

Stresserkennung mit Hilfe von Gesichtsausdrücken aus Videosequenzen

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Paul Angerer, BSc.

Matrikelnummer 01126249

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: PD DI Dr. Martin Kempel

Wien, 5. Oktober 2018

Paul Angerer

Martin Kempel

Stress Detection Using Facial Expressions from Video Sequences

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Medical Informatics

by

Paul Angerer, BSc.

Registration Number 01126249

to the Faculty of Informatics
at the TU Wien

Advisor: PD DI Dr. Martin Kempel

Vienna, 5th October, 2018

Paul Angerer

Martin Kempel

Erklärung zur Verfassung der Arbeit

Paul Angerer, BSc.
Keilgasse 9/22, 1030 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 5. Oktober 2018

Paul Angerer

Kurzfassung

Stress ist Teil des alltäglichen Lebens und beeinflusst die persönliche Gesundheit und das Wohlergehen in ungünstigen emotionalen Zuständen wie innere Unruhe, Furcht oder Zorn. Chronischer und unbehandelter Stress kann zu unheilbaren Krankheiten, Beziehungsverschlechterungen sowie hohen ökonomischen Kosten führen. Unter dem Begriff *Stress* versteht man nach Hans Selye „Die unspezifische Antwort des Körpers auf jede Anforderung nach Veränderung“. Diese unspezifische Antwort erschwert die Quantifizierung von Stress. Stressforschung entwickelte verschiedene computerunterstützte Techniken zur Erkennung von Stress für Vorteile in vielen Bereichen. Hauptsächlich physiologische Parameter werden verwendet um Stress zu diagnostizieren. Die Sammlung dieser Gesundheitsdaten mit Hilfe von Kontaktsensoren kann unpraktisch, besonders für kontinuierliche Stresserkennung, sein. Kürzliche Studien versuchten dieses Problem basierend auf nicht-invasiven Visual Computing-Techniken zu lösen. Die Erkennung von Stress basierend auf Modellen von Gesichtsausdrücken zeigt vielversprechende Resultate.

Diese Diplomarbeit stellt eine Lösung zur Erkennung von Stress basierend auf Gesichtsausdrücken vor. Das entwickelte System ermittelt Stress basierend auf extrahierten Gesichtsmerkmalen von Videos. Stress assoziierte Gesichtsmerkmale von Kopf, Augen und Mund sowie die Herzrate von Gesichtsphtoplethysmographie werden verwendet um Stress mittels Machine Learning Algorithmen zu erkennen. Eine Performance Evaluierung der entwickelten Lösung mit Vergleich von existierenden Herangehensweisen in der Literature sowie an annotierten Daten wird durchgeführt. Des Weiteren werden Anforderungen an Gesichtsbilddaten sowie die Einschränkungen der umgesetzten Lösung diskutiert.

Abstract

Stress is a part of everyday life and impacts a person's health and well-being in less favorable emotional states such as anxiety, fear or anger. Chronic and left untreated stress can lead to incurable diseases, relationship deterioration and high economic costs. Under the term *stress* "the non-specific response of the body to any demand for change", defined by Hans Selye, is understood. This non-specific response makes it difficult to quantify stress conditions. Stress research developed distinct computational techniques to recognize stress for benefits in a wide range of fields. Mainly physiological parameters are used to diagnose stress. Gathering this health data with the help of contact sensors can be impractically, especially when monitoring a person continuously. Recent studies try to solve this problem based on non-invasive visual computing techniques. The detection of stress based on facial expression models offers promising results.

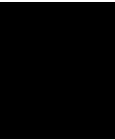
This thesis introduces a solution to detect stress based on a subject's facial expressions. The developed system determines human stress based on extracted facial features from video data. Stress associated facial cues from head, eye, mouth as well as the heart rate from facial photoplethysmography are used to predict stress by machine learning algorithms. A performance evaluation of the developed solution with comparison of existing approaches in literature as well as on annotated data is conducted. Moreover, requirements on facial image data as well as limitations of the implemented system are discussed.

Contents

Kurzfassung	vii
Abstract	ix
Contents	xi
1 Introduction	1
1.0.1 Motivation	2
1.0.2 Aim	3
1.0.3 Contributions	3
1.0.4 Structure	4
2 Related Work	5
2.1 Definition of Stress	5
2.2 Stress Detection Using RGB Imaging	7
2.2.1 Photoplethysmography-based Stress Detection	8
2.2.2 Model-based Stress Detection	14
2.3 Stress Detection Using Thermal and Hyperspectral Imaging	21
2.3.1 Thermal Imaging-based Stress Detection	22
2.3.2 Hyperspectral Imaging-based Stress Detection	28
2.4 Comparison and Limitations	30
3 Stress Data	35
3.1 Data Quality	35
3.2 Data Acquisition	37
3.2.1 Specifications	37
3.2.2 Data Modality	40
3.2.3 Image Resolution	41
3.2.4 Robustness	43
3.3 Databases	44
4 Methodology	47
4.1 Experiment Protocol	48
4.1.1 Experiment Setup	48

xi

4.1.2	Experiment Procedure	48
4.1.3	Stress Phase	49
4.2	Datasets	50
4.3	Data Preprocessing	50
4.4	Face Detection	51
4.4.1	Histogram of oriented gradients	52
4.4.2	Deformable Part Models	53
4.5	Facial Landmark Detection	57
4.6	Eye Related Features	58
4.7	Mouth Related Features	59
4.7.1	Optical Flow	60
4.8	Head Related Features	62
4.9	Heart Rate from Facial Video	63
4.10	Machine Learning	65
4.11	Implementation	65
5	Results & Discussion	67
5.1	Eye Related Features	67
5.2	Mouth Related Features	69
5.3	Head Related Features	72
5.4	Heart Rate from Facial Video	74
5.5	Classification Performance	78
6	Conclusion	83
	List of Figures	85
	List of Tables	89
	Acronyms	91
	Bibliography	95



Introduction

Stress plays a common role in the person's subjective quality of life [1]. When talking about stress, it is associated with the *stressor*. A stressor is an actual or perceived threat to an organism, which then results in the response to the stressor, as the *stress response*. Based on this mechanism, humans try to cope with stressors through different coping responses [2, 3]. The interaction with stress can be either in a positive or a negative way and plays a crucial role [4]. Negative coping with stress shows to significantly alter the human immune responses and leads to a reduction of mental and physical tolerance to diseases [5].

This significant connection between body and mind takes an active part in the manifestation of stress. Physiological diseases as gastrointestinal distress, heart disease and cancer are linked to unresolved lifestyle stresses. Further, it is shown that stressful events can influence longevity through cardiovascular diseases, immunological disorders and consequences of normal aging. [6, 7, 5]

Besides individual negative health effects, stress can have an impact on human relationships [8, 9] as well as on economic costs [10]. The known term "burnout" describes emotional and mental exhaustion which is associated with work stress [11, 12]. Especially in Western countries and in Japan burnout significantly impacts modern society [13].

In order to help individuals, stress conditions and associated issues have to be identified [14, 15]. Distinct stress responses such as body or behavioral signals are used to measure stressful events [16, 17]. The different reactions to stress, due to experience, age, gender etc., results in variable stress responses [18]. Computational techniques such as machine learning (ML) methods, enable to model these stress conditions, despite its complex nature [19, 20]. Furthermore, automated and continuous analysis of stress conditions as well as the extraction of non-invasive stress signals can be performed by computational approaches [18, 21, 17].

Table 1.1: Workers in the EU-27 reporting each individual symptom in percentage [26]

Symptom	
Backache	24.7
Muscular Pain	22.8
Fatigue	22.6
Stress	22.3
Headaches	15.5
Irritability	10.5
Injuries	9.7
Sleeping Problems	8.7
Anxiety	7.8
Eyesight problems	7.8
Hearing problems	7.2
Skin problems	6.6
Stomach ache	5.8
Breathing difficulties	4.8
Allergies	4.0
Heart disease	2.4
Other	1.6

1.0.1 Motivation

Stress at work is present in Western countries, ranking as fourth reported individual symptom impacting health in the European working conditions survey, as illustrated in Table 1.1 [22, 23, 24]. In Austria, a study by an official representation of employees assesses that 30% of all Austrian employees suffer from high mental stress with an increase of mental stress from 9% to 13% in the years from 2010 to 2012 [25].

This situation reveals the importance to interpret stress responses and recognize stress to diminish possible risks [23, 24]. In order to improve the quality of life and prevent diseases, the control and suppression of stress is of interest [5]. Research on stress shows benefits, such as increasing work productivity to benefiting the wider society or improving day-to-day activities [18]. The measurement of stress provides potential for psychological therapies [15], stress management and intervention [27, 28], human interaction and affective computing [19, 29] as well as for stress monitoring [18].

Stress can be determined through the interpretation of distinct signals of the body [18]. Computational analysis of these measured body signals allows to detect stress [30]. With the use of these methods a technology-based evaluation of stress conditions can be conducted without expert knowledge or self-reports [31, 32]. Techniques such as machine analysis enable to analyze variable stress signals [18]. Furthermore, visual computing methods allow to extract stress responses such as physiological signals [33], body language [17] or facial expressions [32]. Hence, this enables the remote assessment of stress conditions without the need of body contact sensors [19, 17, 34].

1.0.2 Aim

The aim of this thesis is to outline an approach to detect stress based on non-invasive visual computing techniques. Focus of the presented approach is set on the use of facial image/video data captured by digital cameras. The detection of stress from these facial data is of particular interest due to the association of facial signs and expressions with classification and analysis of stress [18, 35, 32]. Further, limitations of stress detection approaches utilizing body contact sensors and techniques, such as accelerometer, skin conductance, electrocardiogram, are addressed by the developed solution. Visual stress detection approaches employ techniques to derive physical stress signals without the need for equipment and tools [18]. The analysis of behavior [27], body language [17], facial features and expressions [32, 1] by visual computing techniques enables to detect stress remotely. Possible disadvantages from contact sensors, as obtrusiveness, limited mobility and dexterity, uncomfortable or unnatural feelings for the wearer, can be avoided with the utilization of a visual stress detection approach. Contact sensors measurements such as electroencephalography (EEG), electrocardiogram (ECG), galvanic skin response (GSR), energy expenditure (EE) etc., are in general tethered and require accurate placement as well as knowledge of application [36, 37].

The main part of this thesis is the development of a non-invasive solution for the detection of stress based on facial video data. A focus is set on visual extraction of stress associated facial features, such as eye aperture [38], blinking [39], heart activity [33], mouth [34] and head movements [40]. With the help of machine learning, a model estimating human stress from these facial features is implemented. Limitations and further improvements are discussed based on findings of facial stress signs and on results from the implemented machine analysis. An overall comparison with existing research on stress detection in the field of visual computing and annotated video data evaluates the performance of the developed solution.

Besides the stress detection implementation, the impact of data quality on stress analysis is discussed. Requirements concerning the recording of visual data, such as resolution, lighting, angle of view, can influence the performance of stress detection. To deal with this issue, specifications on data quality regarding visual algorithms are presented.

1.0.3 Contributions

The extraction of stress associated features enables the detection of stress conditions by computational methods [31]. Machine analysis is able to estimate stress based on derived features from body contact sensors as well as from visual sensors [30, 41]. Due to disadvantages of contact sensors, such as obtrusiveness [18], limited mobility and more [19], this thesis focuses on the detection of stress with the help of visual computing techniques. The following approaches are developed and outlined:

- A developed stress detection solution is presented, which enables to detect stress from video sequences. Hence, obtrusive contact sensor measurements for subsequent

stress detection are not required. Conditions of stress are identified in relation to a subject's non-stressed/neutral emotional state.

- A functional experiment protocol is developed to establish a facial stress database. A facial video dataset of subjects in stress and non-stressed emotional states is gathered and annotated.
- Requirements and technical specifications on data for stress and facial analysis are discussed.
- The feasibility and performance of stress detection is assessed on the conducted experiment data as well as on a second dataset. Results of machine analysis are compared with approaches in current literature. Moreover, feature selection highlights most contributing facial attributes for classification.
- A statistical analysis is applied on the extracted facial stress features. Further, extracted features from stressed and non-stressed emotional states are compared.

1.0.4 Structure

The structure of this thesis is the following. In Chapter 2 a definition of stress as well as features associated with stress are presented. Further, state-of-the-art stress detection approaches based on distinct imaging modalities and techniques are outlined. Introduced approaches using RGB imaging are sectioned into photoplethysmography (PPG)-based and model-based solutions. The remaining stress detection approaches include thermal imaging (TI) and hyperspectral imaging (HI)-based solutions.

