate serves as a critical indicator to track changes in an individual's physiological status. It is also commonly used to quantify the intensity of physical exertion during athletic activities. Additionally, variations in heart rate can be indicative of changes in psychological states, with factors like stress or emotional arousal tending to cause an increase in heart rate [Was21].

2.2.4.2 Heart rate variability

The heart rate is regulated by various mechanisms within the human body in response to internal and external conditions. A cardiovascular system that is robust and healthy can rapidly adapt to changes in environment or internal state, ensuring an efficient supply of oxygen and nutrients to meet the body's demands.

One way to assess the effectiveness of this adaptive response is through Heart Rate Variability (HRV). It quantifies the time intervals between successive heartbeats and is regulated by the Autonomic Nervous System (ANS), also known as the vegetative nervous system.

The ANS consists of the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS), which operate in a coordinated yet complementary manner. Sudden changes in the external environment trigger heightened sympathetic nervous system activity, mobilizing the body's resources. Such intensified sympathetic activity elevates the heart rate whilst decreasing heart rate variability. In contrast, heightened parasympathetic activity results in a decrease in heart rate and an increase in heart rate variability. It serves for regeneration and the buildup of the body's own reserves (trophotropic effect) [Ber97]. Heart rate variability can also be influenced by the baroreceptor reflex and respiration. Baroreceptors are sensory nerve endings located on the walls of certain large blood vessels. They detect changes in arterial wall dilation caused by alterations in blood pressure and transmit signals to the brain, which regulates heart rate and maintaining heart rate variability. Respiratory sinus arrhythmia (RSA) is the phenomenon where heart rate is influenced by respiratory activity, primarily regulated by the parasympathetic system.

Since HRV reflects the activity of ANS, it can function as a marker for the body's adaptive capabilities. The analysis of HRV serves as a crucial predictive parameter for cardiac and immune system disorders [Fra22], offering a metric to evaluate the regulatory capability of the autonomic nervous system. Furthermore, HRV analysis yields significant insights in diagnosing mental illnesses, including depression [Jun19], burnout [Lo20], and panic attacks [Coh00].

HRV is characterized not merely by a single metric, but by a set of parameters. These parameters are derived from the Inter-Beat Interval (IBI) signals, and describe the variation in heartbeat from various perspectives. In the case of ECG signals, IBI signals are computed as the duration between two consecutive R-peaks:

$$RR_i = t_i - t_{i-1}\,, \tag{2.1}$$

where the i-th interval RR_i signifies the interval between the $(i-1)$-th and i-th R-peaks. For PPG signals, IBI signals are derived from the peaks of blood volume pulse signals. Technically, the variability parameters obtained from pulse signals are denoted as Pulse Rate Variability (PRV), which, while intrinsically linked to, are fundamentally different from HRV. In this dissertation, the camera sensor measures the blood volume changes, and PRV is analyzed as a proxy for HRV. The PRV/HRV analysis incorporates characteristic values in the time, frequency, and non-linear domains, as shown in Table 2.1. The PRV/HRV parameters adopted in this dissertation will be discussed in greater detail subsequently.

Time domain HRV parameters

Time domain HRV parameters represent the statistical analysis of the inter-beat intervals in the time domain. One of the commonly used parameters is the Standard Deviation of the IBI of Normal Sinus Beats (SDNN), mathematically expressed as:

$$SDNN = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(NN_i - \bar{NN})^2}\,, \tag{2.2}$$

where $\bar{NN}$ and N symbolize the average value and the total count of intervals, respectively. NN intervals refer to the normal-to-normal intervals where any artefacts in the peak-to-peak intervals brought about by cardiac arrhythmia

Table 2.1: Overview of HRV parameters

Domain	Parameter	Unit	Description
Time domain	SDNN	ms	Standard deviation of the IBI of normal sinus beats
	RMSSD	ms	Root mean sum of squared successive distance
	pNN50	%	Percentage of successive differences greater than 50 ms
Frequency domain	Total power	ms^2	Absolute power of the IBI signal (≤0.4 Hz)
	ULF	ms^2	Absolute power of the Ultra Low Frequency band (≤0.003 Hz)
	VLF	ms^2	Absolute power of the Very Low Frequency band (0.003-0.04 Hz)
	LF	ms^2	Absolute power of the Low Frequency band (0.04-0.15 Hz)
	HF	ms^2	Absolute power of the High Frequency band (0.15-0.4 Hz)
	LF/HF		Ratio of LF to HF
	LF norm	n.u.	Relative power of the Low Frequency band (0.04-0.15 Hz)
	HF norm	n.u.	Relative power of the High Frequency band (0.15-0.4 Hz)
Nonlinear	SD_1	ms	Ellipse width of the Poincaré diagram
	SD_2	ms	Ellipse length of the Poincaré diagram
	SD_1/SD_2		Ratio of SD_1 to SD_2

have been filtered out. The customary time frame for SDNN measurement is 5 minutes, although researchers have also proposed short-term measurements ranging from 60 seconds to 240 seconds [Sha17]. Both PNS and SNS activities contribute to SDNN. Furthermore, SDNN shows a high correlation with the low-frequency energy of the IBI signals.

In addition to SDNN, other crucial time domain parameters include Root Mean Sum of Squared Successive Distance (RMSSD) or Percentage of Successive Differences Greater than 50 ms (pNN50). RMSSD is mathematically expressed as:

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(NN_{i+1} - NN_i)^2}\,, \tag{2.3}$$

This formula calculates the average difference between two successive intervals. RMSSD estimates vagally mediated changes in parasympathetic activity and serves as an indicator of organ recovery ability. Lower RMSSD values may point to potential psychological and physical stress or even Sudden Death in Epilepsy (SUDEP) [DeG10].

The pNN50 parameter is the percentage of NN intervals that differ from each other by more than 50 ms:

$$pNN50 = \frac{NN50}{N-1} \cdot 100\%\,. \tag{2.4}$$

Similar to RMSSD, the pNN50 parameter characterizes the high-frequency dynamics of heart rate changes and exhibits a strong correlation with parasympathetic activity. Accurate measurement of pNN50 necessitates a higher level of precision and sampling frequency.

Frequency domain HRV parameters

As the IBIs are represented as a time-discrete signal sequence that may not be uniformly sampled along the time axis, it is necessary to perform interpolation prior to conducting frequency domain analysis. By doing so, the spectrum of the IBI signals can be obtained using methods such as Fast Fourier Transformation (FFT) or Autoregressive (AR) modeling.

The Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology divide the spectrum of IBI into several sub-ranges: Very Low Frequency (VLF), Low Frequency (LF), High Frequency (HF), and occasionally Ultra Low Frequency (ULF) ($\leq$ 0.003 Hz).

The VLF band, ranging from 0.0033 Hz to 0.04 Hz, signifies heart rate variations between 25 seconds and 300 seconds. It gives indications of chronic inflammation [Ste08], cardiovascular disease [Guz05], and shows a strong correlation with the prognosis of metabolic disorders [Ass10] and arrhythmic death [Big92].

The LF band (0.04 Hz to 0.15 Hz) includes modulations with periods between 7 and 25 seconds and is sometimes referred to as the baroreceptor region due to its predominant regulation by baroreceptors during rest [McC15]. This band is subject to both sympathetic and parasympathetic influence and offers a valuable measure for evaluating physical and mental stress levels.

The HF band (0.15 Hz to 0.4 Hz), known as the respiratory band, reflects parasympathetic activity and encapsulates the heart rate variation associated with respiration. The HF power shows high correlation with RMSSD and pNN50, and is associated with negative mental experience such as stress, panic, anxiety, and worry [Sha17].

Since the LF band is regulated by both the sympathetic and parasympathetic systems, and the HF band is chiefly influenced by the parasympathetic system, the ratio of LF to HF power (LF/HF ratio) offers a useful metric for assessing the interplay between these autonomic systems.

The power values are typically expressed either in absolute terms with the unit ms^2 or in normal unit (n.u.). Normalization provides a relative value for each power component in relation to the total power minus the VLF component, thereby facilitating inter-individual comparisons despite individual variations in the power of specific bands or the total power.

Nonlinear HRV parameters

Owing to the intricate regulatory mechanisms involved, HRV exhibits nonlinear characteristics. One of the most widely employed techniques to interpret these non-linear characteristics is the Poincaré Plot, an approach developed by the French mathematician, Jules Henri Poincaré. To construct a Poincaré Plot, pairs of consecutive intervals NN_i and NN_{i+1} are represented in a two-dimensional coordinate system, where NN_i and NN_{i+1} serve as the X and Y coordinates respectively, as depicted in Figure 2.6. For healthy individuals, this usually yields a point cloud approximating an ellipse. The analysis of this elliptical shape can provide valuable insights into the dynamics of HRV.

To calculate the nonlinear parameters, the diagram is first rotated by an angle of $\pi/4$. This operation transforms the coordinates to align with the axes of the ellipse:

$$x_{i,1} = \frac{NN_i - NN_{i+1}}{\sqrt{2}}\,, \tag{2.5}$$

$$x_{i,2} = \frac{NN_i + NN_{i+1}}{\sqrt{2}}\,, \tag{2.6}$$

for $i \in (1, 2, ..., N-1)$.

The parameters are subsequently calculated as the standard deviation of the point distribution along these two new axes:

$$\mathrm{SD}_1 = \sigma(x_{i,1})\,, \tag{2.7}$$

$$\mathrm{SD}_2 = \sigma(x_{i,2})\,. \tag{2.8}$$

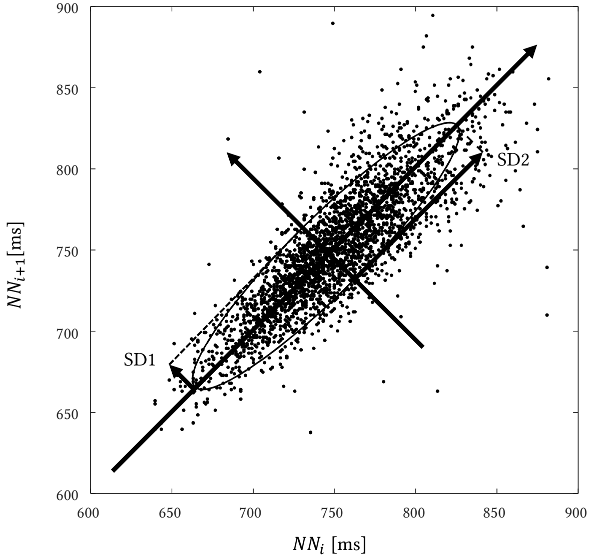


Figure 2.6: Example of a Poincaré diagram. The standard deviations SD_1 and SD_2 stand for the width and length of the ellipse, respectively.

The parameters SD_1 and SD_2 specify the width and length of the ellipse, respectively. Notably, SD_1 is identical to RMSSD and measures the body's short-term adaptation capabilities. On the other hand, SD_2 correlates with both low and high frequency powers, representing short- and long-term HRV characteristics. Analogous to the LF/HF ratio, the SD_1/SD_2 ratio can be employed to assess the balance between the parasympathetic nervous system and sympathetic nervous system.

2.3 Skin

Skin, the largest organ of the human body, constitutes approximately 15% of total body weight [Ric03]. As part of the integumentary system, along with its appendages, it forms the outermost layer of the human body. This system serves as a protective barrier, preventing bacteria and germs from penetrating

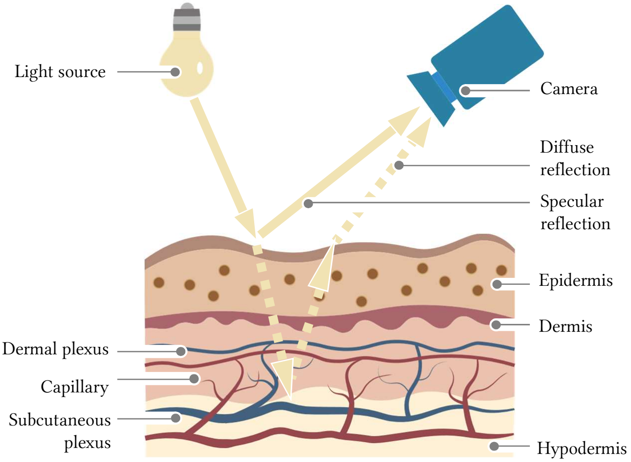


Figure 2.7: Anatomical structure of the human skin and reflection model for rPPG.

the body. The skin provides several critical functions including body temperature regulation, sensory reception (such as pressure, touch, or pain), and maintaining hydration.

As displayed in Figure 2.7, human skin is comprised of three layers: the epidermis, dermis (also known as *corium* in Latin), and hypodermis (or *subcutis* in Latin). The dermis can be further divided into the papillary dermis and the reticular layer. The papillary dermis, the superficial layer, consists of fine, loosely arranged collagen fibers. In contrast, the reticular layer, situated deeper, is composed of dense, irregular connective tissue that imparts firmness and elasticity to the skin.

The skin region features an abundant supply of arteries, veins, and capillaries. Blood vessels that bifurcate and connect to the rest of the body are found in the hypodermis layer. These blood vessels extend their endings into the dermis layer, thereby providing nutrients to the epidermis, which is devoid of blood supply by itself. This network of blood vessels within the hypodermis layer is referred to as the subcutaneous plexus.

The dense network of blood vessels offers an opportunity to non-invasively measure vital signs such as heart rate and oxygen saturation levels by assessing blood volume. Interestingly, the skin on the face is more exposed compared to other body parts and also experiences high blood flow, rendering it an optimal site for such measurements, as will be demonstrated in this study.

2.3.1 Face vascular circulation

The skin on the face exhibits high metabolic activity and is richly vascularized. These vascular tissues facilitate growth and recovery from damage in facial skin. The arterial supply to the face primarily originates from the external carotid artery, from which the facial artery and superficial temporal artery arise directly, and other arteries like the transverse facial artery (from the superficial temporal artery) and the infraorbital artery (from the maxillary artery) are branched out. However, a significant arterial contribution to the forehead is provided by the ophthalmic artery, which arises from the internal carotid artery [Arx18].

Main skin perforators penetrate the deep tissues, forming a dense subdermal plexus that provides vascularization to facial areas [Hou00]. The number of capillary loops beneath the epidermis of the face varies significantly from one region to another, and the diameter of these capillaries also differs across various skin areas. Regions such as the forehead, perioral region, nose, philtrum, lip, chin, and ocular region have abundant capillary blood flow. In contrast, the temporal region, lateral cheeks, and jaw are less vascularized, and the ears display moderate vascularization [Mor59].

2.3.2 Skin blood flow regulation

The regulation of skin blood flow is critical for maintaining thermal homeostasis. It is modulated by sympathetic vasodilation and vasoconstriction mechanisms. At rest, skin blood flow typically falls within the range of 1 to 3 mL/100g/min [How06]. During strenuous physical exercise or exposure to heat, increased temperatures are sensed by the hypothalamus, which further triggers autonomic regulation. More specifically, the process of cutaneous vasodilation is intensified, during which the skin blood flow rate can rise to approximately 60% of total body cardiac output [Dyc75]. Conversely, vasoconstriction reduces skin blood flow and decreases heat dissipation when exposed to cold conditions.

Skin blood flow also responds to other factors, including physiological reactions like a sudden inspiratory gasp, or psychological conditions such as mental stress or emotional stimuli [Ami12]. An inspiratory gasp activates vasoconstriction through a brainstem reflex. Vasoconstriction can also be observed during mental stress or in response to a sudden noise. Empirically, it is known that emotions such as anger and embarrassment can increase facial blood flow, leading to blushing, while fear and disgust may reduce blood flow, resulting in paleness.

2.3.3 Skin color

One of the primary factors contributing to differences in skin color among individuals is the variation in melanin pigmentation. The most common forms of melanin are eumelanin and pheomelanin. Eumelanin imparts coloration that varies from brown to black and is present in the hair and skin. Pheomelanin presents a pink to red hue and can be found in red hair, lips, and nipples.

For individuals with dark skin, skin color is primarily determined by the concentration of melanin. The color of light skin is primarily determined by the bluish-white connective tissue under the dermis, with some influence from the color of blood (due to hemoglobin) in the dermis and subcutaneous tissue. The concentration of indigenous skin melanin correlates with the geographic distribution of ultraviolet radiation (UVR). The Fitzpatrick Scale [Fit75] classifies human skin color into six categories based on the skin's response to UVR, ranging from Type I (palest, always burns) to Type VI (dark brown to darkest brown, never burns).

Owing to the higher concentration of melanin, which absorbs more light, the light intensity captured by a camera is weaker for individuals with dark skin, thus, the quality of rPPG measurements declines. It is a critical consideration within the rPPG research community to develop methods that account for all skin types [Now20]. Changes in skin color can also occur due to variations in blood oxygen saturation. For instance, the lips may appear bluish in color, if blood oxygen saturation levels fall below 70%, a condition also known as cyanosis that can be caused by cold, inadequate gas exchange in the lungs, or heart disease.

2.4 Measurement of cardiac activity

2.4.1 Photoplethysmography

The term *Plethysmography* originates from the Greek word *plethysmos*, meaning "increasing". It is an instrument used to examine volume changes in organs or the entire body. Photoplethysmography specifically examines blood volume changes within the microvascular bed of tissue. It operates on the principle of light absorption by hemoglobin in the blood. When light is emitted from an illumination source, a fraction of it is absorbed by blood, while another fraction is detected by a photo sensor. PPG can operate in either transmissive or reflective modes. In the transmissive mode, the light source and photo-detector are positioned on opposite sides of the tissue to be measured, such as a finger or earlobe. The light then travels through the tissue before it is detected by the photo sensor. In contrast, in the reflective mode, the light source and photo-detector are placed on the same side of the tissue, such as on the arm or forehead. In this setup, the light is reflected back towards the photo sensor for detection after interaction with the tissue.

PPG is a widely employed for the determination of peripheral oxygen saturation (SpO2), a pivotal parameter for patient triage that has been highlighted during the COVID-19 pandemic. The measurement of SpO2 relies on disparities in the light absorption properties of oxygenated and deoxygenated hemoglobin at varying wavelengths, which are shown in Figure 2.8. Notably, the most pronounced distinction in light absorption occurs within the red

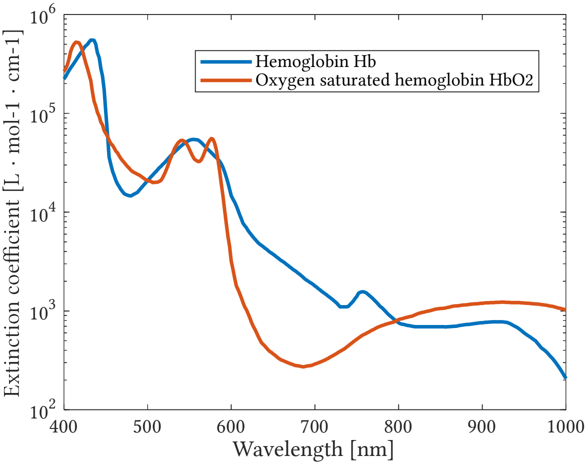


Figure 2.8: Light absorption of blood at various wavelengths. Data source: [Pra23].

spectrum (600-750 nm) and the near-infrared spectrum (850-1000 nm). Specifically, oxygenated hemoglobin exhibits greater absorption of infrared light compared to deoxygenated hemoglobin, whereas deoxygenated hemoglobin displays greater absorption of red light. The property of light absorption variation in the red, green and blue ranges is also adopted for disturbance compensation in pulse signal extraction using RGB cameras, which will be discussed in Chapter 5.

2.4.2 Dichromatic reflection model

The dichromatic reflection model [Sha85] presents the underlying physics governing the reflection properties of heterogeneous materials possessing microscopically irregular surfaces and particles of a colorant that induce scattering and colorization. In the context of pulse signal measurement, the skin can be considered as a heterogeneous medium, and consequently, the reflection mechanism of the skin can be explained using the framework of the dichromatic reflection model:

$$\begin{aligned} C(\lambda, i, e, g) &= C_i(\lambda, i, e, g) + C_b(\lambda, i, e, g) && (2.9) \\ &= m_i(i, e, g) \cdot c_i(\lambda) + m_b(i, e, g) \cdot c_b(\lambda)\,. && (2.10) \end{aligned}$$

In this equation, the light reflected back to the camera sensor C is decomposed into the interface component C_i and body component C_b. The interface component stands for the light that is reflected at the skin surface. The body component is the light component that penetrates into the skin, gets scattered by the under skin tissues or eventually absorbed, then reflected to the camera. Both components are further decomposed into two parts. The composition c_i and c_b refer to the spectral power distributions for the interface and body components which are only dependent on the wavelength λ. The amplitude parts m_i and m_b are independent from the wavelength and only relevant to the geometrical parameters such as the incidence angle (i), emittance angle (e) and the phase angle (g).

2.4.3 Skin reflection model for rPPG

Equation 2.10 introduces a static model that does not account for the dynamic variations of each factor. In reality, the target signal is influenced by the dynamic modulation caused by changes in blood volume. As a result, the composition component c_b should be a function of time as well. Furthermore, when measuring blood volume signals, it is essential to consider potential disturbances such as head motions, as they can alter the geometric parameters for both the body and interface components. Therefore, a comprehensive model should not only incorporate the temporal dynamics of the composition component but also effectively account for the effects of disturbances on the geometric parameters during blood volume measurements.

Based on the dichromatic reflection model, Wang et al. [Wan16b] proposed a model to describe the skin pixel $\mathbf{C}$ captured by the camera:

$$\mathbf{C}(t) = I(t) \cdot (\mathbf{v}_s(t) + \mathbf{v}_d(t)) + \mathbf{v}_n(t)\,, \tag{2.11}$$

where the illumination intensity $I(t)$ is a scalar function representing the intensity of the light, assuming a constant spectral distribution in this model. It is important to note that $I(t)$ is not solely determined by the emitted light intensity from the light source, but also influenced by the relative positioning between the light source, measurement site, and camera sensor. The illumination intensity is further modulated by the specular reflection component $\mathbf{v}_s(t)$ and the diffuse reflection component $\mathbf{v}_d(t)$. These components correspond to the interface and body reflection components in the dichromatic model. Here, the specular and diffuse reflection components are represented as vectors, accommodating both multi-channel and monochrome measurement setups. $\mathbf{v}_n(t)$ represents the quantization noise of the camera system.

The change of the specular component $\mathbf{v}_s(t)$ depends only on the geometrical relationship between the light source, camera sensor and the measurement site on the skin. The specular component can be decomposed into stationary and time-varying components:

$$\mathbf{v}_s(t) = \mathbf{u}_s \cdot (s_0 + s(t))\,, \tag{2.12}$$

where the specular component $\mathbf{v}_s(t)$ is the result of the multiplication of a unit vector $\mathbf{u}_s$ representing the spectral distribution of the illumination, and the sum of the constant term s_0 and the time-varying component $s(t)$. The constant term s_0 represents the stationary part of the specular component, while $s(t)$ accounts for the dynamic changes induced by motion.

The diffuse component can be further expanded as:

$$\mathbf{v}_d(t) = \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t)\,, \tag{2.13}$$

where the diffuse component $\mathbf{v}_d(t)$ is decomposed into two parts. The stationary part is represented by the product of a unit vector $\mathbf{u}_d$ corresponding to the color of the skin tissue, and d_0 which signifies the reflection strength. This model assumes that the time-varying component of the diffuse part is primarily attributed to the changes in blood volume, denoted as $p(t)$. $\mathbf{u}_p$ is a unit vector that characterizes the relative strength of the pulsatile signal in each color channel.

Combining Equation 2.12 and 2.13, Equation 2.11 can be written as:

$$\begin{aligned}\mathbf{C}(t) &= I(t) \cdot (\mathbf{u}_s \cdot (s_0 + s(t)) + \mathbf{u}_d \cdot d_0 + \mathbf{u}_p \cdot p(t)) + \mathbf{v}_n(t) && (2.14)\\ &= I(t) \cdot (\mathbf{u}_c \cdot c_0 + \mathbf{u}_s \cdot s(t) + \mathbf{u}_p \cdot p(t)) + \mathbf{v}_n(t)\,, && (2.15)\end{aligned}$$

where the stationary components are combined into a single term:

$$\mathbf{u}_c \cdot c_0 = \mathbf{u}_s \cdot s_0 + \mathbf{u}_d \cdot d_0\,. \tag{2.16}$$

The illumination intensity can be separated into stationary and time-varying parts as well:

$$I(t) = I_0 \cdot (1 + i(t))\,, \tag{2.17}$$

where I_0 denotes the stationary part, and $I_0 \cdot i(t)$ is the time-varying change in illumination. Thus the model can be further expanded as:

$$\begin{aligned}\mathbf{C}(t) =& I_0 \cdot (1 + i(t)) \cdot (\mathbf{u}_c \cdot c_0 + \mathbf{u}_s \cdot s(t) + \mathbf{u}_p \cdot p(t)) + \mathbf{v}_n(t) && (2.18)\\ =& \mathbf{u}_c \cdot I_0 \cdot c_0 + \mathbf{u}_s \cdot I_0 \cdot s(t) + \mathbf{u}_p \cdot I_0 \cdot p(t) + && (2.19)\\ & \mathbf{u}_c \cdot I_0 \cdot c_0 \cdot i(t) + \mathbf{u}_s \cdot I_0 \cdot s(t) \cdot i(t) + \\ & \mathbf{u}_p \cdot I_0 \cdot p(t) \cdot i(t) + \mathbf{v}_n(t)\,.\end{aligned}$$

By ignoring the noise $\mathbf{v}_n(t)$ and the production of time-varying terms (e.g., $p(t) \cdot i(t)$), the model can be then expressed in a simple linear equation:

$$\mathbf{C}(t) \approx \underbrace{\mathbf{u}_c \cdot I_0 \cdot c_0 + \mathbf{u}_c \cdot I_0 \cdot c_0 \cdot i(t)}_{\text{Intensity}} + \underbrace{\mathbf{u}_s \cdot I_0 \cdot s(t)}_{\text{Specular}} + \underbrace{\mathbf{u}_p \cdot I_0 \cdot p(t)}_{\text{Pulse}}\,. \tag{2.20}$$

This model approximately decomposes the skin color captured by the camera into three main components: intensity, specular, and pulse components. It is important to note that both the specular and pulse components have zero mean values. The only stationary (DC) component is present within the intensity component, which can be easily eliminated through temporal normalization. Consequently, the primary objective of the core remote Photoplethysmography algorithm is to extract or isolate the pulse signals from the time-varying (AC) portions of the intensity and specular reflection components. Advantages and limitations of this model will be discussed in Chapter 5.

3

State of the Art

This dissertation focuses on developing an rPPG measurement system and its application in stress and emotion recognition. Unlike traditional measurement techniques such as ECG, rPPG does not necessitate the attachment of sensors to the human body, allowing for continuous vital parameter measurement without causing physical discomfort. This chapter will introduce the state of the art and related work in the field of rPPG and stress/emotion recognition.

First, Section 3.1 will provide an overview of the current research status of stress and emotion recognition. Then, Section 3.2 will introduce different techniques for contactless vital parameter measurement, and offer a comparative analysis. Finally, Section 3.3 will delve into the state of the art of rPPG in detail.

3.1 State of the art for stress and emotion recognition

3.1.1 Sensor modalities

Data for emotion and stress recognition can be derived from behavioral or physiological modalities. Behavioral modalities include text, tone, speech, body gesture, and facial expression. One drawback of affective status recognition using behavioral data manifests when individuals intentionally regulate their emotional displays or the individuals are naturally more reserved in expressing their emotions [Agr11].

In contrast, physiological modalities record the physiological responses. The physiological responses are regulated by the ANS and not under conscious control. As discussed in Chapter 2, the ANS consists of the sympathetic nervous system and the parasympathetic nervous system. The SNS is active during the alarming stage of the stress response, triggering the fight or flight mechanism. Conversely, the PNS directs the body to relax, promoting the "rest and digest" functions.

The activities of the autonomic nervous system can be reflected by physiological signals, such as the ECG [Bug17, Suz21, Nar15], PPG [Udo17], Electrodermal activity (EDA) [Mar13, Shu19, Udo17], Electromyogram (EMG) [Has19], and EEG [Liu17, Meh17, Lu15, Qin19, Sub21, Zha20b]. Among the parameters measured by these sensors, cardiac and pulmonary parameters like heart rate, respiratory rate and heart rate variability are some of the most widely used for emotion/stress recognition. These parameters can also be measured using non-contact sensors such as radar, WiFi, sonar, or cameras, which will be discussed in detail in Section 3.2.

3.1.2 Processing methods

The standard data processing pipeline for affect recognition comprises data pre-processing, feature calculation/selection, and status recognition. The pre-processing of physiological data for affect recognition includes denoising, detrending, synchronization, resampling, and pruning (of severely contaminated data segments). Feature calculation extracts stress or emotion-related information from the physiological data, which is then input into a recognition module. The recognition module is usually a machine learning model. The selection of features is essential for the recognition performance. An optimal feature set should contain pertinent information about the affect status while avoiding representation redundancy.

While traditional methods focus on manual feature extraction, recent advancements have leveraged deep learning techniques for an end-to-end assessment of affective states. These methods have shown promising performance. Dzieżyc et al. [Dzi20] demonstrated that CNN-based networks can be used for affect recognition from physiological sensors. Martinez et al. [Mar13] used Convolutional Neural Network (CNN) combined with denoising auto-encoders to extract features from EDA and BVP, with the CNN networks outperforming manual ad-hoc feature extraction. Liu et al. [Liu17] proposed a multi-layer Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) with temporal and spectral band attention mechanisms for emotion recognition from videos and EEG signals.

Qiu et al. [Qiu18] introduced the Correlated Attention Network to incorporate the correlation of high-level features between EEG and eye movement signals into the attention mechanism. The Correlated Attention Network achieved higher accuracy than feature-based Support Vector Machines (SVMs). Li et al. [Li20] explored an attention-based LSTM-RNN model to extract emotionally salient information from spectrograms. Harper and Southern [Har20] used a Bayesian deep learning model to classify emotional valence from inter-beat interval series, introducing the confidence analysis to the evaluation results.

Table 3.1: Overview of non-contact methods for measuring cardiac and pulmonary parameters

Method	Modality	Distance	Cost	Multiple subject	Context information	Availability	Motion sensitivity	Illumination sensitivity
Radar	Motion	m	++	>1	+	++	++	o
RFID	Motion	m	+	1	o	+	++	o
WiFi	Motion	m	+	>1	+	+++	++	o
Ultrasound	Motion	m	+	>1	+	+	++	o
Depth camera	Motion	m	++	>1	+	+	++	o
Capacitive ECG	Charge	mm	+	1	o	+	++	o
Thermal camera	Temperature, motion	m	+++	>1	+	+	++	o
RGB/Near-Infrared (NIR) camera	Motion, light absorption	≤50 m [Bla17a]	+	>1	++	+++	+(HR), ++(RR)	++

3.2 Non-contact measurement of cardiac and pulmonary parameters

Cardiac and pulmonary parameters, such as heart rate and respiratory rate, are among the vital signs routinely monitored in the healthcare environment. The gold standard for measurement of heart rate and heart rate variability is ECG, which measures the electrical activity of the heart as reflected in voltage changes on the skin surface. This change can also be measured without direct contact with the skin by capacitive ECG [Heu11]. Capacitive ECG introduces a dielectric layer between the skin and the electrode to build a coupled capacitor. However, it can only provide satisfactory quality at a measurement distance of a few millimeters at most, meaning that this is considered a quasi-contact-based approach. In addition, the signal quality can be affected by the type of clothing worn by the subject.

Cardiac and pulmonary parameters can also be measured by observing motion caused by heartbeats and breathing. Radar is one of the most researched motion-based methods for contactless measurement of vital signs. The first radar-based monitoring system for vital signs was introduced in 1970s [Car71]. Since then, various technical solutions have been developed such as continuous-wave (CW) [Mer17, Gu16], ultra-wideband (UWB) [Gol20, Yan18c] and Frequency Modulated Continuous Wave (FMCW) [Adi15]. FMCW technology allows the separation of signals reflected from different directions based on the time of signal propagation, thus enabling identification of body movement caused by heartbeats and breathing. Radar also enables signal measurement of multiple subjects. Besides measurement of respiratory rate or heart rate, researchers have investigated the application of radars for blood pressure measurement. Zhao et al. [Zha18] presented a system to measure beat-to-beat blood pressure with a continuous wave Doppler radar.

Radio-Frequency Identification (RFID) technique can be utilized to monitor vital signs as well [Hou17, Yan18a, Yan19]. The vital sign signals are captured from the reflected waves by the RFID tags attached to the subjects' clothing. Compared to radar systems which necessitate extra hardware components, RFID-based measurement systems are much cheaper and can produce signals of high accuracy over a broad scanning range.

The WiFi-based solution has also attracted a lot of attention due to the extensive and already existing deployment of WiFi infrastructure. WiFi-based systems have been applied to solve tasks such as activity recognition [Gu17, Xu19, Yan18b], gesture recognition [Pu13, Pu15], fall detection [Din20, Wan16a, Pal18] and indoor localization [Che16a, Du18]. Vital sign measurement using WiFi device relies on the Channel State Information (CSI), which contains amplitude and phase information of subcarriers. This information reflects how human interacts with the electromagnetic waves in the environment, including subtle movements caused by breathing and heartbeats. The algorithms for the CSI data processing can be classified into three groups: model-based methods [Wan17a, Kha17, Wan17c, Zha19], data-driven methods [Kha21, Lee18, Zha17, Wan17d] and hybrid methods [Gu17].

Vital signs can be extracted using acoustic sensing technique as well. The acoustic measurement system contains a speaker sending the inaudible sound waves and a microphone which receives the acoustic waves reflected from the measurement subjects. Nandakumar et al. [Nan15] has used a speaker and microphone embedded in a smartphone to detect apnea. The system was able to detect respiratory events with a distance of 0.7 m. Wang et al. [Wan18b] proposed a correlation-based FMCW algorithm which achieves a ranging resolution of about 0.4 cm. Since heartbeats only incur body movements of 0.3-0.8 mm, measuring heart rates is a much more challenging task for systems based on acoustic sensing. Wang et al. [Wan21] proposed to use FMCW chirp signals to extract cardiac rhythm. The measured R-R intervals exhibited a correlation coefficient of 0.93 compared to the ECG reference. While the acoustic system offers a contactless and cost-effective solution, the measurement performance can be limited by factors such as sound attenuation through thick fabric, or the orientation of the body motion with respect to the acoustic devices.

The utilization of depth sensors for vital sign monitoring has been studied as well, including structured-light sensors [Pu13, Pu15], Time-of-Flight sensors [Che16a, Du18] and stereo vision sensors [Sch20], a more detail introduction of which can be found in [Reh20]. Similar to the above mentioned sensors, the depth sensors measure vital signals by capturing the subtle mechanical movements caused by the heartbeats or breath.

Besides observing changes in electrical charge and motions, cardiac and pulmonary parameters can also be measured through thermography using a thermal camera [Sha12]. For instance, the respiratory rate can be determined by observing temperature changes caused by breath flow around the nostrils. Heart rate can be measured by analyzing temperature changes due to blood perfusion in specific areas, such as the neck. The non-contact measurement techniques for vital parameters are summarized in Table 3.1.

This dissertation will apply camera as the sensor for the parameter measurement. The fundamental principle of camera-based measurement is centered around detecting subtle variations in skin color caused by the underlying blood volume fluctuations, modulated by the rhythmic heartbeat. These color variations can be effectively captured by a camera and subsequently enhanced through signal processing approaches. Compared to other non-contact measurement devices, cameras are more accessible and relatively cost-effective. Additionally, by capturing environmental and background information such as facial expressions, cameras can further enrich the dimensionality of the measurement.