Chapter 3 addresses stress and facial expression data. Data quality and data acquisition are discussed due to the importance for stress detection. Factors influencing data quality as well as face acquisition characteristics, such as specifications, data modality, image resolution and robustness, are taken into consideration. Additionally, a comparison of existing stress and facial expression databases is conducted.

In the methodology chapter 4 an experiment protocol for the acquisition of a stress dataset as well as a stress detection solution is introduced. The setup, procedure and stress inducing task for establishing this dataset is presented. Besides this experiment, the implementation of a stress detection solution based on facial features is described. These facial features characterize head, eye, mouth and heart rate from facial video. Moreover, preprocessing steps such as face and landmark detection are depicted.

The implemented stress detection solution is evaluated by two distinct datasets in Chapter 5. Outcomes of stress detection are presented and compared by classification performance. Additionally, the individual features utilized for stress analysis are discussed. The final chapter 6 summarizes findings of this thesis and possible limitations of the area of research.

Related Work

Stress detection is of interest in distinct research fields such as biology [42], psychology [43], neuroscience [44], medicine [45] or computer science [46]. The detection of stress can be performed with a range of methods, such as questionnaires, hormone measurements, physiological sensors, speech etc. [43, 18]. Obtaining these indicators of stress can require body contact sensors or expert knowledge [47, 48]. As illustrated in Figure 2.2, stress cues can be obtained using distinct measurement sources of the body [18]. Due to the required knowledge to recognize stress states, research strives for an automatic detection of stress by computational analysis [49, 18]. Furthermore, the non-intrusive remote detection of stress with the use of visual methods is preferable over body contact sensors in applications such as vehicle drivers, surgeons, pilots in flight, and more [18]. Current visual approaches focus on stress detection over facial images [50, 20]. Depending on the chosen approach, the optical stress detection solutions utilize RGB, thermal or hyperspectral imaging [19, 20, 51]. The majority of these solutions are designed similar to traditional automated facial expression analysis (AFEA) systems. AFEA systems comprise, as illustrated in Figure 2.1 of face acquisition, facial feature extraction and machine analysis [19, 20, 23]. In contrast to these AFEA systems, solutions implementing machine analysis methods such as deep learning do not require handcrafted features and differ from the common AFEA system [50]. Stress detection approaches using RGB imaging distinguish stress based on photoplethysmography (PPG) or models of the face [19, 20]. Thermal and hyperspectral imaging approaches detect stress based on distinct light radiation patterns of the skin [23, 51].

2.1 Definition of Stress

Stress can be defined in distinct ways [52, 53]. One well known definition is by Hans Selye who understood stress as "the non-specific response of the body to any demand for change" [2]. The term stress is linked with the trigger of stress conditions, the *stressor*,

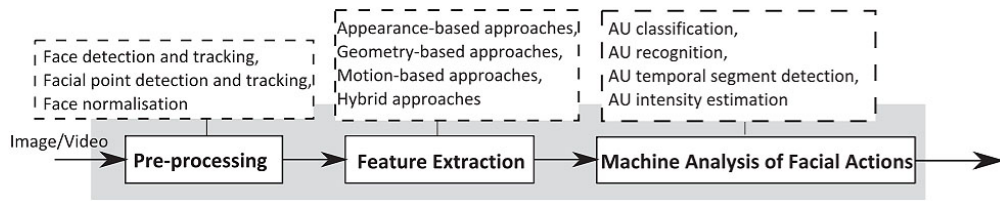


Figure 2.1: Automatic facial expression analysis steps of image/video data. Pre-Processing (Face acquisition) step, Feature Extraction based on distinct approaches, Machine Analysis of Facial Actions by Action Unit (AU, atomic facial muscle actions) processing. [49]

as well as with the answer to the stressor, the *stress response* [54]. When someone experiences stress, a temporarily physiological and/or psychological imbalance is caused by the stressor [4]. The stressor has direct effect on the body and can be categorized into physical/physiological or mental/emotional stressor [5, 54]. A physical/physiological stressor impacts the body over external environmental conditions (e.g. heat, cold, noise) or through internal demands on the body [54]. In contrast to the physical stressor, the mental/emotional stressor effects the cognitive systems (thought processes) or the emotional system through information input [54]. As Selye stated, due to the stressor, the demand for change, a non-specific response is triggered. This response is of interest in order to detect stress conditions and has been researched extensively [46]. Literature of interest for this thesis focuses on the negative stress or *distress*, which can be harmful and can cause negative consequences [46]. The detection of stress is based on the stress response [46]. These responses can be assessed through physiological measurements as well as from facial expressions (see Figure 2.2) [1]. Typical physiological stress responses include:

- sweat production, increased heart rate and muscle activation [55]
- faster respiration and increased blood pressure [56]
- changes in speech characteristic [56]
- skin temperature changes [31]
- decreased heart rate variability [57]
- varying pupil diameter [46]

In addition to the listed physiological responses, facial stress responses can be noticeable. Table 2.1 categorizes facial features associated with stress into head, eyes, mouth, gaze and pupil characteristics. Head movements have shown to be more frequent [31], more rapid [32] and there is greater head motion [40] when under stress. The eye region features such as eye aperture, blink rate, eyelid response, gaze distribution and variation as well as pupil size variation are researched. Blinking can be caused by internal and external stimulus and typically increases with stress and anxiety [58, 32]. Moreover,

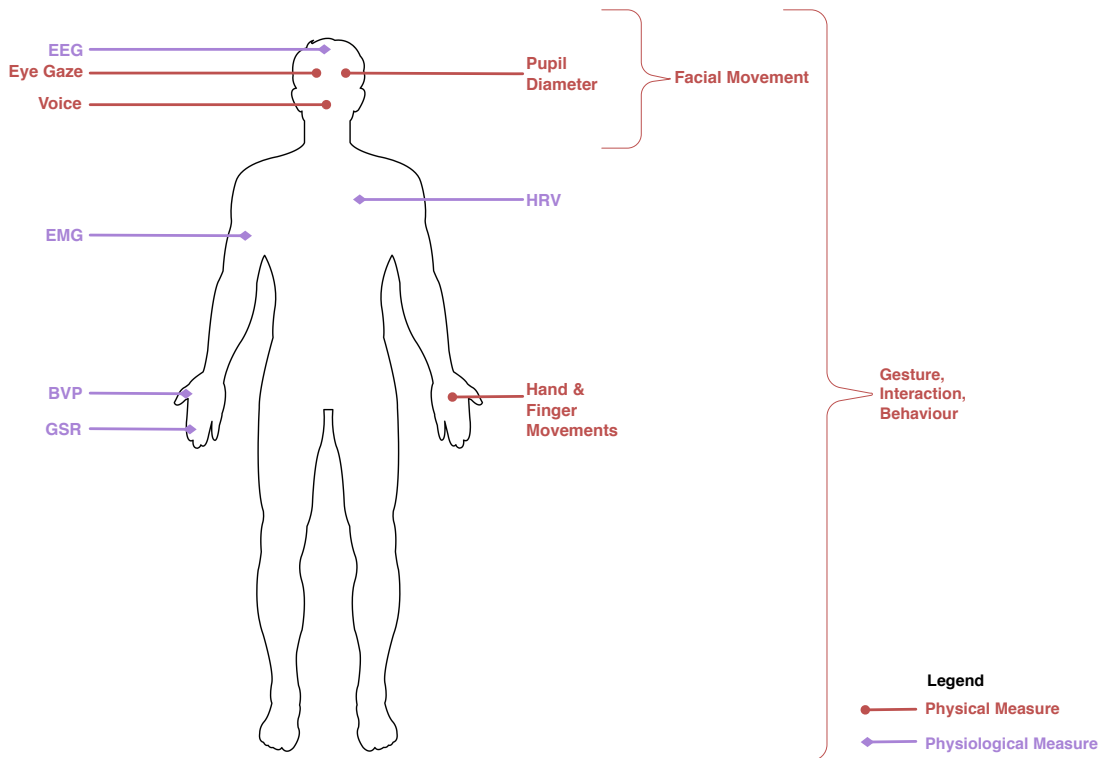


Figure 2.2: Typical physiological and physical measures utilized to detect stress. The usual measurement sources, electroencephalography (EEG), electromyography (EMG), heart rate variability (HRV), blood volume pulse (BVP), galvanic skin reponse (GSR), are shown in the figure. [18]

the gaze direction and gaze congruence changes depending on the level of stress [59]. Pupil size and ratio variations are associated with emotional, sexual or cognitive arousal and are used as stress and anxiety indicator [60]. Lip movements are linked to stress conditions [32], especially asymmetric movements characterize high stress levels [34].

2.2 Stress Detection Using RGB Imaging

Stress detection using RGB imaging as data source is widespread [15, 1]. RGB imaging produces images which can be perceived by the human eye. Video recordings or images for stress detection are mainly captured with the use of RGB cameras [15, 20]. This imaging data is then processed to recognize a person's stress state. As in the case of affective computing and facial action coding system (FACS) ¹ analysis, facial expressions are a major research topic for the detection of stress [49, 20]. In order to gain insight into facial expressions, model-based approaches representing a person's face are used. Primary

¹FACS; definition of 32 atomic facial muscle actions named action units (AUs) [49]

Table 2.1: Facial features associated with stress/anxiety categorized [1].

Head	Eyes	Mouth	Gaze	Pupil
Head movement	Blink rate	Mouth shape	Saccadic eye movements	Pupil size variation
Skin color	Eyelid response	Lip deformation	Gaze spatial distribution	Pupil ratio variation
Heart rate (facial PPG)	Eye aperture	Lip corner puller/depressor	Gaze direction	
	Eyebrow movements	Lip pressor		

facial features, including movements of eyes, lips, eyebrows and head are extracted from these models for subsequent machine analysis [1, 20]. Besides model-based approaches utilizing facial features, photoplethysmography (PPG)-based stress detection is conducted in literature [19, 15]. With the advancements in PPG it is possible to derive a person’s blood volume pulse (BVP) from RGB images [21]. This information can then be used to detect stress states. Both model-based and photoplethysmography-based approaches show promising results in the detection of stress [33, 1]. Depending on the evaluation method, a high agreement with ground truth measurements or a classification accuracy up to 91.68% for machine learning algorithms can be achieved [15, 1].

2.2.1 Photoplethysmography-based Stress Detection

The optical measurement technique for blood volume changes, photoplethysmography (PPG), is used in clinical applications. The use of PPG requires a light source to illuminate a tissue (e.g. skin) and a photo detector to capture subtle light intensity changes associated with volume changes in the tissues blood vessels. With the help of pulse wave analysis techniques, the blood volume pulse (BVP) can then be constructed from the measured light intensity changes. [61]

Further research in PPG show that pulse measurements are also possible with normal ambient light instead of dedicated light sources [62]. Poh et al. [21] illustrates the potential of PPG by the use of ambient light and a low cost RGB camera achieving high agreement with measurements from tested PPG sensors. Through the possibility of measuring blood volume pulse (BVP) remotely in a non-invasive manner PPG gains interest in visual stress detection [15, 33].

McDuff et al. [33] study discusses a remote detection of cognitive load such as measuring stress in a workplace environment. A novel five band digital camera is used in remotely capturing cognitive stress. The classification of stress follows a person-independent algorithm integrating physiological parameters. Measuring physiological

stress indicators, such as the heart rate variability (HRV), is a main part of their study. Their conducted work is based on the PPG advancements by Poh et al. [21]. A camera sensor with five color bands (RGBCO), adding cyan and orange frequency, instead of the traditional red, green and blue (RGB) band sensor, is employed for extraction of the HRV. With the additional color sensor bands, the performance of remote PPG measurements compared to a traditional RGB sensor is increased. This allows the remote capturing of a subject's physiological parameters as heart rate (HR), breathing rate (BR) and HRV. In Figure 2.3 the automated method for capturing the HRV gets illustrated. In the first step, the facial region of interest (ROI) is segmented and the color channel signals are spatially averaged for each frame. An independent component analysis (ICA) then extracts the BVP from the color data. The strongest BVP signal is then selected and used to calculate a HRV spectrogram.

For the evaluation of the developed method, a stress inducing experiment protocol is conducted. Participants are seated in front of a camera and facial images were captured. The experiment comprises of phases at rest as well as under stress. A mental arithmetic task induces stress through cognitive workload. In addition to the arithmetic task, the subjects are told that they were competing against each other to enhance stress conditions. While the experiment is taking place, reference measurement such as the subject's contact PPG signal, respiration as well as electrodermal activity is recorded with the aid of external sensors.