3.3 State of the art for remote Photoplethysmography

In the following sections, the state of the art of rPPG is discussed, with subdivisions according to aspects such as the measurement setup and the signal processing algorithms. Several authors, including McDuff [McD23], Yu et al. [Yu21], Chen et al. [Che18b], and Zaunseder et al. [Zau18], have provided comprehensive overviews of the current state of the art, which serve as foundational references for the subsequent discussion as well.

3.3.1 Measurement setup

The first rPPG system was proposed by Wu et al. [Wu00]. They introduced a Charge-Coupled Device (CCD) near-infrared imaging system capable of visualizing vessel networks and assessing peripheral venous system disorders. Takano and Ohta [Tak07] validated the near-infrared camera system's ability to measure cardiopulmonary parameters. The first rPPG system employing an RGB camera was proposed by Verkruysse et al. [Ver08]. Utilizing a consumer-level camera, he was able to measure heart and respiratory rates under daylight conditions and conventional fluorescent lighting. Although initial efforts in rPPG system development were based on near-infrared setups, most contemporary research prefers RGB cameras due to their wide availability and cost-effectiveness.

Usual camera specifications include sensor type, resolution, pixel size, frame rate and pixel bit depth. Further factors defining imaging performance include quantum efficiency, dark noise, and dynamic range of the imaging sensor. Additionally, software settings such as auto exposure, white balance, and auto focus should be taken into account for imaging quality. The optimal specifications are highly dependent on the intended application scenarios. For instance, in environments with constant illumination, auto exposure is typically turned off to keep the exposure time settings stable. However, in environments with significant changes in light, such as in driver state monitoring, it is desirable for the camera to be able to adaptively adjust exposure time to prevent pixel saturation or underexposure. Consequently, it is essential to validate camera configurations in the intended scenario by examining the signal quality using a benchmark algorithm.

Several researchers have developed more sophisticated setups to enhance measurement accuracy. Hülsbusch and Rembold [Hül08] employed four LED modules in their measurement setup to improve illumination homogeneity. Guazzi et al. [Gua15] used diffuse sheets to create a more uniform light reflection onto the subject's face. Amelard et al. [Ame15] introduced a camera system using a temporally-coded illumination sequence. This sequence measures both active and ambient illumination components, helping to mitigate variations caused by changes in ambient illumination. Researchers have also explored multi-imaging setups. Studies by Estepp et al. [Est14] as well as Blackford and Estepp [Bla17b] indicated that multi-imaging systems can provide superior measurement results during motion artifacts compared to single-camera systems.

Some studies have investigated the use of polarized illumination for rPPG measurement, though findings in the literature are inconsistent. While Hülsbusch and Rembold [Hül08] reported no improvement with the use of polarization, a study by Trumpp et al. [Tru17] found that the perpendicular filter setting could help suppress ballistocardiographic effects and enhance signal quality.

This dissertation aims to maximize the potential applicability of the algorithmic system, with minimal requirements for hardware components. Therefore, it will primarily focus on simple measurement setups with single camera system under lighting conditions such as LED or natural daylight.

3.3.2 Detection and tracking of measurement sites

Regarding the Region of Interest (ROI) for measurement, viable areas include the lower leg, the forearm, the palm, and the face. Compared to other regions, the face is typically uncovered and more readily accessible in video images. Additionally, the facial region is more vascularly supplied, yielding superior quality of measurement signals. Figure 3.1 contrasts the signals extracted from various body parts. The signal derived from the facial region displays the most pronounced periodicity, as observed in the figure.

Initial research efforts obtained the region of interest by manually cropping the facial region or selecting a rectangular area on the forehead in the initial frames of the face video [Ver08]. The selected ROI, once defined, was applied consistently to each video frame, predicated on the assumption of minimal or no significant motion in the video. However, even minor head movement could introduce significant noise or distortions into the target signal. Recognizing these challenges, automatic ROI detection methods were proposed

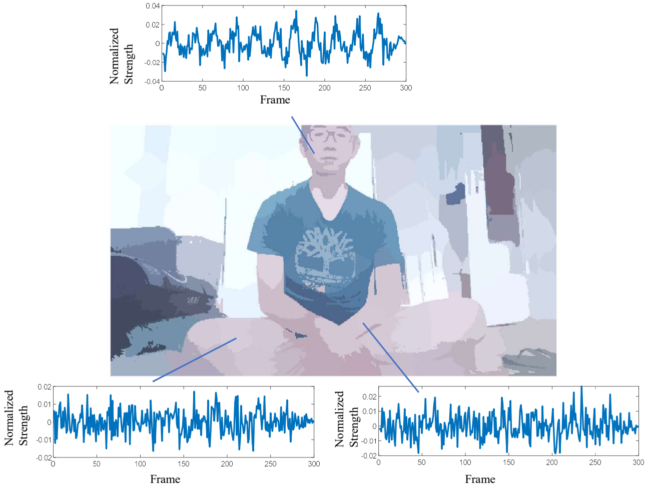


Figure 3.1: Comparision of pulsatility from different measurement sites. The video image was segmented into superpixels using the Simple Linear Iterative Clustering algorithm. The pulse signals were extracted from each superpixel using Plane Orthogonal to Skin (POS) algorithm [Wan16b]. Among all regions, the facial signal exhibits the strongest pulsatility.

to compensate for the head motions. Early methods used an automatic face detection algorithm to localize the face region in the image. The most well-known face detection algorithm is the Viola-Jones algorithm [Vio01]. It detects the face using a cascade of boosted classifiers based on Haar-like features, and then produces a bounding box of the face region. One drawback is that the algorithm searches for the facial region in the entire video image, which slows down the processing, affecting its suitability for real-time applications. Moreover, noise in pixel intensity can lead to variations in the detection results and introduce jittering of box size and location among consecutive frames. Also, the bounding box includes non-skin pixels such as eyes or facial hair, which are not relevant for signal extraction and can introduce disturbances.

With the advancement of computer vision techniques, more efficient and reliable face alignment algorithms have been used for ROI detection, such as Discriminative Response Map Fitting [Ast13], Kanade-Lucas-Tomasi (KLT)

Algorithm [Ast13], Supervised Descent Method [Xio13], and Cascaded Deformable Shape Model [Yu13]. Face alignment not only tracks the face location but also the position of specific facial landmarks, which can be used to localize face regions with stronger pulsatility such as the forehead and cheeks. Similar to the Viola-Jones algorithm, the prediction outputs of the face trackers could also suffer from fluctuations or instability, which further impacts signal extraction. To this end, temporal noise filtering is usually required for the landmark position output by face trackers.

Segmentation of the skin region can be further improved by utilizing color characteristics. This method transforms image pixels into the YCbCr color space, followed by the removal of non-skin pixels by setting a threshold for each color channel. More advanced methods, such as the one proposed by Wang et al. [Wan14], utilize a one-class support vector machines for skin pixel classification. This classification process can also filter out regions with saturated brightness, which is a common cause of measurement accuracy degradation in inhomogeneous illumination scenarios such as driver state monitoring [Blö20].

In addition to using fixed face regions, the region of interest can be adaptively refined by leveraging the spatial redundancy of image sensors. This refining process divides face regions into fine-grained local patches, treating each local patch as an independent sensor [Wan14, Blö20]. Each patch is then individually evaluated based on local pulsating intensity and skin characteristics. A final pulse signal is extracted from the aggregated regions.

One challenge of the multi-patch approach is to align the subregion location in two consecutive frames. Poh et al. [Poh10] utilized the Kanade—Lucas—Tomasi (KLT) feature tracking algorithm to align the subregions based on Speeded Up Robust Features (SURF). Wang et al. [Wan14] proposed the use of the Farnebäck dense optical flow algorithm [Far03] to compensate for local motion. Blöcher [Blö20] defined a dense face tracking model and aligned the local patches using the Supervised Descent Method [Xio13].

The multi-patch approaches demonstrate higher robustness against local disturbances such as facial expressions. Remarkably, multi-patch analysis can even enable signal extraction without face detection. Bobbia et al. [Bob19] used Simple Linear Iterative Clustering (SLIC) to decompose the video image into several superpixels, and then calculated the pulse signal from the superpixels that exhibited pulse-related variations.

3.3.3 Signal processing algorithms

Once the ROI has been determined, the color signals are obtained from the color intensity of pixels inside the area. First studies of rPPG [Ver08] demonstrated that the target pulse signal can be recovered from the color signal of one single channel. Changes in skin color indicating the target signals are very subtle and susceptible to distortions including body motions and illumination changes. To improve the signal quality in presence of disturbances, a vast amount of methods have been proposed. Primarily, this involves the use of detrending methods or bandpass filters. The cut-off frequencies of the bandpass filters are based on the physiological range of the heart rate (e.g. 40-240 bpm in [De 13]). Bousefsaf et al. [Bou13] adopted continuous wavelet filters to process the signal. Wang et al. [Wan17b] proposed the amplitude-selective filter based on the fact that the pulse signal from human skin has only a limited relative pulsatile amplitude. Blöcher et al. [Blö18] proposed to reduce the affect of artefacts for heart rate measurement by adaptively setting the cut-off frequencies of the filters. Huang et al. [Hua17] adopted a Least Mean Squares (LMS) based adaptive filter to measure signals while performing periodical exercises on fitness machines. Li et al. [Li14] employed Normalized Least Mean Squares (NLMS) adaptive filter to rectify the interferences of illumination variations. Additionally, the application of Empirical Mode Decomposition (EMD) [Che16b] and Kalman Filter [Pra18, Jia14] for disturbance compensation were discussed as well.

Spectral redundancy can also be exploited to effectively improve the signal measurement. Blind Source Separation (BSS) is one group of the methods which considers the color signals as a mixture of pulsatile signals with noise and disturbance. Two representative BSS methods are Principal Component Analysis (PCA) [Lew11] and Independent Component Analysis (ICA)

[Poh11]. The BSS methods de-mix the signals based on the statistical characteristics of the underlying signal components. For instance, ICA recovers the pulse component by maximizing the non-gaussianity of the recovered pulse signals. However, several works reported no significant improvement of measurement accuracy when applying BSS methods [Zau18]. Also, the outcome of BSS is dependent on the preselection of input signals. Additionally, since no unique ordering of components is offered after signal separation, the problem of permutation indeterminacy should be addressed for the application of BSS-based methods.

Besides statistical properties, signal extraction can be improved by leveraging the physical and optical characteristics of rPPG signals. One of the critical characteristics used for rPPG algorithm design is the varying relative PPG-amplitude over the spectrum, which has been introduced in Figure 2.8. From the model by Hülsbusch and Rembold [Hül08], De Haan and Van Leest [De 14] identified the normalized BVP vector. This corresponds to the relative PPG-amplitude in temporally normalized color signals. The pulse signal is extracted by projecting the color signal onto a vector, resulting in the relative amplitude on each channel proportional to the Blood Volume Pulse vector. Wang et al. [Wan19] extended the BVP vector by accounting for disturbance signals, improving the measurement in NIR setup. Skin tone is another characteristic used for pulse signal extraction. De Haan and Jeanne [De 13] proposed the Chrominance-based method (CHROM) to extract the blood volume pulse signal by assuming a standardized skin-color to white-balance images. Wang et al. [Wan16b] proposed the Plane Orthogonal to Skin (POS) algorithm, which extracts BVP signals by projecting color traces onto a plane where intensity components cancel out. Both POS and CHROM algorithms project the temporally normalized color signals onto a predefined pair of orthogonal vectors and refine the pulse signal using alpha-tuning [De 13], which will be discussed in detail in Chapter 5. Definition of the projections vectors requires prior knowledge about the color direction of disturbances and pulse components.

Differing from non-supervised methods using linear projection to estimate the physiological signal, supervised methods typically use machine learning methods to regress a non-linear mapping from color signals to pulse signals. Many works have been proposed to use deep networks to extract rPPG signals from image sequences. Usually, these methods take face images [Spe18, Yu19a] or the difference between consecutive frames [Che18a] as the network input. Chen and McDuff [Che18a] introduced an end-to-end Convolutional Attention Network (CAN) that estimates pulse signals from the normalized difference between two consecutive frames, aided by an attention branch with appearance inputs. McDuff et al. [McD20] proposed a method to use synthetic data to train CAN. Spetlík et al. [Spe18] proposed to estimate heart rates with a two-step convolutional network, in which networks for pulse extraction and heart rate prediction were designed separately. Yu et al. [Yu19a] manipulated the spatio-temporal features of face images using a 3D convolutional neural network and recurrent neural network.

In addition to using the face image as the network input, other networks extract signals from stacked color signals. RhythmNet [Niu19] and SynRhythm [Niu18] extract color signals from subregions of faces and concatenate the color signals into a spatial-temporal map. Hsu et al. [Hsu17] applied short-time Fourier transform on the color signals and used the 2D time-frequency representation as input for a VGG network [Sim14]. Several studies utilizing more sophisticated network architectures have been proposed to enhance the measurement accuracy. Yu et al. designed a video-to-video generator to improve the signal quality from compressed videos [Yu19b]. Niu et al. [Niu20] adopted disentanglement representation learning to distill the physiological features from non-physiological information such as head motions and lighting change. Liu et al. [Liu20b] and Lee et al. [Lee20] proposed to use meta-learning to improve cross-dataset performance.

3.3.4 Application of rPPG in stress and emotion recognition

The remote Photoplethysmography (rPPG) technique offers a way to measure BVP signals contactlessly through video cameras. From the derived BVP, PRV can be determined, which can be considered a surrogate of HRV and is related to the activity of the autonomic nervous system.

A limited number of studies have been carried out to explore the possibility of stress detection from rPPG signals. Bousefsaf et al. [Bou13] formulated a stress index to describe stress states considering the trend variations, rhythmic fluctuations and frequency variations from the remote PRV signals. McDuff et al. [McD14, McD16] and Sabour et al. [Sab21] used classical feature-based machine learning methods to recognize cognitive and social stress from face videos.

Such techniques often necessitate expert knowledge to select the classification features manually, a non-trivial process due to the ambiguous relationship between physiological features and affect states. Given that the selected features may not cover the entire feature space for accurate identification of different stress states, the question arises: Could deep networks simplify and enhance the process by analyzing rPPG-derived signals without the need for intricate PRV feature computation?

For emotion recognition, the application of remote Photoplethysmography has been explored to a limited extent in the literature, often exhibiting constrained accuracy or with a focus on specific emotion categories. Burzo et al. [Bur12] attempted to differentiate emotions relying exclusively on heart rate features captured by camera. The resulting accuracy was not ideal, registering at just 53.57% when distinguishing between positive and negative emotions. In another attempt, Benezeth et al. [Ben18] utilized frequency domain features derived from camera-based PRV to detect "disturbed emotions". Notably, both investigations emphasized categorical emotion interpretation.

Since emotions are not always experienced in distinct categories, but rather along continuous dimensions, it would be interesting to explore if rPPG based features can also be utilized for dimensional affect recognition. Taking cues from Bugnon et al. [Bug17]'s work, which demonstrated that HRV features can be used for fine-grained emotional analysis, it stands to reason that delving into the prospects of camera-derived PRV parameters for a dimensional emotion analysis could be a worthwhile endeavor.

3.4 Contributions and distinctions from current works

From current literature, it's clear that stress and emotion recognition through camera-based physiological data is still in its infancy. This dissertation aims to contribute to this area by focusing on the design of a new rPPG algorithm and its application for stress and emotion analysis.

Building upon the notable efficacy of neural networks in diverse computer vision and signal processing domains, this dissertation will first seek to design a deep learning method for the extraction of pulse signals via camera. It begins by dissecting the performance of neural networks in terms of remote BVP signal extraction from two perspectives: temporal processing and spatial operation. Given that current deep learning-based methods utilize only minimal knowledge of the physical rPPG process, as identified in the review of the state of the art, this study aims to bridge the gap between the design process of the deep learning network and traditional rPPG algorithms. For the temporal processing algorithm, this work will investigate if combining the fast response of the traditional framework to changing measurement conditions, with the non-linear temporal modeling ability of neural networks, could facilitate the rPPG signal extraction. Additionally, to leverage the spatial redundancy present in image data, the work will discuss the incorporation of spatial operation into the convolutional network in order to adaptively extract pulsatility-related features from the ROI. Moreover, face normalization is investigated for robustness improvement of the network against motion.

Bridging from the discussions in Section 3.3.4, the dissertation will assess the viability of implementing end-to-end methods for the recognition of cognitive stress using camera-based signals. The investigation aims to determine whether it's possible to detect cognitive stress from camera-derived signals using deep learning methods, without calculating pulse rate variability. This is the first work that explores the application of deep learning methods for stress recognition based on physiological signals derived from cameras.

In addition to stress recognition, the dissertation explores the potential of using parameters derived from camera-captured physiological signals to measure emotions. Specifically, it focuses on measuring dimensional emotions based on measurements through rPPG, which is an area that has not been extensively explored in the current state of the art.

4

System Concept

This chapter provides an overview of the system concept for the development of an affect recognition system based on remote Photoplethysmography (rPPG). The hardware configuration will be specified, drawing from insights gained through the analysis of the current state of the art in Chapter 3. A comparison will be conducted with other measurement setups to validate and justify the configuration of the proposed system. Additionally, the software architecture will be introduced, offering a comprehensive description of the sub-components and their respective interfaces. This detailed exploration will provide a solid foundation for understanding the fundamental functions and interactions of the system subcomponents.

4.1 System specification

The system specification for this dissertation adopts a straightforward measurement setup utilizing RGB color cameras. This selection is driven by various considerations. Firstly, the simplicity of the setup, comprising only a camera and computing component, allows for seamless integration into diverse environments, facilitating the system's prototypical implementation in various application scenarios. In situations where sufficient daylight is available, there is no need for an additional lighting module. In cases where measurements are conducted at night or in environments with limited daylight, a cost-effective LED can be effortlessly augmented to provide adequate illumination.

From the perspective of principles, the measurement using RGB color cameras provides higher signal quality. Firstly, the RGB camera sensors offer the possibility to exploit the spectral redundancy of the signals, allowing for noise compensation by means of multi-channel analysis. Secondly, the absorption maxima of oxygenated hemoglobin lies in the green wavelength range (541nm and 577nm), as shown in Figure 2.8. That means, the blood volume-dependent absorption effect of light can be more sensitively detected by means of RGB cameras. Generally, compared to measurement with near infrared spectral range, measurements in visible range have stronger relative rPPG strength, which gives robuster measurement of the target parameters.

Another notable benefit of employing RGB sensors is the availability of a wide range of low-cost cameras. These off-the-shelf cameras can offer satisfactory measurement accuracy. For this system, webcams (*Logitech c920* and *c922*, priced under 100 euros) were chosen for data collection during the evaluation of all algorithmic components. These cameras have previously proven to meet the minimum requirements for effective signal capture [Blö20]. In more challenging scenarios, it is feasible to replace the image acquisition system with higher-quality sensors to cater to specific application demands.

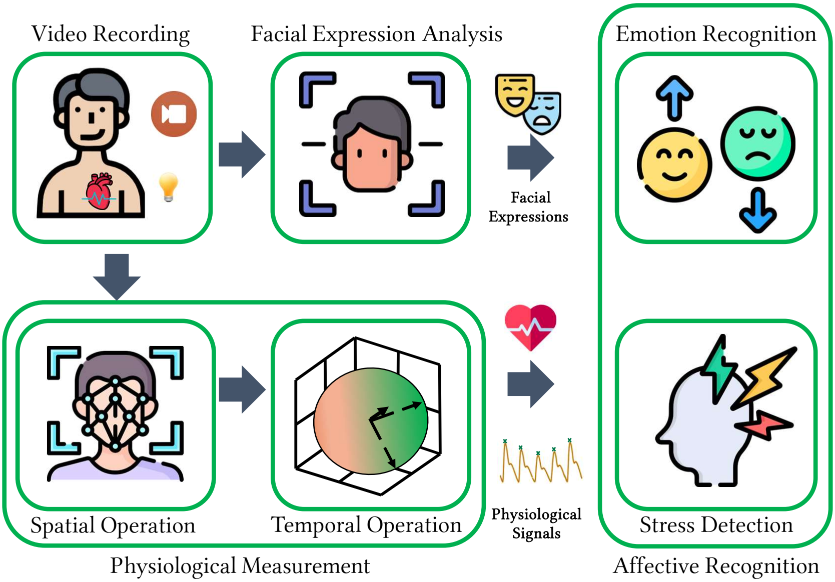


Figure 4.1: Overview of the software architecture. [1]

4.2 Software architecture

The purpose of this section is to develop the concept of software architecture. The software architecture is shown in Figure 4.1 and can be broken down into several sub-components:

- *Read-in component:* Component for loading and storing image data and occasionally reference data for validation
- *Physiological measurement component:* Algorithmic component to extract physiological parameters from images
- *Affective recognition component:* Component for affect status assessment based on the camera-based physiological parameters

[1] This diagram has been designed using icons from Flaticon.com.

The read-in component serves as the interface to the camera system, necessary for the loading and storage of camera images. These stored images are subsequently used for offline analysis. In order to validate the results, the read-in component also registers and synchronizes any available reference data, such as ECG or PPG signals.

The primary focus of this dissertation lies in the development of the algorithmic components. The physiological measurement component is specifically designed to extract physiological signals from videos. Following the standard processing pipeline for rPPG, the physiological measurement component can be further divided into two main parts:

- SO *Spatial operation component:* Evaluation and selection of the measurement site (ROI) for signal extraction
- TO *Temporal operation component:* Signal processing to extract blood volume pulse signals from the selected measurement site

As indicated in Chapter 2, stress and emotion recognition will be addressed separately within this dissertation. Consequently, the affective recognition component can be further divided into two distinct subcomponents:

- SD *Stress detection component:* Stress detection based on physiological data measured using the rPPG measurement component
- ER *Emotion recognition component:* Emotion evaluation based on the parameters measured by the rPPG measurement component

Given that Facial Expression Recognition (FER) is one of the most widely researched approaches for emotion recognition, this work will also compare the effectiveness of rPPG-based recognition with a FER model. The interfaces of these components are further elaborated in terms of input and output data within Table 4.1.

Chapter 5 will provide an introduction to the core function of the rPPG algorithm, namely its temporal signal processing. This chapter will focus on the development of a new temporal operation method, which form the foundation of the algorithm's functionality.

Table 4.1: Component interfaces with input and output data

Component	Input	Output
Read-in component	Raw image data	Image frames in RGB color space ($H \times W \times 3 \times T$)
SO component	Image frames in RGB color space	Color features
TO component	Color features	Signal of length T in BVP Signal Buffer ($1 \times T$)
SD component	Signal of length T in BVP Signal Buffer	Stress level
ER component	Signal of length T in BVP Signal Buffer	Valence and arousal scores

Building upon the temporal processing algorithm, Chapter 6 will present an automatic method that effectively processes spatial pixel information in an adaptive manner. The chapter will delve into the details of this method, highlighting its significance in improving the overall performance and accuracy of the rPPG algorithm.

Upon this, the capabilities of the entire system to assess stress and emotion status will be extensively discussed in Chapter 7 and 8. These chapters will explore the implementation of algorithms in detecting and categorizing stress levels as well as evaluating dimensional emotional states. The comprehensive analysis will shed light on the potential applications and benefits of the rPPG technology in affective recognition.

5

Short Window Network for Remote Photoplethysmography

Accurate extraction of physiological signals is fundamental for recognizing individuals' physical and psychological status. As introduced in Chapter 4, the signal extraction is accomplished by the spatial and temporal operations, with the temporal operation being the core of the rPPG algorithm. This chapter will discuss the temporal operation method proposed in this dissertation. The work presented in this chapter has been published previously in the paper: "Short Window Network for Remote Heart Rate Measurement" [Zho21].

As discussed in Chapter 3, the core remote Photoplethysmography (rPPG) algorithms can be broadly classified into non-supervised and supervised

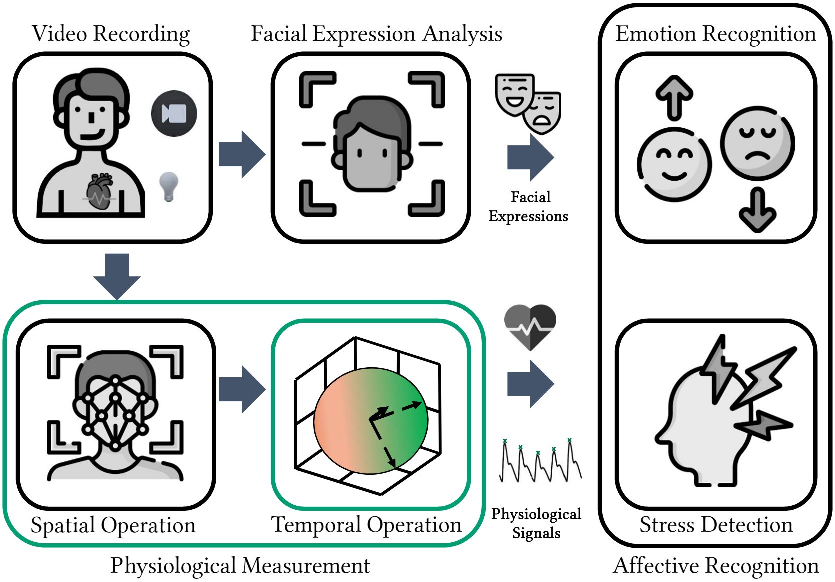


Figure 5.1: Overview of the software architecture - Temporal Operation.

methods. Supervised and non-supervised methods follow different processing paradigms. Unlike non-supervised methods that use linear projection for estimating the physiological signal, most supervised methods utilize neural networks to establish a non-linear mapping from color signals to pulse signals. While non-supervised core rPPG algorithms typically focus on temporal operations and require additional spatial operations such as pixel averaging of a selected skin region or using a more sophisticated multi-site strategy [Wan14], supervised methods simultaneously extract spatial and temporal features, from cropped face images [Che18a, Yu19a, Liu20b] or stacked color signals that are extracted from different facial locations [Niu20, Niu19]. In this regard, it is challenging to determine the extent to which the improved performance of deep learning algorithms is attributed to their temporal modeling capabilities.

Furthermore, the design of supervised methods often deviates from the framework of classical rPPG algorithms. On the other hand, insights gained from the design of classical rPPG algorithms can still be applied to supervised methods. Specifically, the short-window and overlap-adding strategy [De 13], a classical processing pipeline that allows algorithms to adapt more effectively to changing measurement conditions, could also benefit deep learning algorithms. Thus, we propose integrating a deep network into the short-window and overlap-adding pipeline as a replacement for conventional rPPG algorithms. By sharing the same pre- and post-processing steps with classical methods, this approach enables a direct and fair comparison of temporal operation ability between deep networks and conventional linear projection algorithms.

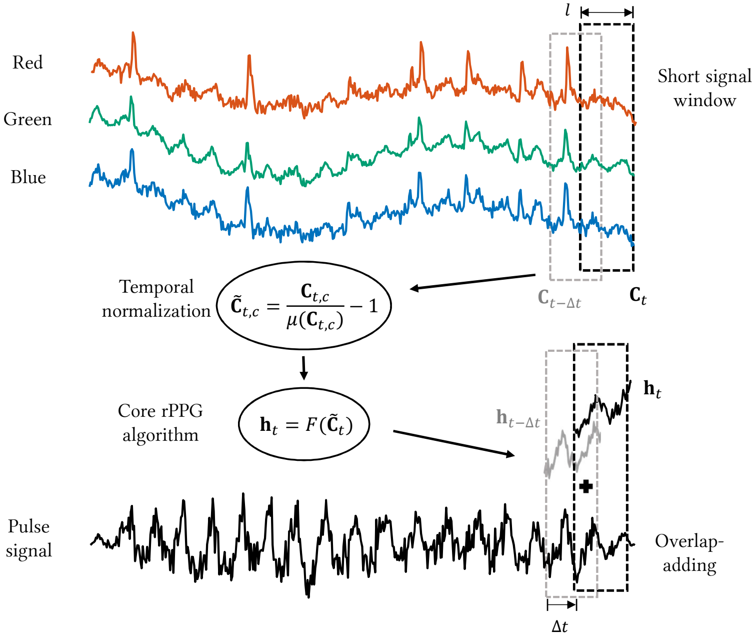
The remainder of this chapter is structured as follows. In Section 5.1, we will introduce the proposed method in detail. In Section 5.2, a benchmark experiment will be conducted to evaluate the measurement performance of the proposed method. In the final Section 5.3, we draw conclusions.

5.1 Method

In Section 5.1.1, we commence our discussion by examining the short-window and overlap-adding pipeline employed in the model-based algorithms. This pipeline is the foundation for the subsequent discussions. We then discuss about the limitation of traditional projection methods for rPPG. Following this, the network architecture will be introduced in Section 5.1.2.

5.1.1 Short observation window and overlap-add

The main task of rPPG is to derive blood volume pulse signals from face video recording. In classical rPPG algorithms, this process involves extracting raw signals from each color channel by spatially averaging pixel values inside the ROIs for each individual video frame. The time-varying raw color signals are typically expressed as a matrix $\mathbf{C} \in \mathbb{R}^{3\times T}$ with T being the signal length (with unit in frame). The model-based rPPG algorithms extract the pulse signal using the short-window and overlap-adding schema, in which pulse signals are first retrieved from short color signals and then overlap-added into a long time pulse signal. Unlike the BSS-based methods which require a relatively longer observation interval to ensure sufficient resolution in the frequency domain, methods using the short-window and overlap-adding pipeline have faster adaptation to disturbances and changing measurement conditions [De 13]. The short-window and overlap-adding pipeline is illustrated in Figure 5.2. The red, green and blue traces are raw color signals extracted from a face video. At a given time point t, sampling short color signals from the long



l
Red
Green
Blue
Short signal window
Temporal normalization
$\tilde{\mathbf{C}}_{t,c} = \frac{\mathbf{C}_{t,c}}{\mu(\mathbf{C}_{t,c})} - 1$
$\mathbf{C}_{t-\Delta t}$
$\mathbf{C}_t$
Core rPPG algorithm
$\mathbf{h}_t = F(\tilde{\mathbf{C}}_t)$
$\mathbf{h}_{t-\Delta t}$
$\mathbf{h}_t$
Pulse signal
Overlap-adding
Δt

Figure 5.2: The long color signals are sampled by a short time window. After being temporally normalized, the short window signals are mapped into a short pulse signal using a core rPPG algorithm. The short pulse signals are gluing together into a long pulse signal using overlap-adding.

signal traces $\mathbf{C}$ can be expressed as:

$$\mathbf{C}_t = \mathbf{C}\mathbf{L}_t, t \in \{0, 1, 2, ..., T - l\}, \tag{5.1}$$

with the entry $\mathbf{L}_{t,ij}$ of the sampling matrix $\mathbf{L}_t \in \mathbb{R}^{T \times l}$ being 1 only if $i - j = t$, otherwise zero, which can be expressed using the Kronecker delta function:

$$\mathbf{L}_{t,ij} = \delta(t - i + j). \tag{5.2}$$

The sampling Equation 5.1 yields short color signals $\mathbf{C}_t \in \mathbb{R}^{3 \times l}$ with the length l being equal to the observation window length. The observation window moves along the time axis with a step size of Δt, which is set as 1 frame in this work.

Time-varying components of color signals are of greater interest in the context of rPPG. Since the pulse component has a relatively small amplitude in the color signals, temporal normalization is required to remove the stationary components before mapping the color signals to pulse signals. Temporal normalization is achieved via dividing the color signals by their stationary components and then removing the average (which is equal to $\mathbf{1}$):

$$\tilde{\mathbf{C}}_{t,c} = \frac{\mathbf{C}_{t,c}}{\mu(\mathbf{C}_{t,c})} - 1, c \in \{r, g, b\} \tag{5.3}$$

where c denotes the index of the color channel and $\mu(\cdot)$ calculates the stationary components. Temporal normalization eliminates the dependency of the signals on the stationary color as well as brightness level of the light source, yielding zero-centered color signals. The pulse signal $\mathbf{h}_t$ is extracted using a core rPPG algorithm on $\tilde{\mathbf{C}}_t$:

$$\mathbf{h}_t = F(\tilde{\mathbf{C}}_t), \tag{5.4}$$

where $\mathbf{h}_t$ has a length of l. Then, overlap-adding the short signals $\mathbf{h}_t$ gives the pulse signal $\mathbf{H}$ with the original length T:

$$\mathbf{H} = \sum_{t=0}^{T-l} \mathbf{h}_t \mathbf{L}_t^\top . \tag{5.5}$$

Linear projection in classical methods

Conventional methods for rPPG signal extraction can be considered as looking for a projection vector $\mathbf{p} \in \mathbb{R}^{3\times1}$ in the color space, on which the projected signals have the highest pulsatility. Thus, the core function in Equation 5.4 can be then expressed as:

$$\mathbf{h}_t = F(\tilde{\mathbf{C}}_t) = \mathbf{p}^\top \tilde{\mathbf{C}}_t . \tag{5.6}$$

Various rPPG methods select the projection vector $\mathbf{p}$ by utilizing signal characteristics from various aspects. For example, ICA treats the color signals as a mixture of plethysmographic signals and other disturbances caused by artifacts. It assumes that the underlying signal sources are independent of each other and seeks to determine a de-mixing matrix that maximizes the non-Gaussian of each source [Poh10].

In addition to utilizing statistical characteristics, the determination of the projection vector can also be based on the optical properties of the signal components. The model-based algorithms, for instance POS and CHROM, decompose the time-varying components of color signals into a combination of pulsatile components along with intensity and specular disturbances, as discussed in Chapter 2. These algorithms initially project the normalized color signals onto two pre-defined vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, resulting in two signal traces $\mathbf{s}_1$ and $\mathbf{s}_2$:

$$[\mathbf{s}_1^\top, \mathbf{s}_2^\top]^\top = [\mathbf{p}_1, \mathbf{p}_2]^\top \tilde{\mathbf{C}}_t . \tag{5.7}$$

Subsequently, the signals are refined using the alpha-tuning method. The resulting short rPPG signal can be written as:

$$\mathbf{h}_t = \mathbf{s}_1 + \beta \cdot \frac{\sigma(\mathbf{s}_1)}{\sigma(\mathbf{s}_2)} \cdot \mathbf{s}_2 \,, \tag{5.8}$$

where $\beta = -1$ for CHROM and 1 for POS. $\sigma(\cdot)$ calculates the standard deviation of the signals.

The projection of color signals onto two vectors can be seen as a projection onto a plane within the RGB space. In model-based methods, this plane is specifically designed to compensate for certain disturbances, such as intensity disturbance in the case of the POS algorithm and specular disturbance in the case of the CHROM algorithm. The alpha-tuning process is based on the assumption that if the target signal components are in-phase in $\mathbf{s}_1$ and $\mathbf{s}_2$, the disturbance components will be anti-phase in the two signal traces, and vice versa [Wan16b].

Therefore, for POS and CHROM algorithms, the signal extraction process can be conceptualized as initially constraining the projection vector onto a predefined plane, followed by further refinement using alpha-tuning. For POS algorithm, the final projection vector can be expressed as:

$$\mathbf{p}_{\mathrm{POS}} = \frac{1}{\sqrt{1 + \sigma^2(\mathbf{s}_1)/\sigma^2(\mathbf{s}_2)}} (\mathbf{p}_{\mathrm{POS},1} + \frac{\sigma(\mathbf{s}_1)}{\sigma(\mathbf{s}_2)} \mathbf{p}_{\mathrm{POS},2}) \,, \tag{5.9}$$

where $\mathbf{p}_{\mathrm{POS},1} = 1/\sqrt{2} \cdot [0, 1, -1]^\top$ and $\mathbf{p}_{\mathrm{POS},2} = 1/\sqrt{6} \cdot [-2, 1, 1]^\top$. Since $\mathbf{p}_{\mathrm{POS},1}$ and $\mathbf{p}_{\mathrm{POS},2}$ are in unit length and orthogonal to each other, the coefficient $1/\sqrt{1 + \sigma^2(\mathbf{s}_1)/\sigma^2(\mathbf{s}_2)}$ ensures that $\mathbf{p}_{\mathrm{POS}}$ is a unit vector. The projection vector for CHROM can be determined in an analogous fashion.

The projection vector can also be determined based on frequency characteristics of the signals. The Projection using Spectral Characteristics (PSC) [Zho20] algorithm, for example, extracts the pulse signal by searching for the projection vector that maximizes the ratio of LF to HF energy in stationary scenarios, and minimizes the LF energy in the presence of disturbances. Figure 5.3 illustrates signals obtained by projecting a set of color signals contaminated by disturbances onto different vectors on a predefined plane. It is demonstrated that the signal projected onto different vectors shows various frequency distribution and different levels of pulsatility, with the highest level of pulsatility exhibited by the signal with the minimal HF energy.

While the process to determine the projection vector **p** is non-linear, such as using Equation 5.9 for the model-based algorithms or through an optimization step for the PSC algorithm, the pulse signal is reconstructed by Equation 5.6 in these methods, which means that the nature of mapping from color signals to pulse signals is still linear. Figure 5.4 provides a comparison of signals

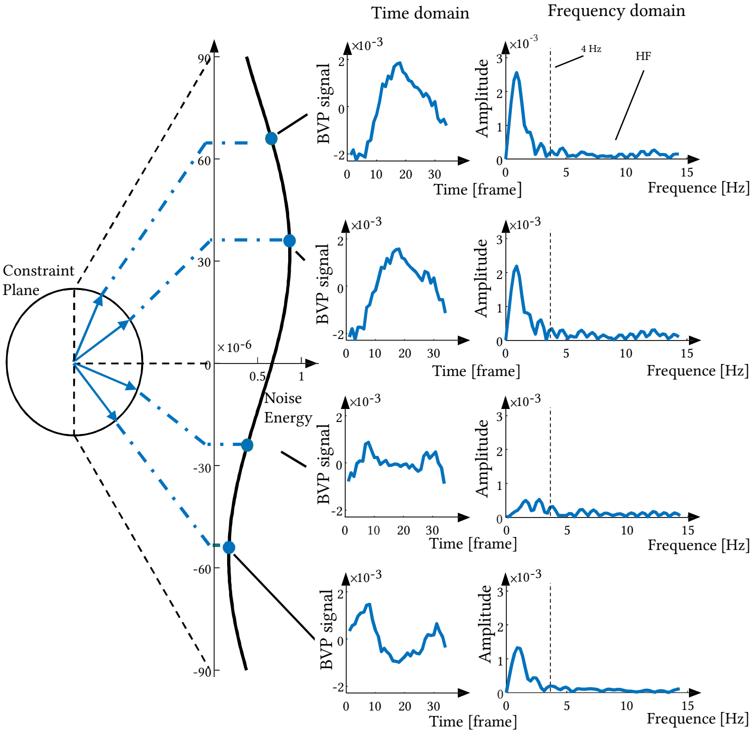


Figure 5.3: Projecting the color signals onto four distinct vectors on the constraint plane yields signals with varying pulsatility and HF energy. The signal obtained with the minimal HF energy (the lowest plotted) exhibits the highest level of pulsatility.

extracted using the POS and PSC algorithms. Although these methods use different approaches to determine the projection vectors and there is a considerable distance between the vectors in the color space, the signals obtained using these methods exhibit a high level of morphological similarity.

The linear projection methods also share certain common limitations. In scenarios where disturbances are present, the linear projection methods determine the projection direction by approximately minimizing disturbance energy. Since the energy of disturbances is usually much higher than the target signal, the resulted projection vector depends heavily on the relative strength of the disturbance in the color channel. If the disturbance has a smaller angle relative to the target components in the color space, the projection will also cancel out the energy of pulsatile components. This explains exactly the reduced signal amplitude observed in segments (II), (IV), and (VI) in Figure 5.4, where disturbances are present.

Moreover, the determination of the projection vector in the classical methods is usually based on certain assumptions or approximations about the physical properties of the rPPG process, for example, the linearity of the description model in Equation 2.20, or the in/anti-phase properties, upon which the alpha-tuning was developed. In situations where these assumptions are invalid, the pulse signal could then not be properly derived.

It is also important to note that linear projection methods have only explored a limited region within the overall function space, as indicated by the morphological similarity of signals extracted using the POS and PSC algorithms. Given that the skin-light interaction involved in rPPG is a highly non-linear process, there is a motivation to extend the core rPPG algorithms to a more complex function space. In light of this, we propose replacing the linear core rPPG algorithm with a deep neural network, allowing for the learning of a non-linear mapping from skin color to pulse signals.

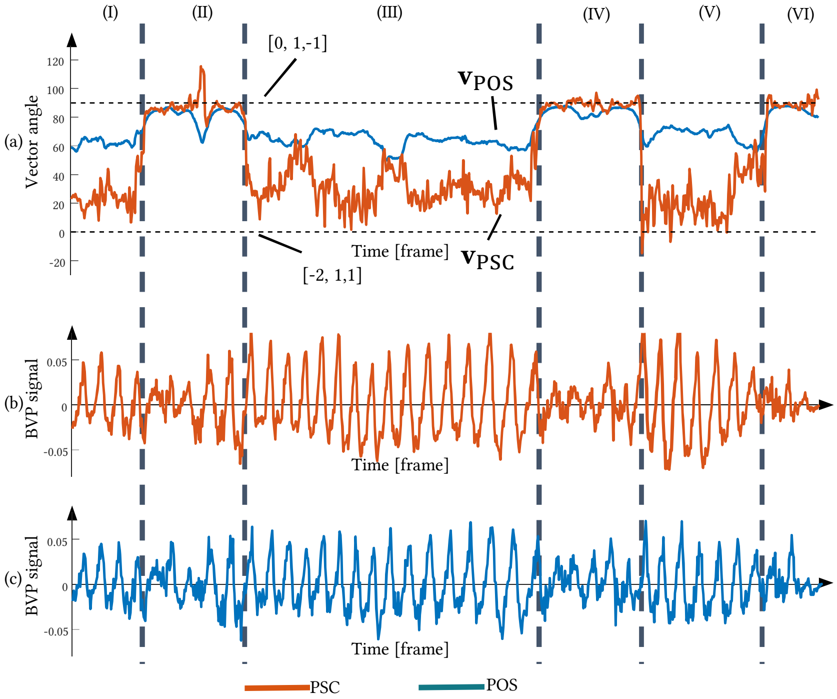
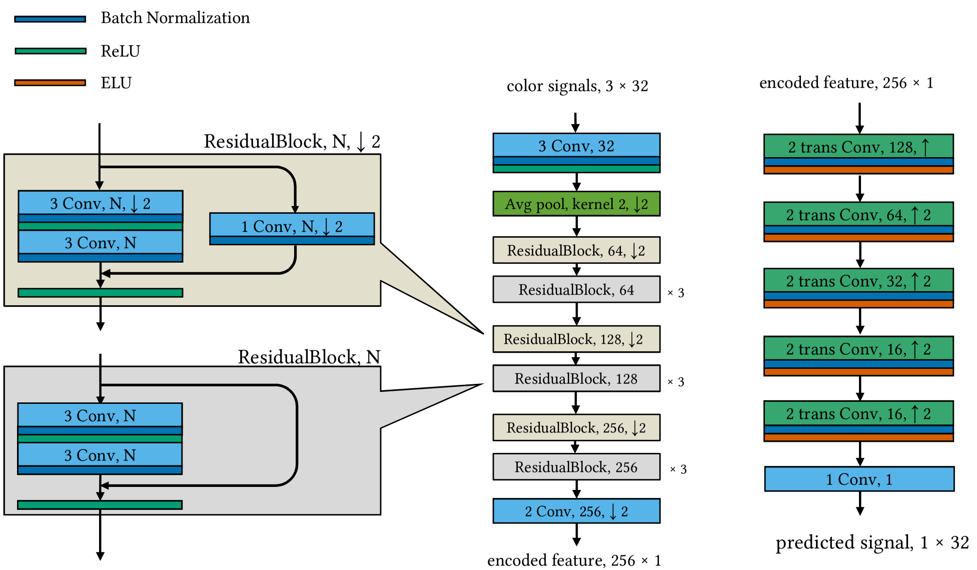


Figure 5.4: Comparison of projection vectors obtained by POS and PSC. The projection vectors in (a) are constrained on the plane orthogonal to $[1, 1, 1]^{\top}$. The directions of the projection vectors are close when a large disturbance occurs in (II), (IV) and (VI). Despite differences in the projection angle in stationary scenarios (I, III and V), the pulse signals extracted by PSC (b) and POS (c) are morphologically similar to each other.

5.1.2 Short window network

The proposed neural network is specifically designed to integrate into the processing pipeline of classical linear methods, where the neural network is selected as the core function represented by Equation 5.4, while the remaining steps of the processing pipeline are kept unchanged. The utilization of the classical short-window overlap-adding pipeline provides several noteworthy advantages in this particular context. As previously discussed, this approach significantly enhances the algorithm's adaptability to distortions by



(a) Residual Blocks

(b) Encoder

(c) Decoder

Figure 5.5: The network consists of an encoder (b) and a decoder (c). The main architecture of the encoder is a cascaded connection of residual blocks (a). The number after "↓" stands for the stride value in the convolutional and average pooling layers, while the one after "↑" represents the stride value in the transposed convolutional layers. The setting of the stride controls the up-sampling and down-sampling behavior of the network. The encoder projects color signals into the high dimensional feature space, while the decoder transforms the features back into a temporal signal.

effectively limiting the temporal impact range of disturbances. Moreover, the implementation of a short sampling window not only reduces the size of input and output data, but also contributes to the reduction in the overall size of the network model, thereby enhancing its generalization capability. Furthermore, the network's repetitive signal processing on each frame within the overlap-adding pipeline introduces the bagging mechanism, enhancing the overall robustness of its predictions.

Inspired by Niu et al. [Niu20], we choose the encoder-decoder structure as the network backbone. The network architecture is shown in Figure 5.5. The encoder E temporally down-samples the input color signals into a high-dimensional feature space. Down-sampling is controlled by the stride step in the average pooling and 1-D convolutional layers. The decoder D transforms the high-dimensional features back to a temporal signal through a series of transposed convolutional layers. The pulse signal can be written as:

$$\mathbf{h}_t = D(E(\tilde{\mathbf{C}}_t, \boldsymbol{\alpha}_E), \boldsymbol{\alpha}_D)\,, \tag{5.10}$$

where the core function $F(\cdot)$ in Equation 5.6 for linear projection methods is replaced by the encoder-decoder $D(E(\cdot, \boldsymbol{\alpha}_E), \boldsymbol{\alpha}_D)$. $\boldsymbol{\alpha}_E$ and $\boldsymbol{\alpha}_D$ are the concatenations of all parameters of the encoder and decoder respectively. Combining with Equation 5.5, the long-interval pulse signal extracted by the network is expressed as:

$$\mathbf{H} = \sum_{t=0}^{T-l} D(E(\tilde{\mathbf{C}}_t, \boldsymbol{\alpha}_E), \boldsymbol{\alpha}_D)\mathbf{L}_t^\top\,. \tag{5.11}$$

The network is trained by minimizing the negative Pearson correlation [Yu19a] between the prediction $\mathbf{s}_{pre}$ and ground truth signal $\mathbf{s}_{gt}$:

$$L_{nP}(\mathbf{s}_{pre}, \mathbf{s}_{gt}) = -\frac{Cov(\mathbf{s}_{pre}, \mathbf{s}_{gt})}{\sqrt{Cov(\mathbf{s}_{pre}, \mathbf{s}_{pre})}\sqrt{Cov(\mathbf{s}_{gt}, \mathbf{s}_{gt})}}\,, \tag{5.12}$$

which drives the network to generate a signal correlated to the target signal. In our context, $\mathbf{s}_{pre}$ is the predicted pulse signal $\mathbf{h}_t$ and the **PPG** signal is the target ground truth $\mathbf{s}_{gt}$. The loss function of the whole length predicted signal is represented as the penalty summation of the short window signals $\mathbf{h}_t$. Considering Equation 5.10, it can be expressed as:

$$\begin{aligned} L_{rppg}(\boldsymbol{\alpha}_E, \boldsymbol{\alpha}_D) &= \sum_{t=0}^{T-l} L_{nP}(\mathbf{h}_t(\boldsymbol{\alpha}_E, \boldsymbol{\alpha}_D), \mathbf{PPG}_t) \\ &= \sum_{t=0}^{T-l} L_{nP}(D(E(\tilde{\mathbf{C}}_t, \boldsymbol{\alpha}_E), \boldsymbol{\alpha}_D), \mathbf{PPG}_t)\,, \end{aligned} \tag{5.13}$$

where $\mathbf{PPG}_t = \mathbf{PPG} \cdot \mathbf{L}_t$ is the cropped PPG signal.

As we can see, the network operates in the identical pipeline of model-based methods. It takes the same input signals cropped by a short interval window, and the output signals are further processed using the same steps as in CHROM and POS. Since the network extracts pulse from short color signals, we name the Short Window Network (SWN).

5.2 Experiment

In this section we evaluate measurement performance of the proposed SWN. The evaluation dataset is introduced in Section 5.2.1. Section 5.2.2 describes the implementation details. The experiment results are reported in Section 5.2.3.

5.2.1 Dataset

The VitalCamSet [Blö19], which contains videos of 26 participants, was used for the evaluation. Each video was recorded at 30 frames per second (fps) and lasts for 120 seconds. PPG signals were recorded as ground truth signals using an oximeter. The experiment was conducted on RGB videos from 7 scenarios. The first scenario is a stationary scenario without considerable head motion and illumination change. Scenarios 103 - 104 are scenarios with various illumination changes. The illumination was controlled by means of a smart

Table 5.1: Scenario overview of VitalCamset

Disturbance	Nr.	Scenario	
		Name	**Description**
Still	101	Natural lighting	Daylight without movement
Lighting change	103	Abrupt changing lighting	Scenario with rapid and smaller lighting changes
	104	Slowly changing lighting	Scenario with slow and larger lighting changes
Motion	201	Rotatory movement	Motion scenario with rotatory head movement
	202	Scaling movement	Motion scenario with scaling head movement
	203	Translatory movement	Motion scenario with translatory head movement
	204	Text writing	Motion scenario by writing a text

home light shutter. Scenarios 201 - 204 simulate situations with movement disturbances, where the participants were instructed to conduct various head motions. An overview of the scenarios is shown in Table 5.1.

5.2.2 Implementation details

We extracted the face landmarks using the open-source face tracker OpenFace [Bal16], as it has less jittering effect and more stable alignment for large head motions. A region of interest was defined based on the landmarks output by the face tracker, as shown in Figure 5.6. We calculated raw color signals by averaging pixels inside the defined region and then extracted pulse signals via different core rPPG algorithms.

The model-based core algorithms CHROM and POS, and the proposed SWN were implemented in the short-window overlap-adding pipeline. The window length l was set as 32 frames. As BSS-methods require a sufficient observation

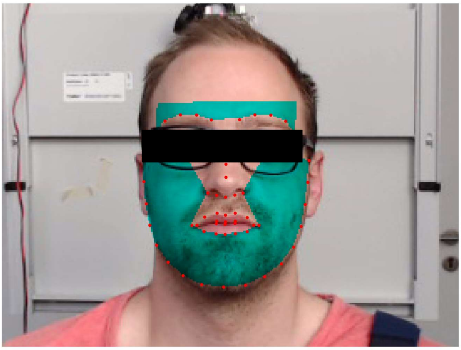


Figure 5.6: The ROI is defined based on the landmark points (denoted as red points) output by the face tracker. Raw color signals are extracted from the defined ROI.

length for an effective spectral resolution, ICA extracted pulse signals from the long color signals (10 seconds) in the experiment. Joint Approximation Diagonalization of Eigen-matrices (JADE) [Car99] was adopted as the core implementation for ICA.

To train the neural network, the ground truth PPG signals were first downsampled into 30 Hz corresponding to the videos' frame rate. We eliminated the time delay between the video and oximeter sensor by aligning the PPG signal with the remote pulse signal extracted using POS. The neural network was implemented in PyTorch. We used an Adam optimizer [Kin14] to train the network model. The learning rate was set to 0.005 and the network was trained for 30 epochs.

We set the length of the long pulse signal as 10 seconds (300 frames). A band-pass filter was used to reduce noise in signals extracted by the non-supervised methods. The cutting frequencies of the band-pass filter were set as 0.6 and 4 Hz. We calculated pulse rates from the extracted pulse signals by performing peak detection in the frequency domain. In order to compare SWN with other benchmark methods, we calculated Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Success Rate (SR), Signal-to-Noise Ratio (SNR)

Table 5.2: Benchmark performance of pulse measurement on stationary and motion scenarios of VitalCamSet

Metric	Scenario	ICA	CHROM	POS	SWN
MAE (bpm)↓	101	1.23	0.75	0.64	**0.42**
	201	6.27	4.98	3.94	**1.13**
	202	2.83	1.02	0.97	**0.47**
	203	5.32	2.86	2.38	**0.72**
	204	3.85	2.53	2.09	**1.02**
	Arg.	3.87	2.42	2.00	**0.75**
RMSE (bpm)↓	101	4.52	3.11	2.71	**1.55**
	201	11.76	11.92	10.30	**3.03**
	202	8.13	3.51	3.50	**1.27**
	203	11.91	8.05	7.09	**2.00**
	204	10.09	8.27	8.25	**3.49**
	Arg.	9.64	7.71	7.00	**2.44**
SR (%) ↑	101	94.42	96.31	96.83	**98.00**
	201	65.77	73.78	78.28	**91.87**
	202	88.31	95.02	95.28	**98.00**
	203	78.02	86.70	88.27	**95.75**
	204	83.57	87.50	90.48	**94.13**
	Arg.	82.12	87.88	89.86	**95.53**
SNR (dB) ↑	101	4.28	6.41	7.37	**9.10**
	201	-4.26	-2.20	-1.05	**4.69**
	202	1.34	4.16	5.06	**7.57**
	203	-1.20	0.95	2.67	**6.38**
	204	0.62	2.27	3.68	**7.08**
	Arg.	0.20	2.35	3.57	**6.98**
ρ ↑	101	0.93	0.97	0.97	**0.99**
	201	0.52	0.61	0.66	**0.95**
	202	0.73	0.94	0.94	**0.99**
	203	0.49	0.72	0.77	**0.98**
	204	0.63	0.72	0.72	**0.95**
	Arg.	0.67	0.78	0.81	**0.97**

and Pearson's correlation coefficient (ρ) as evaluation metrics. We define the Success Rate as the percentile number of predictions within a tolerance of ±3 bpm as in [De 13].

5.2.3 Benchmark experiment

5.2.3.1 Cross-participant evaluation

We conducted the cross-participant experiment using the VitalCamSet, dividing the dataset into four folds. Each fold comprised 6 - 7 participants. Following a four-fold cross-validation schema, we obtained the evaluation results, which are presented in Table 5.2.

From the tables, we can see that all methods achieve the best performance on the stationary scenario 101 and have the worst results in Scenario 201 with rotatory head movement. Comparing the proposed method with the linear projection approaches, SWN shows the best results for all metrics (MAE: 0.75 bpm, RMSE: 2.44 bpm, SR: 95.53%, SNR: 6.98 dB, ρ: 0.97) and outperforms the linear-projection methods significantly (with over 60% reduction in MAE compared with POS). It can also be seen that the model based methods have higher accuracy than the BSS-based ICA; among the non-supervised methods, POS provides the second best results. We illustrate the correlation between the ground truth and predicted PR given by POS and SWN in Figure 5.7. We can see that SWN provides accurate predictions in the heart rate range between 40 and 110 bpm, and has a higher correlation with the ground truth HR than POS. The failing predictions of POS tend to give a lower estimate of the pulse rate, which can be explained by the fact that the head motions in these scenarios have relatively lower frequencies than pulse rates.

5.2.3.2 Cross-scenario evaluation

We also conducted a cross-scenario evaluation on the changing illumination scenarios. The evaluation was kept participant-independent as before and the

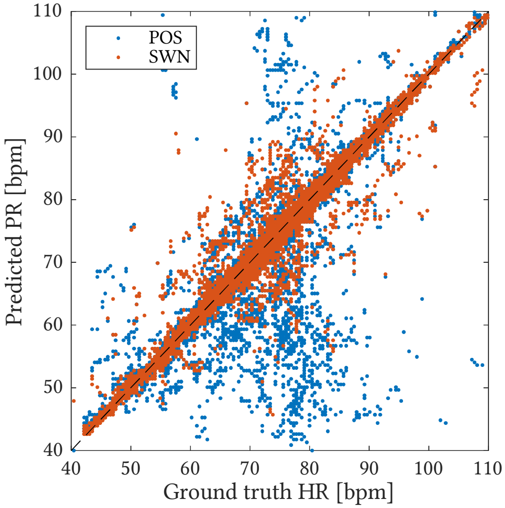


Figure 5.7: The proposed SWN gives proper PR estimates in the range between 40 bpm to 110 bpm on VitalCamSet. In terms of correlation with the ground truth HRs, the predicted PRs given by SWN (red points) outperform those given by the POS algorithm (blue points).

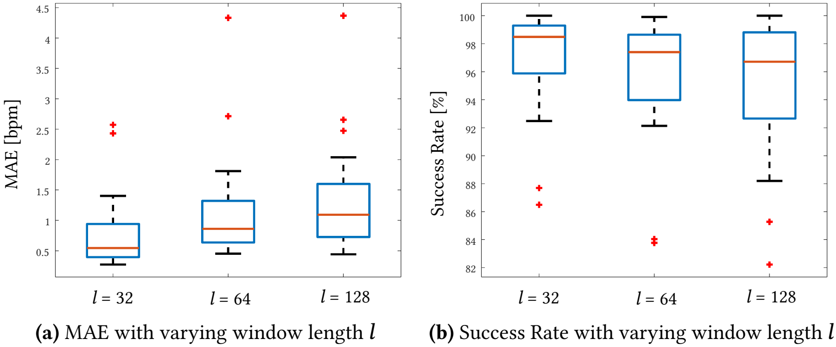
network model was not fine-tuned on the videos of the changing illumination scenarios. The results are listed in Table 5.3. We can see that all methods provide accurate PR predictions in the scenario with slowly changing lighting, while the measurement performance drops significantly in the scenario with abrupt illumination changes. Besides, even though the network model was only trained on videos from stationary and motion scenarios, it generalized well to the slowly changing illumination scenario 104 and outperformed the linear-projection approaches in Scenario 103. This can be explained by the fact that both head motion and illumination change induce signal components with different amplitudes and morphology to pulse signals. The network has learned to differentiate between color variations caused by pulsatile components and disturbances, and will try to suppress the disturbance strength if head motion or lighting change is present.

Table 5.3: Benchmark performance of pulse measurement on changing lighting scenarios of VitalCamSet

Metric	Scenario	ICA	CHROM	POS	SWN
MAE (bpm)↓	103	6.44	5.04	5.41	**3.69**
	104	2.32	1.06	0.99	**0.38**
	Arg.	4.41	3.08	3.22	**2.06**
RMSE (bpm)↓	103	17.45	16.00	16.24	**14.93**
	104	6.93	4.46	3.73	**1.42**
	Arg.	13.34	11.81	11.85	**10.67**
SR (%) ↑	103	76.93	82.26	80.99	**89.90**
	104	88.90	94.88	95.13	**98.54**
	Arg.	82.84	88.49	87.96	**94.16**
SNR (dB) ↑	103	-1.91	-0.97	-0.48	**5.01**
	104	2.71	4.75	5.61	**8.68**
	Arg.	0.37	1.85	2.52	**6.82**
ρ ↑	103	0.46	0.53	0.51	**0.61**
	104	0.84	0.93	0.95	**0.99**
	Arg.	0.58	0.66	0.66	**0.73**

5.2.4 Experiment for the window length

Since SWN works with a short window length, it is necessary to discuss how the window length could impact the extraction performance. To this end, we ran an experiment with varying window length l from 32 to 128 frames. In this experiment, we fixed the network architecture and only change the input length. The varying length of the input signals gives rise to changing feature dimensions after down-sampling by the encoder. Since the temporal up-sampling in the decoder has the same step number as down-sampling in the encoder, the output will keep the same length with the input signal. The experiment results are displayed in Figure 5.8. It shows that SWN has a drop in performance as the window length increases. With a smaller l, the estimated pulse rates have higher accuracy in terms of both MAE and Success Rate. Analogously to the model-based methods, a shorter observation window can improve the adaptability of SWN to distortions as well, and thereby boost the measurement performance of the algorithm.



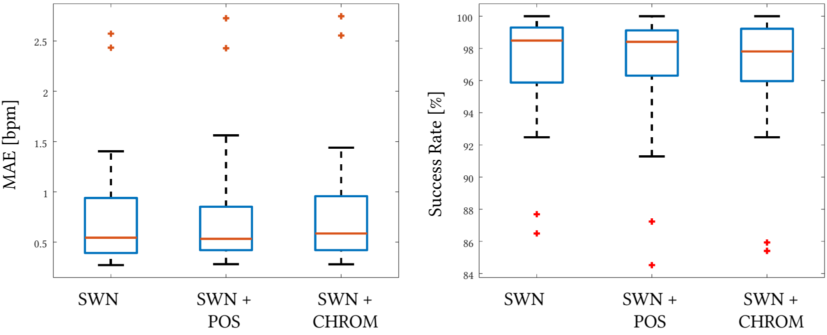
(a) MAE with varying window length l

(b) Success Rate with varying window length l

Figure 5.8: Experiment results for varying window length l. The estimation performance degrades for an increasing window length.

5.2.5 Combination with prior knowledge

We followed the analysis approach in [Zha20a] to investigate if combining SWN with prior knowledge of the model-based rPPG algorithms could benefit the measurement. We projected the temporally normalized color signals onto the vectors predefined in CHROM and POS before inputting them into the networks. The vectors in POS are defined as $[0, 1, -1; -2, 1, 1]^\top$, and in CHROM as $[3, -2, 0; 1.5, 1, -1.5]^\top$. Since projection of color signals onto two vectors gives two signal traces, we adapted the channel number of the first convolution layer in the encoder into two. Networks with inputs projected on CHROM and POS vectors are denoted as SWN + CHROM and SWN + POS respectively. We compare the performance of the modified networks with the original SWN in the box plots in Figure 5.9. The box plots illustrate the statistics of MAE and Success Rates across participants. It shows that SWN + POS has a marginal performance improvement against the original SWN, while SWN + CHROM shows lower accuracy than the original SWN. This may be due to the invalid assumption of the skin reflection vectors in CHROM. The



(a) MAE combining different prior knowledge **(b)** Success Rate combining different prior knowledge

Figure 5.9: Experiment results combining with prior knowledge. Combining with the POS vectors can slightly improve the accuracy of SWN, but projecting input signals onto the CHROM vectors gives worse results than the original network.

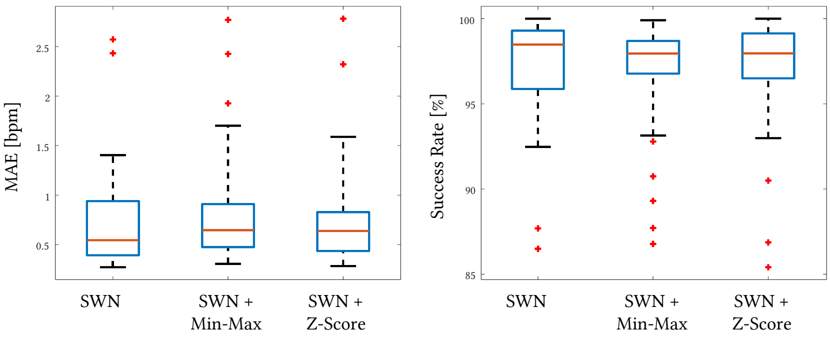
projection onto the inaccurately defined skin reflection vectors could enhance the disturbance and reduce the relative strength of pulsatile components in the projected signals. The experiment results suggest that combining SWN with the prior knowledge can potentially improve the measurement accuracy, but a less robust physical assumption could harm the performance of SWN.

We also investigated effectiveness of temporal normalization. We simply replaced the temporal normalization defined in Equation 5.3 with two general normalization approaches, i.e., Min-Max normalization and Z-Score normalization. Min-Max normalization is defined as:

$$\tilde{\mathbf{C}}_{t,c} = 2 \cdot \frac{\mathbf{C}_{t,c} - \min(\mathbf{C}_{t,c})}{\max(\mathbf{C}_{t,c}) - \min(\mathbf{C}_{t,c})} - 1\,, \tag{5.14}$$

while Z-Score normalization is expressed as:

$$\tilde{\mathbf{C}}_{t,c} = \frac{\mathbf{C}_{t,c} - \mu(\mathbf{C}_{t,c})}{\sigma(\mathbf{C}_{t,c})}\,. \tag{5.15}$$



(a) MAE with different input normalizations

(b) Success Rate with different input normalizations

Figure 5.10: Experiment results with different input normalization approaches. The original SWN with temporal normalization slightly outperforms the Min-Max and Z-Score normalization.

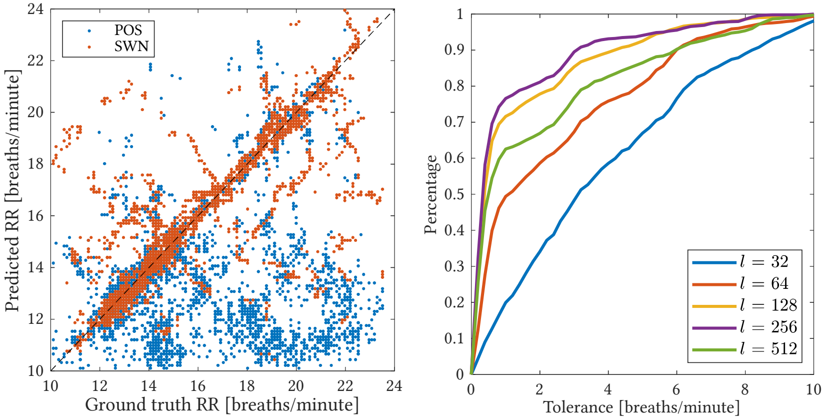
The network with the min-max normalized input is denoted as SWN - Min-Max and that with the Z-Score normalized input as SWN - Z-Score. We mention that both Min-Max and Z-Score approaches distort the cross-channel relativity of pulse signal strength which was the foundation for the design of the model-based methods.

The experiment results for temporal normalization are shown in Figure 5.10. We can see that SWN - Min-Max and SWN - Z-Score work properly despite the distorted relative rPPG strength in the color channels and both give better measurement results (with MAE < 1 bpm) than the non-supervised methods. Besides, SWN with the original temporal normalization outperforms the general normalization approaches slightly. We note that methods with general normalization approaches display more outliers in the box plot, though they have more compact 25th/75th percentiles than the original SWN. The results indicate that temporal normalization, which conserves the relative rPPG strength in the color channels, could help the network extract more accurate signals.

5.2.6 Respiratory rate measurement

This section evaluates the measurement performance of the network for respiratory rate (RR), which is reported in [Gen22]. The facial color change is modulated by the respiration in three ways. First, the intra-thoracic pressure variations cause the change in the baseline color intensity under the face skin. Secondly, the variation amplitude of the pulsatile components in the skin color signals can be modulated by the cardiac output, which decreases due to reduced ventricular filling during inspiration. Thirdly, the breathing frequency also causes the rhythmical fluctuations in heart rates (RSA) due to the regulation of the autonomous nervous system. The change in the baseline color is the primary source for photoplethysmographic respiratory measurement [Van16].

The network for respiratory rate measurement was also trained on the VitalCamSet. In the dataset, the reference breathing signals (from the abdomen and thorax) were recorded using the vital sign monitoring system, *SomnoScreen Plus*. Since the reference signal recording was affected by body movement as



(a) Comparison between SWN and POS in terms of accuracy in RR measurement

(b) AUCs of RR measurement using SWN with different window lengths

Figure 5.11: Experiment results for respiratory rate measurement.

well, only the data from the stationary scenario were used for training. Any measurement segments with artifacts were pruned from the data.

Figure 5.11a compares the network and the POS algorithm for their ability to extract respiratory rate. It can be seen that the POS algorithm is prone to be affected by the LF noise, which is potentially caused by the involuntary body motions. The network has a much lower measurement error (MAE: 1.06 breaths/minute) compared to the linear method (MAE: 3.12 breaths/minute).

Here it should be mentioned that the window size for respiratory measurement is larger than the pulse signal measurement, since the respiratory rates (typically in the range 10-40 breaths/min) are lower than heart rates. A window that is too short cannot appropriately capture dynamics of the breathing signals. An experiment involving various window lengths for respiratory measurement is depicted in 5.11b, where the plot illustrates the Area under the Curve (AUC) of the measurements. The figure shows that the best measurement results were achieved with a window length of 256 frames, which is used for the evaluation in Figure 5.11a as well.

5.3 Conclusion

In this chapter, we proposed a deep network for blood pulse signal extraction, which is integrated into the framework of the classic rPPG algorithms. The design of the network was aimed to combine the advantages of both the deep learning approaches and the traditional rPPG algorithms. In the evaluation experiment, the proposed method exhibited more robust performance than the benchmark algorithms on the large-scale dataset VitalCamSet. The experiment also discussed the impact of the observation window length on the measurement performance of the proposed network. The results indicate that the short-window overlap-adding pipeline benefits signal extraction for the SWN. Moreover, it was demonstrated that the proposed network can be used to extract respiratory signals as well, and exhibited better results than the linear POS algorithm.

6

Face Normalization and Spatial Feature Extraction for Short Window Network

Chapter 5 introduced the concept of the short window network. This network processes signals along the temporal dimension, converting color signals from face videos into Blood Volume Pulse (BVP) signal. In classical methods, the color signals are calculated by averaging pixel intensities within a fixed, predefined face region. This naive averaging approach assigns equal weights to all face sub-regions, disregarding the variations in the relative strengths of pulse signals across different regions. These variations can arise due to factors such as inhomogeneity in illumination, presence of facial hair, local facial motions, and individual anatomical differences.

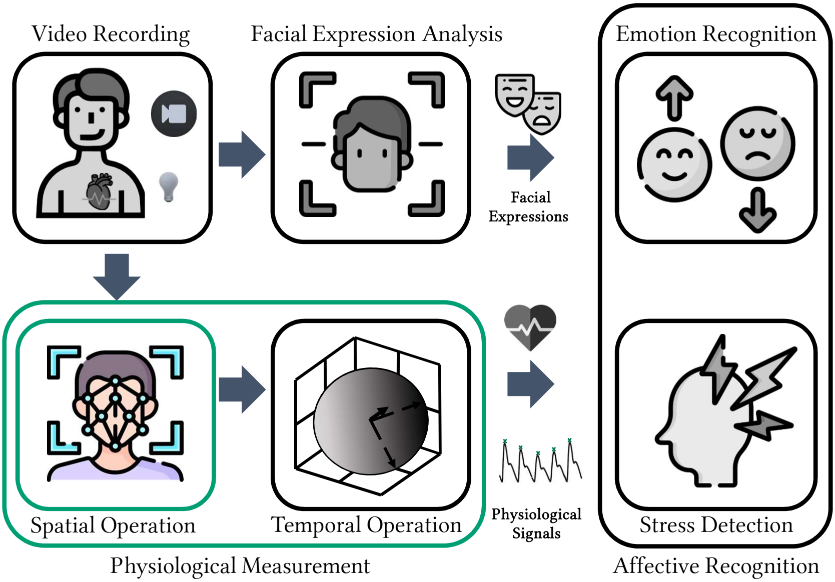


Figure 6.1: Overview of the software architecture - Spatial Operation.