The prediction of stress is done using the physiological parameters measured from the camera PPG. As stress classifiers a Naive Bayes Model and a linear support vector machine (SVM) are chosen. The input features consist of mean HR, mean BR and distinct modes of the HRV. Testing is done after training the models with features from 9 out of 10 video sequences, which left one test sequence. Furthermore, training and testing is done 10 times (person-independent), once for each participant. Results of the predictions show an accuracy of 85% for the SVM and 80% for the Naive Bayes Model. The physiological information used as input features influence depending on the parameter classification. Whereas the BR is higher for 90% of the participant during cognitive stress, the HR alone is not a strong predictor of stress or resting conditions. Comparing all health input parameters (see Table 2.2), the SVM classifier provides with 85% the best accuracy. The best predictors for cognitive stress are BR and HRV.

In conclusion, McDuff et al. solution shows to remotely (at a distance of 3m) measure physiological parameters with the use of a digital camera. Remote measurements close to contact measurement and high classification accuracy suggest useful applications of their technique. However rigid head movements, facial expressions and additional stressors provide further research opportunities.

McDuff et al. [19] recent work evaluates the cognitive stress detection using a digital camera. Motivation for their study is building a non-intrusive system, the so called *COGCAM*, determining stress during computer work. In their conducted study a test set-up with a digital camera as well as demanding computer tasks was defined (see

Table 2.2: Overview of the classification accuracy for Naive Bayes Model and SVM with different health input features [33].

Accuracy (%)	Random	HR	BR	HRV	All
Naive Bayes	50	65	75	70	80
SVM	50	60	75	70	85

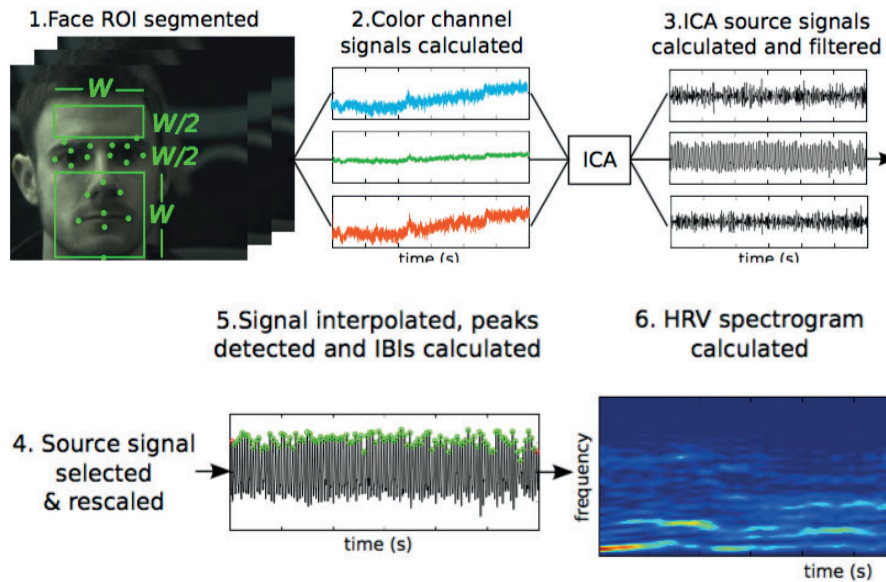


Figure 2.3: Illustration of an automated method used to recover a HRV spectrogram from video sequences of a human face. 1) Detection of facial landmarks and segmentation of face region of interest (ROI) (excluding the region around the eyes), 2) Spatial averaging over time of each color channel from the ROI, 3) Calculating source signal via Independent Component Analysis, 4) Selection of strongest blood volume pulse (BVP) signal and inverting if necessary, 5) Interpolation of BVP signal, BVP peak detection, inter-beat interval (IBI) calculation, 6) HRV spectrogram calculation. [33]

Figure 2.4). Participants sit in front of a camera, while they are recorded completing mentally demanding tasks such as a ball control task and a berg card sorting task. In the ball control task the user's goal is to keep the ball centered on the screen with the challenge that the ball is drawn on it's own to the edges of the screen. If a user failed centering the ball and it reaches an edge, a loud audio buzzer rang. Additionally participants are told that their performance is measured and compared to others for increasing stress levels. The second berg card sorting task is a problem solving exercise by sorting cards into one of four piles based on rules that are not explained to the participants. While sorting cards, rules could randomly change without indication. Both, the berg card sorting task and the ball control task, last three minutes ending automatically and are sufficient challenging for participants.

A PPG signal from each video of the experiment is recovered and physiological data as HR, BR and HRV (in different modalities) are calculated using the previous approach from McDuff et al. [16]. The determination of stress is performed with a Naive Bayes classifier, which is trained using independent physiological data from a preliminary study [33]. Furthermore, the following seven physiological features are used in the classification model: i) HR, ii) BR, iii) HRV low frequency (LF) normalized power, iv) HRV high frequency (HF) normalized power, v) HRV LF/HF ratio, vi) HRV LF total power, and vii) HRV HF total power. With the use of the extracted features the Naive Bayes classifier reaches an accuracy of 86% for distinguishing between rest vs. cognitive load state. Evaluating the different input parameters results suggest that the HRV was the best predictor for cognitive load in comparison to HR and BR. Despite that the preliminary study [33] states the BR as a powerful predictor, the different type of task suggests to influence the prediction outcomes. Moreover, the mean predicted cognitive stress of participants is assessed. As illustrated in Figure 2.5, 70% of the participants show a higher mean predicted stress during both experiment tasks compared to the resting periods. Especially during the ball task, stress conditions are significant higher compared to the resting state.

Overall McDuff et al. work demonstrates the remote measurement of HR, BR and HRV from data gathered by a digital camera during computerized tasks. With the use of this data and person-independent prediction models, stress between rest periods and cognitive demanding exercises can be differentiated. Furthermore, results suggest a correlation between distinct stress inducing tasks and feature types on prediction outcomes.

Bousefsaf et al. [15] introduces a framework for the detection of mental stress. A lion's share of their work builds on the non-invasive remote measurements of the heart rate variability (HRV) via PPG. A subsequently conducted study by them shows the feasibility of the detection of mental stress. Their developed framework consists of image and signal processing (as illustrated in Figure 2.6). In the first step of their presented framework, video data gets recorded using a HD camera capturing a subject's face. Afterwards, an automatic face detection with pan, tilt and zoom parameter (PTZ)

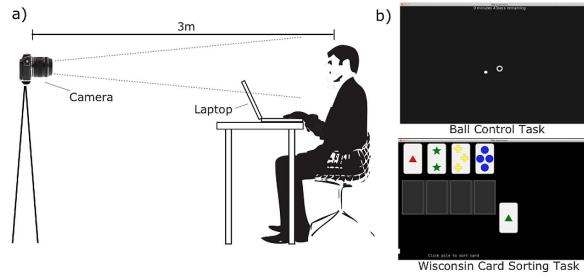


Figure 2.4: a) Experiment set-up. Participants are completing tasks on a laptop while sitting in front of a camera recording them. b) Screenshots of cognitive stress inducing tasks performed on a laptop. [19]

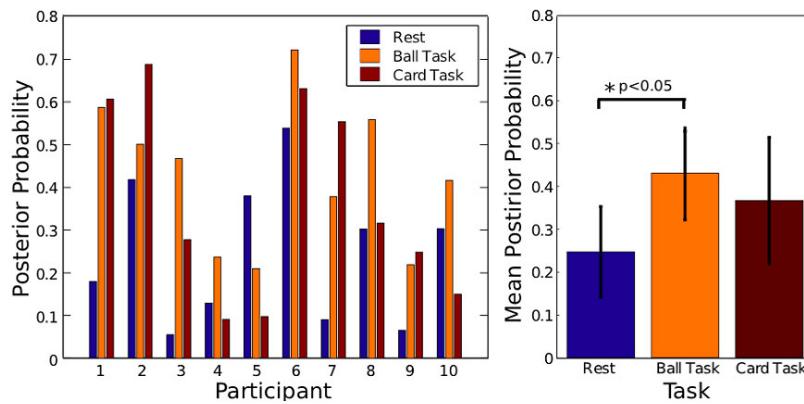


Figure 2.5: Posterior probability prediction of cognitive stress. i) For each task and participant. ii) Average probability (with 95% confidence intervals) for each task across all participants. During the ball task the predicted stress is significant higher compared to the rest period. Rest task, N=20; Ball task, N=30; Card task, N=30. [19]

calculation is performed for the following preprocessing. In the preprocessing step, a skin detection mask for collecting pixels is employed and the pixels' color space is converted from RGB to CIE LUV². The conversion of from the RGB color space to LUV provides a better PPG signal due to reduction of light variations and head movements. With color converted pixels then the U component, representing a red to green color indicator, is calculated. A spatial averaging over the U pixel intensities from the skin detection with further noise reduction and a custom algorithm results in an instantaneous heart rate trace.

In an experiment with twelve students Bousefsaf et al. evaluate their proposed framework. The participants are filmed during completion of the stress inducing Stroop color word test [64]. Each stress test consists of an introduction session, two stress

²CIE LUV [63]

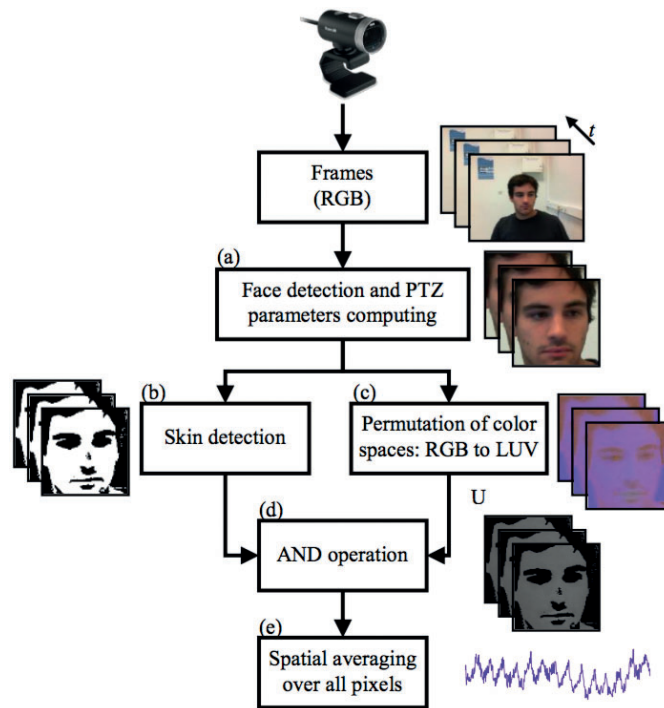


Figure 2.6: Framework overview. (a) Face detection and pan, tilt, zoom (PTZ) computation is performed. (b) Skin detection isolating pixels containing PPG information. (c) Conversion of color space RGB to LUV. (d) U component calculation combined (AND) with skin detection information. (e) Spatial averaging over U pixel intensities of a set of frames, resulting into a single raw signal. [15]

sessions (SS) and relaxation sessions (RS) between. After the data is gathered, stress analysis is applied on the relaxation session and stress session videos. The computed HRV signals are then compared against electrodermal response ground truth measurements. A typical example of their computed results is illustrated in Figure 2.7. The plotted HRV stress curve is in close range with the electrodermal response (EDR) measurements. In general, the experience shows that in stress conditions the HR increases, which also influences the HRV. During relaxation the HRV inclines to be rhythmic, whereas under stress HRV tends to be chaotic and disordered. To detect these stress vs. relaxation rhythm fluctuations in HRV, Bousefsaf et al. compute the third derivative of the captured HRV signal. The conducted remote measurements of HR and HRV indicate a powerful approach for monitoring and evaluating mental stress of a person. Their use of an affordable technology, a low-cost camera, for measuring physiological parameters provides a feasible method to assess mental stress.

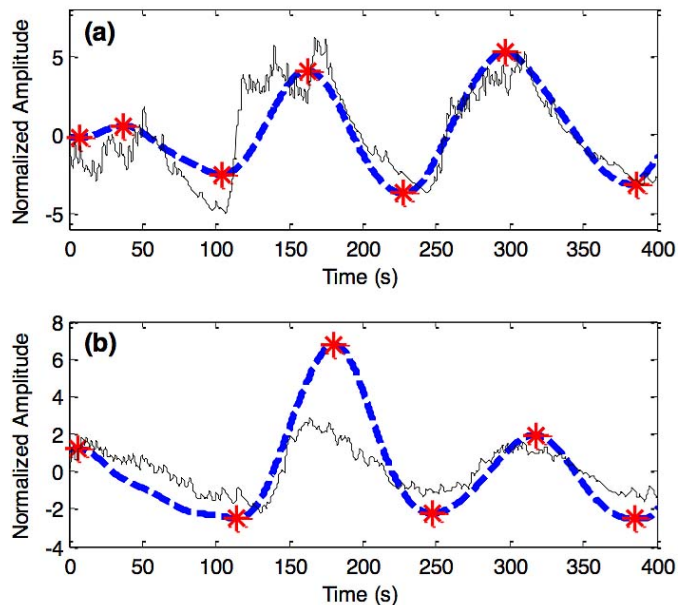


Figure 2.7: Results of stress detection for two participants. The dashed lined plot corresponds to the low-cost camera derived stress signal. The solid-line plot represents the EDR trace. [15]

2.2.2 Model-based Stress Detection

Model-based approaches are particularly suited for the interpretation of faces in images and are used in state-of-the-art literature [65, 66]. The goal of these approaches is to create or fit a model in order to represent faces in video or image data optimally [65]. With the use of a 2D or 3D facial model, the appearance of a face can be described. Further, parameters derived from the model can be helpful to characterize pose, expression or identity of the face [65]. These parameters can be obtained by facial feature extraction methods such as deformation extraction or motion extraction methods [67]. Deformation extraction methods (e.g. Active Appearance Model (AAM) [65]) rely on a reference image to detect the facial changes (the deformations) [67]. Motion extraction methods directly focus on the changes in the face due to facial expressions [67]. Facial modeling and the derived information is used in a wide range of applications from video surveillance, virtual reality, training programs, facial expression analysis and more [68, 69]. Solutions in stress detection also employ model-based approaches to link facial expressions with stress states [34, 32]. Moreover, extracted facial features such as movements of eyebrows, lips and head as well as blinking, gaze etc., are of interest for stress detection [70].

In the work of Giannakakis et al. [1] a framework for the detection and analysis of stress and anxiety states is introduced. Detection and analysis of these states are

conducted through video-recorded facial cues and heart rate estimation. Further, an experiment protocol for the collection of video material is defined. Participants are filmed while being exposed to external and internal stressors visible on a computer monitor in front of them. Distinct phases and tasks such as social exposure phase, emotion recall phase, stressful images/mental tasks and a stressful videos phase, are completed by subjects. After capturing neutral, stressed and relaxed video data from the participants stress detection and analysis is performed. As a first step of stress detection, input data gets preprocessed. Histogram equalization is applied for enhancing contrast before face detection and facial region of interest (ROI) determination took place. For the facial ROI detection active appearance models (AAMs) [65] are chosen.

With the AAM, stress indicators as eye and mouth related features as well as head movements can be extracted from the video data. In the case of the eyes, a time series of eye-aperture changes is detected using landmarks around the eyes, resulting in a calculated mean aperture value as a feature. Additionally, as a second eye feature, blinks per minute is extracted from the same eyes aperture time series. Mouth related features are calculated using the optical flow algorithm. Stress indicators such as reduced rhythmicity and fast mouth movements can be extracted from the optical flow maximum magnitude signal. Head movement and head velocity are tracked using AAM with landmarks preserving information under translation and rotation. As an additional feature, the HR is calculated through facial PPG [21].

After feature extraction, stress detection with the use of machine learning algorithms is performed. Distinct algorithms such as k-nearest neighbors (k-NN), Generalized Likelihood Ratio, SVMs, Naive Bayes classifier and AdaBoost classify the data. The classification is performed task-wise by differentiating between feature sets during neutral and stress phases. The highest classifier accuracy with 91.68% is reached by the Adaboost classifier, whereas other tested algorithms show an accuracy between 80% to 90%. These results show a stress classification compared to related studies and approaches using bio-signals. Giannakakis et al. conducted research shows that facial cues such as head movements, heart activity, eye as well as mouth related features can be used to determine stress and anxiety from video data.

Aigrain et al. [17] develop an automatic stress detection approach based on facial and body activity features. Their solution allows to detect stress from RGB video and depth data. The classification of stress and non-stress states is performed with the help of SVMs. Feature extraction is applied on the data (Skeleton, Video of the body) provided from a Kinect ³ and on additional RGB data (Video of the face) from HD camera. The depth data enables the extraction of emotion associated features such as the quantity of movement, high body activity and posture changes as well as the detection of self-touching. The quantity of movement (QoM) is calculated as the displacement of each joint from the skeleton (SQoM) and as RGB vector from the Kinect image data (IQoM). Both calculations are applied frame-wise. For the extraction of the body activity and posture

³Microsoft Kinect; motion sensing input device

changes the peaks of the QoM calculations are used. Self-touching is detected with the help of the skeleton data and a custom skin detection algorithm for precise location of the hands. A difference is made between hand self-touching and head self-touching. Facial features include 12 AUs defined in the facial action coding system (FACS). These AUs describe movements of eyebrows, lips, cheek, nose, chin and jaw. The detection of the facial AUs is done with the method outlined by Nicolle et al. [71]. In the proposed method, AUs are detected using regression based classification with SVMs. Features for AU classification are shape-based with 49 facial landmarks as well as appearance-based by extracting histogram of oriented gradients (HoG) descriptors [72]. The HoG descriptor features provide additional information and even out possible landmark tracking error in challenging conditions.

The evaluation of Airgrain et al. solution is performed on data from a stress experiment conducted by them. Fourteen subjects are filmed while being in stressed and non-stressed states. Afterwards the videos are labeled based on the subject's self-assessments. The classification of the videos is performed with the help of SVMs. Different classifier parameters (kernel function, etc.) and features (body activity vs. facial) are evaluated. A classifier accuracy of 77% shows the ability of their developed solution to detect stress.

Metaxas et al. [34] implements a model-based dynamic face tracking system for the detection of stress. Evaluation and development of their presented system is performed with video data from a stress experiment of University of Pennsylvania NSBRI center. Participants, taking place in the psychology study, experience high and low stress situations while they are video recorded. The presented tracking system of Metaxas et al. uses deformable model tracking, which fits a generic model to a subject's face. In general the initialized deformable face model updates accordingly to the movement of a given subject's face. These deformations are tracked by a set of n parameters q . A function F_i that takes deformation parameters q , then finds from every point i of the surface model, the according position p_i 2.1 in the approach's world frame.

$$p_i = F_i(q) \tag{2.1}$$

For fitting and tracking of the deformable model a 2D displacement algorithm, called "image forces", is used and transforms to n-dimensional displacement parameter space f_g . Image forces's displacement can be described as the difference between where a point of the model is situated and where it actually is on the image. Displacements in the new displacement space can then be expressed as following 2.2:

$$q = f_g + F_{internal}(q) \tag{2.2}$$

where $F_{internal}(q)$ is the result of internal forces or the elasticity of the model. Further calculations of displacement and projection into image space results in face models as illustrated in Figure 2.8.



Figure 2.8: Illustration of the face model with left and right mouth corner movements [34]

With the help of the established model, stress detection can then be performed within two stages. In the first stage, movement patterns in the region of eyebrows, mouth and eyes are recognized tracking the nonrigid deformation components of a model's parameter vectors. Especially finding stress response patterns such as head-, eye- and mouth- movements, blinking as well as negative facial expressions is necessary. One important indicator is eye blinking, which is harder to detect due to missing model information. A custom grayscale algorithm is implemented over the eye regions, the holes in the deformable model, and observes the grayscale change rates of opened and closed eye lids. For the second stress detection stage hidden Markov models (HMMs) process the video signals for finding stress patterns. Hidden Markov models (HMMs) provide benefits in segmenting input data implicitly, state-based detection of signals as well as providing an inherent degree of variation. Training of the HMMs is done on hand-labeled examples of stress video sequences e.g. rapid head movements, eye blinking and more.

The HMM classifies test data and detected stress responses from the facial tracking data. Evaluation of 25 datasets with one half of low-level stress sequences and the other half with high-level stress sequences is done with two HMMs. Each level of stress is analyzed by one specific HMM. Training and test data are split by 75% to 25%, which results in a correct classification of low/high stress conditions in all test cases. Changing the training and test data ratio to 50% - 50% results in 12 out of 13 correct classifications. Generally speaking, the approach by Metaxas et al. provides a method for detection of stress from dynamic data by using deformable models and HMMs.

Another work in the field of stress detection is conducted by Dinges et al. [32]. In their study an optical computer recognition (OCR) for the detection of facial expressions related with stress is developed. Main part of the OCR algorithm is based on a similar approach published by Metaxas et al. [34] using deformable models. These deformable models allow a representation of a subject's face with the use of a statistical technique called cue integration [73]. In the cue integration method numerous low-level visual computing procedures extract two-dimensional information from a video-recording which can then be transformed into the three-dimensional space. Tracking the created three-dimensional surface, the deformable model (see Figure 2.9), translation and orientation

of the face as well as movements of eyebrows, mouth etc. is possible. Further, with the use of a deformable model, it is feasible to recognize facial features when under stress.

For the evaluation of their approach facial video data from participants exposed to low-stressor and high-stressor scenarios is gathered. Distinct psychological test variations, mainly workload tasks, are designed for triggering a subject's stress. After each test, a rating and assessment of alertness, physical and mental fatigue, exhaustion, stress and other factors is done. In addition, stress reactions are registered using self reports, salivary cortisol as well as HR. Main characteristic for the detection of stress through the deformable model is the nonrigid deformation component of the model's parameter vector. This parameter vector represents movements of the eyebrow, mouth and nearby regions. Furthermore, eye blinking is detected using the change of grayscale values when opening or closing the eyes. Mouth movements, especially asymmetric lip movements, are recognized using an image-based method instead of the deformable model.

The classification of stress signal is done with the use of HMMs. Based on the frequency of emerging facial stress features the HMM decides when a subject is under stress or not. The presence of stress is identified by which specific patterns e.g. eye blinking together with rapid head movements occurs simultaneously. Not only the frequency is decisive for the classification of stress, also the correlation was from importance. Results of classification are compared to classification results from a human scorer. The scorer identifies facial stress responses of the identical 60 video-recorded participants which are provided to the OCR. 85% of the 60 study subjects' facial expressions are categorized correctly as low- vs. high-stressor conditions. In comparison the OCR reaches an initial classification of 75% for the first 20 participants. After a second evaluation, due to falsifying lighting, 15 of 17 (88%) participants' stress states are identified correctly. As a result of this, preliminary results suggest a possible accuracy of 75% to 88% of the OCR algorithm. The described approach by Dingens et al. shows a robust 3D tracking of facial expressions under stress inducing workload demands. Accuracy of OCR classification shows a promising approach for detecting stress in comparison to human classification.

In the paper of Liao et al. [31], a human stress monitoring system using a dynamic Bayesian network gets presented. The proposed dynamic Bayesian network (DBN) is chosen due to the volatile, dynamic and user-dependent nature of human stress. Physiological and behavioral parameters are considered recognizing stress states. A developed monitoring system, visible in Figure 2.10, performs the automated stress monitoring task.