Figure 6.2 illustrates an example from the Talking scenario in the PURE dataset [Str14], a dataset that will be utilized for cross-dataset evaluation in subsequent sections. The raw color signal was computed by averaging the pixels inside three face regions - the forehead, the chin, and the overall selected face region - while the participant was talking. The figure clearly shows that the illumination on the skin is not homogeneous across the face region, with the forehead exhibiting lighter color and the chin region appearing darker. Furthermore, the skin color signals exhibit different levels of periodicity across different face regions. The pulsatile variation is most pronounced in the signal extracted from the forehead, while the signal from the chin is dominated by local movements. Given the larger energy of disturbance compared to the pulsatile variation, simply averaging the pixels in the selected face regions will result in color signals significantly contaminated by the motion. This is illustrated in the second plot of the figure.

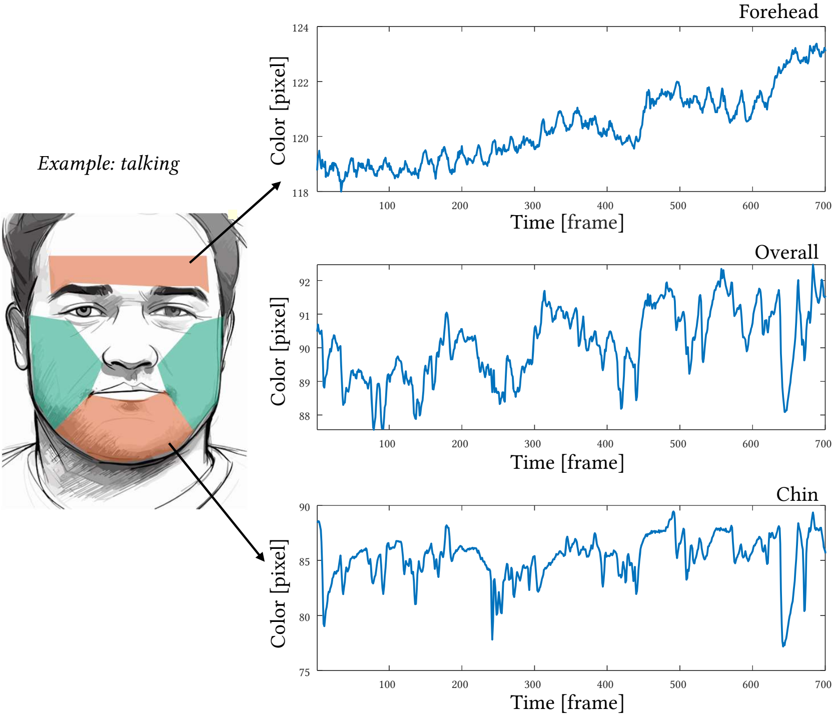


Figure 6.2: Signals of different quality exhibited from different facial regions. In this example, local facial motions such as talking and making facial expressions introduce disturbance components into the color signal.

Several works have proposed more adaptive spatial operation strategies to leverage the spatial redundancy of video images. For instance, Wang et al. [Wan14] proposed a method that generates a set of pixel trajectories by connecting pairs of pixels from two consecutive frames using the forward-backward optical flow tracking [Kal10]. They subsequently performed global motion compensation by combining the Viola-Jones face detection algorithm [Vio01] and tracking-by-detection using Circulant Structure with Kernels (CSK) [Hen12]. The pixel trajectories were pruned based on pixel color and signal amplitude in the motion direction. Another approach was presented by Blöcher [Blö20], who defined local patches based on the Active

Appearance Model (AAM) [Coo98] and selected ROI according to the local signal-to-noise ratio.

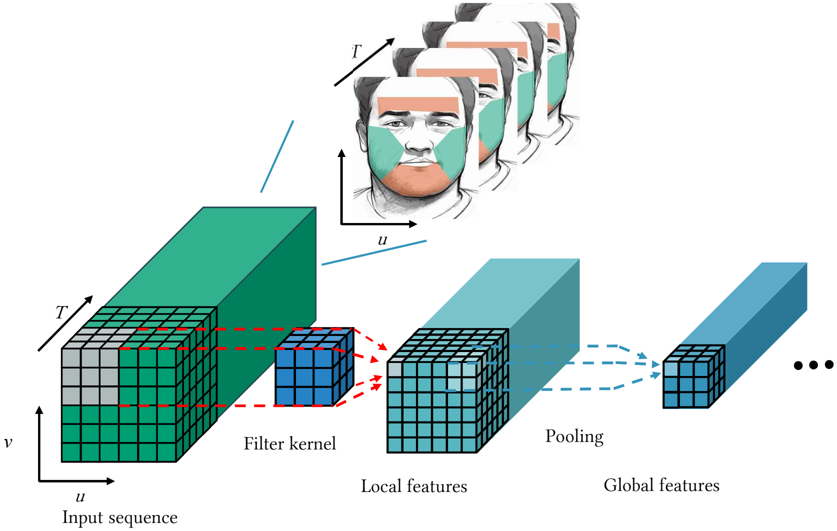


Figure 6.3: The face images are stacked along the time dimension and fed into a neural network. The network employs spatial convolution to extract local features, which are then aggregated via the pooling operation.

Neural network-based approaches can exploit spatial redundancy in various ways. One approach involves extracting raw signals from subregions of the face, then stacking them to build a signal map that represents the temporal color changes of each face region [Niu19, Niu20]. However, this has a disadvantage of losing the relative spatial relationship between adjacent skin subregions due to the one-dimensional indexing of the signal map.

Another approach is to use cropped face images directly as inputs for neural networks [Che18a, Liu20b, Niu20, Spe18, Yu19a, Yu19b]. The cropped face images are stacked along the time dimension and input into a three dimensional (3D) convolutional network. The local features can be extracted in both spatial and temporal dimensions, as shown in Figure 6.3. In the temporal dimension, the network learns patterns and changes that occur from one frame to the next. The operation in the spatial dimensions extracts information across different measurement sites. However, the spatial operation usually necessitates an attention mechanism to guide the network to assign weights to the measurement regions since the input image also encompasses background noise and areas with scarce physiological information. One potential factor that could degrade the performance of the attention mechanism is the discrepancies in data distribution between training and testing conditions [Liu20b]. The attention mechanism might not generalize effectively across varying subjects and recording environments, possibly leading to diminished measurement accuracy by overemphasizing image regions with less physiological information.

Furthermore, utilizing cropped raw images as input for neural networks introduces challenges related to addressing the spatial shift of the ROI in adjacent frames, especially in scenarios involving head motions. The same pixel in two consecutive frames often corresponds to different locations on the face, leading to spatial misalignment. This spatial shift can introduce noise into the color signal and affect measurement accuracy, particularly given that skin color on the face is not perfectly homogeneous.

In this chapter, we propose using a three dimensional deep network to extract pulse signals from face image sequences that are normalized onto a standardized coordinate system. The face normalization can remove face shape variations and irrelevant information from backgrounds, and further reduce distribution discrepancy between training and testing data. The core contributions of this chapter are to: (1) propose the use of face normalization for remote Photoplethysmography measurement, (2) propose a 3D convolutional network (3D-SWN) which works in the classical short-window overlap-adding scheme for rPPG measurement, (3) evaluate the proposed network on public rPPG datasets (VitalCamSet [Blö19], PURE [Str14], UBFC [Bob17] and NIRP [Mag18]).

6.1 Method

This section will elaborate on the method proposed in this chapter. Subsection 6.1.1 will detail an approach for face normalization, which can present the pixel information of face images in a standardized coordinate system. The architecture and specific training details of the network will be discussed in Subsection 6.1.2 and 6.1.3, respectively.

6.1.1 Face normalization

Facial shape variation caused by head motion and facial expressions is a major challenge in various face analysis tasks. Several studies have employed face normalization techniques to improve the accuracy of these tasks, such as for facial expression recognition [Yao16, Hu17], face recognition [Hua17, Tra17] and face alignment [Zhu17, Tzi17]. These techniques include texture warping [Has15] and generative model-based methods [Sag15]. The texture warping methods synthesized the normalized images by projecting the input face image into a reference coordinate system, while the generative model based methods construct the normalized image using a statistical model trained from data.

While texture warping methods translate pixel intensities directly from the input face image, generative model-based methods may introduce deviations in intensity into the normalized image. In the context of rPPG signal extraction, where the detection of subtle intensity changes is critical, any introduced deviation could lead to undesirable noise in the signal extraction process. Thus, texture warping is more suitable for the rPPG task compared to generative model-based methods and will be adopted in this work.

The definition of the face shape is fundamental for achieving effective face normalization. While a common approach involves using a Procrustes transformation on a face shape dataset to obtain a 2D front-view face model [Coo01, Sag15, Yin17, Has15], this method may not be suitable for the rPPG task, particularly when dealing with lateral head poses. It fails to adequately account for pixels from lateral facial regions, such as the cheekbone (zygomatic) area and the area around the jaw muscle (parotid-masseteric), which are key regions of interest for the lateral head pose in rPPG analysis.

In this work, we propose a unified model that can represent faces in both front and lateral poses. The model is based on a dense 3D Morphable Model (3DMM) defined using the 300W-LP [Zhu16] dataset. The 3D face model $\mathbb{M}_{3D}$ consists of a set of points with their 3D locations $\{(x_i, y_i, z_i), i \in [0, N-1]\}$, where N represents the number of model points. As shown in Figure 6.4 (a), the 3D face model can be approximately considered as a cylinder with the axis $\{(x, y, z)|x = 0, z = 0\}$. To represent the face surface in a 2D plane, face images are projected onto a cylinder surface that shares the same axis as the face model. The coordinate of the projected point i can be expressed as:

$$u_i = \arctan(\frac{z_i}{x_i}), \tag{6.1}$$

$$v_i = y_i \, . \tag{6.2}$$

The projected face model, denoted as $\mathbb{M}$, is equal to $\{(u_i, v_i), i \in [0, N-1]\}$, as shown in Figure 6.4 (b). Next, the face model is rescaled to the network's input size $H \times W$. To achieve this, the u_i and v_i coordinates of the projected face model are normalized to be within the range $[1, W]$ and $[1, H]$, respectively, using the min-max normalization. We denote the normalized 2D image canvas as $\mathbf{N}$ and set H and W as 36, which is a compromise between averaging camera noise and preserving spatial resolution, a choice validated by previous works [Liu20a, Che18a, Liu20b].

The projected face model $\mathbb{M}$ is highly dense. In order to apply this model on the 36×36 canvas $\mathbf{N}$ more efficiently, a subset of model points is selected, which will serve as vertices for face region subdivision. To ensure that the selected vertices are evenly distributed on the plane, we define a set of anchor points $\mathbb{A}$. The anchor points are equidistantly sampled on the 2D plane $\mathbf{N}$ with a fixed distance of d pixels in both u and v dimensions. The horizontal coordinate of each second row is offset by $d/2$ pixels:

$$\mathbb{A} = \{(dm + (n \bmod 2) \cdot d/2, dn)|m \in [0, ..., W/d], n \in [0, ..., H/d]\} \, . \tag{6.3}$$

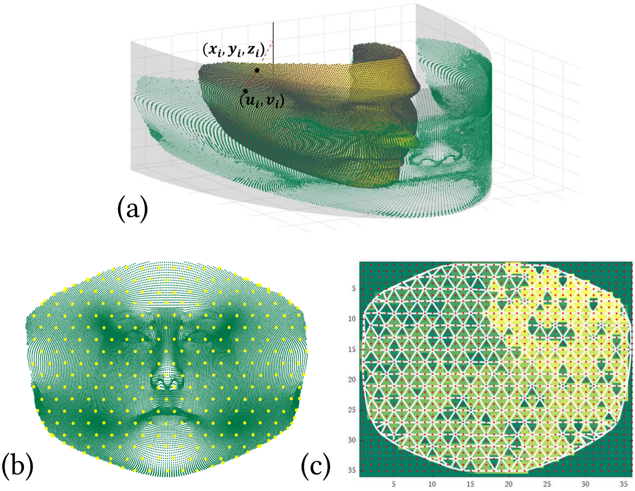


Figure 6.4: Definition of the face model for face normalization: (a) The 3D Morphable Model is projected onto a cylinder surface with the same axis as the face model. (b) Vertices for triangulation are sampled equidistantly from the projected model points on the 2D surface. (c) Delaunay triangulation divides the face surface into subregions.

The offset is used to generate isosceles triangles where the base and height are equal, as opposed to right-angled isosceles triangles without the offset. This creates a more balanced subdivision of the facial region into smaller, more symmetric triangles, which decreases the maximal distance between two arbitrary pixels inside the subregion. The anchor points $\mathbb{A}$ are depicted as points of lighter color in Figure 6.4 (b).

The closest point $\mathbf{v}$ in the face model $\mathbb{M}$ to each anchor point is subsequently selected as a vertex for the face region subdivision. Mathematically, this can be expressed as follows:

$$\mathbb{V} = \{argmin_{\mathbf{v}\in\mathbb{M}}(|\mathbf{v}-\mathbf{a}|)|\mathbf{a}\in\mathbb{A}\}. \tag{6.4}$$

Having selected the vertices, we then use the Delaunay Triangulation [Bor34] to subdivide the face surface. The triangulated face region is denoted as $\mathbb{T}$, with each subregion represented by $\mathbf{T}_k$ ($k \in [0, ..., |\mathbb{T}| - 1]$), as shown in Figure 6.4 (c).

Given an input image frame $\mathbf{I}$ during runtime, we first use a face alignment algorithm $FA(\cdot)$ to localize the triangle vertices. Here it should be noted that the triangle position $\mathbf{T}_k$ is in the coordinate system of the normalization canvas $\mathbf{N}$, while $FA(\mathbf{T}_k)$ gives the triangle position in the coordinate system of the image $\mathbf{I}$. For each pixel $\mathbf{o}$ in the triangle $\mathbf{T}_k$, the pixel color in the normalization canvas $\mathbf{N}$ is determined as the average value within the corresponding triangle region $FA(\mathbf{T}_k)$ in the input image $\mathbf{I}$. The average color is computed using the following equation:

$$\mathbf{N}(\mathbf{o}) = \frac{1}{|FA(\mathbf{T}_k)|} \sum_{j \in FA(\mathbf{T}_k)} \mathbf{I}(j), \quad \forall \mathbf{o} \in \mathbf{T}_k\,, \tag{6.5}$$

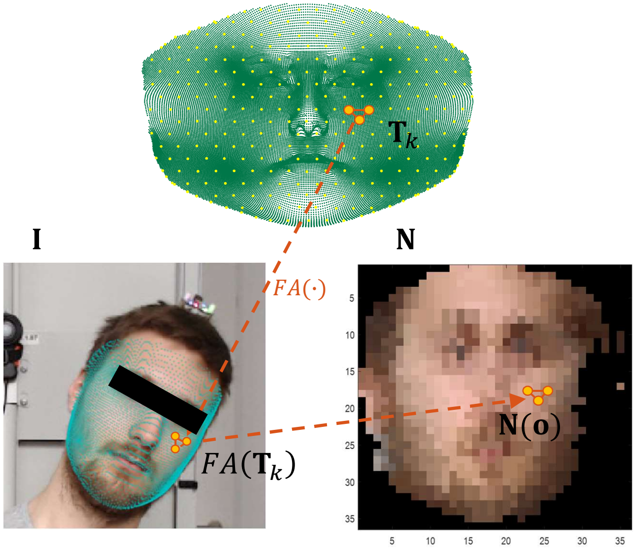


Figure 6.5: Processing steps for face normalization: For each subregion $\mathbf{T}_k$, the location within the face image $\mathbf{I}$ is determined using the alignment algorithm $FA(\cdot)$; the pixel intensity on the normalization canvas $\mathbf{N}$ is calculated as the average color within the triangle $FA(\mathbf{T}_k)$.

where $|FA(\mathbf{T}_k)|$ is the pixel number of the subregion $\mathbf{T}_k$ in the image $\mathbf{I}$. The normalization steps are illustrated in Figure 6.5.

Figure 6.6 compares face images generated using the normalization method (in the third column) with those simply cropped from the input images (in the second column). The first and last columns depict input images with different head motions, specifically translation and rotation. In the normalized face image, the target facial point (such as the left eye corner) is brought to a consistent coordinate. In contrast, in the cropped face images, there is a positional offset of this point. Additionally, to reduce noise from the background, the face normalization process masks out pixels outside the face region by setting them zero, which eliminates the disturbance in the background pixels and can benefit the signal extraction further.

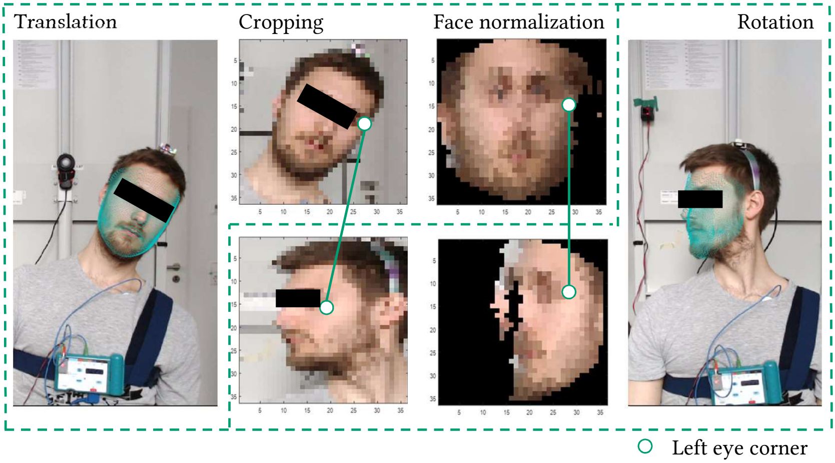


Figure 6.6: Comparison between cropping and face normalization: (1) Face normalization provides shape invariant face images for the network; (2) pixel changes originating from the background are filtered out.

6.1.2 Network architecture

Similar to Section 5.1.2, the 3D network follows the short-window overlap-adding pipeline. Given an RGB video with a length of T, the normalized face images $\mathbf{N}$ are stacked in the temporal dimension to form an image sequence $\mathbf{M} \in \mathbb{R}^{T\times3\times36\times36}$. The image sequence is segmented using a time window, resulting in a short sequence $\mathbf{M}_t \in \mathbb{R}^{l\times3\times36\times36}$ of length l, where $t \in \{0, 1, 2, ..., T - l\}$ represents the time index of frame sequences. Next, to remove the non-varying temporal components of image sequence as in the classic one-dimensional (1D) methods, a channel-wise temporal normalization is performed on the image sequence:

$$\tilde{\mathbf{M}}_t^{c,u,v} = \frac{\mathbf{M}_t^{c,u,v}}{\mu(\mathbf{M}_t^{c,\cdot,\cdot})} - 1\,, \tag{6.6}$$

where $\mu(\mathbf{M}_t^{c,\cdot,\cdot})$ stands for the average value across the channel c. u and v represent spatial locations of the pixel.

Following the temporal normalization of the image sequence, an encoder-decoder network architecture is employed for processing the input data. In contrast to Section 5.1.2, where the input was a one-dimensional temporal signal, the network input in this chapter is a three-dimensional matrix with three color channels. Given the effectiveness of temporal processing of the model demonstrated in Section 5.1.2, the operation along the time dimension is maintained unchanged in the 3D network. To account for the spatial characteristics of the data, we introduce spatial processing by replacing the corresponding layers in the 1D network with three-dimensional blocks, while preserving the parameters for the temporal dimension.

After passing through the final pooling layer of the encoder, the spatial dimensions of each feature map are reduced to 1×1, meaning that the feature map is condensed to a single point in the spatial dimensions, but still retains a full set of feature channels. This implies that the features obtained using the 3D encoder have the same shape as those obtained with the 1D version.

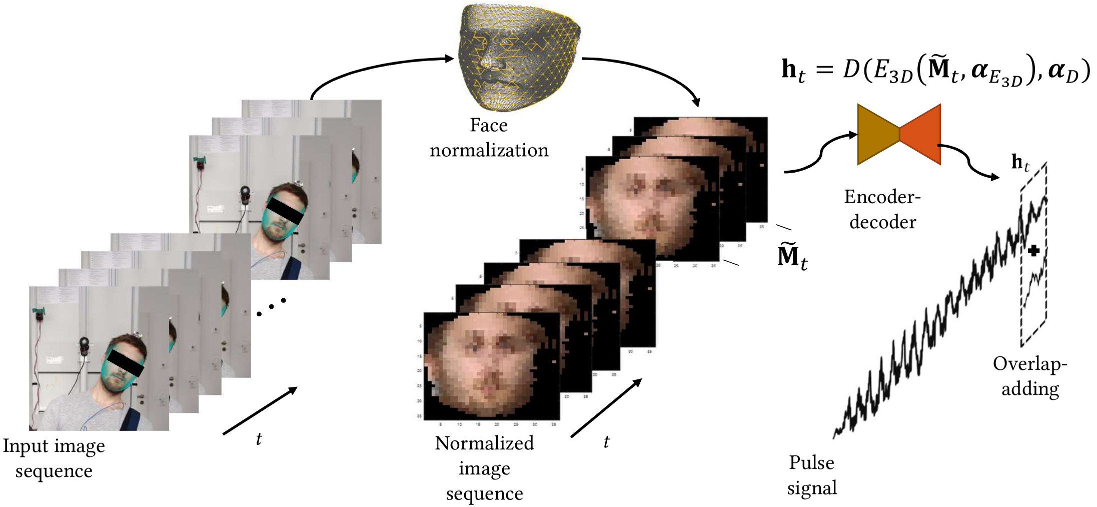

Face normalization
$\mathbf{h}_t = D(E_{3D}(\widetilde{\mathbf{M}}_t, \boldsymbol{\alpha}_{E_{3D}}), \boldsymbol{\alpha}_D)$
Encoder-decoder
$\mathbf{h}_t$
$\widetilde{\mathbf{M}}_t$
Overlap-adding
Input image sequence
t
Normalized image sequence
t
Pulse signal

Figure 6.7: The network's input is a three-dimensional matrix, consisting of three color channels. To introduce spatial operation in the 3D network, we substitute specific layers in the 1D model with three-dimensional blocks, while maintaining the parameters for the temporal dimension unchanged. We retain the original temporal operation of the 1D network in the 3D counterpart. The structure of the decoder remains the same.

Consequently, the decoder architecture remains unchanged, as described in Section 5.1.2.

The pulse signal, calculated using the three-dimensional network, can be represented as:

$$\mathbf{h}_t = D(E_{3D}(\tilde{\mathbf{M}}_t, \boldsymbol{\alpha}_{E_{3D}}), \boldsymbol{\alpha}_D), \tag{6.7}$$

where E_{3D} represents the 3D encoder architecture and $\boldsymbol{\alpha}_{E_{3D}}$ its parameters. After performing overlap-adding, we obtained the full-length signal:

$$\mathbf{H} = \sum_{t=0}^{T-l} D(E_{3D}(\tilde{\mathbf{M}}_t, \boldsymbol{\alpha}_{E_{3D}}), \boldsymbol{\alpha}_D)\mathbf{L}_t^\top . \tag{6.8}$$

Similar to Equation 5.13, for each data pair $(\mathbf{M}, \mathbf{PPG})$ in the training set, consisting of an image sequence and a corresponding ground truth PPG signal, we define the loss function as the sum of losses computed for each individual short window of the signals:

$$\begin{aligned} L_{rppg}(\boldsymbol{\alpha}_{E_{3D}}, \boldsymbol{\alpha}_D) &= \sum_{t=0}^{T-l} L_{nP}(\mathbf{h}_t, \mathbf{PPG}_t) && (6.9) \\ &= \sum_{t=0}^{T-l} L_{nP}(D(E_{3D}(\tilde{\mathbf{M}}_t, \boldsymbol{\alpha}_{E_{3D}}), \boldsymbol{\alpha}_D), \mathbf{PPG}_t), && (6.10) \end{aligned}$$

where L_{nP} is the negative Pearson correlation defined in Equation 5.12, which aims to maximize the correlation (or minimizing the negative correlation) between the predicted and ground-truth PPG signals.

6.1.3 Training details

The training process for the 3D network followed the same procedure as described in Chapter 5. An Adam optimizer [Kin14] with a learning rate of 0.005

was used to train the network for 30 epochs. To ensure a fair comparison between the 1D and 3D networks, both were trained using the same dataset with identical pre- and post-processing steps in the ensuing evaluations.

6.2 Evaluation

This section evaluates the performance of the proposed 3D short window network. First, we will explore the enhancements achieved through the spatial operation of the 3D convolutional network. Secondly, the improvements resulting from face normalization will be examined. Lastly, a comparison will be conducted between the performance of the proposed network and other state-of-the-art methods.

6.2.1 Dataset

Besides the VitalCamSet adopted in Chapter 5, we evaluated the algorithm using three further public datasets: PURE [Str14], UBFC [Bob17], and NIRP [Mag18].

PURE

The PURE dataset contains videos from 10 participants, recorded at a distance of 1.1 meters with a resolution of 640x480 and a frame rate of 30 fps. This dataset simulates scenarios with global head movements, including translation and rotation, as well as local facial motions, like talking. The diverse motion scenarios within the dataset enable a more thorough evaluation of the proposed method's performance.

UBFC

The UBFC dataset is one of the most widely utilized datasets for evaluating rPPG algorithms. It consists of 42 RGB videos, recorded at a frame rate of

30 fps and a resolution of 640x480. During the recording, participants were instructed to engage in a mathematical game designed to increase their heart rates. Reference PPG data were collected using a transmissive pulse oximeter. The ground truth heart rates within the dataset cover a range of 60 to 140 bpm.

NIRP

The NIRP dataset is a public dataset designed to evaluate the rPPG algorithm within a NIR setup. It comprises eight participants aged between 20-40 years, including four Indians, three Caucasians, and one East Asian individual. Illumination within the measurement setup is provided by two Bosch illuminators (EX12LED-3BD-9W) with diffusers, for both horizontal and vertical orientations. The NIR videos were captured using a Grasshopper camera (GS3-U3-41C6NIR-C), equipped with a narrow-band 940 nm bandpass filter and a 10 nm passband.

6.2.2 Results

Spatial operation

The evaluation results of the 3D network are presented in Table 6.1. The table compares the performance of the one-dimensional and three-dimensional networks. It also discusses performance differences resulting from adopting various pre-processing steps on the face images, including cropping (SWN-Crop), masking (SWN-Mask), and face normalization (SWN-FN). During pixel masking, image regions outside the face areas are masked as zero, without the face shape in the input image normalized. This help us to discern the advantages of negating the impact of pixel intensity changes from the background.

Similar to the findings in Chapter 5, all methods achieved the best results in the stationary scenario (Scenario 101) and exhibited lower performance in the motion scenarios. In terms of all evaluation metrics, the 3D network with face

Table 6.1: Comparison of pulse measurement between 1D and 3D networks on the VitalCamSet

Metric	Scenario	SWN-1D	SWN-FN	SWN-Crop	SWN-Mask
MAE (bpm)↓	101	0.41	**0.33**	0.38	0.37
	201	1.11	**0.87**	1.18	1.03
	202	0.54	**0.51**	0.67	0.61
	203	0.77	**0.69**	0.98	0.91
	204	1.02	**0.78**	0.95	0.88
	Arg.	0.77	**0.64**	0.83	0.76
RMSE (bpm)↓	101	1.48	**1.17**	1.32	1.24
	201	2.92	**2.65**	3.93	2.58
	202	2.01	**1.53**	2.69	2.27
	203	**2.29**	2.48	3.46	3.07
	204	3.37	**3.10**	3.62	3.36
	Arg.	2.51	**2.31**	3.14	2.61
SR (%) ↑	101	98.39	**99.14**	98.61	98.78
	201	93.51	**95.48**	94.46	93.96
	202	98.14	**98.26**	97.74	97.82
	203	96.23	**97.20**	95.69	95.94
	204	94.81	**97.15**	96.05	96.59
	Arg.	96.21	**97.45**	96.52	96.63
SNR↑	101	11.72	**12.34**	12.01	12.04
	201	6.47	**8.14**	7.65	7.30
	202	9.80	**10.23**	9.86	9.86
	203	8.48	**9.50**	8.91	8.95
	204	8.92	**10.11**	8.68	9.08
	Arg.	9.10	**10.08**	9.43	9.46
ρ ↑	101	0.99	**1.00**	0.99	0.99
	201	**0.96**	**0.96**	0.92	**0.96**
	202	0.98	**0.99**	0.96	0.97
	203	**0.97**	**0.97**	0.94	0.95
	204	0.95	**0.96**	0.94	0.95
	Arg.	0.97	**0.98**	0.96	0.97

normalization (SWN-FN) consistently outperformed the 1D network (SWN-1D). However, it is worth noting that the 3D network did not always outperform the one-dimensional network when using simply cropped face images as input (SWN-Crop). For instance, in the motion scenarios, the SWN-Crop method had higher MAE values (1.18 bpm, 0.67 bpm, 0.98 bpm, and 0.95 bpm for the motion scenarios 201, 202, 203 and 204, respectively) compared to the 1D-SWN method (1.11 bpm, 0.54 bpm, 0.77 bpm and 1.02 bpm). This can be attributed to the changes in the background pixels and face shape variations during motion. The face region masking operation improved performance (MAE 1.03 bpm, 0.61 bpm, 0.91 bpm and 0.88 bpm for Scenario 201, 202, 203 and 204), proving more effective than cropping alone. This is due to its ability to effectively remove the impact of background changes. The most significant improvement was observed with face normalization (SWN-FN: 0.87 bpm, 0.51 bpm, 0.69 bpm and 0.78 bpm for Scenario 201, 202, 203 and 204), which addressed both background changes and face shape variations. Averaged across all scenarios on the VitalCamSet dataset, the 3D network with face normalization reduced the MAE by approximately 20% compared to the 1D method.

To further demonstrate the superior performance of the proposed three-dimensional network compared to the one-dimensional approach, an evaluation was conducted using the public PURE dataset. Table 6.2 presents the MAE, RMSE, and Pearson coefficient for each approach. The results indicate that the 3D network achieved lower errors in pulse rate estimation

Table 6.2: Benchmark performance of pulse measurement on the PURE dataset

Method	**MAE↓**	**RMSE↓**	ρ↑
2SR [Wan15]	2.44	3.06	0.98
CHROM [De 14]	2.07	2.50	0.99
HR-CNN [Niu18]	1.84	2.37	0.98
SynRhythm [Niu18]	1.88	2.45	0.98
NAS-HR[Lu21]	1.65	2.02	0.99
SWN-1D	0.55	2.21	**1.00**
SWN-FN	**0.45**	**1.10**	**1.00**

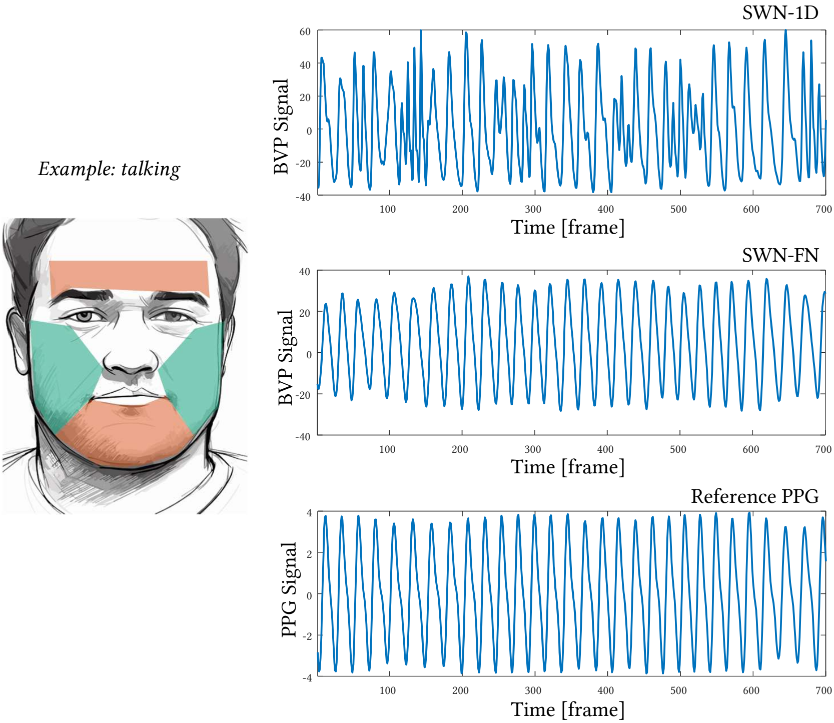


Figure 6.8: Comparison of pulse signal extraction using the 1D and 3D networks. The 1D network suffers from contamination by local facial movement, while the 3D network successfully recovers the blood volume pulse signal by leveraging spatial redundancy.

(MAE: 0.45 bpm, RMSE: 1.10 bpm) compared to the 1D network (MAE: 0.55 bpm, RMSE: 2.21 bpm).

Referring back to the example shown in Figure 6.2, the pulse signals calculated using the 1D and 3D networks are depicted in Figure 6.8. It is evident that the pulse signal extracted using the 1D network is heavily contaminated by local facial movement. However, the 3D network successfully recovers the blood volume pulse signal despite the presence of disturbances. This underscores the ability of the 3D network to extract pulse-related features from face regions by leveraging spatial redundancy.

Cross-dataset evaluation

The proposed network's generalization ability is further evaluated by comparing it to other state-of-the-art approaches on the public datasets listed in Section 6.2.1. The network model was trained on the VitalCamSet dataset, which has a heart rate distribution of up to about 110 bpm. To ensure that the training data encompasses a wider range of heart rates, we augmented the training data by downsampling videos by a factor of 2. More specifically, the augmented videos were generated by selecting every second frame from the original video data. This approach allowed the training data to cover heart rates up to 220 bpm. The evaluation was conducted on 12-second sequences extracted from the video data, following the parameter setting adopted in [Liu20b].

Results in Table 6.2 demonstrate that the 1D (MAE: 0.55 bpm, RMSE: 2.21 bpm, ρ: 1.00) and 3D (MAE: 0.45 bpm, RMSE: 1.10 bpm, ρ: 1.00) SWNs achieve the highest measurement accuracy on the PURE dataset. Compared to other methods that extract signals from cropped face images, the proposed networks process signals from a cleaner region of interest with less noise from irrelevant pixels. Furthermore, compared to the best benchmark algorithm (NAS-HR [Lu21]) listed in the table, the 3D network reduces the MAE from 1.65 bpm to 0.45 bpm and the RMSE from 2.02 bpm to 1.10 bpm, demonstrating its superior performance.

Table 6.3: Benchmark performance of pulse measurement on the UBFC dataset

Method	**MAE ↓**	**RMSE ↓**	**SNR ↑**	ρ **↑**
3D CNN [Bou19]	5.45	8.64	-	-
Meta-rPPG [Lee20]	5.97	7.42	-	0.53
CAN [McD20]	5.16	10.1	-2.83	0.80
MetaPhys [Liu20b]	1.90	2.62	3.84	0.96
SWN-1D	0.59	**1.91**	3.95	**0.99**
SWN-FN	**0.55**	1.98	**4.34**	**0.99**

Table 6.4: Benchmark performance of pulse measurement on the NIRP dataset

Metric	**SWN-FN**	**CAN+ Distraction** [Now21]	**CAN** [Now21]	**Sparse-PPG** [Mag18]	**Distance-PPG** [Kum15]
MAE ↓	**0.65**	2.34	7.78	-	-
RMSE ↓	1.95	4.46	16.81	**1.06**	6.23
SNR ↑	**4.75**	2.27	-3.24	-	-
ρ ↑	**0.976**	0.85	-0.03	-	-
SR ↑	**96.41**	-	-	95.18	82.32

The proposed network also demonstrates the best performance on the UBFC dataset, as presented in Table 6.3. The 3D network significantly reduces the MAE by over 70%, from 1.90 bpm achieved by MetaPhys [Liu20b] to 0.55 bpm.