As a starting point their proposed monitoring system extracts features from different sensors monitoring participants sitting in front of a computer. Visual data in the form of real-time videos is used to extract nine different physiological parameters as Blinking Frequency, Average Eye Closure Speed, Percentage of Saccadic Eye Movement (PerSac), Gaze Spatial Distribution (GazeDis), Percentage of Large Pupil Dilation (PerLPD), Pupil Ratio Variation (PRV), Head Movement, Mouth Openness and Eyebrow Movement. All these listed parameters are gathered by a sensor consisting of one wide-angle camera focusing on the face, and two eye focusing narrow-angle cameras. Four

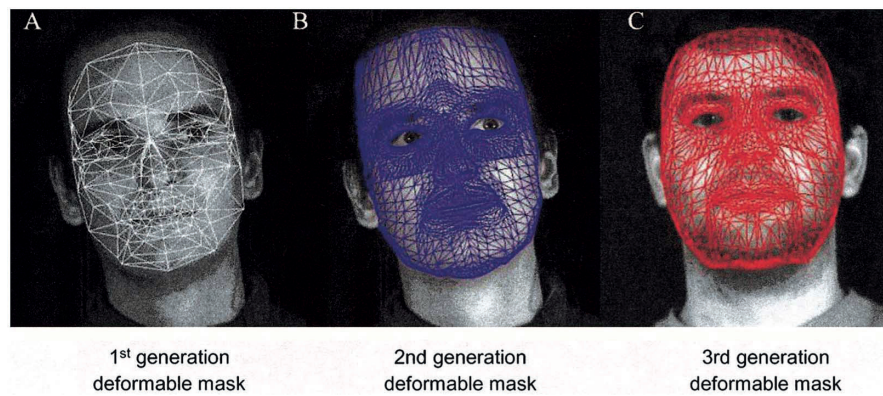


Figure 2.9: Development of the deformable mask from generation 1 to 3 (A-C) used for tracking facial features [32]

separate components analyze the captured video data by applying distinct techniques. Eye detection and tracking is done using a combination of appearance based mean-shift tracking and the bright pupil effect under infrared light sources [74]. With successful eye tracking, eye lid movement parameters can be established. Further, eye-related methods as an eye gaze estimation is performed by utilizing a computational dynamic head compensation model [75]. This model allows an automated gaze tracking, while moving the head in front of the camera. Besides tracking features of the eye, a facial expression analysis is conducted. Facial features around the eyes and mouth are selected and represented by Gabor Wavelets [76]. Detection of facial expressions is done with the use of a flexible global shape model based on active shape models (ASMs) [77]. Face poses can be identified by dynamically deforming the flexible shape model. Further, the introduction of a multi-state face shape model and a confidence verification results in a robust facial expression detection under variable conditions. In addition, to the mentioned visual evidences of stress, physiological, behavioral and performance measurements are gathered. An "emotional" computer mouse (a normal mouse equipped with physiological sensors) measures heart rate, skin temperature, Galvanic skin response (GSR) and finger pressure. For the behavioral evidences mouse clicks and mouse pressure from fingers in a time interval is measured during a user's computer interaction. Performance data is quantified by collecting math error rate, math response time, audio error rate, audio response rate of a user's responses from provided tasks.

After detection of facial features, a dynamic bayesian network (DBN) is integrated. The DBN is built to consider different portions of features and dependencies to determine the stress of a subject. One portion, the predictive portion, are the factors that alter human stress which depends on the previous stress level, workload, environmental context, individual subject's traits as well as the goal a subject is pursuing. Another one, the diagnostic portion, is represented by quantifiable measurements on physical appearance, physiology, behaviors and performance. Both of these portions enable the DBN to inference stress after a learning and active sensing step. Whereas the prediction of stress

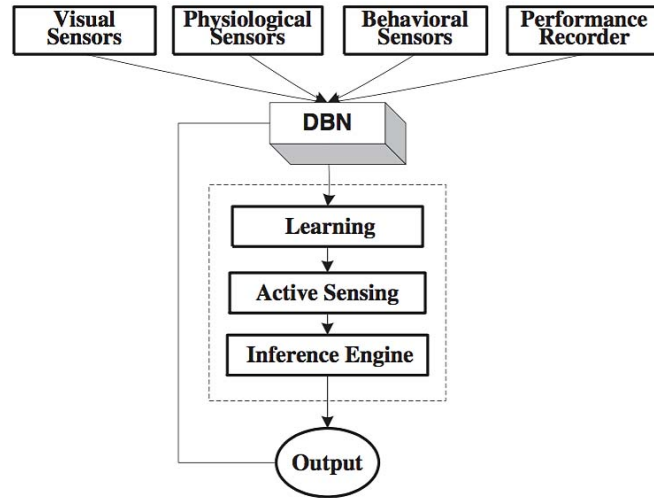


Figure 2.10: Components of the stress monitoring system [31].

derives from the predictive portion and the inference correction of the diagnostic portion. In the learning step a domain expert is consulted to initialize the DBN's conditional probabilities, before the "EM" learning algorithm [78] is applied. After the learning phase, an active sensing technique, which rules out non-informative or contrary stress evidences, is utilized. In the last step the inference engine classifies a subject's stress as output.

A conducted experiment by Liao et al. shows that there system can successfully monitor human stress compared to a ground-truth. Five participants' physiological states are evaluated through Liao et al. monitoring system. Varying workload shall impact a subject's stress state throughout the experiment. During the experiment it is possible to identify distinct signs of a person's stress condition. With increasing stress levels, participants show less blinking, closing the eyes faster, dilating the pupils more often as well as focusing the eye gaze on the screen more often and remaining longer. Also head movements and opening the mouth are more frequent when under stress. Results suggest that the presented non-invasive approach shows a consistent inference of stress. Output of their DBN using evidences of different modalities indicates predictions comparable to psychological theories. However, the employed physiological sensors besides visual sensors distinguish Liao et al. approach from previous outlined literature.

Bevilacqua et al. [20] research tries to distinguish between boring and stressful emotions from facial cues. Their facial analysis tool shall be used as an assessment tool of boredom and stress while playing games. The evaluation of their developed facial analysis approach is done on video recordings from participants playing three different video games. Facial analysis is mainly based on Euclidian distances between automatically detected facial points. These distances are calculated from 68 facial landmarks resulting

in seven features. Face Detection and the placement of landmarks are performed with a constrained local neural field (CLNF) model [79]. Extracted features included mouth, eye and head related features. In Figure 2.11 these features calculated from facial landmarks are illustrated. Features F_1 , F_2 are mouth related and describe the movement of the mouth contour and mouth corners. Eye related features characterize the eye opening F_3 as well as the eyebrow activity F_4 . The remaining head related features comprising of face area F_5 , face motion F_6 and facial COM F_7 in general represent the movement of the head as well as the overall movement of all facial landmarks (facial COM). Calculation of the described features is based on the Euclidian distance between landmarks and on area calculation with the use of Green's Theorem [80].

In order to evaluate their developed approach an experiment inducing boredom and stress is conducted. Participants are video recorded while playing video games. These games provoke boredom at the beginning and stress at the end of the game. While playing, the heart rate of participant is measured by a sensor. Depending on higher changes of a participant's heart rate, a game session is categorized as stressful. With this information, the videos can be divided into boring (H_0) and stressful (H_1) game interactions. For each video of both categories H_0 , H_1 facial features are extracted. Afterwards, an empirical analysis is conducted on these features with the use of the paired two-tail t-test. Hypotheses for all features are formulated stating that the mean value difference of a given feature for the sessions H_0 , H_1 are greater than zero. Results show a decrease of all features from H_0 to H_1 . This decrease, the mean difference to it's mean value, is 10.7% for the outer mouth (F_1), 11.8% for the mouth corner (F_2), 10.4% for the eye area (F_3), 8.1% for eyebrow activity (F_4), 9.4% for the face area (F_5), 8.2% for face motion (F_6) and 11% for facial COM (F_7). All feature except F_6 and F_7 show statistically significant changes. These results support the initial hypotheses that the mean value difference is greater than zero and as a consequence it can be distinguished between boring and stressful states. In conclusion, the approach of Bevilacqua et al. shows promising results in distinguishing between boredom and stress of players. Further research of their solution consists of a new experiment with additional participants, adding remote PPG [21] as a feature and creating a model to classify the emotional states of boredom and stress of players.

2.3 Stress Detection Using Thermal and Hyperspectral Imaging

In contrast to RGB imaging techniques, thermal and hyperspectral imaging captures images in the non-visible spectrum. Depending on the used technology, light with a bandwidth of 10nm for HI and $0.8 - 14\mu m$ for TI is captured [51, 81]. Both of these techniques are used in a wide range of fields. TI devices are available for commercial and industrial usage and can even be attached to smartphones [81, 50]. HI is primarily used for material discrimination [51]. Application scenarios of HI range from food quality control [82, 83] to detection/localization of cancer, diabetes or vascular diseases [84]. The possibilities of these imaging techniques are also of interest in psychology [85],

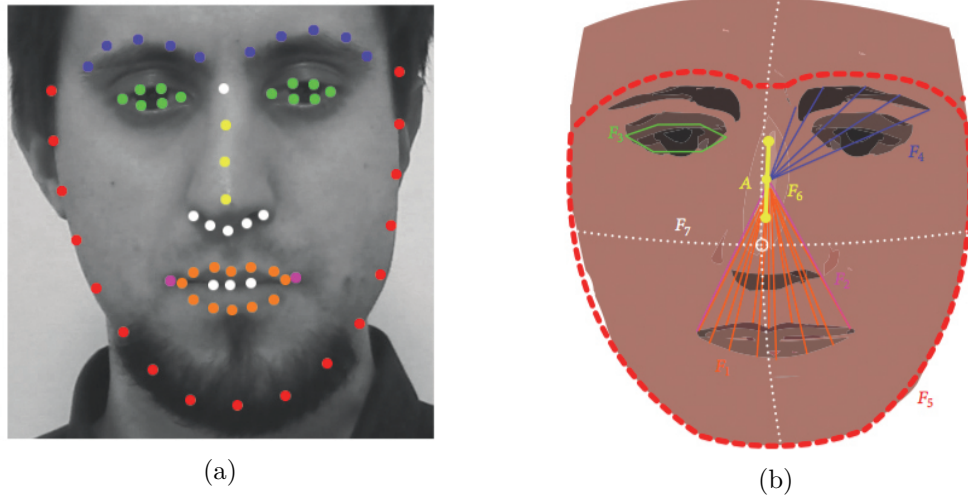


Figure 2.11: Facial landmarks and features. (a) The 68 detected facial landmarks. (b) Facial features F_1 to F_6 visually represented. [20]

affective computing [86, 87] as well as in stress detection [23, 51]. As in the same case for RGB imaging, the properties of being non-intrusive and being able to remotely measure physiological signals makes TI and HI attractive for stress detection [23, 51]. Further, with the measurement of the different light spectra non-visible signals, such as thermal imprints as well as the oxygenation level of the blood, can be used to detect stress [88, 23]. Approaches in stress detection utilizing TI/HI mainly focus on detecting stress based on these signal from facial data [85, 51]. Results of TI and HI approaches prove the feasibility of detecting stress with classification accuracy ranging from 56.52% to 88.9% [50, 51]. At the time of writing most research is found on stress detection with TI (5 references), whereas stress detection with HI shows to be a novel field of research with only two references.

2.3.1 Thermal Imaging-based Stress Detection

The human temperature is of particular interest to medicine [89, 85]. Biological and psychological triggers such as environmental changes, virus infections, are responded by the body's temperature control. In addition to these triggers, emotional reactions are associated with the temperature control. However, for emotional reactions the temperature control is more complex due to their different purpose and it has been shown that emotional reactions carry their own thermal imprints. With the help of TI these thermal imprints can be recorded and provide research opportunities for the detection of stress, anxiety, empathy and more. Primary source of the captured thermal imprints is the face due to the accessibility and ability to communicate. [85]

Research in the field of stress detection with the use of TI focuses primarily on facial heat patterns [23, 41]. Especially, heat signatures in the upper face such as on the

forehead, nostril and eyebrow regions are of interest [29, 50, 23]. The detection of stress from these heat patterns is done mainly by handcrafted features [23, 41]. Solely in the case of Cho et al. [50] a direct data analysis with the use of deep learning, instead of traditional machine learning approaches, can be found at the time of writing.

Sharma et al. [23] work addresses a computer vision-based stress detection with the use of temporal thermal spectrum (TS) and visible spectrum (VS) from video data. Compared to approaches introduced in the previous Chapter 2.2, their approach makes use of thermal patterns radiated from superficial blood vessels situated under the skin of a person's face. An afterward evaluation of their measurements is done using a SVM with genetic algorithms (GAs).

The detection of stress by Sharma et al. is done by the analysis of video data. Before data can be analyzed, a preprocessing is applied. In VS videos, faces are detected using the Viola-Jones face detector [90]. Whereas for TS videos a face detection based on eye coordinates and template matching algorithm is used. Facial regions (3x3 blocks) are extracted from each frame in VS and corresponding TS modality. These extracted regions are then used as facial blocks, illustrated in Figure 2.12. Each block consists of a X, Y and T component, representing width, height, and time. With the video data in block format, a feature extraction using local binary patterns top (LBP-TOP) on VS data is applied. Local binary patterns (LBPs) show promising results for detecting facial expressions [91]. In comparison to the VS, performing LBP-TOP on TS videos do not offer as much information as in the VS case. Therefore a LBP-TOP-inspired method, capturing dynamical thermal patterns in histograms helps to extract suitable features out of the TS data.