Furthermore, the network can be utilized for pulse extraction in a NIR setup. As discussed in Chapter 4, the measurement in NIR setup presents additional challenges compared to the RGB setup. In the NIRP dataset, the images consist of only one color channel, necessitating adaptation of the network's input channel to 1. The network was trained on the data from the NIR subset of the VitalCamSet. Evaluation results displayed in Table 6.4, show that the 3D network with the proposed face normalization achieves the best results in terms of the most metrics, suggesting that the network is able to extract reliable pulse signals in the NIR setup as well.

6.3 Discussion

This chapter proposes a novel approach for accurate pulse signal extraction using a three-dimensional convolutional neural network. The network architecture was designed to operate in the short-window overlap-adding pipeline as the one dimensional network proposed in Chapter 5, to reserve the temporal processing ability of the model. The 3D network addresses the limitations of the 1D model by adaptively incorporating the spatial information of face images. Moreover, a method for face normalization was proposed to eliminate the impact of face shape variation and disturbances from backgrounds to improve the measurement robustness.

The effectiveness of the proposed method was evaluated on multiple public datasets, including the VitalCamSet, PURE, UBFC, and NIRP. The results demonstrated that the proposed method outperformed the 1D network in terms of measurement accuracy, especially in scenarios involving head motions and facial shape variations. The face normalization proved beneficial for signal extraction.

Comparison with state-of-the-art methods showcased the superior performance of the proposed network in terms of MAE, RMSE, and Pearson correlation coefficient. The 3D network exhibited better accuracy and robustness, reducing errors by a significant margin compared to existing benchmark algorithms.

7

End-to-End Deep Learning for Stress Recognition Using Remote Photoplethysmography

The algorithm described for pulse signal extraction in previous discussions forms the foundation for subsequent chapters focused on affective status recognition. This chapter provides a detailed analysis of the system's ability to detect stress, specifically stress resulting from cognitive workloads that mirror common workplace challenges. By utilizing camera-based measurements, the system aims to determine if an individual is under stress or in a baseline state. As highlighted in Section 3.4, this discussion emphasizes end-to-end approaches for rPPG-based stress detection, distinguishing it from traditional feature extraction methods.

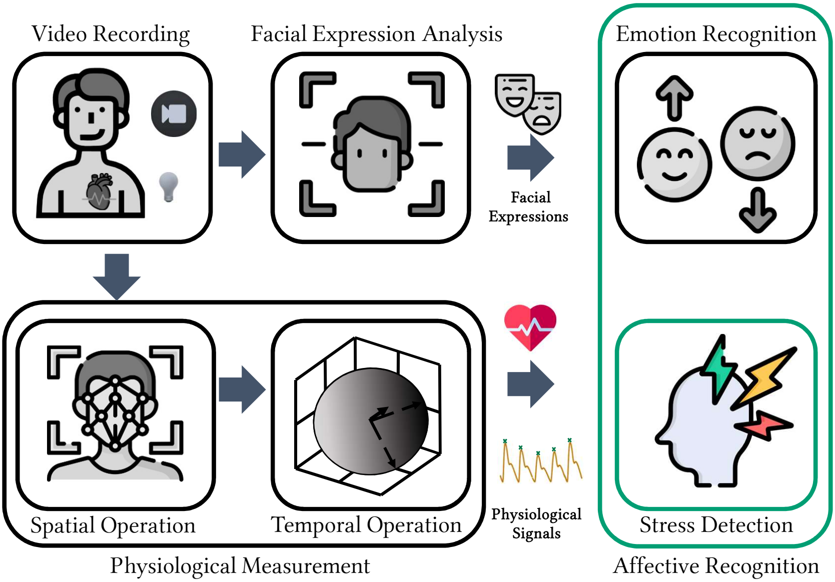


Figure 7.1: Overview of the software architecture - Stress Detection.

The remainder of this chapter is structured as follows. Section 7.1 introduces the collection protocol and pre-processing steps of the experiment data. Classification methods are described in detail in Section 7.2. Section 7.3 discusses the the experiment results. The work presented in this chapter has been published previously in the paper: [Zho22].

7.1 Data collection

This section outlines the procedure for data collection adopted in the experiment. The details of the recording setup, including how measurements were made, are discussed in Section 7.1.1. Section 7.1.2 presents the cognitive tasks employed during the experiment to induce stress. The steps involved in processing the recorded data are further explicated in Subsection 7.1.3.

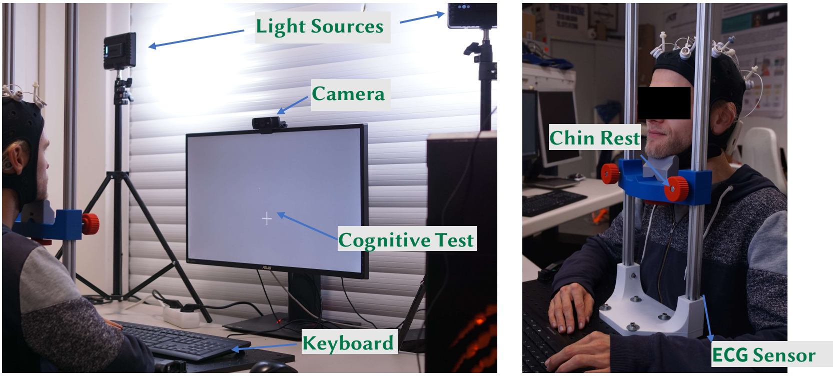


Figure 7.2: The experiment measurement setup. The head was stabilized by a headrest; an ECG sensor was used to record the reference signal.

7.1.1 Recording setup

The experimental data were recorded in a laboratory [Wei22], where the blinds were closed. In the experiment 15 participants were recruited, including male and female, different ages (Mean ± SD: 26.71 ± 2.58) and ethnicities. Lighting was provided by two LED light sources reflected from the wall, as shown in Figure 7.2. This ensures that a constant lighting condition was maintained during all measurements and that the lighting of the external environment was negligible.

Participants sat in a height-adjustable chair. In addition to face videos, brain activity was recorded by an EEG device for later multimodal analysis. Based on the previous work [Sto20], a headrest was used to stabilize the head in order to minimize disturbance of head movements. Experimental instructions and visual stimuli were displayed on a monitor, on which sat a webcam recording the participant's face. The auditory stimuli were played by speakers with the volume set to the most comfortable level for each participant. A chest strap (*Movisens EcgMove 4*) recorded the reference ECG signal for the validation of the camera-based measurements.

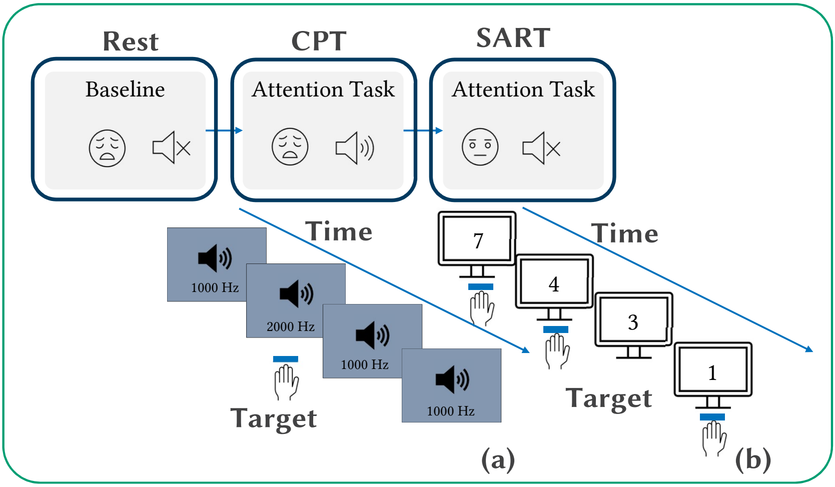


Figure 7.3: Structure of the measurement, including one rest session and two neurocognitive tests: (a) Continuous Performance Task (CPT), (b) Sustained Attention Response Task (SART).

7.1.2 Neurocoginitive tests

Continuous Performance Task (CPT) and Sustained Attention Response Task (SART) were adopted as the neurocognitive tests in the experiment. Prior to performing the neurocognitive tests, participants were asked to take a two-minute break to achieve a relaxed state. Then, the cognitive tasks were performed to generate cognitive load. At the beginning of each task, the participants were asked to read instructions displayed on the screen and familiarize themselves with the tasks through the given examples. The neurocognitive tests were implemented using the Python package *PsychoPy* [Pei19].

Continuous Performance Task

The CPT test is an auditory task with eyes closed. The general structure is shown in Figure 7.3(a). During the task, two auditory stimuli with different tones were played: a non-target tone stimulus with a frequency of 1000 Hz and a target tone stimulus with a frequency of 2000 Hz. The stimuli were randomly distributed during each test with a temporal distance between 1.5 and 2 seconds. Participants had to focus on the tone being played and respond to the target stimulus as quickly as possible by pressing the space bar. The duration of each test session was 10 minutes.

Sustained Attention to Response Task

The SART test was implemented as a visual task. In this task, the digits 1 to 9 were displayed randomly in white on the black screen with a fixed period of 800 ms. If a digit other than 3 was displayed on the screen, participants had to press the space bar as quickly as possible, as shown in Figure 7.3(b). The probability of the number 3 appearing was 20%. All digits were displayed for a brief period of 200 ms only. The test lasted for approximately 6 minutes.

7.1.3 Data processing

An overview of the data processing steps is shown in Figure 7.4. We used the Short Window Network introduced in this work to extract pulse signals as shown in Figure 7.4 (a). Then the pulse peaks were detected in the pulse signals, as illustrated in Figure 7.4 (b). After that, the temporal distances between two adjacent peaks were calculated as the IBI signals.

Despite using a headrest during recording, slight head motions and facial expressions were observed for several participants. Additionally, light from the monitor might introduce disturbances into the face videos. Consequently, these factors can contaminate the pulse signals, leading to false detection of pulse peaks. We used the *PhysioNet-Toolbox* [Ves18] to reduce artifacts in the IBI signals caused by incorrectly detected pulse peaks. Absolute and relative thresholds were defined to identify non-normal intervals. The absolute thresholds formed a confidence range, beyond which the intervals were considered artifacts. The upper and lower thresholds were set as 0.375 and 2 seconds, respectively, corresponding to 160 and 30 bpm in heart rate. The relative threshold was defined based on the median value of the previous and next five

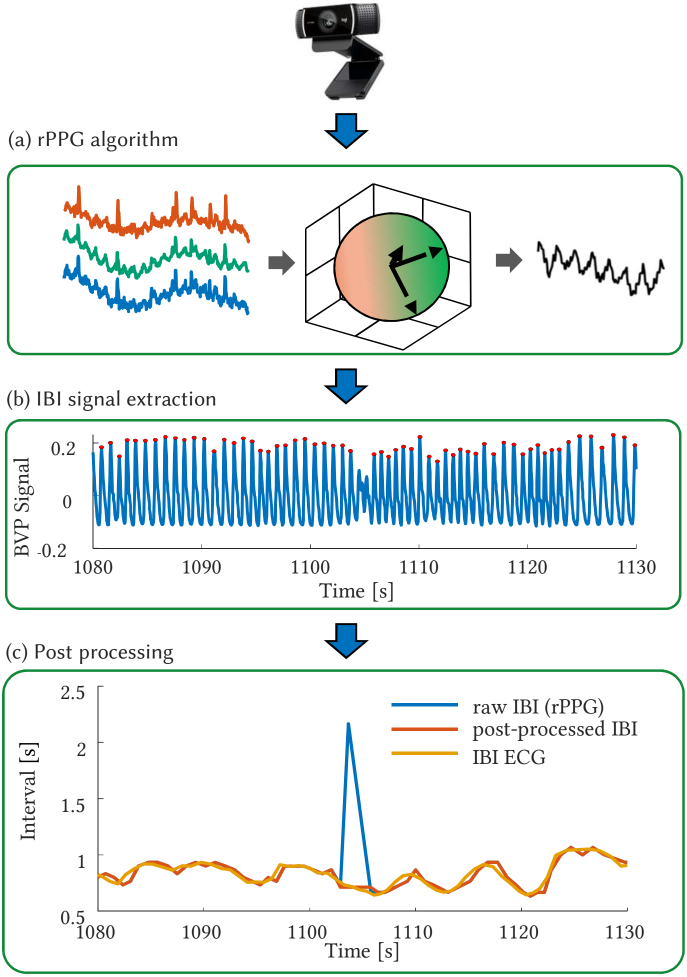


Figure 7.4: Overview of the data processing: (a) blood volume pulse signals extracted from face region, (b) peaks detected in pulse signals, (c) IBI signals after artefacts pruned.

neighboring intervals. We pruned the intervals that change more than a relative threshold with respect to the median value. The relative threshold was set as 0.2 in this work. The pruned intervals were then interpolated using the nearest valid values. The red line in Figure 7.4(c) shows the post-processed IBI signal, in which the non-normal interval observed around 1105 s in Figure 7.4(b) was filtered out.

7.2 Methods

This section will explore the methodologies employed for stress recognition. Initially, the data collected during the experiment will be validated, as discussed in Section 7.2.1. Following this, the specific methods under investigation will be detailed in Section 7.2.2. Lastly, the process of training and testing the models will be described in Section 7.2.3.

7.2.1 Measurement validation

We validate the non-contact measurement by comparing the camera-based PRV features with HRV measurements using the reference ECG sensor. PRV and HRV features were calculated from the extracted IBI signals. The time-domain features considered include the Average Interval between Normal Heart Beats (AVNN), SDNN, RMSSD and pNN50. Notably, the AVNN is essentially the reciprocal of the Heart Rate (HR).

For frequency-domain features, we calculated the power spectral density (PSD) from the Lomb-scargle periodogram of IBI signals. The VLF, LF and HF powers were calculated as the area under PSD curve corresponding to the frequency bands respectively. The total power was obtained by summing the power in all three bands. In addition to absolute spectrum powers, normalized powers and LF/HF were also calculated. Moreover, the nonlinear features SD_1, SD_2 and S from Poincaré plots were extracted as well.

We investigated the relative errors and correlation between camera-based PRV features and the reference ECG sensor. The relative error was defined as the ratio of the measurement difference to the reference value. Relative errors of all sessions are visualized in Figure 7.5. Pearson correlations between the camera and reference sensor are shown in Figure 7.6.

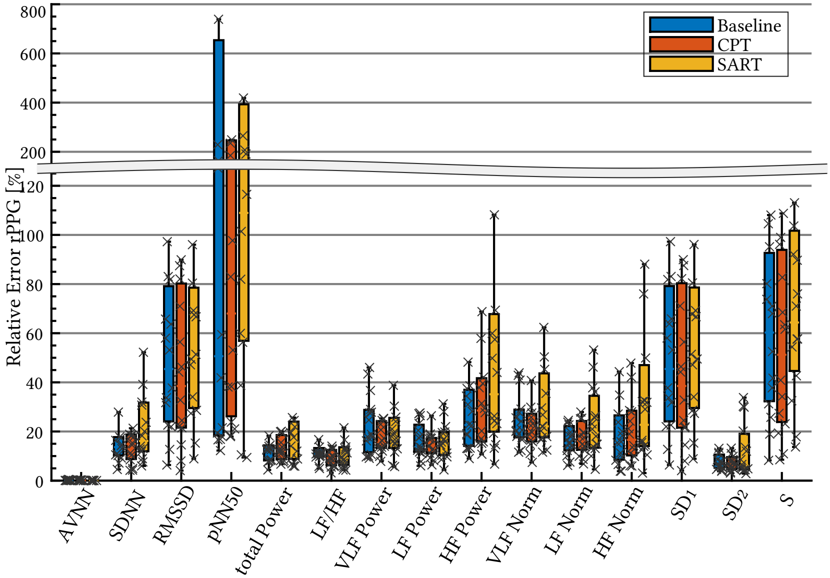


Figure 7.5: The relative error of the rPPG-based PRV features compared to the reference ECG sensor. The PRV features were measured with different accuracy: pNN50 had significant measurement errors for all sessions, while features such as AVNN can be accurately measured.

Comparing the results across different sessions in Figure 7.5, we can find that the measurement accuracy of the SART test was generally lower than that of the baseline and the CPT test. This is due to the fact that more head motions occurred during the SART test. We can see that the relative error for AVNN was negligible during all sessions, demonstrating that pulse rates can be measured with a very high accuracy. SDNN and SD_2 showed lower relative errors than RMSSD, SD_1 and S. pNN50 exhibited the most significant measurement error, which was caused by the camera's relatively low frame rate (30 frames per second). The shift of the peak position in just one frame leads to a change of 67 ms in the difference between two corresponding adjacent intervals, making the pNN50 measurements very sensitive to disturbances. The same reason can explain the significant inaccuracy in the measurement of RMSSD.

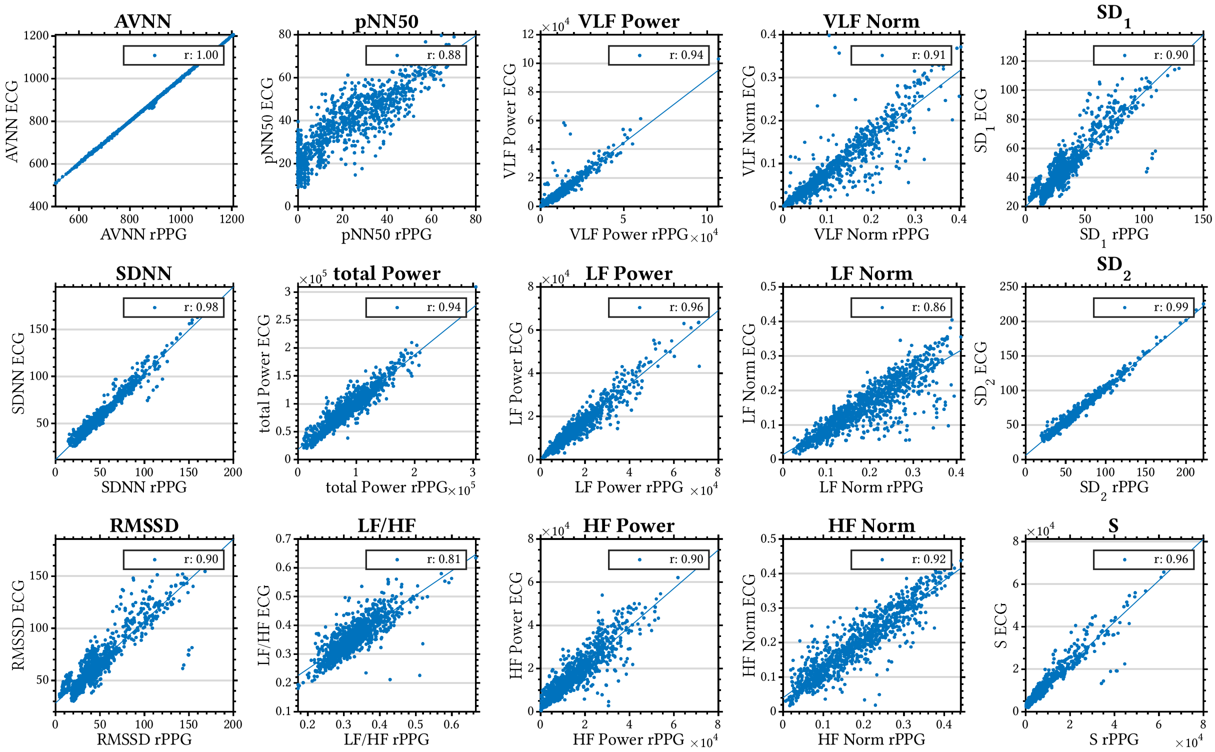


Figure 7.6: The correlation between rPPG and ECG-based PRV/HRV features for the resting and stress conditions. All PRV features had correlations over 0.80 between the camera-based and reference ECG measurement; the LF/HF ratio had the lowest correlation of 0.81 and AVNN had a correlation of 1.00.

Except the LF/HF ratio and normalized LF, all frequency domain features had a correlation greater than 0.9 between camera and ECG-based measurements. The correlation for the LF/HF ratio was 0.81, though it had a low relative error. The outliers of LF/HF measurements have significantly reduced the correlation between the measurement and reference, as shown in Figure 7.6.

Unlike the use of expensive industrial cameras or digital single-lens reflex (DSLR) cameras in [McD16, Sab21], a low-cost webcam was used in the measurement setup of this work. While some PRV parameters like AVNN could be accurately measured in the experiment, other widely used PRV features such as LF/HF ratio and RMSSD had a only limited accuracy. It should be investigated if the accuracy of the low-cost measurement setup is sufficient in recognizing stress states and if it is possible to obsolete the calculation of the PRV features for stress state recognition.

7.2.2 Classification methods

We investigated two groups of classification methods. The first group of approaches recognizes the stress state based on the handcrafted PRV features discussed above. The second group of approaches skips the calculation of PRV features and identifies the stress state directly from IBI signals in an end-to-end manner.

Feature-based methods take PRV features as input and use a classical machine learning model to predict whether an individual is at rest or in a state of stress. For these methods, the features were first normalized to zero mean and unit deviation based on the normalization parameters (feature mean and deviation) obtained from the training data. Nine machine learning methods were considered in this experiment:

- Nearest Neighbors
- Support Vector Machines with linear kernels (Linear SVM)
- Support Vector Machines with Radial Basis Function kernels (RBF SVM)

- Gaussian process
- Decision Trees
- Random Forest
- Neural Network
- AdaBoost
- Naive Bayes.

In the end-to-end methods, the IBI signals were resampled at 4 Hz and then normalized into a range between zero and one. Various deep learning network architectures were discussed for the end-to-end classification:

- five architectures for time-series classification such as:
  - Fully Convolutional Network (FCN) [Wan18a]
  - Residual Network (RESNET) [Wan17e]
  - Multi-Layer Perception (MLP) [Wan18a]
  - Time Convolutional Neural Network (CNN)
  - Multichannel Deep Convolutional Neural Network (MCDCNN)) [Zha21]
- two Long Short-Term memory (LSTM) networks such as:
  - Convolutional Neural Network with LSTM (CNN-LSTM)
  - Multi-Layer Perception with LSTM (MLP-LSTM)
- plus two additional networks:
  - Spectrotemporal Residual Network (STRESNET) [Gjo20]
  - INCEPTIONTime [Ism20].

Classification methods were implemented in Python. The feature-based methods were implemented using the *Scikit-learn* toolbox [Ped11], while the implementation of the end-to-end methods was based on the code provided by [Dzi20].

Table 7.1: Summary of the classification methods and hyper-parameters

Classifier		Optimized hyperparameters
End-to-end	MLP	Output dense layer size, dense layer depth
	MLPLSTM	
	MCDCNN	Convolutional filter size, number of convolutional filters
	CNN	
	FCN	
	CNNLSTM	Convolutional filter size, number of convolutional filters, LSTM unit number
	RESNET	Kernel size, filter number, number of blocks
	INCEPTION	
	STRESNET	
Feature-based	Nearest Neighbors	Number of neighbors
	Linear SVM	Regularization parameter
	RBF SVM	Regularization parameter, kernel coefficient
	Gaussian Process	Scale coefficient, RBF kernel coefficient
	Decision Tree	Maximum depth of tree
	Random Forest	Maximum depth of tree, number of trees
	Neural Net	Strength of the L2 regularization term
	AdaBoost	Maximum number of estimators

7.2.3 Training and validation

The end-to-end methods were trained from scratch, without using any pre-trained weights. We followed the more challenging Leave-One-Subject-Out (LOSO) validation scheme to compare the recognition performance of the methods. The data were cropped into one-minute segments and divided into the training set (11 participants), validation set (3 participants), and test set (1 participant). We resampled the training data to compensate for the imbalance caused by the different session lengths. Since the test data were unseen by the model, results of the LOSO validation were subject-independent. In order to ensure a fair comparison of all classification methods, the two groups of methods adopted the same pre- and post-processing steps.

As in [Dzi20], hyperparameters were tuned using the *Hyperopt* package [Ber13] in this work. Tree-structured Parzen Estimator Approach (TPE) [Ber11], one of the Sequential Model-based Global Optimization (SMBO) [Hut11] algorithms, was chosen as the optimization strategy. SMBO algorithms build a model to approximate the performance score of hyperparameters based on historical measurements and suggest the next set of hyperparameters to evaluate based on the model. The considered hyperparameters for each classification model are listed in Table 7.1.

7.3 Results

This section analyzes the experiment results. We first compare the recognition accuracy of all classification methods. Four metrics were used to measure the performance: precision, accuracy, recall and F1-score. The results are shown in Table 7.2, with best results represented in bold.

For classification with handcrafted features, the best performance was achieved by Neural Net in the SART test (Precision: 0.76, Accuracy: 0.70, Recall: 0.72, F1: 0.74) and SVM with linear kernel function in the CPT test (Precision: 0.75, Accuracy: 0.62, Recall: 0.71, F1: 0.73).

The end-to-end methods showed better results than the feature-based methods in general. In agreement with the results reported in [Dzi20], CNN-based architectures outperformed other approaches, with FCN (Precision: 0.86, Accuracy: 0.82, Recall: 0.82, F1: 0.84 for SART, Precision: 0.84; Accuracy: 0.72, Recall: 0.75, F1: 0.79 for CPT) achieving the best results in terms of the most metrics for both test tasks. Architectures with LSTM-unit did not improve the recognition accuracy, suggesting that convolutional layers may be more robust than LSTM in feature representation. Poor performance of STRESNET indicates that spectro-temporal networks could be unsuitable for stress detection from IBI signals.

Table 7.2: Rest vs stress state classification results

Method		SART				CPT			
		Pre	Acc	Rec	F1	Pre	Acc	Rec	F1
Feature-based	Nearest Neighbors	0.72	0.65	0.65	0.69	0.71	0.48	0.48	0.57
	Linear SVM	0.74	0.69	0.72	0.73	0.75	0.62	0.71	0.73
	RBF SVM	0.69	0.64	0.69	0.69	0.74	0.58	0.64	0.69
	Gaussian Process	0.75	0.68	0.69	0.72	0.73	0.50	0.50	0.59
	Decision Tree	0.75	0.68	0.69	0.72	0.74	0.53	0.55	0.63
	Random Forest	0.71	0.63	0.63	0.67	0.75	0.52	0.52	0.62
	Neural Net	0.76	0.70	0.72	0.74	0.75	0.53	0.54	0.63
	AdaBoost	0.70	0.63	0.63	0.66	0.73	0.49	0.49	0.58
	Naive Bayes	0.68	0.63	0.72	0.70	0.72	0.52	0.54	0.62
End-to-end	FCN	0.86	**0.82**	0.82	**0.84**	0.84	**0.72**	**0.75**	**0.79**
	STRESNET	0.66	0.58	0.58	0.62	0.74	0.59	0.67	0.70
	RESNET	0.86	0.81	0.80	0.83	0.84	0.65	0.64	0.72
	MCDCNN	0.87	**0.82**	0.82	**0.84**	0.84	0.64	0.63	0.72
	CNN	0.83	0.80	0.82	0.82	0.82	0.64	0.65	0.73
	MLP	0.67	0.65	0.79	0.73	0.78	0.59	0.62	0.69
	MLPLSTM	0.70	0.70	**0.84**	0.76	0.80	0.58	0.57	0.67
	CNNLSTM	0.80	0.75	0.77	0.79	0.78	0.61	0.64	0.70
	INCEPTION	**0.88**	0.80	0.77	0.82	**0.85**	0.71	0.72	0.78

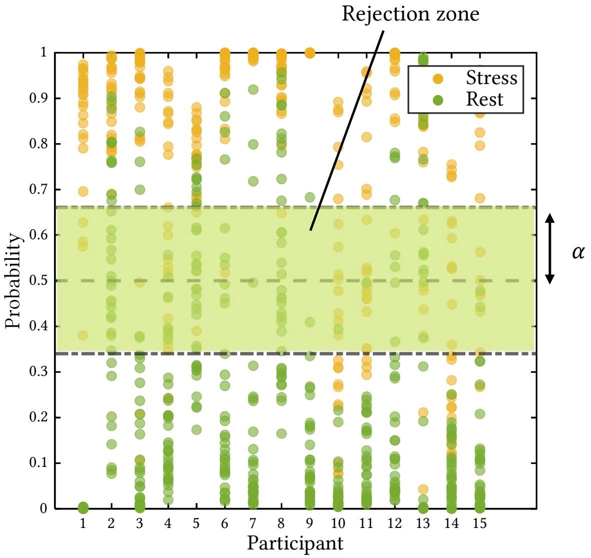


Figure 7.7: Acceptance or rejection of predictions. The predictions further away from the 0.5 boundary are more likely to be accurate. Predictions are rejected if inside the range [0.5 - α, 0.5 + α].

The classification performance was generally worse on the CPT test than on the SART test. All classifiers had better recognition results than a random guess for the SART test, while for the CPT test accuracy and recall of some feature-based methods were only about 50%. This is due to the fact that the test task in CPT is easier than in SART, thus less cognitive load was elicited in the CPT test.

The deep networks output a continuous prediction of probability through a softmax layer. The predictions of the Fully Convolutional Network, which showed the best results in the experiment, are displayed in Figure 7.7, where green points stand for rest class and yellow points for stress. It can be seen that the predictions close to 0 and 1 were more likely to be classified correctly. To integrate confidence into the result analysis, Harper and Southern [Har20] utilized Monte Carlo dropout to approximate posterior distribution over model predictions, in which the network needs to run an efficient number of times to generate the posterior distribution for a single input sample. The repeated runs of the network increase the execution time for the prediction. We simply considered the distance of the output to the 0.5-threshold as a measure of certainty and defined a rejection zone by a boundary coefficient α. For all outputs in the range [0.5 - α, 0.5 + α], the classifier considered it uncertain whether the measurement was in a restful or stress state and made therefore no prediction. Measurements outside the range were considered reliable and classified into either rest or stress state.

The relationship between the boundary coefficient α and the classification performance is demonstrated in Figure 7.8 and 7.9 for the SART and CPT test, respectively.

We can see that the classification performance increased with the growth of the boundary coefficient α, suggesting that the model provided more reliable classification results with a larger boundary coefficient. On the other hand, the coverage rate decreased with the increase of the coefficient α as more outputs below the threshold were discarded. We see that with the boundary threshold set as 0.2, our model achieved 0.9 for the F1-Score and 0.89 for the

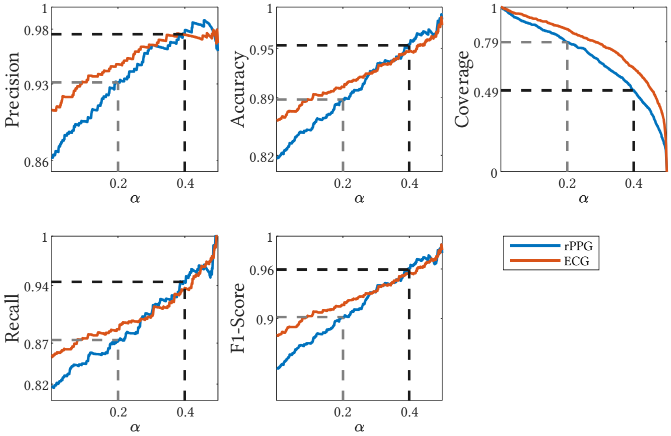


Figure 7.8: Metrics as functions of α on SART test.

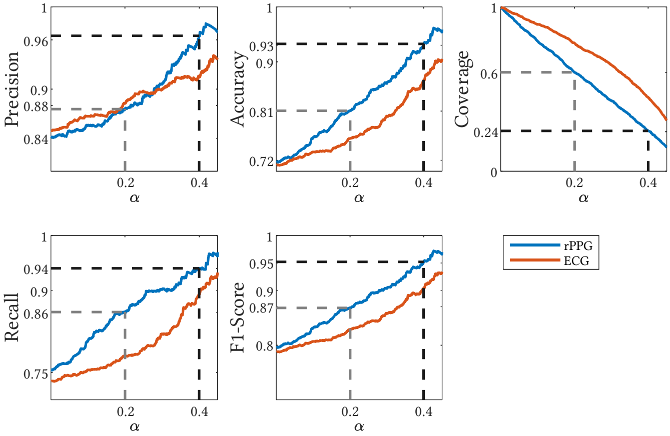


Figure 7.9: Metrics as functions of α on CPT.

accuracy on the SART test, with 79% of the outputs being accepted. If the boundary was set as 0.4, 49% of measurements were reliable, with the F1-Score and accuracy being 0.96 and 0.95, respectively. A similar tendency was observed for the results on the CPT-test: the F1-Score and accuracy were 0.87 and 0.81 respectively, if α was set as 0.2, and increased to 0.95 and 0.93 with the boundary coefficient set as 0.4.

The classification performance with the camera and the ECG sensor were compared as well. In both tests, ECG-based recognition had a higher recovery rate than the camera-based recognition at the same boundary coefficient. In the SART test, ECG-based detection outperformed the camera, while the camera-based measurement achieved higher performance scores in the CPT test.

Outperforming of the deep networks suggests that the networks were able to extract characteristics of different cognitive states from the IBI signals. To obtain a deeper insight, an experiment was conducted to visualize the patterns learned by the network using the activation-maximization technique [Sim13]. We froze the network model's parameters and calculated the output's gradient with respect to the input signal. An optimal pattern for each

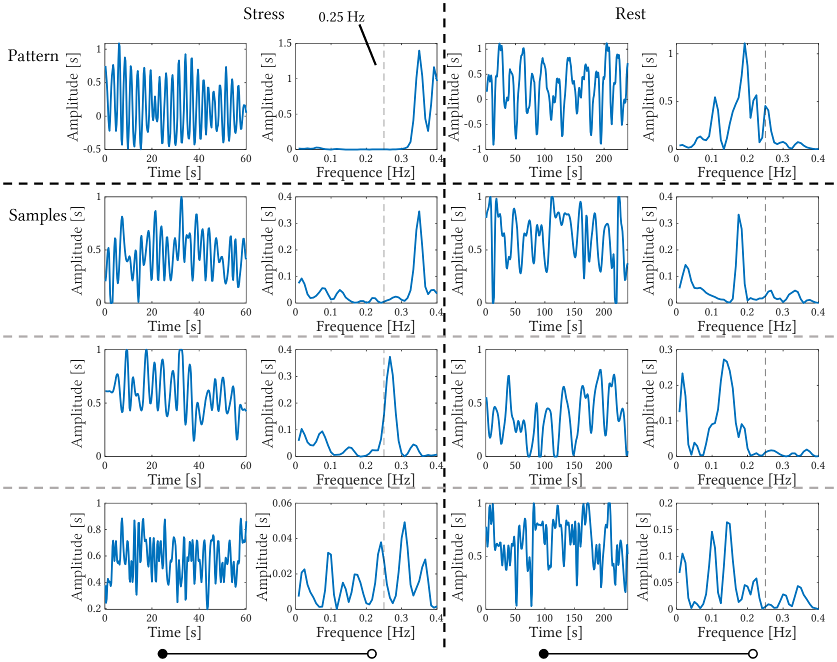


Figure 7.10: Row 1 are input patterns that maximize the output activations. Rows 2-4 are samples from the training data, where samples in the same row correspond to the same participant. Samples in the stress class have a higher frequency peak than that in the rest class. This characteristic is also presented in the activation patterns learned by the network.

class was obtained by updating the input signal using the gradients until the corresponding output unit of the softmax layer was maximized.

The obtained patterns for both stress and rest class are displayed in the first row of Figure 7.10. The patterns are presented in the time and frequency domains. Both patterns had a clear frequency peak, while the pattern for the stress class had a peak of higher frequency than the rest class. Some input signal samples corresponding to both classes from the dataset are presented in rows 2-4. We can see that the samples demonstrated the same characteristics learned by the network: samples in the stress class had a frequency peak higher than 0.25 Hz, and those in the rest class had a lower one. It should be

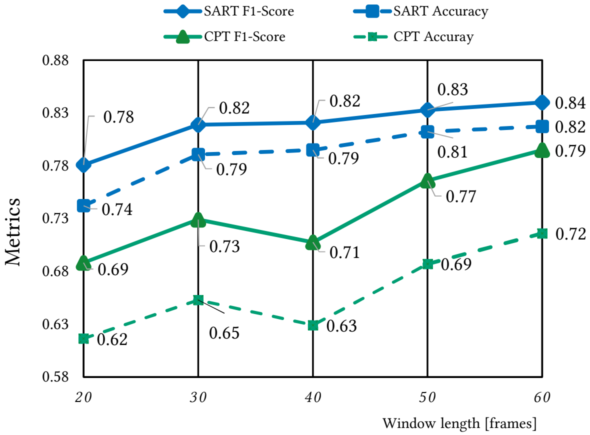


Figure 7.11: Model accuracy as a function of the window length. Accuracy metrics improve with the increasing window length for both tests.

mentioned that 0.25 Hz is not the typical frequency boundary to define PRV features. Instead, the boundary between the LF and HF power is usually set as 0.15 Hz. The examples demonstrate that the network can extract characteristics of different cognitive states without prior knowledge about the features.

The relationship between the classification performance of FCN and the window length is shown in Figure 7.11. Even though the performance degraded with a decreasing window length, FCN could achieve a F1-score of 78% on the SART test and 69% on the CPT test for the 20-second window length. This indicates that the deep network can extract features from a relatively short time interval, which allows cognitive state measurement in situations with a requirement of low latency.

7.4 Conclusion

This chapter investigates the ability of the remote measurement system for cognitive stress recognition. The vital parameters measured by the system were first evaluated by being compared with a reference ECG sensor. For selection of the recognition method, classification performance of ten end-to-end deep networks was compared with the feature-based classification methods. The CNN-based networks showed the best performance in the "Leave-One-Subject-Out" validation. Visualizing the features learned by the network using activation maximization has shown that the network can extract characteristics of signals for different stress states, indicating that computing handcrafted features is not necessary for camera-based stress recognition.

8

Dimensional Emotion Recognition from Camera-based PRV Features

This chapter delves into the potential of rPPG in the realm of emotion recognition. While rPPG-based emotion recognition has been explored, as highlighted in Section 3.3.4, the prevailing focus remains on categorical affect interpretation. Since emotions are not always experienced in distinct categories, but rather along continuous dimensions, dimensional representation of emotion status holds significant interest. Drawing insights from the study of Bugnon et al. [Bug17], which used self-organizing models and extreme learning machines for emotion recognition with ECG-based HRV features, this chapter seeks to assess the applicability of rPPG for dimensional emotion recognition. The work presented in this chapter has been published previously in the paper: [Zho23].

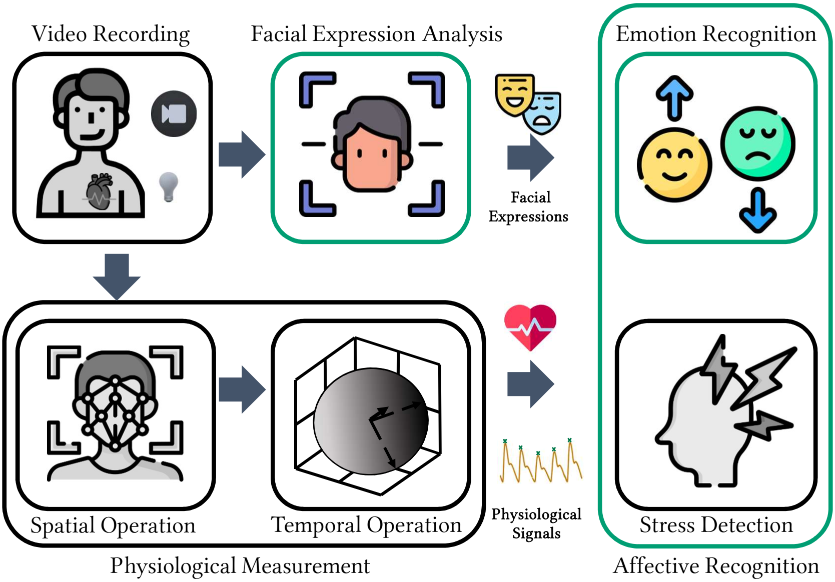


Figure 8.1: Overview of the software architecture - Emotion Recognition.

Section 8.1 provides an introduction to a FER-based benchmark method for emotion recognition, with which the rPPG-based approach will be compared in the later sections. Thereafter, Section 8.2 outlines the data acquisition process, followed by a discussion on data annotation in Section 8.3. Insights into the features and recognition methods used in this Chapter can be found in Section 8.4 and Section 8.5.The analysis results of the methods are discussed in 8.6. The chapter concludes with findings summarized in Section 8.7.

8.1 Benchmark facial expression analysis method

Facial expression analysis is one of the most widely researched solutions for vision-based emotion analysis, with state-of-the-art performance achieved by

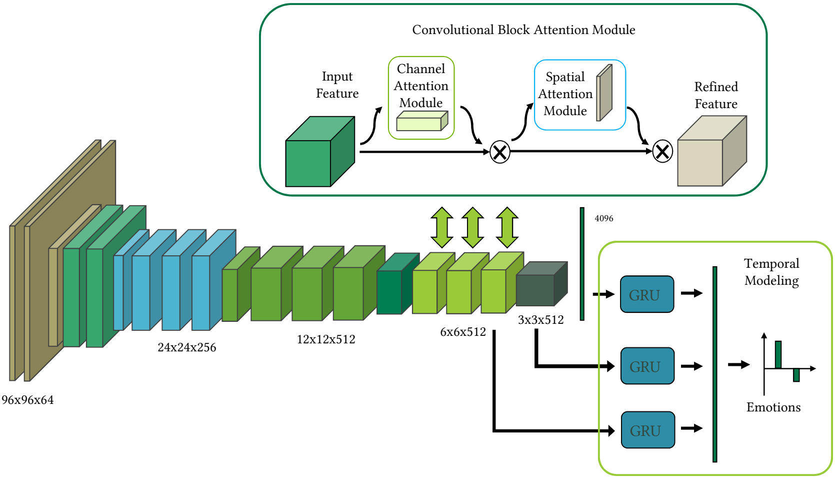


Figure 8.2: The architecture of the baseline network used for facial expression analysis. The network is derived from the VGG-Face network. CBAM modules are integrated into the last three convolutional layers. Additionally, three Gated Recurrent Units (GRUs) are connected to the last three layers of the network.

various deep learning approaches. This section introduces a VGG-based network [Sim14] for Facial Expression Recognition (FER), which will serve as a benchmark for evaluating the measurement method using camera-based physiological signals. The architecture of the network is shown in Figure 8.2.

In the network architecture, Convolutional Block Attention Module (CBAM) [Woo18] modules are integrated into the last three convolutional layers to refine the intermediate features along the channel and spatial dimensions. Additionally, three Gated Recurrent Units (GRUs) [Cho14] are connected to the last three layers (the last convolutional layer, the last pooling layer, and the fully connected layer) respectively in order to extract the temporal features of the sequential inputs.

The affect-in-the-wild (Aff-Wild) dataset and the one-minute general emotion behavior (OMG-Emotion) dataset were utilized for network training and test. The Aff-Wild dataset contains 298 videos with a total length of more than 30 hours, and the OMG-Emotion dataset contains 420 videos with a total length of around 10 hours. We trained the network on the Aff-Wild and the training set of the OMG-Emotion dataset, and then tested it on the test set of the OMG-Emotion dataset.

Figure 8.3 shows the Concordance Correlation Coefficient (CCC) of the network on the test set. It can be seen that the attention modules and the GRUs have notably improved the prediction accuracy of the network, with the best results achieved by the model *CNN + CBAM + GRU* (Valence: 0.40, Arousal: 0.19). More details of the benchmark FER method can be found in [Wan20].

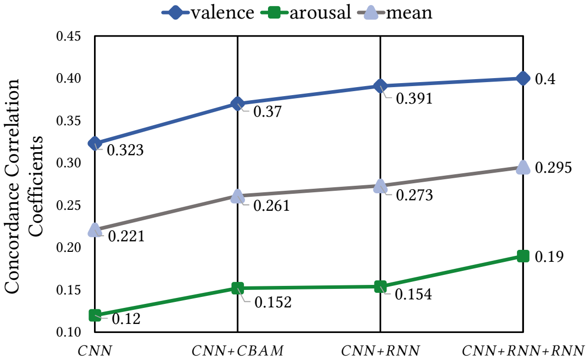


Figure 8.3: Performance comparison upon the ablation of CBAM and GRU modules. The best results were achieved with the CNN+CBAM +GRU model.

8.2 Data acquisition

Thirteen participants (6 females, 7 males) between the ages of 22 and 35, from four different nations and with no history of mental illness, were recruited for this experiment [Ard19]. Prior to the study, the participants were fully informed about the experimental procedures, including the recording of ECG data and face videos. They were requested to sign a consent form, which allowed the recordings to be used for this study and future research purposes.

Unlike in many datasets that were collected for facial expression analysis, where the expressions were usually posed, this study aimed to recognize emotions that were naturally elicited. To ensure the study could be adapted to real-life monitoring conditions, measurements were conducted in a simple setup with daylight and no use of additional light sources. We employed the same camera system and processing algorithm as used in the previous work in Chapter 7, where the reliability of the PRV measurements was validated.

Participants were presented with audio-video stimuli selected to elicit High-Arousal High-Valence (HAHV) and Low-Arousal Low-Valence (LALV) emotions (specifically amusement and sadness), while seated approximately 0.6 m away from a screen. The audio-visual stimuli utilized in the study were short film clips selected from a list validated by Gross and Levenson [Gro95]. The video used in the HAHV session lasted for about 65 seconds, while the one for the LALV session lasted for about 120 seconds. Including the baseline neutral state, the recording sessions involved three emotion classes.

Before the experiment began, participants were instructed to sit quietly for a period of 2 minutes to return to the neutral emotional state. This duration was determined to be adequate, as it yielded no significant variance in either pulse rates or the model's emotion output at the end of this pre-experiment resting phase when compared to baseline levels. After that, each audio-video stimulus was then played on the screen while the participants' faces were recorded simultaneously. To minimize the carry-over of emotions from one stimulus to the next, participants received a 2 minute rest period after each stimulus. The reference ECG signal was recorded using a chest strap (*EcgMove, Movisens* 512 Hz). Additionally, the sensor also registered signals for acceleration, rotation rate, air pressure, and temperature in parallel. The signals measured by the contact sensor were synchronized with the facial video based on the system timestamp.

8.3 Annotation

8.3.1 Annotation tool

The Self-Assessment Manikin (SAM) is commonly used in affective computing for obtaining emotion labels. However, it may not be suitable for continuous labeling, since it requires frequent inquiries about emotional status, which can distract participants from viewing stimuli and potentially fail to induce the target emotions. To this end, we developed an annotation tool in MATLAB for external annotators to label emotions after video recording. A screenshot of the tool is shown in Figure 8.4. The interface of the annotation tool is divided into four regions. The top left region displays the face video, which can be paused or resumed by clicking the play button located in the bottom left corner of the player window. At the top right of the annotation tool, two sets of reference labels are displayed separately in two coordinate systems. The reference labels were generated automatically through the use of two models: the FER model introduced in Section 8.1 and a HRV-feature-based model. The HRV-featured-based emotion classifier was provided by [Bug17], where the HRV features were extracted from the ECG sensor. The labels generated by the FER model and the HRV-feature-based model are intended to serve as basic guides, offering a general direction for emotion tracking.

A dotted line is plotted in both coordinates to represent the current image frame in the time axis. Annotators can drag the dotted line along the time axis to select any time point for annotation input. The current values of the reference labels are shown in the valence-arousal-plane at the lower right of the interface, represented in red and blue colors. The labeling values for both arousal and valence dimensions were limited within the range [-0.5, 0.5]. The annotators input an label by clicking on a point in the valence-arousal plane. The annotated label sequences for the entire video sequence are displayed at the lower left section of the interface.

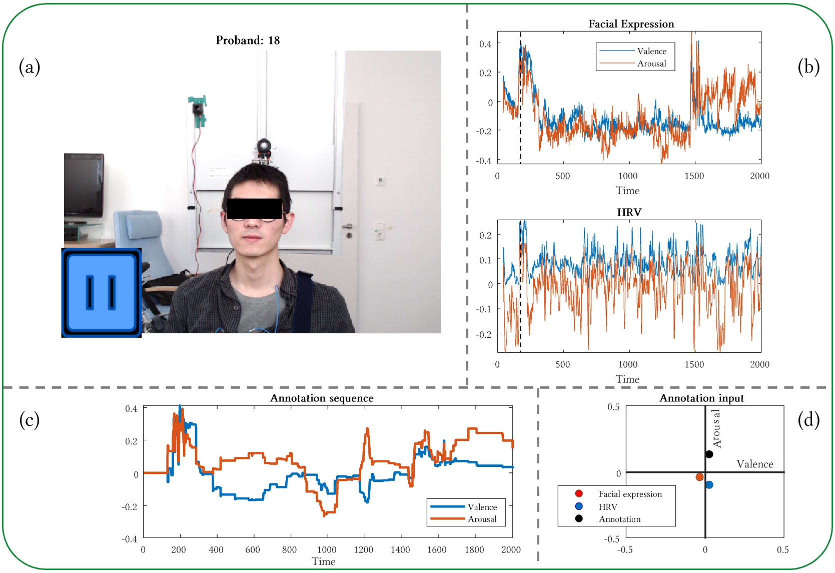


Figure 8.4: Annotation tool for emotional labeling: (a) video player, (b) reference labels given by a facial expression recognition network and a HRV-based recognition model, (c) label sequence over time, (d) annotation input.

8.3.2 Annotation validation

Three annotators were recruited for the annotation task and were given an introduction to the dimensional emotion model, along with guidance on how to use the annotation tool, prior to the annotation process. The annotators were not provided with any information about the content of the video stimuli. It is crucial to acknowledge that the labels given by the benchmark FER model and the HRV-based model are not to be regarded as gold standards, due to the inherent imperfections that come with algorithmic outputs. Our annotators have been thoroughly informed about these limitations and have been explicitly directed to prioritize their own subjective assessments, rather than relying exclusively on these algorithmically generated labels for their annotation tasks.

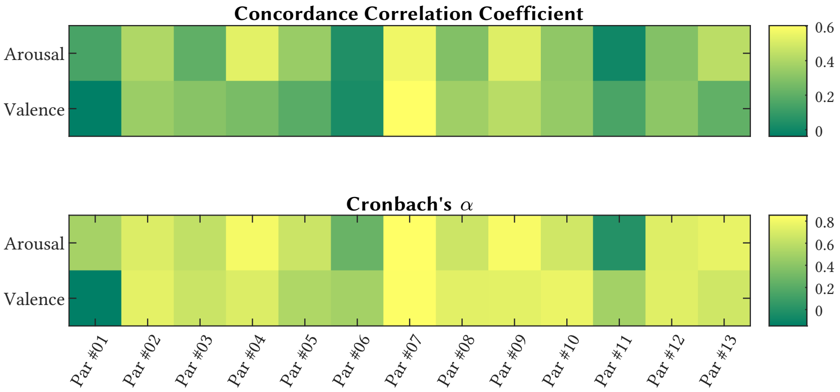


Figure 8.5: Labeling consistency among the annotators. Annotators rated valence and arousal of the participant by viewing the face videos. The Cronbach's alpha and Concordance Correlation Coefficient for each video are indicated by the scale color: from green to yellow, the higher the consistency of the labels.

Since the emotions develop relatively slowly and do not change in every frame, it is unnecessary to manually annotate each frame. Instead, the annotators were asked to input the annotations every second based on the observation of the video image displayed in the tool. The annotations were then interpolated at 30 fps to produce a label for each frame of the video.

Table 8.1: Statistics of labeling consistency

		CCC	MAE	Cronbach's α
Arousal	mean	0.31	0.09	0.62
	std	0.18	0.04	0.24
Valence	mean	0.27	0.11	0.61
	std	0.17	0.03	0.25

Since the annotation was based on subjective judgments, there may be variances between the labels provided by different annotators. To assess the quality of the annotations, we used three metrics to describe the consistency of the annotations: MAE, CCC, and Cronbach's alpha. We calculated the value of each metric by taking the average of all three pairs of annotators-to-annotators.

The mean and standard variance of the labeling consistency metrics across all 13 participants are listed in Table 8.1. The mean MAE values for arousal and valence were 0.09 and 0.11, respectively, which are considered acceptable given the value range of [-0.5, 0.5]. The CCC and Cronbach's alpha also showed positive correlations among the annotators. Figure 8.5 visualizes the CCC and Cronbach's alpha for all participants, where brighter colors represent higher values and darker colors represent lower values. The figure displays that labeling consistency varies among the annotators for different participants, with Participant 1, 6 and 11 exhibiting the lowest correlation. The valence label for Participant 1 and the arousal label for Participant 11 even showed negative correlation, which may have resulted from the lack of distinct facial expressions displayed by these participants during watching stimuli, resulting in the inaccessibility of facial cues for the annotators. These annotation analysis results demonstrate the challenges of acquisition of ground-truth emotional label. Although there were differences in annotation for some participants, the labels were considered to be largely consistent among the annotators.

8.4 Feature calculation

As in Chapter 7, the features for emotion recognition were extracted from the recorded video data. The pulse signal was extracted from the skin color signals using the Short Window Network [Zho21]. A peak detection algorithm was then utilized to determine the positions of each heart beat in the pulse signal, and the distances between two adjacent peaks were calculated as the IBI signals. To mitigate disturbances caused by local facial motions and light changes from the monitor, artifacts were removed from the IBI signals using *PhysioNet-Toolbox* [Ves18].

As in the implementation provided [Bug17], a window size of 20 seconds and a sliding step of 0.5 seconds were used to extract the features. The features extracted included statistics of Pulse Rate such as mean (MeanPR), variation (ampPR), skewness, and kurtosis. Additionally, frequency domain features such as LF and HF power, LF/HF ratio, and spectral decay slope (Pregr3) were considered. Five equidistant bands P0-4 in the range of [0.04-1] Hz were included, along with two commonly used time domain features SDNN and RMSSD. In total, 18 PRV features were extracted from the camera-based IBIs signals.

8.4.1 Feature analysis

This section analyzes whether there is a significant relationship between the features and emotion classes. Prior to analysis, the features were rescaled using min-max normalization. A visualization of the normalized features from different recording sessions is presented in Figure 8.6.

The figure shows that the PR features alone demonstrate significant differences between the amusement and the other two emotion sessions. Participants tended to have a higher heart rate while viewing the stimuli of the amusement emotion. Similarly, other features such as ampPR, LF, SDNN, and the equidistant frequency bands (P0-4) also exhibit similar tendencies. High order statistical features of heart rate, such as skewness and kurtosis, show the least relevance to the target emotions compared to other features.

Furthermore, it can be observed that the feature difference between the neutral emotion and the sadness session is much less significant. This could be due to the induced sadness being less intense compared to the amusement emotion, which leads to a less pronounced physiological response.

8.4.2 Feature selection

It was verified in [Suz21] that the ensemble feature selection can improve the accuracy of emotion estimation compared to using a single feature selection method. In the ensemble approach, the features are first ranked by multiple feature selection methods, and then selected based on the results of all feature selection approaches. In this study, we employed an ensemble strategy that

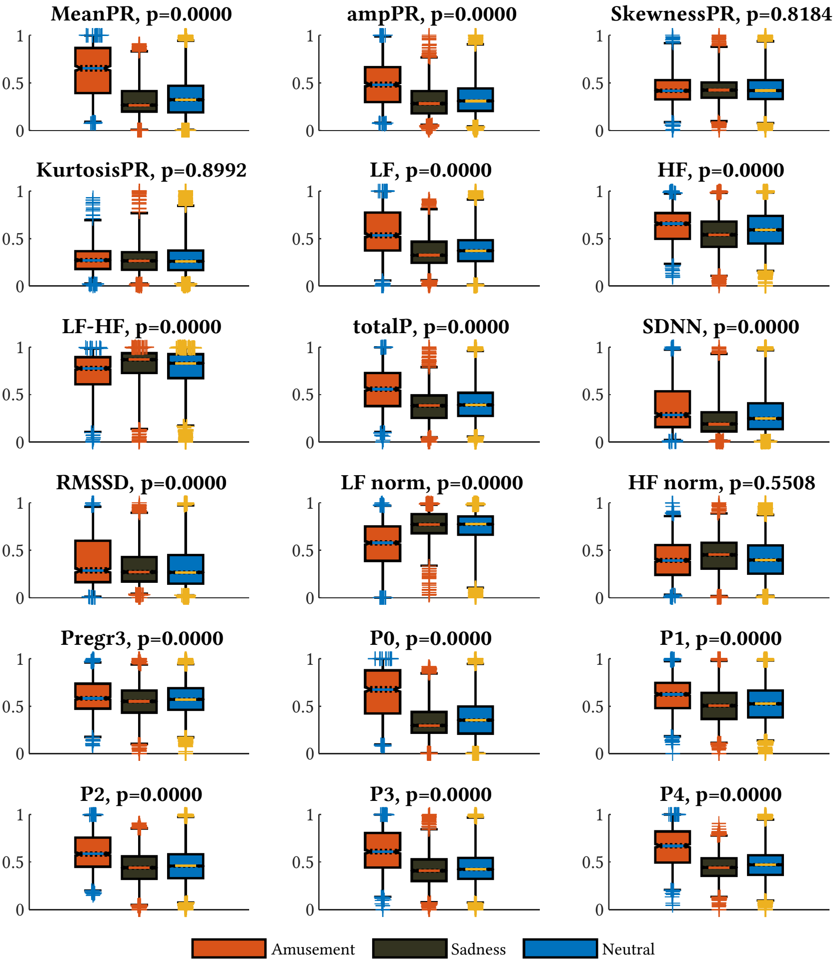


Figure 8.6: Feature comparison across the experiment sessions. Remarkable differences are shown between the feature of amusement and other recording sessions; the differences between the neutral emotion and sadness session are less pronounced. Skewness and kurtosis show the least relevance among all features.

incorporated four feature selection methods: Minimum Redundancy Maximum Relevance (MRMR) [Din05], F-test, Chi-square test, and ReliefF [Rob03].

The MRMR algorithm selects the optimal feature set by maximizing the relevance between the selected features and the targeted variables while concurrently minimizing the information redundancy among the features. The F-test and Chi-square-test are statistical hypothesis tests employed to ascertain whether the features are independent of the targets. Specifically, one-way analysis of variance (ANOVA) is applied for the F-test. The ReliefF algorithm evaluates and ranks the importance of the features by penalizing the features that exhibit differing values in closely related samples, while rewarding features with varying values in samples that have contrasting targets.

The weights of the features were analyzed with respect to both arousal and valence labels. To conduct the F-test and Chi-square-test methods, the features were first discretized into 15 bins. Then, the features were ranked based on their importance as determined by the algorithms, and represented in the

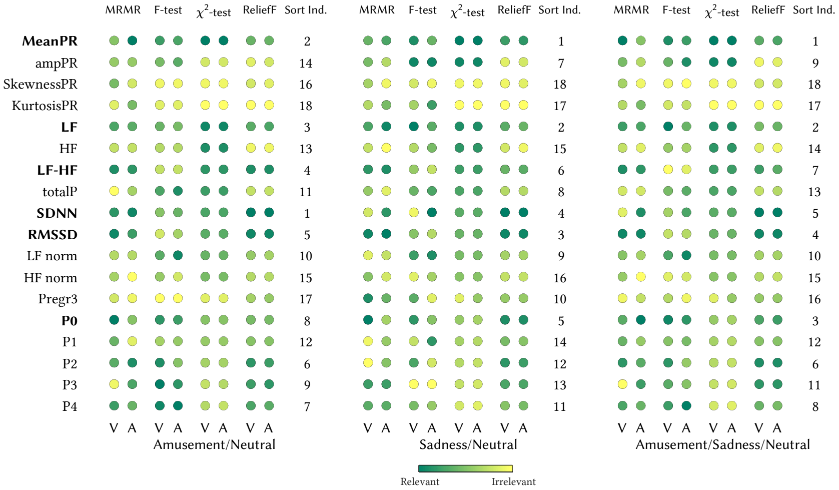


Figure 8.7: Feature importance using four feature selection methods: Minimum Redundancy Maximum Relevance (MRMR), F-test, Chi-square test, and ReliefF. V: Valence label; A: Arousal label. The sorting index stands for the average ranking obtained by all methods.

colormap in Figure 8.7. Features with a higher rank are shown in dark green, while those with a lower rank are represented in lighter colors. The sorting index was obtained by averaging the results of all four methods.

The figure indicates that there is a significant correlation between the feature weights for arousal and valence labels, since this work is focused on the recognition of HAHV and LALV emotions. To investigate whether the identification of amusement and sadness is sensitive to the same set of features, feature importance analysis was performed on segmented data, i.e., separately for amusement *vs.* neutral and sadness *vs.* neutral sessions, in addition to the analysis of the entire measurement session.

The results show that MeanPR, LF, LF/HF, SDNN, RMSSD, and P0 demonstrated high correlations with the labels of both emotions. AmpPR had a high importance only for sadness recognition, while P2 and P4 had a relatively high importance only for the identification of amusement. Among the results for all feature selection algorithms, PR skewness and kurtosis were found to have the lowest correlations with the labels, indicating a low dependence between these two features and emotional states, which is consistent with the results illustrated in Figure 8.6. Features that demonstrated high correlation with both emotions were selected for the recognition task later, as shown in bold in Figure 8.7.

8.5 Recognition methods

Totally ten machine learning methods were discussed for the dimensional emotion recognition task:

- Supervised Self-organizing Maps (sSOM)
- Extreme Learning Machines with neural network (nELM)
- Extreme Learning Machines based on kernels (kELM)
- Nearest Neighbors
- Linear SVM

- RBF SVM
- Decision Tree
- Random Forest
- Neural Network
- AdaBoost

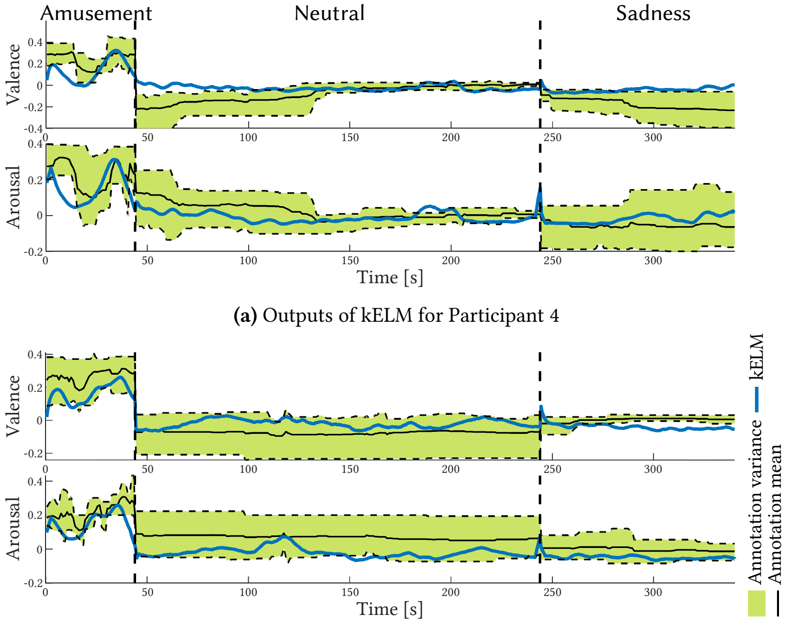
The evaluation of the sSOM [Koh13], nELM and kELM [Hua14] models was performed using the implementation provided in [Bug17], while for other methods, the *Scikit-learn* toolbox [Ped11] was employed. To validate the performance of the models, the Leave-One-Subject-Out (LOSO) cross-validation approach was adopted. Specifically, the data was split into N (13 in this work) partitions, with each partition representing a single participant. For each model, the training was performed on the data from $N-1$ partitions and tested on the remaining partition. This ensured that the test data was completely unseen by the model and the results were thus subject-independent.

8.6 Results and discussion

8.6.1 Correlation analysis

Table 8.2 presents the evaluation results of the recognition models, which are compared based on their predictive performance using CCCs and F-statistics as evaluation metrics for prediction quality.

CCCs assess the correlation between model predictions and labels. The CCCs were computed for each participant's valence and arousal labels and then averaged across all 13 participants. The table shows that all recognition methods yielded positive CCC values for both valence and arousal estimates. Among the methods, kELM achieved the highest CCC values with 0.34 for valence prediction and 0.36 for arousal prediction. For valence, sSOM had the second highest correlation with a CCC of 0.26, while for arousal, NeuralNet had the second highest CCC with a value of 0.3.



(a) Outputs of kELM for Participant 4

(b) Outputs of kELM for Participant 13

Figure 8.8: Outputs of kELM that achieved the best Concordance Correlation Coefficient. The black solid line represents the averaged annotations, while the shadowed region reflects the variations in annotations.

In Figure 8.8, the predictions of kELM are displayed for two participants. The black solid lines represent the average label, and the green shaded areas indicate the labeling variation among the annotators. Overall, it can be observed that the labels have positive values in the amusement session for both arousal and valence dimensions. However, the change from the neutral emotion is less noticeable in the sadness session. For instance, the average valence labeling in the sadness session of Participant 13 is even slightly higher than the neutral emotion, highlighting the challenge of distinguishing sad emotions from neutral one.

The kELM model's predictions (represented in blue lines) exhibit a similar trend as the labels, with high valence and arousal values for the amusement

Table 8.2: Mean CCCs and F-statistics of the prediction models

Method	Valence			Arousal		
	CCC	F_A	F_S	CCC	F_A	F_S
sSOM	0.26	14.46	9.79	0.17	16.46	3.94
nELM	0.23	17.15	2.35	0.26	11.23	4.60
kELM	**0.34**	18.10	3.62	**0.36**	17.51	**8.19**
Nearest Neighbors	0.22	9.44	2.67	0.14	9.71	3.73
Linear SVM	0.23	5.66	1.83	0.13	12.54	5.72
RBF SVM	0.24	17.38	5.93	0.23	12.97	3.60
Decision Tree	0.15	**19.84**	3.76	0.28	15.20	4.32
Random Forest	0.17	13.95	**13.39**	0.28	**20.40**	2.43
Neural Net	0.19	17.82	5.78	0.30	11.24	0.58
AdaBoost	0.23	14.80	1.70	0.21	16.40	1.88

session in the data from both participants. The model is even able to track the dynamics of mood change; for example, the labels and model predictions showed a valley value around 20 seconds for both participants in the amusement session. However, the transition from a neutral emotional state to a sad one was very gradual and subtle, making it difficult to discern noticeable shifts from the neutral emotion to sadness in the output curves. In order to gain more insight into the models' performance, a statistical analysis will be conducted in the next section.

8.6.2 Statistical analysis

We further analyse the separability of measurements from different recording sessions using the one-way ANOVA F-statistic. The one-way ANOVA F-statistic is defined as the ratio between the group variability and the within-group variability:

$$F = \frac{V_{between}/(c-1)}{V_{within}/(P-c)} \sim F_{c-1,P-c}\,, \tag{8.1}$$

where

$$V_{between} = \sum_{j=0}^{c} (\overline{y}_j - \overline{y})^2 , \tag{8.2}$$

and

$$V_{within} = \sum_{i=1}^{c} \sum_{j=1}^{P} (y_{ij} - \overline{y}_j)^2 , \tag{8.3}$$

with y_{ij} denoting the averaged prediction value for the i-th participant and the j-th emotion. $\overline{y}_j$ is the average output for the j-th emotion and $\overline{y}$ represents the overall average.

Under the null hypothesis that all group means are equal, the F-statistic follows an F-distribution with degrees of freedom of $(c-1, P-c)$, where c is the number of groups (i.e., sessions in this case) and P is the total number of samples (i.e., participants). The F-statistic can be interpreted as a measure of separability among the emotion sessions, where a larger F-value indicates that the distribution among the emotions are less likely to be the same, indicating more significant differences in the predictions for different emotions.

The F-statistics were calculated with the neutral session as reference. The separability of amusement F_A and sadness F_S with respect to the neutral session was assessed independently. The results can be found in Table 8.2 as well.

For the amusement session, all methods achieved high F-statistics for both arousal and valence dimensions. Among the methods, Decision Tree achieved the highest F-statistic of 19.84 for valence prediction, while Random Forest performed the best with a F-statistic of 20.4 for arousal prediction. As expected, the results show a lower separability of the sadness session from the neutral emotion. For valence prediction, sSOM, RBF SVM, Random Forest, and Neural Net had F-statistics greater than 5.0. The highest separability of the prediction of the valence for the sadness session was obtained by Random Forest with an F statistic of 13.39. Regarding the prediction of arousal, high separability was obtained by kELM and linear SVM, with kELM achieving the best results with an F-statistic of 8.19.

The scatter plot in Figure 8.9 - (rPPG) displays the measurements obtained by kELM, which has the highest correlation coefficient. The differently colored dots represent measurements from the three emotional sessions. It is apparent that the measurements for the amusement session are mainly located in the positive quadrant of the arousal-valence plane, and are distinctly separated from the measurements of the other sessions. However, the distinction between the sadness and neutral sessions is less noticeable, consistent with the observations in Figure 8.8. Nonetheless, it is still noticeable that the measurements for sadness tend to have lower arousal and valence values than neutral emotions. The plot also depicts an outlier from the sadness session (Valence: 0.060, Arousal:0.013) with a higher valence value than all neutral measurements, which can explain the lower F-statistic of kELM (3.62) for valence prediction presented in Table 8.2.

8.6.3 Comparison with facial expression analysis

Table 8.3 compares the recognition using different modalities. The scatter plot of each modality is illustrated in Figure 8.9. The results for facial expression analysis were given by the model described in 8.1. The measurements shown in Figure 8.9 - (*rPPG + Facial Expression*) were obtained by concatenating the PRV features with the output of the facial expression analysis network. In Table 8.3, it can be seen that, both facial expression (F_a: 35.26 for valence and 35.51 for arousal) and rPPG-based measurements (F_a: 18.10 for valence and 17.51 for arousal) show high F-statistics for the amusement session, indicating that both modalities can differentiate between amusement and neutral emotions. In terms of sadness recognition, the rPPG-based measurements (F_s: 3.62 for valence and 8.19 for arousal) gave higher F-statistics than facial expressions (F_s: 3.15 for valence and 3.94 for arousal). It is worth noting that the annotation label demonstrated less pronounced statistical differences between neutral and sadness than between neutral and amusement as well.