Feature extraction of the TS videos is done using the user-independent histogram of the dynamical thermal pattern (HDTP) method. As each stress response can be individual due to more or less subject's tolerance for stressors, a normalization of stress data was required. HDTP provides this kind of normalization by considering the overall thermal state of a person in order to minimize the individual-bias in stress analysis. The calculation of HDTP features requires at first a statistic (i.e. the standard deviation) calculated for each facial region frame for a participant for a particular block for all the videos. All statistics are then used to declare bins partitioning all these frames. A bin can have continuous value ranges with a defined location from the statistic values. With the help of the bins, it is possible to partition statistics for the facial region blocks. For each bin there is a value representing the frequency of statistic values that falls within the bin range interval. From these frequencies a histogram for each block can be formed for subsequent machine analysis.

The classification of stress is conducted with the use of machine learning. Stress is categorized using SVMs and GAs. Evaluating the classification of the SVMs, different results for input features of VS and/or TS videos are gathered. Depending on input features derived from the LBP-TOP or HDTP algorithm the SVM perform better or worse classifying. HDTP features for TS videos as input for the SVM improves

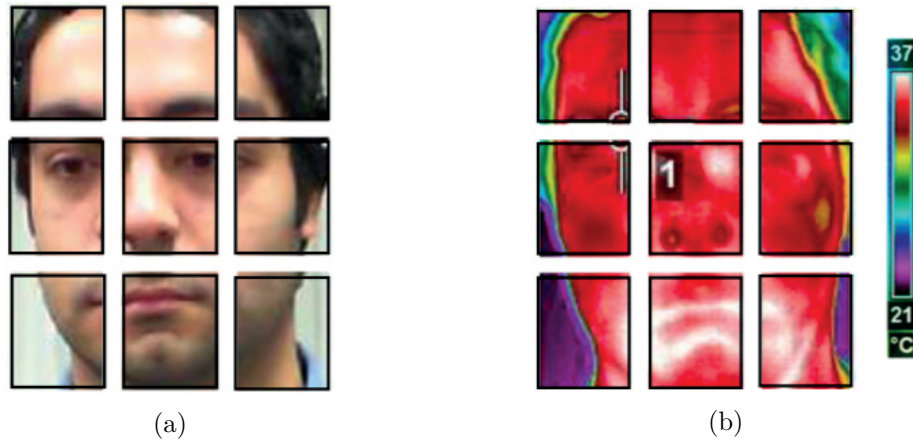


Figure 2.12: A facial region segmented into 3x3 blocks. (a) Block of the frame in the VS. (b) Blocks of the corresponding frame in the TS [23]

stress detection measures. Whereas $TS_{LBP-TOP}$ features show the lowest classification rate when provided as input for the SVM. Best classification results are reached using $VS_{LBP-TOP} + TS_{HDTP}$ features. Comparing classification results, obtained by the use of distinct input features for the SVM, it occurs that the most impact on classification have TS_{HDTP} input features. Furthermore, providing features as input to the GA-SVM show significantly better stress detection measurements as to the classic SVM. Overall a classification rate of 86% could be reached with the use of GAs processing HDTP input features for the SVM.

The paper of Mohd et al. [41] describes a vision-based measurement for recognizing mental stress. Different parameters such as eyes blinking, thermal measurements from three region of interests (ROIs) and blood vessel volume at supraorbital⁴ area, are considered in their approach. A new feature detection method using thermal and visual imaging is presented by Mohd et al.. Subsequent stress detection of obtained features is conducted with a SVM.

In their non-invasive stress detection, a measurement of mental stress through thermal infrared and visual camera sensors is conducted. Temperature measurements are considered based on research linking stress with increased blood flow in facial areas and resulting heat. Thermal patterns in facial areas such as periocular⁵, cheeks, nasal and neck regions occur under different activities or emotions. Especially during mental stress three ROIs are most affected. The presented feature extraction focuses on these face regions by using distinct algorithms.

Initially their performed feature extraction selects faces of thermal images with the Viola and Jones's boosting algorithm [90] and a set of Cascade features as Haar-like

⁴supraorbital; region immediately above the eye sockets

⁵periocular; surrounding the eyeball within the orbit

features. Afterwards, three ROIs are chosen from the detected faces based on a measured face ratio. A mapping of temperature values to a specific ROI is done based on a relationship between brightness of image values and temperature. In addition, a graph cut algorithm removes falsifying ROI areas such as hair and glasses. Besides ROI selection, a blood vessel detection is applied. Sustained stress shows an increased frowning of the eyebrows, which again increases blood flow in vessels caused by contraction of eyebrows related muscles. The detection of vessels is done after the use of a top-hat segmentation method and a bilateral filter. Top-hat segmentation uses erosion and dilation operations to segment the vessels. A bilateral filter then isolates the red range of pixels associated with blood.

Facial feature extraction from visual image data is performed after face detection. The Viola and Jones algorithm is executed and a set of Cascade structures with Haar-like features are calculated. As a next step, the nose area (nostril mask) features are detected by the assumption that the nose is in the center of the face. After that, eyes are located by calculating the edges of the eyes. Eyes are considered as the brightest areas of the face. With determination of the exact locations of eyes and iris, blinking as eye-related feature is extracted. The ratio of white and black pixel in the eye selection mask determines, based on a specific threshold, if the eye is opened or closed. Other features than thermal and visual features are calculated using a developed method introduced by Mohd et al.. Further, with the use of scale invariant feature transform (SIFT) facial features of thermal and visual data are extracted. SIFT allows to map different views of the same object. In the case of Mohd et al. method, the nostril area features are detected using thermal and visual data and then match based on SIFT.

Obtained facial features through the presented techniques are evaluated using a SVM. A preliminary experiment providing facial video data is conducted before evaluation. In their experiment protocol participants undergo a stress test for two times. The second stress test series shall provide more significant results. Further, a relaxation period is planned at the beginning of each test. A Stroop color word test [64] is chosen to raise a subject's mental stress level. During completion of the stress test in front of a monitor, facial video data is gathered. Moreover, ground truth in the form of heart rate and salivary measurements is recorded. A SVM then analyzes the correlation between stress and physiological features. Results of the SVM show an accuracy of 88.9% for the detection of stress. The paper of Mohd et al. demonstrates the correlation of mental stress with a person's blood flow utilizing TI. Especially eyebrow movements corresponding to the activation of the muscle on the forehead (corrugator muscle) can be proven.

In the approach by Puri et al. [29] the so called *StressCam*, a non-contact measurement of users' emotional states through TI, is introduced. *StressCam* tries to recognize stress through capturing heat patterns of the face. A thermal camera, connected to a computer and pointed towards the face of the user, is used for acquiring thermal images. The use of a camera is suitable as a continuous monitoring device because it requires no physical contact to the user. Physiological variables can be extracted from the facial thermal

video through bio-heat modeling. *StressCam* is based on the knowledge that a person's blood flow in the forehead region is increased during stress. Typically the blood flow is centered on the forehead at and above the corrugator or "frowning muscle", shown in Figure 2.13). As a result of increased blood flow, a temperature change on the forehead can be measured through a thermal imaging sensor. Due to that thermal video processing is applied on a subject's frontal vessels. The ROI on the forehead, including the frontal vessels, are selected from a subject's face. With the help of a tracking algorithm, this ROI can then be tracked throughout a conducted experiment. A mean temperature calculation for the 10% of hottest pixels in the tracked ROI, results in a forehead temperature signal. An afterward bio-heat modeling allows to compute the blood flow of the frontal vessels based on the forehead temperature signal.

The evaluation of the *StressCam* solution is done by inducing stress in twelve participants. A computerized variant of the Stroop color word test [64] as stress test is chosen. Participants undergo two sequential test sessions. In the first session, a participant is equipped with a metabolic rate measurement device for estimating EE. Measurements of EE is selected for validation of their stress detection approach. In the second session of the experiment, a baseline part at rest and the Stroop stress test is carried out. During this session thermal images of the subject's face are captured. Recording thermal images and measuring EE simultaneously is impracticable due to the gas masks participants wear. Oxygen consumption, quantified through the gas masks, indicate the EE.

Experiment results of the *StressCam* state that the presented thermal imaging method correlates with ground truth EE data specifying stress. Puri et al. TI method shows to be a viable method for monitoring a user during computer work. Further applications of the *StressCam* can be in distinct areas e.g. identifying stress-inducing questions in computerized testing, identify user interfaces that increase stress levels and more.

In Cho et al. [50] paper a solution to detect stress from breathing patterns using thermal imaging is presented. Temperature changes of the nostril are tracked with a low cost thermal camera which then were used to identify the stress status of a person. The evaluation of their solution is performed on a dataset collected from an experiment inducing mental stress. The basis for their research is the connection between stress and breathing [27, 92]. With the use of captured thermal videos insights into this connection shall be gained. In the first step of their approach, a one-dimensional respiration variability spectrogram (RVS) signal from the nostril ROI on the thermal videos is recovered. The RVS is created by transforming a two-dimensional spectrogram into one-dimensional sequences. This transformation is performed with the use of a power spectral density (PSD) function [93], which identifies similarities between neighboring signal patterns and creates a one-dimensional PSD vector. Stress detection with the RVS signal is performed with the use of a convolution neural network (CNN). The CNN enables direct analysis of the data without hand-crafting features [50]. However, a

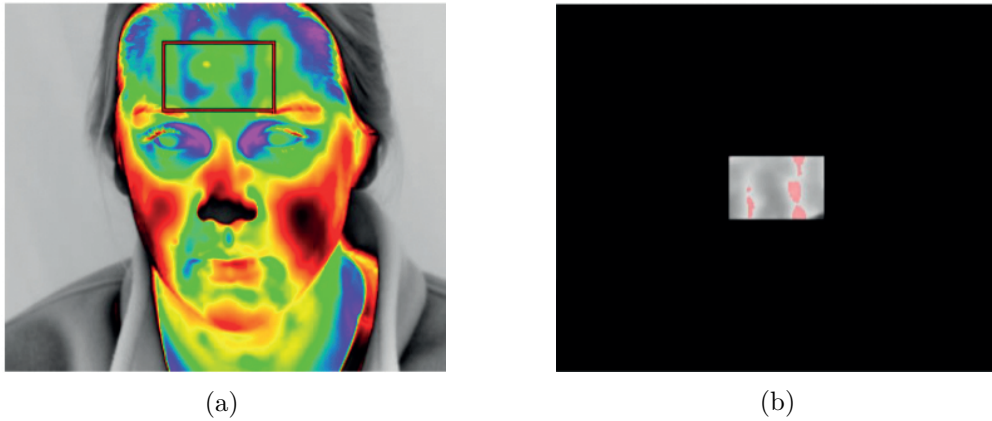


Figure 2.13: (a) Subject's forehead and ROI. (b) The frontal vessels ($\sim 10\%$ hottest pixels in ROI) marked in pink. [29]

problem in affective computing are generally the small dataset, which do not easily allow the use of deep learning approaches [50]. To this end the RVS data is augmented, which artificially enlarges the dataset by applying label-preserving transformations [94]. The architecture of the CNN is illustrated in Figure 2.14. In the first step of the CNN the RVS image patches of 120×120 are resized to 28×28 with the use of a bi-cubic interpolation. Afterwards, the interpolated patches are fed forward to the first convolution layer, which filters the image using n kernels of size 5×5 . The second convolution layer consists of j kernels of size 5×5 . The number of kernels is set along with the number of stress levels to distinguish. Between the convolution layers two pooling layers with 2×2 averaging filters are present. As activation function a sigmoid function is connected to each convolution and fully connected layer. The target class is identified from each output neuron of the final fully connected layer.

The evaluation of Cho et al. work is done with the data collected by a stress experiment. Participants complete the stress inducing Stroop Color Word Test [64] as well as mathematics tests with varying difficulty. Based on videos recorded from these stress inducing experiment sessions and non-stressful sessions their developed solution is evaluated. Ground truth in the form of stress level self reports are used as reference to measure the classification performance. Target classes of the CNN algorithm for multi-level stress detection are *no-stress*, *low-stress level* and *high-stress level*. For the binary case it is *no-stress* and *stress*. The classification performance differs depending on the multi-level or binary case. In the case of the multi-level stress detection, the highest accuracy with 56.52% is reached with the CNN. In the binary case, the CNN reaches an accuracy of 84.59%. Further, an evaluation is done on shallow learning methods such as single layer neural network (NN). All accuracy measurements are performed using leave-one-out cross validation. Overall the results are best for the CNN but not significantly higher than compared to the single layer NN (84.59% vs. 77.31% (binary), 56.52% vs. 53.65% (multi-class)). All in all, Cho et al. approach shows the use of a novel

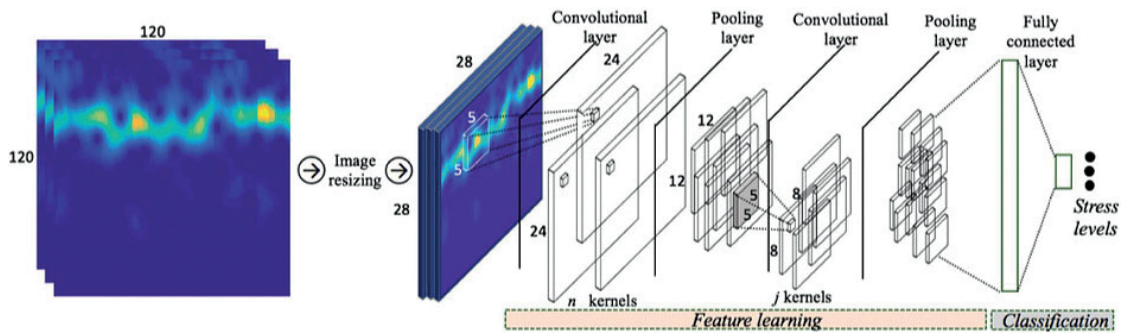


Figure 2.14: Cho et al. [50] CNN architecture comprising of two convolution layers, two pooling layer and one fully connected layer.

low-cost thermography based stress detection solution. Moreover, the system implements automatic feature learning with the use of a CNN.

2.3.2 Hyperspectral Imaging-based Stress Detection

Similar to thermal imaging approaches, hyperspectral imaging-based stress detection utilizes a distinct non-visible electro-magnetic spectrum [23, 50]. Hyperspectral imaging is operating at a bandwidth of approximately 10nm and its power lies in material discrimination. The property of material discrimination allows to determine for instance the blood oxygenation by distinguishing between blood chromophores and body tissue. These changes of the tissue oxygen saturation (StO₂) values can be measured by hyperspectral imaging (HSI). The StO₂ correlates with the bound oxygen to the blood as well as with the hormone adrenaline. In a response to the stressor adrenaline is secreted and facilitates body reactions which lead to a substantial increase of the StO₂. As a consequence, these changes of oxygen content in the blood can then be measured from HSI as absorption changes of the light. At the time of writing, only few approaches [88, 51] using HSI for stress detection are found. [51]

Yuen et al. [88] pre-study to Chen et al. [51] work investigates the use of HSI for the detection of stress. Their developed solution to detect stress is based on the oxygenation of the blood in the human face tissue. The evaluation of their approach is performed with a physical and an emotional stressor. At the beginning of the physical experiment the subject sit on a chair while HSI baseline measurements were recorded. After baseline measurements, the subject are asked to breath deep and slow for 10 times. The breathing exercise are used to test the sensitivity of HSI. As a last step of the physical experiment, the subject run extensively for five minutes. After each physical task, HSI measurements, illustrated in Figure 2.15, are taken. In the case of the emotional stressor, the subject undergo well-practiced experiment protocols such as interviews, public speeches or quizzes. The evaluation of HSI for stress detection is conducted in comparison to the baseline measurements. Results show a positive identification of stress of facial oxygen

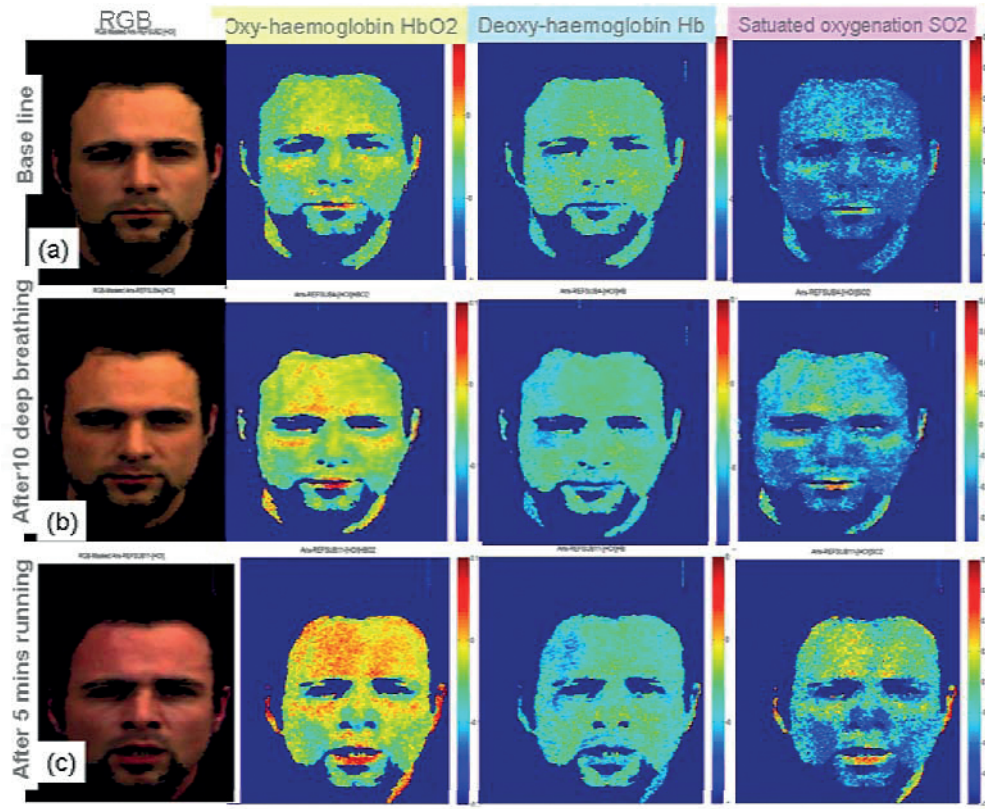


Figure 2.15: Results of stress detection by blood oxygenation measurements from HSI. The experiment consists of task (a) sitting on a chair for baseline measurements (b) deep and slow breaths for HSI sensitivity reference (c) subject ran for five minutes. The illustrated HSI maps are shown in false colors with high (red) and low (blue) concentrations of blood oxygenation. [88]

saturation measurements from HSI. Especially in the case of the physical stressors a more prominently effect than the emotional one can be experienced.

The approach of Chen et al. [51] utilizes HSI for stress detection. With HSI the StO2 value are extracted as primary physiological feature. An afterward binary classification determines the stress status of a person. In order to evaluate the HSI stress detection method a Trier Social Stress Test (TSST) [95] study is performed. Besides stress detection, Chen et al. evaluates HSI in regard of changing ambient temperature and perspiration. For the measurement of stress Chen et al. uses the previously outlined content of StO2 in the blood of a person. Based on these measurements a spatial and spectral optical absorption model is created.

The detection of stress is performed on TSST experiment data from 21 participants. During the experiment HSI data is collected and represented as an image cube consisting

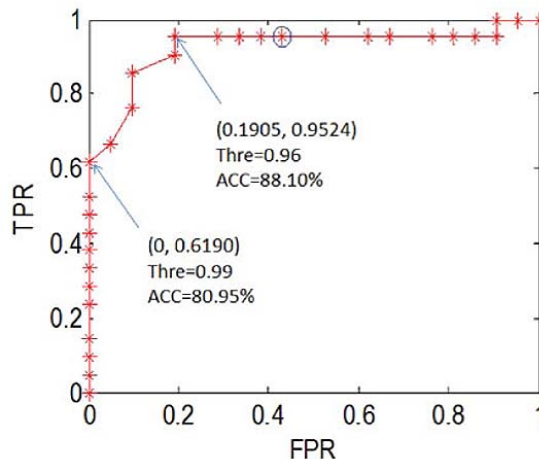


Figure 2.16: ROC curve of a binary classifier [51].

of spatial coordinates/pixels (x, y) and the wavelength (λ) as spectral coordinate (z) . In addition to the image data, cortisol and HR ground truth measurements are conducted. The gathered HSI data is not recorded continuously but rather sampled based on changes of the HR. From each participant a map of StO₂ is produced. Each map, similar to Figure 2.15, is represented by the tissue oxygen saturation. In order to extract StO₂ features, eleven ROIs are studied. Similar to the extracted forehead ROI in Puri et al. [29] approach, the ROIs average StO₂ levels are calculated and normalized for each participant. After StO₂ feature extraction a binary classifier is employed to determine the stress state. As classifier scoring measurement the receiver operating characteristic (ROC) curve with selection of different thresholds is chosen (see Figure 2.16). The classification accuracy range from most conservative at $(0, 0.619)$ with 80.95% to 88.1% at $(0.1905, 0.9524)$.

All in all, Chen et al. conducts the first pilot study employing HSI for stress detection. The strong material-discriminating ability of HSI allows to measure the tissue oxygenation saturation values of the face. These StO₂ values are proven as features to detect stress with a binary classification accuracy up to 88.1%. Furthermore, the robustness of HSI StO₂ values in comparison to thermal imaging (TI) is proven by its independence of perspiration and sudden changes in ambient temperature. Drawbacks or limitations of their approach are high demands on processing power when obtaining data continuously in real-time without a HR reference as recording indicator.

2.4 Comparison and Limitations

All of the presented approaches utilize facial images to remotely detect stress. Facial images are used due to the fact that the face is not obscured and is open to interaction and social communication [85]. Further, facial expressions are of interest for stress

detection [96, 32]. Differences in the outlined stress detection approaches can be found in the imaging modalities comprising of RGB, thermal and hyperspectral imaging. The majority of visual stress detection approaches in the presented literature uses RGB imaging with nine references, followed by thermal imaging with four references and hyperspectral with two references. In Table 2.3 the different approaches with the conducted assessments and outcomes are summarized.

The RGB imaging stress detection solutions use PPG and facial feature extraction based on models of the face. PPG-based approaches show high classification accuracy with 80% and higher as well as high agreement with ground truth measurements. All of the PPG solutions are based on the research of Poh et al. [21] which presents improvements in the PPG signal extraction. Bousefsaf et al. [15] PPG-based solution converts RGB to CIE LUV [63] to reduce noise. McDuff et al. [33] utilizes a RGBCO sensor, an extended RGB sensor with cyan and orange color band, for an improved PPG signal. Additional imaging approaches, detect stress with the help of facial models. These models allow to track deformations and movements of the face [65]. Depending on the solution, mainly eyes, eyebrows, mouth, gaze and the pupils are chosen as stress indicators. Differences can be found in the methods for feature extraction. Giannakakis et al. [1] and Bevilacqua et al. [20] use facial landmark tracking and Euclidian distances as features, whereas Metaxas et al. [34] and Dinges et al [32] use deformable models to derive their features. Liao et al. [31] extract facial features with the help of Gabor wavelets⁶, AAMs and physiological sensors. Face detection is performed in the two-dimensional image space, only Dinges et al. [32] implements three-dimensional tracking of facial expressions. Aigrain et al. [17] takes advantage of depth data in combination with facial video data. Body activity features, posture changes as well as self-touching are detected with the help of these depth data. AUs associates with stress Aigrain et al. extract with the captured video data. The performance of the model-based approaches are assessed by classification algorithms, comparison to ground truth and empirical analysis. Results show a classifier accuracy ranging from 75% to perfect classification, high correlation with ground truth measurements as well as statistically proven differences between stress and non-stress states.

The thermal imaging (TI) and hyperspectral imaging (HI) approaches are operating in the non-visible light spectrum and measure facial heat patterns and the tissue oxygen saturation. Cho et al. [50] and Puri et al. [29] detect stress solely through TI, whereas Sharma et al. [23] and Mohd et al. [41] also extract additional features from RGB imaging. Overall, the TI solutions show a classification performance of 56.52% to 88.9% and correlation with ground truth measurements. The HI technique for the detection of stress is a novel approach at the time of writing. In current literature only two references can be found. Yuen et al. [88] and Chen et al. [51] show the feasibility of TI for stress detection with agreements with ground truth measurement and classification accuracy above 80.95%.