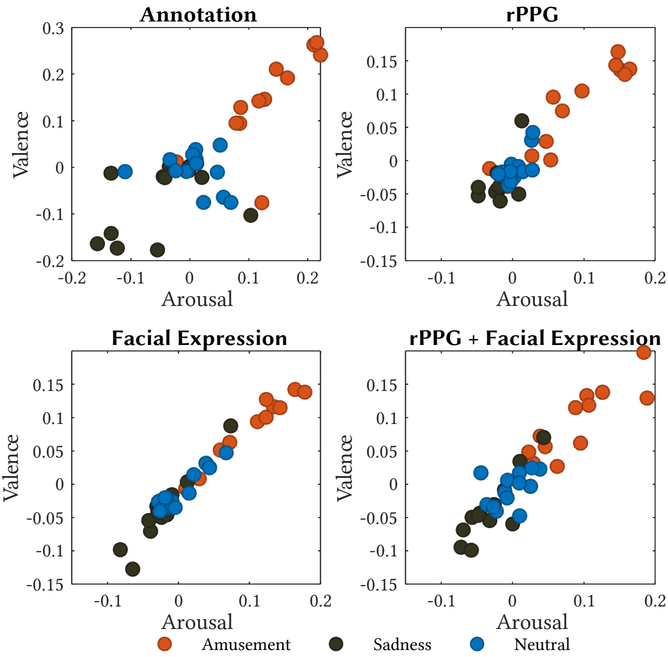


Figure 8.9: Distribution of the averaged output of each session on the valence-arousal plane. For all modalities, the outputs for amusement are mainly located in the HAHV region. For the sadness session, more samples are located in the LALV region when using both rPPG and facial expression features.

However, in Figure 8.9 - (*Annotation*) it can be seen that, the annotation assigned much lower values to the samples from the sadness sessions compared to the model predictions.

Furthermore, the combination of facial expression features with rPPG has been shown to improve the detection of negative emotions compared to analysing facial expressions alone. Specifically, our analysis revealed that the F-statistic for valence and arousal increased from 3.15 and 3.94 to 3.96 and 5.41, respectively, when using both modalities as opposed to only facial expression data. Moreover, although the arousal recognition for the sadness emotion did not outperform the rPPG-based measurement (F_s: 8.19) in terms of F-score when combining both modalities (F_s: 5.41), there were more

Table 8.3: Mean CCCs and F-statistics of different modalities

Modality	Valence			Arousal		
	CCC	F_A	F_S	CCC	F_A	F_S
Annotation	–	20.93	5.67	–	20.65	4.80
rPPG	0.34	18.10	3.62	0.36	17.51	**8.19**
Facial Expression	**0.42**	35.26	3.15	0.47	**35.51**	3.94
rPPG + Facial Expression	0.41	**39.34**	**3.96**	**0.49**	30.66	5.41

samples from the sadness emotion located in the LALV region shown in Figure 8.9 - (*rPPG + Facial Expression*) than in Figure 8.9 - (*rPPG*), indicating that the combination of facial expression features and rPPG can still provide valuable insights for detecting the negative emotions. However, although rPPG-based measurements showed better recognition results in detecting LALV emotions compared to facial expression analysis, recognizing emotions with less noticeable changes remains a challenge.

8.7 Conclusion

This chapter examined the ability of the rPPG system for dimensional emotion recognition. Ten machine learning methods were investigated, along with camera-based PRV features, to carry out the recognition task. The results of the analysis have confirmed that the system is capable of detecting high-arousal high-valence emotions, such as amusement. Furthermore, for emotions with less noticeable changes, such as sadness, rPPG-based measurements exhibited superior performance compared to the benchmark method for facial expression analysis. To improve the accuracy of affective labels, future studies may benefit from using an interactive experimental protocol. Additionally, it is important to expand the sample size to increase generalizability of the findings. Furthermore, investigating the recognition of emotions in other valence-arousal quadrants, such as anger, would be valuable, despite the challenges associated with eliciting and labeling data for these emotions.

9

Summary and prospects

WITH the significant advancements of measurement technologies, the acquisition of physiological parameters has moved beyond the conventional clinical sphere, spreading into everyday life. Beyond the fundamental medical applications of diagnosis, monitoring, and treatment of diseases, these measurements now serve as tools for enhancing physical, psychological, and emotional performance in daily life. In recent years, camera-based measurement of physiological parameters using rPPG has attracted a tremendous interest in the research community. In comparison to measurements using contact sensors, the camera-based measurement approach eliminates the need for specialized apparatus and negates the requirement of professional guidance for sensor probe placement. Another key advantage of rPPG is its non-invasive nature. It avoids direct interaction between the skin and the sensor, which significantly improves user comfort.

The central point of this work is the development of robust algorithms for a camera-based emotion/stress recognition system. The algorithm development can be divided into two distinct parts: the remote Photoplethysmography (rPPG) algorithm for measurement of physiological signals, and affective assessment algorithms. The accurate measurement of physiological signals forms the foundation of this measurement system. To ensure scalability and cost-effectiveness, the study utilizes low-cost cameras exclusively. The signal quality obtained from the blood volume pulse measurement is rigorously examined, with a particular emphasis on its sufficiency for stress/emotion recognition.

The development of the rPPG algorithms begins with the discussion of the short-window overlap-adding processing pipeline used in classical rPPG algorithms. The new algorithm was developed by iteratively investigating and improving the processing steps of the traditional methods. The algorithm consists of two integral components:

- *Spatial operation component:* Identifies the measurement sites for signal extraction based on the temporal and spectral features of each skin region.
- *Temporal operation component:* Processes the colors signals to extract blood volume pulse signals from the selected measurement sites.

In the temporal operation component, the core linear projection used in traditional rPPG algorithms was replaced by an encoder-decoder network. The network's performance was then evaluated on the VitalCamSet dataset, which comprises data from 26 participants. The evaluation experiment showed that the network benefits from using the short window overlap-add pipeline. Moreover, the proposed method has been proven to extract respiratory parameters with higher accuracy than traditional linear methods.

For the spatial operation component, the algorithm incorporates spatial convolution within the three-dimensional neural network to leverage the inherent spatial redundancy in image data. In contrast to traditional multi-trace approaches that select the ROIs through numerous hard thresholds, the neural network can autonomously retrieve the pulse-related features across spatial dimensions. Additionally, this research introduces a technique to normalize facial images, thereby mitigating the impact of disturbance factors such as different head poses, facial expressions, and background disruptions. Evaluations on the VitalCamSet dataset demonstrated that this face normalization method enhanced signal measurement compared to the fixed ROI approach.

The proposed methods have achieved state-of-the-art performance. An evaluation was conducted to compare with the state-of-the-art approaches using three public datasets. On the PURE dataset, the 3D network developed in this research outperformed others, achieving the best results (MAE: 0.45 bpm, RMSE: 1.10 bpm, ρ: 1.00). When compared to the second-best results, the proposed network significantly reduced the MAE from 1.65 to 0.45 bpm, and the RMSE from 2.02 bpm to 1.10 bpm. A similar superior performance of the network was observed on the UBFC dataset, where the 3D network reduced the MAE by over 70%, from 1.90 to 0.55 bpm. The network was also evaluated for measure pulse signals in Near-Infrared setups, and demonstrated competitive results on the NIRP dataset.

The algorithm developed for physiological measurement is subsequently applied to the tasks of stress and emotion recognition.

In the context of stress recognition, this research introduced an end-to-end strategy for stress detection, bypassing the conventional calculation of pulse rate variability parameters and directly determining stress levels from inter-beat-interval signals derived from the rPPG measurement. An evaluation experiment was conducted to evaluate the end-to-end algorithm. In the experiment, 15 participants were instructed to perform cognitive stress tasks. Their reactions were captured using a cost-effective webcam. In the data processing steps, the accuracy of PRV measurement was assessed by comparing it with the measurement using an ECG sensor. Although there were disparities in the absolute measurement values, rPPG-based measurements of most PRV parameters demonstrated a correlation over 0.9 with ECG-based HRV measurement. The HF features exhibited lower accuracy due to the low frame rate of the camera. Subsequently, ten feature-based approaches and ten end-to-end deep learning networks were tested using the recorded data. The outcomes indicated that, among all the tested methods, the fully convolutional network achieved the highest performance (Precision: 0.86, Accuracy: 0.82, Recall: 0.82, F1: 0.84 for SART test).

In the emotion analysis task, an experiment was conducted with 13 participants, who were subjected to audio-visual stimuli designed to elicit various emotions. The rPPG-based measurements were compared with a network for facial expression analysis, which was developed based on the VGG-Face network, supplemented with Convolutional Block Attention Module (CBAM) and Gated Recurrent Unit (GRU) modules to enhance its feature extraction capabilities. The results indicated that the rPPG-based emotion analysis achieved a Concordance Correlation Coefficient (CCC) of 0.34 for valence prediction and 0.36 for arousal prediction. The statistical analysis confirmed that for subtler emotional shifts, such as sadness, rPPG-based measurements showed superior performance compared to the benchmark facial expression analysis method.

Prospects

Although this study has shown the potential viability of rPPG technology for affective evaluation, there remain obstacles to address before it can be widely deployed in real-world systems. First, this dissertation focused on measurements obtained in an indoor setting, where illumination conditions are relatively controllable. In application scenarios such as driver state monitoring, factors such as head movements, shadows projected from buildings or trees alongside the street, and light reflection from other traffic participants, could pose challenges for signal extraction. Furthermore, the parameters of interest may differ across scenarios. For instance, there are differences between the stress experienced during driving and the stress elicited in cognitive tests. It is important to investigate whether the stress induced by real-world driving conditions can be accurately detected in vehicle applications as well. These practical considerations include additional variables such as traffic congestion, driving speed, or even the stress level from interacting with other road users. Addressing these factors requires the technology to discern more nuanced or complex affective states.

Aside from software and algorithm, improvement related to hardware, such as structured illumination or camera parameter regulation, can augment the robustness of measurement, especially in scenarios with high illumination dynamics.

In addition, investigating other physiological parameters such as blood pressure or Pulse Transit Time (PTT) also holds great promise. Apart from the intrinsic significance of these parameters themselves, such investigation can also enrich the field of emotion and stress recognition.

Moreover, the incorporation of multi-modal setups, which leverage other contactless sensor technologies such as radar or WiFi, could bring about innovative solutions that enhance the overall performance of the system. Such advancements could offer a more comprehensive, accurate, and robust assessment, thereby propelling the field of remote physiological measurement to new heights.

Bibliography

[Aba15] Abadi, Mojtaba Khomami; Subramanian, Ramanathan; Kia, Seyed Mostafa; Avesani, Paolo; Patras, Ioannis and Sebe, Nicu: “DECAF: MEG-based multimodal database for decoding affective physiological responses”. In: *IEEE Transactions on Affective Computing* 6.3 (2015), pp. 209–222 (cit. on p. 10).

[Adi15] Adib, Fadel; Mao, Hongzi; Kabelac, Zachary; Katabi, Dina and Miller, Robert C: “Smart homes that monitor breathing and heart rate”. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* 2015, pp. 837–846 (cit. on p. 37).

[Agr11] Agrafioti, Foteini; Hatzinakos, Dimitris and Anderson, Adam K: “ECG pattern analysis for emotion detection”. In: *IEEE Transactions on affective computing* 3.1 (2011), pp. 102–115 (cit. on p. 34).

[Alb16] Alberdi, Ane; Aztiria, Asier and Basarab, Adrian: "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review". In: *Journal of biomedical informatics* 59 (2016), pp. 49–75 (cit. on p. 3).

[Ame15] Amelard, Robert; Scharfenberger, Christian; Wong, Alexander and Clausi, David A: "Illumination-compensated non-contact imaging photoplethysmography via dual-mode temporally coded illumination". In: *Multimodal Biomedical Imaging X*. Vol. 9316. SPIE. 2015, pp. 35–39 (cit. on p. 40).

[Ami12] Aminoff, Michael J: Aminoff's Electrodiagnosis in Clinical Neurology: Expert Consult-Online and Print. Elsevier Health Sciences, 2012 (cit. on p. 25).

[Ard19] Ardinski, Tatjana: "Ertellen einer Datenbasis für die Emotions und Stressanalyse auf Basis der kamerabasierten Vitalparametermessung". Master's Thesis. Karlsruhe Institute of Technology, 2019 (cit. on p. 126).

[Arx18] Arx, Thomas von; Tamura, Kaori; Yukiya, Oba and Lozanoff, Scott: "The face–a vascular perspective. A literature review". In: *Swiss dental journal* 128.5 (2018), pp. 382–392 (cit. on p. 24).

[Ass10] Assoumou, HG Ntougou; Pichot, V; Barthelemy, JC; Dauphinot, V; Celle, S; Gosse, P; Kossovsky, M; Gaspoz, JM and Roche, F: "Metabolic syndrome and short-term and long-term heart rate variability in elderly free of clinical cardiovascular disease: the PROOF study". In: *Rejuvenation research* 13.6 (2010), pp. 653–663 (cit. on p. 20).

[Ast13] Asthana, Akshay; Zafeiriou, Stefanos; Cheng, Shiyang and Pantic, Maja: "Robust discriminative response map fitting with constrained local models". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3444–3451 (cit. on pp. 42, 43).

[Ave74] Averill, James R: “An analysis of psychophysiological symbolism and its influence on theories of emotion.” In: *Journal for the Theory of Social Behaviour* (1974) (cit. on p. 9).

[Bal16] Baltrušaitis, Tadas; Robinson, Peter and Morency, Louis-Philippe: “Openface: an open source facial behavior analysis toolkit”. In: *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2016, pp. 1–10 (cit. on p. 70).

[Ben18] Benezeth, Yannick; Li, Peixi; Macwan, Richard; Nakamura, Keisuke; Gomez, Randy and Yang, Fan: “Remote heart rate variability for emotional state monitoring”. In: *2018 IEEE EMBS international conference on biomedical & health informatics (BHI)*. IEEE. 2018, pp. 153–156 (cit. on p. 47).

[Ber11] Bergstra, James; Bardenet, Rémi; Bengio, Yoshua and Kégl, Balázs: “Algorithms for hyper-parameter optimization”. In: *Advances in neural information processing systems* 24 (2011) (cit. on p. 115).

[Ber13] Bergstra, James; Yamins, Daniel and Cox, David: “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures”. In: *International conference on machine learning*. PMLR. 2013, pp. 115–123 (cit. on p. 115).

[Ber97] Berntson, Gary G; Thomas Bigger Jr, J; Eckberg, Dwain L; Grossman, Paul; Kaufmann, Peter G; Malik, Marek; Nagaraja, Haikady N; Porges, Stephen W; Saul, J Philip; Stone, Peter H et al.: “Heart rate variability: origins, methods, and interpretive caveats”. In: *Psychophysiology* 34.6 (1997), pp. 623–648 (cit. on p. 16).

[Bet13] Betts, J Gordon; Desaix, P; Johnson, E; Johnson, JE; Korol, O; Kruse, D; Poe, B; Wise, JA; Womble, Mark and Young, KA: “Anatomy and physiology. OpenStax”. In: *Rice University: Houston, TX, USA* (2013) (cit. on p. 13).

[Big92] Bigger Jr, J Thomas; Fleiss, Joseph L; Steinman, Richard C; Rolnitzky, Linda M; Kleiger, Robert E and Rottman, Jeffrey N: “Frequency domain measures of heart period variability and mortality after myocardial infarction.” In: *Circulation* 85.1 (1992), pp. 164–171 (cit. on p. 20).

[Bla17a] Blackford, Ethan B and Estepp, Justin R: “Measurements of pulse rate using long-range imaging photoplethysmography and sunlight illumination outdoors”. In: *Optical Diagnostics and Sensing XVII: Toward Point-of-Care Diagnostics*. Vol. 10072. SPIE. 2017, pp. 122–134 (cit. on p. 36).

[Bla17b] Blackford, Ethan B and Estepp, Justin R: “Using consumer-grade devices for multi-imager non-contact imaging photoplethysmography”. In: *Optical Diagnostics and Sensing XVII: Toward Point-of-Care Diagnostics*. Vol. 10072. SPIE. 2017, pp. 96–104 (cit. on p. 41).

[Blö17] Blöcher, Timon; Schneider, Johannes; Schinle, Markus and Stork, Wilhelm: “An online PPGI approach for camera based heart rate monitoring using beat-to-beat detection”. In: *2017 IEEE Sensors Applications Symposium (SAS)*. IEEE. 2017, pp. 1–6 (cit. on p. 2).

[Blö18] Blöcher, Timon; Zhou, Kai; Krause, Simon and Stork, Wilhelm: “An Adaptive Bandpass Filter Based on Temporal Spectrogram Analysis for Photoplethysmography Imaging”. In: *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2018, pp. 1–6 (cit. on p. 44).

[Blö19] Blöcher, Timon; Krause, Simon; Zhou, Kai; Zeilfelder, Jennifer and Stork, Wilhelm: “VitalCamSet - a dataset for Photoplethysmography Imaging”. In: *2019 IEEE Sensors Applications Symposium (SAS)*. 2019, pp. 1–6 (cit. on pp. 69, 85).

[Blö20] Blöcher, Timon: “Kamerabasiertes System zur kontaktlosen Messung der momentanen Herzfrequenz für den Einsatz unter realen Umgebungsbedingungen”. In: (2020) (cit. on pp. 43, 52, 83).

[Bob17] Bobbia, Serge; Macwan, Richard; Benezeth, Yannick; Mansouri, Alamin and Dubois, Julien: “Unsupervised skin tissue segmentation for remote photoplethysmography”. In: *Pattern Recognition Letters* 124 (2017), pp. 82–90 (cit. on pp. 85, 94).

[Bob19] Bobbia, Serge; Macwan, Richard; Benezeth, Yannick; Mansouri, Alamin and Dubois, Julien: “Unsupervised skin tissue segmentation for remote photoplethysmography”. In: *Pattern Recognition Letters* 124 (2019), pp. 82–90 (cit. on p. 44).

[Bor34] Boris, N: “Delaunay. Sur la sphere vide”. In: *Izvestia Akademia Nauk SSSR, VII Seria, Otdelenie Matematicheskii i Estestvennyka Nauk* 7 (1934), pp. 793–800 (cit. on p. 88).

[Bou13] Bousefsaf, Frédéric; Maaoui, Choubeila and Pruski, Alain: “Remote assessment of the heart rate variability to detect mental stress”. In: *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE. 2013, pp. 348–351 (cit. on pp. 44, 47).

[Bou19] Bousefsaf, Frédéric; Pruski, Alain and Maaoui, Choubeila: “3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video”. In: *Applied Sciences* 9.20 (2019), p. 4364 (cit. on p. 99).

[Bug17] Bugnon, Leandro A; Calvo, Rafael A and Milone, Diego H: “Dimensional affect recognition from HRV: An approach based on supervised SOM and ELM”. In: *IEEE Transactions on Affective Computing* 11.1 (2017), pp. 32–44 (cit. on pp. 4, 34, 48, 123, 128, 131, 136).

[Bun23] Bundesamt, Statistisches: Cost-Illness. https://www.destatis.de/EN/Themes/Society-Environment/Health/Cost-Illness/_node.html. Accessed: 2023-03-18. 2023 (cit. on pp. 3, 4).

[Bur12] Burzo, Mihai; McDuff, Daniel; Mihalcea, Rada; Morency, Louis-Philippe; Narvaez, Alexis and Pérez-Rosas, Verónica: “Towards sensing the influence of visual narratives on human affect”. In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. 2012, pp. 153–160 (cit. on p. 47).

[Cam02] Campbell, Neil A and Reece, JB: Biology: International Edition. 2002 (cit. on p. 14).

[Can53] Cannon, Walter B: “Bodily changes in pain, hunger, fear, and rage, ed2”. In: *Boston, CH Branford* (1953), pp. 20–36 (cit. on p. 10).

[Car71] Caro, CG and Bloice, JA: “Contactless apnoea detector based on radar”. In: *The Lancet* 298.7731 (1971), pp. 959–961 (cit. on pp. 2, 37).

[Car99] Cardoso, Jean-François: “High-order contrasts for independent component analysis”. In: *Neural Computation* 11.1 (1999), pp. 157–192 (cit. on p. 71).

[Che16a] Chen, Chen; Chen, Yan; Han, Yi; Lai, Hung-Quoc and Liu, KJ Ray: “Achieving centimeter-accuracy indoor localization on WiFi platforms: A frequency hopping approach”. In: *IEEE Internet of Things Journal* 4.1 (2016), pp. 111–121 (cit. on p. 38).

[Che16b] Cheng, Juan; Chen, Xun; Xu, Lingxi and Wang, Z Jane: “Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition”. In: *IEEE journal of biomedical and health informatics* 21.5 (2016), pp. 1422–1433 (cit. on p. 44).

[Che18a] Chen, Weixuan and McDuff, Daniel J.: “DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 356–373 (cit. on pp. 46, 58, 84, 87).

[Che18b] Chen, Xun; Cheng, Juan; Song, Rencheng; Liu, Yu; Ward, Rabab and Wang, Z Jane: “Video-based heart rate measurement: Recent advances and future prospects”. In: *IEEE Transactions on Instrumentation and Measurement* 68.10 (2018), pp. 3600–3615 (cit. on p. 39).

[Cho14] Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger and Bengio, Yoshua: “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014) (cit. on p. 125).

[Coh00] Cohen, Hagit; Benjamin, Jonathan; Geva, Amir B; Matar, Mike A; Kaplan, Zeev and Kotler, Moshe: “Autonomic dysregulation in panic disorder and in post-traumatic stress disorder: application of power spectrum analysis of heart rate variability at rest and in response to recollection of trauma or panic attacks”. In: *Psychiatry research* 96.1 (2000), pp. 1–13 (cit. on p. 17).

[Con18] Conti, Daniela; Trubia, Grazia; Buono, Serafino; Di Nuovo, Santo and Di Nuovo, Alessandro: “Evaluation of a robot-assisted therapy for children with autism and intellectual disability”. In: *Towards Autonomous Robotic Systems: 19th Annual Conference, TAROS 2018, Bristol, UK July 25-27, 2018, Proceedings 19*. Springer. 2018, pp. 405–415 (cit. on p. 3).

[Coo01] Cootes, T.F.; Edwards, G.J. and Taylor, C.J.: “Active appearance models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), pp. 681–685 (cit. on p. 86).

[Coo98] Cootes, Timothy F; Edwards, Gareth J and Taylor, Christopher J: “Active appearance models”. In: *Computer Vision—ECCV' 98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*. Springer. 1998, pp. 484–498 (cit. on p. 84).

[Cro20] Crosswell, Alexandra D and Lockwood, Kimberly G: “Best practices for stress measurement: How to measure psychological stress in health research”. In: *Health psychology open* 7.2 (2020), p. 2055102920933072 (cit. on p. 3).

[Dar98] Darwin, Charles and Prodger, Phillip: The expression of the emotions in man and animals. Oxford University Press, USA, 1998 (cit. on p. 8).

[De 13] De Haan, Gerard and Jeanne, Vincent: “Robust pulse rate from chrominance-based rPPG”. In: *IEEE Transactions on Biomedical Engineering* 60.10 (2013), pp. 2878–2886 (cit. on pp. 44, 45, 59, 60, 73).

[De 14] De Haan, Gerard and Van Leest, Arno: “Improved motion robustness of remote-PPG by using the blood volume pulse signature”. In: *Physiological Measurement* 35.9 (2014), pp. 1913–1926 (cit. on pp. 45, 97).

[DeG10] DeGiorgio, Christopher M; Miller, Patrick; Meymandi, Sheba; Chin, Alex; Epps, Jordan; Gordon, Steven; Gornbein, Jeffrey and Harper, Ronald M: “RMSSD, a measure of vagus-mediated heart rate variability, is associated with risk factors for SUDEP: the SUDEP-7 Inventory”. In: *Epilepsy & Behavior* 19.1 (2010), pp. 78–81 (cit. on p. 19).

[Din05] Ding, Chris and Peng, Hanchuan: “Minimum redundancy feature selection from microarray gene expression data”. In: *Journal of bioinformatics and computational biology* 3.02 (2005), pp. 185–205 (cit. on p. 134).

[Din20] Ding, Jianyang and Wang, Yong: “A WiFi-based smart home fall detection system using recurrent neural network”. In: *IEEE Transactions on Consumer Electronics* 66.4 (2020), pp. 308–317 (cit. on p. 38).

[Du18] Du, Xuan; Yang, Kun and Zhou, Dongdai: “MapSense: Mitigating inconsistent WiFi signals using signal patterns and pathway map for indoor positioning”. In: *IEEE Internet of Things Journal* 5.6 (2018), pp. 4652–4662 (cit. on p. 38).

[Dyc75] Dyck, Peter James; Thomas, Peter Kynaston and Lambert, Edward Howard: Peripheral neuropathy. Vol. 2. Saunders, 1975 (cit. on p. 25).

[Dzi20] Dziężyc, Maciej; Gjoreski, Martin; Kazienko, Przemysław; Saganowski, Stanisław and Gams, Matjaž: “Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data”. In: *Sensors* 20.22 (2020), p. 6535 (cit. on pp. 35, 113, 115).

[Ekm73] Ekmann, P: “Universal facial expressions in emotion”. In: *Studia Psychologica* 15.2 (1973), p. 140 (cit. on p. 8).

[Est14] Estepp, Justin R; Blackford, Ethan B and Meier, Christopher M: “Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography”. In: *2014 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE. 2014, pp. 1462–1469 (cit. on p. 41).

[Far03] Farnebäck, Gunnar: “Two-frame motion estimation based on polynomial expansion”. In: *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer. 2003, pp. 363–370 (cit. on p. 43).

[Fit75] Fitzpatrick, TB: “Sun and skin”. In: *Journal de Medecine Esthetique* 2 (1975), pp. 33–34 (cit. on p. 25).

[Fra22] Frasch, Martin G: “Heart Rate as a Non-Invasive Biomarker of Inflammation: Implications for Digital Health”. In: *Frontiers in Immunology* 13 (2022), p. 930445 (cit. on p. 17).

[Gen22] Geng, Yexing: “Einsatz des Short-Window Network Verfahrens für kontaktlose Atemfrequenzmessung”. Bachelor’s Thesis. Karlsruhe University of Applied Sciences, 2022 (cit. on p. 79).

[Gjo20] Gjoreski, Martin; Gams, Matja Ž; Luštrek, Mitja; Genc, Pelin; Garbas, Jens-U and Hassan, Teena: “Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals”. In: *IEEE Access* 8 (2020), pp. 70590–70603 (cit. on p. 113).

[Gol20] Goldfine, Charlotte E; Oshim, Md Farhan Tasnim; Carreiro, Stephanie P; Chapman, Brittany P; Ganesan, Deepak and Rahman, Tauhidur: "Respiratory rate monitoring in clinical environments with a contactless ultra-wideband impulse radar-based sensor system". In: *Proceedings of the... Annual Hawaii International Conference on System Sciences. Annual Hawaii International Conference on System Sciences*. Vol. 2020. NIH Public Access. 2020, p. 3366 (cit. on p. 37).

[Gri06] Grillon, Christian; Pine, Daniel S; Baas, Johanna MP; Lawley, Megan; Ellis, Valerie and Charney, Dennis S: "Cortisol and DHEA-S are associated with startle potentiation during aversive conditioning in humans". In: *Psychopharmacology* 186 (2006), pp. 434–441 (cit. on p. 4).

[Gro95] Gross, James J and Levenson, Robert W: "Emotion elicitation using films". In: *Cognition & emotion* 9.1 (1995), pp. 87–108 (cit. on p. 127).

[Gu16] Gu, Changzhan; Peng, Zhengyu and Li, Changzhi: "High-precision motion detection using low-complexity Doppler radar with digital post-distortion technique". In: *IEEE Transactions on Microwave Theory and Techniques* 64.3 (2016), pp. 961–971 (cit. on p. 37).

[Gu17] Gu, Yu; Zhan, Jinhai; Ji, Yusheng; Li, Jie; Ren, Fuji and Gao, Shangbing: "MoSense: An RF-based motion detection system via off-the-shelf WiFi devices". In: *IEEE Internet of Things Journal* 4.6 (2017), pp. 2326–2341 (cit. on pp. 2, 38).

[Gua15] Guazzi, Alessandro R; Villarroel, Mauricio; Jorge, Joao; Daly, Jonathan; Frise, Matthew C; Robbins, Peter A and Tarassenko, Lionel: "Non-contact measurement of oxygen saturation with an RGB camera". In: *Biomedical optics express* 6.9 (2015), pp. 3320–3338 (cit. on p. 40).

[Guz05] Guzzetti, Stefano; Rovere, Maria Teresa La; Pinna, Gian Domenico; Maestri, Roberto; Borroni, Ester; Porta, Alberto; Mortara, Andrea and Malliani, Alberto: "Different spectral components of 24 h heart rate variability are related to different modes of death in chronic heart failure". In: *European heart journal* 26.4 (2005), pp. 357–362 (cit. on p. 20).

[Har20] Harper, Ross and Southern, Joshua: "A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat". In: *IEEE transactions on affective computing* 13.2 (2020), pp. 985–991 (cit. on pp. 35, 117).

[Has15] Hassner, Tal; Harel, Shai; Paz, Eran and Enbar, Roee: "Effective face frontalization in unconstrained images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4295–4304 (cit. on p. 86).

[Has19] Hassan, Mohammad Mehedi; Alam, Md Golam Rabiul; Uddin, Md Zia; Huda, Shamsul; Almogren, Ahmad and Fortino, Giancarlo: "Human emotion recognition using deep belief network architecture". In: *Information Fusion* 51 (2019), pp. 10–18 (cit. on p. 34).

[Hen12] Henriques, Joao F; Caseiro, Rui; Martins, Pedro and Batista, Jorge: "Exploiting the circulant structure of tracking-by-detection with kernels". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*. Springer. 2012, pp. 702–715 (cit. on p. 83).

[Heu11] Heuer, Stephan: "Ambiente kapazitive EKG-Messung: Elektroden, Systeme und Konzepte". PhD thesis. Karlsruhe, Karlsruher Institut für Technologie (KIT), Diss., 2011, 2011 (cit. on p. 37).

[Hou00] Houseman, Nicholas D; Taylor, G Ian; Pan, Wei-Ren et al.: "The angiosomes of the head and neck: anatomic study and clinical applications". In: *Plastic and reconstructive surgery* 105.7 (2000), pp. 2287–2313 (cit. on p. 24).

[Hou17] Hou, Yuxiao; Wang, Yanwen and Zheng, Yuanqing: "TagBreathe: Monitor breathing with commodity RFID systems". In: *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2017, pp. 404–413 (cit. on p. 37).

[How06] Howell, Simon and Dodman, James: “The peripheral circulation”. In: *Foundations of anesthesia*. Elsevier, 2006, pp. 525–536 (cit. on p. 24).

[Hsu17] Hsu, Gee-Sern; Ambikapathi, ArulMurugan and Chen, Ming-Shiang: “Deep learning with time-frequency representation for pulse estimation from facial videos”. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. 2017, pp. 383–389 (cit. on p. 46).

[Hu17] Hu, Ping; Cai, Dongqi; Wang, Shandong; Yao, Anbang and Chen, Yurong: “Learning supervised scoring ensemble for emotion recognition in the wild”. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, pp. 553–560 (cit. on p. 86).

[Hua14] Huang, Guang-Bin: “An insight into extreme learning machines: random neurons, random features and kernels”. In: *Cognitive Computation* 6 (2014), pp. 376–390 (cit. on p. 136).

[Hua17] Huang, Rui; Zhang, Shu; Li, Tianyu and He, Ran: “Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2439–2448 (cit. on pp. 44, 86).

[Hül08] Hülsbusch, Markus and Rembold, Bernhard: Ein bildgestütztes, funktionelles Verfahren zur optoelektronischen Erfassung der Hautperfusion. Tech. rep. Lehrstuhl und Institut für Hochfrequenztechnik, 2008 (cit. on pp. 40, 41, 45).

[Hut11] Hutter, Frank; Hoos, Holger H and Leyton-Brown, Kevin: “Sequential model-based optimization for general algorithm configuration”. In: *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers 5*. Springer. 2011, pp. 507–523 (cit. on p. 115).

[Ism20] Ismail Fawaz, Hassan; Lucas, Benjamin; Forestier, Germain; Pelletier, Charlotte; Schmidt, Daniel F; Weber, Jonathan; Webb, Geoffrey I; Idoumghar, Lhassane; Muller, Pierre-Alain and Petitjean, François: “Inceptiontime: Finding alexnet for time series classification”. In: *Data Mining and Knowledge Discovery* 34.6 (2020), pp. 1936–1962 (cit. on p. 113).

[Jia14] Jiang, Wen Jun; Gao, Shi Chao; Wittek, Peter and Zhao, Li: “Real-time quantifying heart beat rate from facial video recording on a smart phone using Kalman filters”. In: *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).* IEEE. 2014, pp. 393–396 (cit. on p. 44).

[Jun19] Jung, Wookyoung; Jang, Kuk-In and Lee, Seung-Hwan: “Heart and brain interaction of psychiatric illness: a review focused on heart rate variability, cognitive function, and quantitative electroencephalography”. In: *Clinical Psychopharmacology and Neuroscience* 17.4 (2019), p. 459 (cit. on p. 17).

[Kal10] Kalal, Zdenek; Mikolajczyk, Krystian and Matas, Jiri: “Forward-backward error: Automatic detection of tracking failures”. In: *2010 20th international conference on pattern recognition.* IEEE. 2010, pp. 2756–2759 (cit. on p. 83).

[Kev17] Kevat, Ajay C; Bullen, Denise VR; Davis, Peter G and Kamlin, C Omar F: “A systematic review of novel technology for monitoring infant and newborn heart rate”. In: *Acta Paediatrica* 106.5 (2017), pp. 710–720 (cit. on p. 2).

[Kha17] Khan, Usman Mahmood; Kabir, Zain and Hassan, Syed Ali: “Wireless health monitoring using passive WiFi sensing”. In: *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC).* IEEE. 2017, pp. 1771–1776 (cit. on p. 38).

[Kha21] Khan, Muhammad Imran; Jan, Mian Ahmad; Muhammad, Yar; Do, Dinh-Thuan; Rehman, Ateeq ur; Mavromoustakis, Constandinos X and Pallis, Evangelos: “Tracking vital signs of a patient using channel state information and machine learning for a smart healthcare system". In: *Neural Computing and Applications* (2021), pp. 1–15 (cit. on pp. 2, 38).

[Kin14] Kingma, Diederik P and Ba, Jimmy: "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 71, 93).

[Koe11] Koelstra, Sander; Muhl, Christian; Soleymani, Mohammad; Lee, Jong-Seok; Yazdani, Ashkan; Ebrahimi, Touradj; Pun, Thierry; Nijholt, Anton and Patras, Ioannis: "Deap: A database for emotion analysis; using physiological signals". In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31 (cit. on p. 10).

[Koh13] Kohonen, Teuvo: "Essentials of the self-organizing map". In: *Neural networks* 37 (2013), pp. 52–65 (cit. on p. 136).

[Kor09] Korpas, David; Halek, J and Doležal, L: "Parameters describing the pulse wave." In: *Physiological research* 58.4 (2009) (cit. on p. 14).

[Kum15] Kumar, Mayank; Veeraraghavan, Ashok and Sabharwal, Ashutosh: "DistancePPG: Robust non-contact vital signs monitoring using a camera". In: *Biomedical optics express* 6.5 (2015), pp. 1565–1588 (cit. on p. 100).

[Kur10] Kurian, Thomas; Ambrosi, Christina; Hucker, William; Fedorov, Vadim V and Efimov, Igor R: "Anatomy and electrophysiology of the human AV node". In: *Pacing and clinical electrophysiology* 33.6 (2010), pp. 754–762 (cit. on p. 13).

[Lee18] Lee, Sangyoun; Park, Young-Deok; Suh, Young-Joo and Jeon, Seokseong: "Design and implementation of monitoring system for breathing and heart rate pattern using WiFi signals". In: *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE. 2018, pp. 1–7 (cit. on p. 38).

[Lee20] Lee, Eugene; Chen, Evan and Lee, Chen-Yi: “Meta-rPPG: Remote Heart Rate Estimation Using a Transductive Meta-learner.” In: *European Conference on Computer Vision.* 2020, pp. 392–409 (cit. on pp. 46, 99).

[Lew11] Lewandowska, Magdalena; Rumiński, Jacek; Kocejko, Tomasz and Nowak, Jędrzej: “Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity”. In: *2011 federated conference on computer science and information systems (FedCSIS).* IEEE. 2011, pp. 405–410 (cit. on p. 44).

[Li14] Li, Xiaobai; Chen, Jie; Zhao, Guoying and Pietikainen, Matti: “Remote heart rate measurement from face videos under realistic situations”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2014, pp. 4264–4271 (cit. on p. 44).

[Li20] Li, Chao; Bao, Zhongtian; Li, Linhao and Zhao, Ziping: “Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition”. In: *Information Processing & Management* 57.3 (2020), p. 102185 (cit. on p. 35).

[Liu17] Liu, Jiamin; Su, Yuanqi and Liu, Yuehu: “Multi-modal emotion recognition with temporal-band attention based on LSTM-RNN”. In: *Pacific rim conference on multimedia.* Springer. 2017, pp. 194–204 (cit. on pp. 4, 34, 35).

[Liu20a] Liu, Xin; Fromm, Josh; Patel, Shwetak N. and McDuff, Daniel J.: “Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement”. In: *Advances in Neural Information Processing Systems.* Vol. 33. 2020, pp. 19400–19411 (cit. on p. 87).

[Liu20b] Liu, Xin; Jiang, Ziheng; Fromm, Josh; Xu, Xuhai; Patel, Shwetak N. and McDuff, Daniel J.: “MetaPhys: Unsupervised Few-Shot Adaptation for Non-Contact Physiological Measurement.” In: *arXiv preprint arXiv:2010.01773* (2020) (cit. on pp. 46, 58, 84, 85, 87, 99, 100).

[Lo20] Lo, Ei-Wen Victor; Wei, Yin-Hsuan and Hwang, Bing-Fang: “Association between occupational burnout and heart rate variability: a pilot study in a high-tech company in Taiwan”. In: *Medicine* 99.2 (2020) (cit. on p. 17).

[Lu15] Lu, Yifei; Zheng, Wei-Long; Li, Binbin and Lu, Bao-Liang: “Combining eye movements and eeg to enhance emotion recognition.” In: *IJCAI*. Vol. 15. Buenos Aires. 2015, pp. 1170–1176 (cit. on pp. 4, 34).

[Lu21] Lu, Hao and Han, Hu: “NAS-HR: Neural architecture search for heart rate estimation from face videos”. In: *Virtual Reality & Intelligent Hardware* 3.1 (2021), pp. 33–42 (cit. on pp. 97, 99).

[Mag18] Magdalena Nowara, Ewa; Marks, Tim K; Mansour, Hassan and Veeraraghavan, Ashok: “SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 1272–1281 (cit. on pp. 85, 94, 100).

[Mar13] Martinez, Hector P; Bengio, Yoshua and Yannakakis, Georgios N: “Learning deep physiological models of affect”. In: *IEEE Computational intelligence magazine* 8.2 (2013), pp. 20–33 (cit. on pp. 5, 34, 35).

[McC15] McCraty, Rollin and Shaffer, Fred: “Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk”. In: *Global advances in health and medicine* 4.1 (2015), pp. 46–61 (cit. on p. 20).

[McD14] McDuff, Daniel; Gontarek, Sarah and Picard, Rosalind: “Remote measurement of cognitive stress via heart rate variability”. In: *2014 36th annual international conference of the IEEE engineering in medicine and biology society*. IEEE. 2014, pp. 2957–2960 (cit. on p. 47).

[McD16] McDuff, Daniel J; Hernandez, Javier; Gontarek, Sarah and Picard, Rosalind W: "Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 4000–4004 (cit. on pp. 47, 112).

[McD20] McDuff, Daniel J.; Hernandez, Javier; Wood, Erroll; Liu, Xin and Baltrusaitis, Tadas: "Advancing Non-Contact Vital Sign Measurement using Synthetic Avatars." In: *arXiv preprint arXiv:2010.12949* (2020) (cit. on pp. 46, 99).

[McD23] McDuff, Daniel: "Camera measurement of physiological vital signs". In: *ACM Computing Surveys* 55.9 (2023), pp. 1–40 (cit. on p. 39).

[Meh17] Mehmood, Raja Majid; Du, Ruoyu and Lee, Hyo Jong: "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors". In: *Ieee Access* 5 (2017), pp. 14797–14806 (cit. on pp. 4, 34).

[Mer17] Mercuri, Marco; Liu, Yao-Hong; Lorato, Ilde; Torfs, Tom; Bourdoux, André and Van Hoof, Chris: "Frequency-tracking CW Doppler radar solving small-angle approximation and null point issues in non-contact vital signs monitoring". In: *IEEE transactions on biomedical circuits and systems* 11.3 (2017), pp. 671–680 (cit. on p. 37).

[Mor59] Moretti, Giuseppe; Ellis, Richard A and Mescon, Herbert: "Vascular patterns in the skin of the face". In: *Journal of Investigative Dermatology* 33.3 (1959), pp. 103–112 (cit. on p. 24).

[Nan15] Nandakumar, Rajalakshmi; Gollakota, Shyamnath and Watson, Nathaniel: "Contactless sleep apnea detection on smartphones". In: *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 2015, pp. 45–57 (cit. on pp. 2, 38).

[Nar15] Nardelli, Mimma; Valenza, Gaetano; Greco, Alberto; Lanata, Antonio and Scilingo, Enzo Pasquale: “Recognizing emotions induced by affective sounds through heart rate variability”. In: *IEEE Transactions on Affective Computing* 6.4 (2015), pp. 385–394 (cit. on pp. 4, 34).

[Nir14] Nirmalan, Mahesh and Dark, Paul M: “Broader applications of arterial pressure wave form analysis”. In: *Continuing Education in Anaesthesia, Critical Care & Pain* 14.6 (2014), pp. 285–290 (cit. on p. 14).

[Niu18] Niu, Xuesong; Han, Hu; Shan, Shiguang and Chen, Xilin: “SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. 2018, pp. 3580–3585 (cit. on pp. 46, 97).

[Niu19] Niu, Xuesong; Shan, Shiguang; Han, Hu and Chen, Xilin: “RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 2409–2423 (cit. on pp. 46, 58, 84).

[Niu20] Niu, Xuesong; Yu, Zitong; Han, Hu; Li, Xiaobai; Shan, Shiguang and Zhao, Guoying: “Video-based Remote Physiological Measurement via Cross-verified Feature Disentangling”. In: *ECCV (2)*. 2020, pp. 295–310 (cit. on pp. 46, 58, 68, 84).

[Now20] Nowara, Ewa M; McDuff, Daniel and Veeraraghavan, Ashok: “A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 284–285 (cit. on p. 26).

[Now21] Nowara, Ewa M; McDuff, Daniel and Veeraraghavan, Ashok: “The benefit of distraction: Denoising camera-based physiological measurements using inverse attention”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4955–4964 (cit. on p. 100).

[ORo01] O'Rourke, Michael F; Pauca, Alfredo and Jiang, Xiong-Jing: "Pulse wave analysis". In: *British journal of clinical pharmacology* 51.6 (2001), p. 507 (cit. on p. 15).

[Pal18] Palipana, Sameera; Rojas, David; Agrawal, Piyush and Pesch, Dirk: "FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (2018), pp. 1–25 (cit. on p. 38).

[Ped11] Pedregosa, F. et al.: "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 113, 136).

[Pei19] Peirce, Jonathan; Gray, Jeremy R; Simpson, Sol; MacAskill, Michael; Höchenberger, Richard; Sogo, Hiroyuki; Kastman, Erik and Lindeløv, Jonas Kristoffer: "PsychoPy2: Experiments in behavior made easy". In: *Behavior research methods* 51.1 (2019), pp. 195–203 (cit. on p. 106).

[Poh10] Poh, Ming-Zher; McDuff, Daniel J and Picard, Rosalind W: "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." In: *Optics express* 18.10 (2010), pp. 10762–10774 (cit. on pp. 43, 62).

[Poh11] Poh, Ming-Zher; McDuff, Daniel J and Picard, Rosalind W: "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam". In: *IEEE Transactions on Biomedical Engineering* 58.1 (2011), pp. 7–11 (cit. on p. 45).

[Por17] Poria, Soujanya; Cambria, Erik; Bajpai, Rajiv and Hussain, Amir: "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information fusion* 37 (2017), pp. 98–125 (cit. on p. 10).

[Pra18] Prakash, Sakthi Kumar Arul and Tucker, Conrad S: "Bounded Kalman filter method for motion-robust, non-contact heart rate estimation". In: *Biomedical optics express* 9.2 (2018), pp. 873–897 (cit. on p. 44).

[Pra23] PRAHL, Scott: Extinction Coefficient for Hemoglobin. https://omlc.org/spectra/hemoglobin/summary.html. Accessed: 2023-07-28. 2023 (cit. on p. 27).

[Pu13] PU, Qifan; GUPTA, Sidhant; GOLLAKOTA, Shyamnath and PATEL, Shwetak: "Whole-home gesture recognition using wireless signals". In: *Proceedings of the 19th annual international conference on Mobile computing & networking*. 2013, pp. 27–38 (cit. on p. 38).

[Pu15] PU, Qifan; GUPTA, Sidhant; GOLLAKOTA, Shyamnath and PATEL, Shwetak: "Gesture recognition using wireless signals". In: *GetMobile: Mobile Computing and Communications* 18.4 (2015), pp. 15–18 (cit. on p. 38).

[Pud19] PUDANE, Mara; PETROVICA, Sintija; LAVENDELIS, Egons and EKENEL, Hazım Kemal: "Towards truly affective AAL systems". In: *Enhanced Living Environments: Algorithms, Architectures, Platforms, and Systems* (2019), pp. 152–176 (cit. on p. 3).

[Qin19] QING, Chunmei; QIAO, Rui; XU, Xiangmin and CHENG, Yongqiang: "Interpretable emotion recognition using EEG signals". In: *Ieee Access* 7 (2019), pp. 94160–94170 (cit. on pp. 4, 34).

[Qiu18] QIU, Jie-Lin; LI, Xiao-Yu and HU, Kai: "Correlated attention networks for multimodal emotion recognition". In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 2656–2660 (cit. on p. 35).

[Reh20] REHOUMA, Haythem; NOUMEIR, Rita; ESSOURI, Sandrine and JOUVET, Philippe: "Advancements in methods and camera-based sensors for the quantification of respiration". In: *Sensors* 20.24 (2020), p. 7252 (cit. on p. 38).

[Ric03] RICHARDSON, Marion: "Understanding the structure and function of the skin." In: *Nursing times* 99.31 (2003), pp. 46–48 (cit. on p. 22).

[Rob03] Robnik-Šikonja, Marko and Kononenko, Igor: “Theoretical and empirical analysis of ReliefF and RReliefF”. In: *Machine learning* 53 (2003), pp. 23–69 (cit. on p. 134).

[Rus79] Russell, James A: “Affective space is bipolar.” In: *Journal of personality and social psychology* 37.3 (1979), p. 345 (cit. on p. 9).

[Sab21] Sabour, Rita Meziati; Benezeth, Yannick; De Oliveira, Pierre; Chappe, Julien and Yang, Fan: “Ubfc-phys: A multimodal database for psychophysiological studies of social stress”. In: *IEEE Transactions on Affective Computing* (2021) (cit. on pp. 47, 112).

[Sag15] Sagonas, Christos; Panagakis, Yannis; Zafeiriou, Stefanos and Pantic, Maja: “Robust Statistical Face Frontalization”. In: *2015 IEEE International Conference on Computer Vision (ICCV).* 2015, pp. 3871–3879 (cit. on p. 86).

[San10] Sanches, Pedro; Höök, Kristina; Vaara, Elsa; Weymann, Claus; Bylund, Markus; Ferreira, Pedro; Peira, Nathalie and Sjölinder, Marie: “Mind the body! Designing a mobile stress management application encouraging personal reflection”. In: *Proceedings of the 8th ACM conference on designing interactive systems.* 2010, pp. 47–56 (cit. on p. 11).

[San17] Sancarlo, Daniele; D’ Onofrio, Grazia; Oscar, James; Ricciardi, Francesco; Casey, Dympna; Murphy, Keith; Giuliani, Francesco and Greco, Antonio: “Mario project: A multicenter survey about companion robot acceptability in caregivers of patients with dementia”. In: *Ambient Assisted Living: Italian Forum 2016 7*. Springer. 2017, pp. 311–336 (cit. on p. 3).

[Sch02] Schimmack, Ulrich and Rainer, Reisenzein: “Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation.” In: *Emotion* 2.4 (2002), p. 412 (cit. on p. 11).

[Sch19] Schmidt, Philip; Reiss, Attila; Dürichen, Robert and Van Laerhoven, Kristof: “Wearable-based affect recognition—A review”. In: *Sensors* 19.19 (2019), p. 4079 (cit. on p. 11).

[Sch20] Schätz, Martin; Procházka, Aleš; Kuchyňka, Jiří and Vyšata, Oldřich: “Sleep apnea detection with polysomnography and depth sensors”. In: *Sensors* 20.5 (2020), p. 1360 (cit. on p. 38).

[Sel36] Selye, Hans: “A syndrome produced by diverse nocuous agents”. In: *Nature* 138.3479 (1936), pp. 32–32 (cit. on p. 10).

[Sel76] Selye, Hans: Stress without distress. Springer, 1976 (cit. on p. 10).

[Sha12] Shastri, Dvijesh; Papadakis, Manos; Tsiamyrtzis, Panagiotis; Bass, Barbara and Pavlidis, Ioannis: “Perinasal imaging of physiological stress and its affective potential”. In: *IEEE Transactions on Affective Computing* 3.3 (2012), pp. 366–378 (cit. on pp. 2, 39).

[Sha17] Shaffer, Fred and Ginsberg, Jay P: “An overview of heart rate variability metrics and norms”. In: *Frontiers in public health* (2017), p. 258 (cit. on pp. 19, 20).

[Sha85] Shafer, Steven A: “Using color to separate reflection components”. In: *Color Research & Application* 10.4 (1985), pp. 210–218 (cit. on p. 27).

[Shu19] Shukla, Jainendra; Barreda-Angeles, Miguel; Oliver, Joan; Nandi, Gora Chand and Puig, Domenec: “Feature extraction and selection for emotion recognition from electrodermal activity”. In: *IEEE Transactions on Affective Computing* 12.4 (2019), pp. 857–869 (cit. on pp. 5, 34).

[Sim13] Simonyan, Karen; Vedaldi, Andrea and Zisserman, Andrew: “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013) (cit. on p. 119).

[Sim14] Simonyan, Karen and Zisserman, Andrew: “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 46, 125).

[Spe18] Spetlík, Radim; Franc, Vojtech; Cech, Jan and Matas, Jiri: “Visual Heart Rate Estimation with Convolutional Neural Network.” In: *BMVC*. 2018, p. 84 (cit. on pp. 46, 84).

[Ste08] Stein, Phyllis K; Barzilay, Joshua I; Chaves, Paulo HM; Traber, Jennifer; Domitrovich, Peter P; Heckbert, Susan R and Gottdiener, John S: “Higher levels of inflammation factors and greater insulin resistance are independently associated with higher heart rate and lower heart rate variability in normoglycemic older individuals: the Cardiovascular Health Study”. In: *Journal of the American Geriatrics Society* 56.2 (2008), pp. 315–321 (cit. on p. 20).

[Sto20] Stock, Simon C; Armengol-Urpi, Alexandre; Kovács, Bálint; Maier, Heiko; Gerdes, Marius; Stork, Wilhelm and Sarma, Sanjay E: “A system approach for closed-loop assessment of neuro-visual function based on convolutional neural network analysis of EEG signals”. In: *Neurophotonics*. Vol. 11360. SPIE. 2020, pp. 25–42 (cit. on p. 105).

[Str14] Stricker, Ronny; Müller, Steffen and Gross, Horst-Michael: “Non-contact video-based pulse rate measurement on a mobile service robot”. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2014, pp. 1056–1062 (cit. on pp. 82, 85, 94).

[Sub19] Subasi, Abdulhamit: “Chapter 1 - Introduction and Background”. In: *Practical Guide for Biomedical Signals Analysis Using Machine Learning Techniques*. Ed. by Subasi, Abdulhamit. Academic Press, 2019, pp. 1–26. doi: https://doi.org/10.1016/B978-0-12-817444-9.00001-5. url: https://www.sciencedirect.com/science/article/pii/B9780128174449000015 (cit. on p. 14).

[Sub21] Subasi, Abdulhamit; Tuncer, Turker; Dogan, Sengul; Tanko, Dahiru and Sakoglu, Unal: “EEG-based emotion recognition using tunable Q wavelet transform and rotation forest ensemble classifier”. In: *Biomedical Signal Processing and Control* 68 (2021), p. 102648 (cit. on pp. 4, 34).

[Suz21] Suzuki, Kei; Laohakangvalvit, Tipporn; Matsubara, Ryota and Sugaya, Midori: “Constructing an emotion estimation model based on eeg/hrv indexes using feature extraction and feature selection algorithms”. In: *Sensors* 21.9 (2021), p. 2910 (cit. on pp. 4, 34, 132).

[Tak07] Takano, Chihiro and Ohta, Yuji: “Heart rate measurement based on a time-lapse image”. In: *Medical engineering & physics* 29.8 (2007), pp. 853–857 (cit. on p. 40).

[Tha90] Thayer, Robert E: The biopsychology of mood and arousal. Oxford University Press, 1990 (cit. on p. 11).

[Tra17] Tran, Luan; Yin, Xi and Liu, Xiaoming: “Disentangled representation learning gan for pose-invariant face recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1415–1424 (cit. on p. 86).

[Tru17] Trumpp, Alexander; Bauer, Philipp L; Rasche, Stefan; Malberg, Hagen and Zaunseder, Sebastian: “The value of polarization in camera-based photoplethysmography”. In: *Biomedical Optics Express* 8.6 (2017), pp. 2822–2834 (cit. on p. 41).

[Tzi17] Tzimiropoulos, Georgios and Pantic, Maja: “Fast Algorithms for Fitting Active Appearance Models to Unconstrained Images”. In: *International Journal of Computer Vision* 122.1 (2017), pp. 17–33 (cit. on p. 86).

[Udo17] Udovičić, Goran; Đerek, Jurica; Russo, Mladen and Sikora, Marjan: “Wearable emotion recognition system based on GSR and PPG signals”. In: *Proceedings of the 2nd international workshop on multimedia for personal health and health care*. 2017, pp. 53–59 (cit. on pp. 5, 34).

[Van16] Van Gastel, Mark; Stuijk, Sander and Haan, Gerard de: “Robust respiration detection from remote photoplethysmography”. In: *Biomedical optics express* 7.12 (2016), pp. 4941–4957 (cit. on p. 79).

[Ver08] Verkruysse, Wim; Svaasand, Lars O and Nelson, J Stuart: "Remote plethysmographic imaging using ambient light." In: *Optics express* 16.26 (2008), pp. 21434–21445 (cit. on pp. 40, 41, 44).

[Ves18] Vest, Adriana N; Da Poian, Giulia; Li, Qiao; Liu, Chengyu; Nemati, Shamim; Shah, Amit J and Clifford, Gari D: "An open source benchmarked toolbox for cardiovascular waveform and interval analysis". In: *Physiological measurement* 39.10 (2018), p. 105004 (cit. on pp. 107, 131).

[Vio01] Viola, Paul and Jones, Michael: "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001.* Vol. 1. Ieee. 2001, pp. I–I (cit. on pp. 42, 83).

[Wan14] Wang, Wenjin; Stuijk, Sander and De Haan, Gerard: "Exploiting spatial redundancy of image sensor for motion robust rPPG". In: *IEEE transactions on Biomedical Engineering* 62.2 (2014), pp. 415–425 (cit. on pp. 43, 58, 83).

[Wan15] Wang, Wenjin; Stuijk, Sander and De Haan, Gerard: "A novel algorithm for remote photoplethysmography: Spatial subspace rotation". In: *IEEE transactions on biomedical engineering* 63.9 (2015), pp. 1974–1984 (cit. on p. 97).

[Wan16a] Wang, Hao; Zhang, Daqing; Wang, Yasha; Ma, Junyi; Wang, Yuxiang and Li, Shengjie: "RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices". In: *IEEE Transactions on Mobile Computing* 16.2 (2016), pp. 511–526 (cit. on p. 38).

[Wan16b] Wang, Wenjin; Den Brinker, Albertus C; Stuijk, Sander and De Haan, Gerard: "Algorithmic principles of remote PPG". In: *IEEE Transactions on Biomedical Engineering* 64.7 (2016), pp. 1479–1491 (cit. on pp. 2, 29, 42, 45, 63).

[Wan17a] Wang, Pei; Guo, Bin; Xin, Tong; Wang, Zhu and Yu, Zhiwen: "TinySense: Multi-user respiration detection using Wi-Fi CSI signals". In: *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE. 2017, pp. 1–6 (cit. on pp. 2, 38).

[Wan17b] Wang, W.; Brinker, A.C. den; Stuijk, S. and Haan, G. de: "Amplitude-selective filtering for remote-PPG". In: *Biomedical Optics Express* 8.3 (2017), pp. 1965–1980 (cit. on p. 44).

[Wan17c] Wang, Xuyu; Yang, Chao and Mao, Shiwen: "PhaseBeat: Exploiting CSI phase data for vital sign monitoring with commodity WiFi devices". In: *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2017, pp. 1230–1239 (cit. on p. 38).

[Wan17d] Wang, Xuyu; Yang, Chao and Mao, Shiwen: "TensorBeat: Tensor decomposition for monitoring multiperson breathing beats with commodity WiFi". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.1 (2017), pp. 1–27 (cit. on p. 38).

[Wan17e] Wang, Zhiguang; Yan, Weizhong and Oates, Tim: "Time series classification from scratch with deep neural networks: A strong baseline". In: *2017 International joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 1578–1585 (cit. on p. 113).

[Wan18a] Wang, Jingyuan; Wang, Ze; Li, Jianfeng and Wu, Junjie: "Multilevel wavelet decomposition network for interpretable time series analysis". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2437–2446 (cit. on p. 113).

[Wan18b] Wang, Tianben; Zhang, Daqing; Zheng, Yuanqing; Gu, Tao; Zhou, Xingshe and Dorizzi, Bernadette: "C-FMCW based contactless respiration detection using acoustic signal". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (2018), pp. 1–20 (cit. on pp. 2, 38).

[Wan19] Wang, Wenjin; Brinker, Albertus C den and De Haan, Gerard: “Discriminative signatures for remote-PPG”. In: *IEEE Transactions on Biomedical Engineering* 67.5 (2019), pp. 1462–1473 (cit. on p. 45).

[Wan20] Wang, Ruochen: “Development of Attention-based Neural Network for Facial Expression Recognition”. Master’s Thesis. Karlsruhe Institute of Technology, 2020 (cit. on p. 126).

[Wan21] Wang, Anran; Nguyen, Dan; Sridhar, Arun R and Gollakota, Shyamnath: “Using smart speakers to contactlessly monitor heart rhythms”. In: *Communications biology* 4.1 (2021), pp. 1–12 (cit. on pp. 2, 38).

[Was21] Wascher, Claudia AF: “Heart rate as a measure of emotional arousal in evolutionary biology”. In: *Philosophical Transactions of the Royal Society B* 376.1831 (2021), p. 20200479 (cit. on p. 16).

[Wei22] Weimar, Sascha: “Development of a System to Support the Diagnosis of Dementia based on Digital Biomarkers”. Master’s Thesis. Karlsruhe Institute of Technology, 2022 (cit. on p. 105).

[Woo18] Woo, Sanghyun; Park, Jongchan; Lee, Joon-Young and Kweon, In So: “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV).* 2018, pp. 3–19 (cit. on p. 125).

[Wu00] Wu, Ting; Blazek, Vladimir and Schmitt, Hans Juergen: “Photoplethysmography imaging: a new noninvasive and noncontact method for mapping of the dermal perfusion changes”. In: *Optical techniques and instrumentation for the measurement of blood composition, structure, and dynamics.* Vol. 4163. SPIE. 2000, pp. 62–70 (cit. on p. 40).

[Wu17] Wu, Bing-Fei; Chu, Yun-Wei; Huang, Po-Wei; Chung, Meng-Liang and Lin, Tzu-Min: “A motion robust remote-PPG approach to driver’ s health state monitoring”. In: *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13.* Springer. 2017, pp. 463–476 (cit. on p. 2).

[Xio13] Xiong, Xuehan and De la Torre, Fernando: “Supervised descent method and its applications to face alignment”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 532–539 (cit. on p. 43).

[Xu19] Xu, Qinyi; Han, Yi; Wang, Beibei; Wu, Min and Liu, KJ Ray: “Indoor events monitoring using channel state information time series”. In: *IEEE Internet of Things Journal* 6.3 (2019), pp. 4977–4990 (cit. on p. 38).

[Yan18a] Yang, Chao; Wang, Xuyu and Mao, Shiwen: “AutoTag: Recurrent variational autoencoder for unsupervised apnea detection with RFID tags”. In: *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2018, pp. 1–7 (cit. on p. 37).

[Yan18b] Yang, Jianfei; Zou, Han; Jiang, Hao and Xie, Lihua: “Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes”. In: *IEEE Internet of Things Journal* 5.5 (2018), pp. 3991–4002 (cit. on p. 38).

[Yan18c] Yang, Zhicheng; Bocca, Maurizio; Jain, Vivek and Mohapatra, Prasant: “Contactless breathing rate monitoring in vehicle using UWB radar”. In: *Proceedings of the 7th international workshop on real-world embedded wireless systems and networks*. 2018, pp. 13–18 (cit. on p. 37).

[Yan19] Yang, Chao; Wang, Xuyu and Mao, Shiwen: “RFID-based driving fatigue detection”. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2019, pp. 1–6 (cit. on p. 37).

[Yao16] Yao, Anbang; Cai, Dongqi; Hu, Ping; Wang, Shandong; Sha, Liang and Chen, Yurong: “HoloNet: towards robust emotion recognition in the wild”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 472–478 (cit. on p. 86).

[Yin17] Yin, Xi; Yu, Xiang; Sohn, Kihyuk; Liu, Xiaoming and Chandraker, Manmohan: “Towards large-pose face frontalization in the wild”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3990–3999 (cit. on p. 86).

[Yu13] Yu, Xiang; Huang, Junzhou; Zhang, Shaoting; Yan, Wang and Metaxas, Dimitris N: "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model". In: *Proceedings of the IEEE international conference on computer vision.* 2013, pp. 1944–1951 (cit. on p. 43).

[Yu19a] Yu, Zitong; Li, Xiaobai and Zhao, Guoying: "Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks." In: *BMVC.* 2019, p. 277 (cit. on pp. 46, 58, 68, 84).

[Yu19b] Yu, Zitong; Peng, Wei; Li, Xiaobai; Hong, Xiaopeng and Zhao, Guoying: "Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* 2019, pp. 151–160 (cit. on pp. 46, 84).

[Yu21] Yu, Zitong; Li, Xiaobai and Zhao, Guoying: "Facial-video-based physiological signal measurement: Recent advances and affective applications". In: *IEEE Signal Processing Magazine* 38.6 (2021), pp. 50–58 (cit. on p. 39).

[Zau18] Zaunseder, Sebastian; Trumpp, Alexander; Wedekind, Daniel and Malberg, Hagen: "Cardiovascular assessment by imaging photoplethysmography–a review". In: *Biomedical Engineering/Biomedizinische Technik* 63.5 (2018), pp. 617–634 (cit. on pp. 39, 45).

[Zha17] Zhang, Jin; Xu, Weitao; Hu, Wen and Kanhere, Salil S: "Wicare: Towards in-situ breath monitoring". In: *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.* 2017, pp. 126–135 (cit. on p. 38).

[Zha18] Zhao, Heng; Gu, Xu; Hong, Hong; Li, Yusheng; Zhu, Xiaohua and Li, Changzhi: "Non-contact beat-to-beat blood pressure measurement using continuous wave Doppler radar". In: *2018*

IEEE/MTT-S International Microwave Symposium-IMS. IEEE. 2018, pp. 1413–1415 (cit. on pp. 2, 37).

[Zha19] Zhang, Dongheng; Hu, Yang; Chen, Yan and Zeng, Bing: “BreathTrack: Tracking indoor human breath status via commodity WiFi”. In: *IEEE Internet of Things Journal* 6.2 (2019), pp. 3899–3911 (cit. on p. 38).

[Zha20a] Zhan, Qi; Wang, Wenjin and Haan, Gerard de: “Analysis of CNN-based remote-PPG to understand limitations and sensitivities”. In: *Biomedical optics express* 11.3 (2020), pp. 1268–1283 (cit. on p. 76).

[Zha20b] Zhang, Hongli: “Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder”. In: *IEEE Access* 8 (2020), pp. 164130–164143 (cit. on pp. 4, 34).

[Zha21] Zhang, Chiyuan; Bengio, Samy; Hardt, Moritz; Recht, Benjamin and Vinyals, Oriol: “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115 (cit. on p. 113).

[Zho20] Zhou, Kai; Krause, Simon; Blöcher, Timon and Stork, Wilhelm: “Enhancing Remote-PPG Pulse Extraction in Disturbance Scenarios Utilizing Spectral Characteristics”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 280–281 (cit. on p. 63).

[Zho21] Zhou, Kai; Krause, Simon; Blöcher, Timon and Stork, Wilhelm: “Short Window Network for Remote Heart Rate Measurement”. In: *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2021, pp. 200–208 (cit. on pp. 57, 131).

[Zho22] Zhou, Kai; Schinle, Markus; Weimar, Sascha; Gerdes, Marius; Stock, Simon and Stork, Wilhelm: “End-to-End Deep Learning for Stress Recognition Using Remote Photoplethysmography”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 1435–1442 (cit. on p. 104).

[Zho23] Zhou, Kai; Schinle, Markus and Stork, Wilhelm: “Dimensional emotion recognition from camera-based PRV features”. In: *Methods* 218 (2023), pp. 224–232 (cit. on p. 123).

[Zhu16] Zhu, Xiangyu; Lei, Zhen; Liu, Xiaoming; Shi, Hailin and Li, Stan Z: “Face alignment across large poses: A 3d solution”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 146–155 (cit. on p. 87).

[Zhu17] Zhu, Xiangyu; Liu, Xiaoming; Lei, Zhen and Li, Stan Z: “Face alignment in full pose range: A 3d total solution”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 78–92 (cit. on p. 86).

Own publications

This section contains a complete list of own publications. The publications [4], [5] address the subject of the rPPG algorithm development, while the publications [6], [7] deal with the topic stress and emotion recognition. Other listed publications are considered relevant for the presented work in a broader sense.

[1] Blöcher, Timon; Zhou, Kai; Krause, Simon and Stork, Wilhelm: "An Adaptive Bandpass Filter Based on Temporal Spectrogram Analysis for Photoplethysmography Imaging". In: *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2018, pp. 1–6.

[2] Blöcher, Timon; Krause, Simon; Zhou, Kai; Zeilfelder, Jennifer and Stork, Wilhelm: "VitalCamSet - a dataset for Photoplethysmography Imaging". In: *2019 IEEE Sensors Applications Symposium (SAS)*. 2019, pp. 1–6.

[3] Busch, Tobias; Zeilfelder, Jennifer; Zhou, Kai and Stork, Wilhelm: “A jaw based human-machine interface with machine learning”. In: *2019 IEEE Sensors Applications Symposium (SAS)*. IEEE. 2019, pp. 1–6.

[4] Zhou, Kai; Krause, Simon; Blöcher, Timon and Stork, Wilhelm: “Enhancing Remote-PPG Pulse Extraction in Disturbance Scenarios Utilizing Spectral Characteristics”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.* 2020, pp. 280–281.

[5] Zhou, Kai; Krause, Simon; Blöcher, Timon and Stork, Wilhelm: “Short Window Network for Remote Heart Rate Measurement”. In: *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2021, pp. 200–208.

[6] Zhou, Kai; Schinle, Markus; Weimar, Sascha; Gerdes, Marius; Stock, Simon and Stork, Wilhelm: “End-to-End Deep Learning for Stress Recognition Using Remote Photoplethysmography”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 1435–1442.

[7] Zhou, Kai; Schinle, Markus and Stork, Wilhelm: “Dimensional emotion recognition from camera-based PRV features”. In: *Methods* 218 (2023), pp. 224–232.

Cited student theses

[1] Ardinski, Tatjana: “Ertellen einer Datenbasis für die Emotions und Stressanalyse auf Basis der kamerabasierten Vitalparametermessung”. Master’s Thesis. Karlsruhe Institute of Technology, 2019.

[2] Wang, Ruochen: “Development of Attention-based Neural Network for Facial Expression Recognition”. Master’s Thesis. Karlsruhe Institute of Technology, 2020.

[3] Geng, Yexing: “Einsatz des Short-Window Network Verfahrens für kontaktlose Atemfrequenzmessung”. Bachelor’s Thesis. Karlsruhe University of Applied Sciences, 2022.

[4] Weimar, Sascha: “Development of a System to Support the Diagnosis of Dementia based on Digital Biomarkers”. Master’s Thesis. Karlsruhe Institute of Technology, 2022.

List of Figures

List of Tables

Acronyms

3DMM	3D Morphable Model
ANS	Autonomic Nervous System
AR	Autoregressive
AVNN	Average Interval between Normal Heart Beats
BP	Blood Pressure
bpm	beats per minutes
BSS	Blind Source Separation
BVP	Blood Volume Pulse
CAN	Convolutional Attention Network

CBAM	Convolutional Block Attention Module
CCC	Concordance Correlation Coefficient
CHROM	Chrominance-based method
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network with LSTM
CPT	Continuous Performance Task
ECG	Electrocardiography
EDA	Electrodermal activity
EEG	Electroencephalogram
EMG	Electromyogram
FCN	Fully Convolutional Network
FER	Facial Expression Recognition
FFT	Fast Fourier Transformation
FMCW	Frequency Modulated Continuous Wave
fps	frames per second
GRU	Gated Recurrent Unit
GSR	Galvanic Skin Response
HAHV	High-Arousal High-Valence
HF	High Frequency
HMI	human-machine interface
HR	Heart Rate

HRV	Heart Rate Variability
IBI	Inter-Beat Interval
ICA	Independent Component Analysis
LALV	Low-Arousal Low-Valence
LF	Low Frequency
LSTM	Long Short-Term memory
MAE	Mean Absolute Error
MCDCNN	Multichannel Deep Convolutional Neural Network
MLP	Multi-Layer Perception
MLP-LSTM	Multi-Layer Perception with LSTM
NIR	Near-Infrared
n.u.	normal unit
PCA	Principal Component Analysis
pNN50	Percentage of Successive Differences Greater than 50 ms
PNS	parasympathetic nervous system
POS	Plane Orthogonal to Skin
PPG	Photoplethysmography
PR	Pulse Rate
PRV	Pulse Rate Variability
PSC	Projection using Spectral Characteristics
RESNET	Residual Network

RFID	Radio-Frequency Identification
RMSE	Root Mean Square Error
RMSSD	Root Mean Sum of Squared Successive Distance
ROI	Region of Interest
rPPG	remote Photoplethysmography
RR	Respiratory Rate
RSA	Respiratory sinus arrhythmia
SART	Sustained Attention Response Task
SDNN	Standard Deviation of the IBI of Normal Sinus Beats
SNR	Signal-to-Noise Ratio
SNS	sympathetic nervous system
SpO2	blood oxygen saturation
SR	Success Rate
STRESNET	Spectrotemporal Residual Network
SVM	Support Vector Machine
SWN	Short Window Network
ULF	Ultra Low Frequency
VLF	Very Low Frequency