When comparing the distinct stress detection approaches by the outcomes, a high stress detection performance is reached. Solutions utilizing RGB imaging such as PPG and

⁶Gabor wavelets [76]

model based approaches, can easily be implemented due to the low-costs and widespread use of RGB cameras. Recent developments of TI technologies lead to low costs of these devices and make it attractive for stress detection [93, 50]. HI solutions for the detection of stress in current literature are novel, this could be due to the high costs of TI devices in comparison to RGB and thermal imaging devices [97, 98]. Besides the costs and availability of imaging devices, PPG, TI and HI solutions do not require the facial movement to detect stress. Whereas, model-based approaches identifying stress from facial features are based on facial motions. Further, RGB-based approaches work best in moderate stable ambient light conditions and struggle in varying conditions [93]. TI and HI are not depending on lighting conditions. However, environmental changes can impact the performance [51].

Limitations of the distinct approaches include data related problems such as noise or artifacts, inconclusive labeling as well as the drawbacks of the different algorithms and methods [1]. In the case of PPG, noise in the form of unusual BVP changes can occur due to lighting variations or movements of the face [19]. Further, a low video sampling rate can cause less precise PPG measurements [21]. The model-based solutions can generate noise and artifacts due to misfitting of the model as well as pose and illumination changes [99]. Further, the fitting of the model can vary depending on the used algorithm as well as on the pre-trained face model [99]. Thermal imaging (TI) techniques can suffer from sudden ambient and body temperature changes, which can bias the results [51]. Besides technical limitations, the comparison of the different approaches may be difficult due to diverse datasets and conducted stress experiments. A common, open dataset on which benchmarking tests can be performed is not found in literature at the time of writing. Moreover, the sample size of datasets can be too small to conclude universal facial stress signs as Bevilacqua et al. [20] states.

Table 2.3: Summary of the outcomes for each assessed approach. The approaches are based on photoplethysmography (PPG), models, thermal imaging (TI) and hyperspectral imaging (HI). Abbreviations: Electrodermal response (EDR), energy expenditure (EE), tissue oxygen saturation (StO₂).

Reference	Approach	Assessment	Outcome
Bousefsaf et al. [15]	PPG-based	comparison to ground truth	high agreement with EDR
McDuff et al. [33]	PPG-based	Bayes, SVM	accuracy of 85% (SVM) and 80% (Bayes)
McDuff et al. [19]	PPG-based	Bayes	accuracy of 86%
Giannakakis et al. [1]	model-based	Generalized likelihood ratio, k-NN, Bayes, SVM, AdaBoost	accuracy of 80 to 90%
Aigrain et al. [17]	model-based	SVM	accuracy of 77%
Metaxas et al. [34]	model-based	HMM	accuracy of 92% and perfect classification
Dinges et al. [32]	model-based	HMM and comparison with human classifier	accuracy of 75% to 88%
Liao et al. [31]	model-based	comparison to ground truth	high correlation
Bevilacqua et al. [20]	model-based	empirical analysis of extracted features	statistically significant differences between boredom and stress
Sharma et al. [23]	TI-based	SVM	accuracy of 86%
Mohd et al. [41]	TI-based	SVM	accuracy of 88.9%
Puri et al. [29]	TI-based	comparison to ground truth	correlation with EE
Cho et al. [50]	TI-based	CNN	accuracy of 56.52% to 84.59%
Yuen et al. [88]	HI-based	comparison to ground truth	agreement with StO ₂ measurements
Chen et al. [51]	HI-based	binary classifier	accuracy of 80.95% to 88.1%

Stress Data

When talking about stress data physiological features of persons under stress are meant. With these features the human stress response shall be represented. Stress data can consist of captured facial features or other physiological signals recorded over distinct sensors. Facial stress features can include movements of the head, eyebrows or lips as well as blinking, gaze etc.. Physiological signals can comprise of ECG [30], EEG [100], electrodermal activity (EDA) [31, 101], skin temperature [102], BVP [103] and more [46]. In the case of stress detection from facial data mainly video data is used [15, 34, 1, 23]. This video data allows to capture a person's facial expressions over a time series establishing insights into the facial stress response [104]. Stress data plays an important role in the detection of stress [46, 105]. Factors such as data acquisition and data quality can have an impact on performance of stress analysis solutions [106]. Differences in stress data can be found in the modality ranging from RGB [15], depth [17], thermal [23] or hyperspectral imaging [51]. Furthermore, the data specifications can vary depending on the chosen approach and settings. A standardized data definition for visual facial expression analysis is not found at the time of writing. Besides image data, sensor information can be gathered as ground truth measurement or additional data source. Depending on the type of sensor, these data can vary in a wide range from previous mentioned ECG to skin temperature, etc. [33]. In addition to the distinct feature types, the properties, quality and acquisition characterize stress data.

3.1 Data Quality

Data quality in computer science influences the quality of the computed results and the conclusions drawn from it [106]. The colloquial phrase "garbage in, garbage out" summarizes this principle by stating that flawed or nonsense input data produces nonsense output data [107]. In order to evaluate the overall quality of data, the question: "*What is data quality?*", has to be answered. An accepted definition of data quality describes it

as a "fitness for purpose" measurement [106]. To assess the quality of data metrics can be used for evaluation. Pipino et al. [108] defines objective and subjective data quality assessments. Subjective assessments are characterized by the needs and experiences of stakeholders working with the data. In contrast objective assessments, e.g. free-of-error-, completeness- or consistency-metric, can be performed without the contextual knowledge of the application. The ideal high quality data fulfills the criteria of being accurate, complete, relevant, timely, sufficiently detailed, appropriately represented, and retains sufficient contextual information to support decision making [109].

For the purpose of stress detection with the use of facial features we can estimate the data quality by the needs of the chosen methods as well as the field of application. Furthermore, the resulting data quality can be estimated by the process of data collection [46]. In the case of stress detection from facial data, the collection step consists (on a technical sight) mainly of video recording and ground truth measurements. Despite this simply sounding processes, aspects ranging from general technical specifications to appearance of participants have to be taken into consideration. Additionally, the recording environment can have an impact on quality. General speaking, controlled laboratory environments provide advantages in terms of captured data quality over more uncertain real world application scenarios [110, 111]. The quality of facial video data in stress detection we can subjectively assess by data collection in existing literature as well as in regard of chosen methods. As an ideal facial video data quality, we can argue that requirements as lighting, view on the area of interest as well as clean and crisp images are met. Moreover, for other imaging modalities e.g. thermal imaging other factors such as the ambient temperature have to be taken into account [51]. Besides data collection settings, the afterwards data processing methods can determine if the data quality is sufficient. Robust computational methods may produce significant results on bad data quality than sensitive methods [110]. Depending on the chosen method, this factors can be taken into account when collecting data. For stress detection approaches based on e.g. PPG highly varying lighting conditions or motion of the area of interest can influence the outcome in an unwanted manner [16]. Other solutions such as AAMs and landmark points can produce artifacts or noise due to fitting degradation on low resolution images [112].

Further, the data quality from additional data signals, e.g. BVP, EDA etc., has to be assessed. Due to distinct sensor quality the produced signals can differ. Especially the accuracy of consumer devices is often unknown due to missing validation studies [113, 114]. To outline this in more detail, the study of Shcherbina et al. [114] evaluates commercial wearable wristband devices in regard of HR and EE measurements for different physical tasks. As these types of devices gain more and more popularity for physiological monitoring this can be of interest [115, 116, 117]. Available consumer devices from the year 2017 such as the Apple Watch, Samsung Gear S2 etc. show an median error rate below 5% for HR and above 20% for EE measurements against the gold standard ECG and gas analysis from indirect calorimetry. This example elaborates the varying sensor accuracy and the significance on data quality. For the acquisition of data a trade-off between factors such as the sensor accuracy and the impact on data quality can be made.

In addition to the influence of monitoring devices properties, the overall background or motivation of data collection can affect data quality. Research in other disciplines, e.g. psychology or medicine, collect data for different purpose [118, 119]. This data can also be of interest for other fields of research such as computer science. Despite that the gathered data may be sufficient for the original aim of research, it can be insufficient for the different fields of research due to the lack of precision, accuracy or other requirements for technical analysis [120]. A more interdisciplinary approach of data acquisition for collaborate research may result in better data quality for all stakeholders as our experience shows with acquired data.

Another approach which can affect the data quality is data preprocessing. This process can have an significant impact on the performance of data analysis algorithms such as machine learning [121, 122]. Typical data preprocessing can include data cleaning, normalization, transformation, feature extraction and selection, etc.. Depending on the dataset, distinct preprocessing steps may be required to improve data quality and achieve better results. Especially on real world non-academic datasets preprocessing can be required [123].

3.2 Data Acquisition

In order to recognize stress by an automated computing approach data characterizing stress conditions is necessary [46, 105]. These data is crucial in techniques such as machine learning to find underlying information or patterns in it [124]. Due to the dependency on this information, requirements on the design of a stress experiment protocol and the resulting data quality have to be taken into consideration [106]. Testing procedures, technical details, inducement of stress states as well as the ground truth measurement of stress are main aspects of test specifications [33]. Based on these variables the quality of captured video data can influence the performance of an implemented solution [107]. Moreover, the robustness of the developed system to detect facial expressions and acquire the necessary data for stress detection is crucial. Environment aspects such as lighting, shadows as well as the person's appearance can influence results [21, 125]. In the following sections, a definition of data and data quality in the context of this thesis, an overview of requirements for the design of a stress test as well as on the quality of data is provided. A special focus is set on technical specifications regarding the detection of facial expressions through visual sensors. Limitations and comparison of different technologies shall help to highlight possible benefits and drawbacks. Evaluation of technical details is conducted based on available literature in the domain of visual computing and facial expression analysis.

3.2.1 Specifications

Research on facial expressions focuses on establishing facial expression databases [126, 127, 128, 129]. These databases are established by either inducing emotional states e.g. stress or by giving the subject's instructions to perform facial displays [126, 19]. The

technical specifications of the captured image material however varies. In Table 3.1 these differences are outlined. A comparison of data characteristics from facial stress and facial action coding system (FACS) databases, such as sample size, data modality, resolution, frames per second (FPS), ground truth measurements, gives an oversight. Facial databases differ in the number of subjects recorded, which may influence the validity of outcomes as Bevilacqua et al. [20] states. Stress databases, outlined in Table 3.1, are smaller in comparison to the FACS databases with 5 to 35 participants vs. 19 to 182 participants. Data modalities of the sets include RGB, depth data (3D), thermal and grayscale. The majority of stress analysis solutions performs detection on RGB imaging. Image resolution of the different datasets varies from thermal spectrograms in the case of Cho et al. [50] with a spatial resolution from 120×120 pixels to Bevilacqua et al. [20] with a resolution of 1920×1080 pixels. Further, in datasets [17, 126] multiple images and/or videos are captured with different resolution. The impact of image resolution is discussed in Chapter 3.2.3. Further, the temporal resolution or frames per second (FPS) column describes the sampling rate at which the videos are recorded. These sampling rate determines how accurate a signal or facial expression can be reconstructed [130, 128]. Features such as Blinking, which is of interest in facial analysis, lasts about a third of a second [131, 31]. Further, micro-expressions, a short facial movement revealing genuine emotion, can last less than 0.04 seconds and be of very low intensity [132]. In order to detect these subtle movements and get more details of the movement, a high sampling rate (greater than 30 FPS) is preferred [128, 132, 133]. When comparing the temporal sampling rate of the databases in Table 3.1, the FPS rate ranges from 20 to 200 FPS. The high FPS rates, e.g. 200 FPS, are chosen to provide more detailed information of facial muscle movements [128].

Besides video data, ground truth measurements are recorded to evaluate facial analysis results or provide additional information for analysis [15, 1]. Typically, these measurements are captured with the use of body contact sensors, psychological measurements or self-reports [29, 50, 31]. Physiological signals, such as the GSR or electrodermal response (EDR), BVP, electromyograms, electroencephalograms, are captured by body contact sensors (see also Figure 2.2) [18]. Bousefsaf et al. [15] and McDuff et al. [33] compare their results against ground truth measurements from proven PPG devices and/or the EDR. As illustrated in Figure 2.7, the results of PPG measurements are compared against the EDR measurements from a body contact sensor [15]. For the comparison of results against ground truth measurements the temporal resolution has to be taken into account. Without synchronizing the data, e.g. interpolation, sampling, the signals of two sensors with distinct temporal resolution can not be compared with the same precision, e.g. 1 Hz contact PPG sensor measurements can not be compared with 30 Hz facial PPG measurements. In addition to ground truth measurement from sensors, self-reports, psychology measurements and knowledge of experts (FACS coder) is part of datasets. These measurements are mainly used for annotation of the data for machine analysis [50, 128].

Table 3.1: Facial expression databases and corresponding characteristics of image/video recordings. Upper part of the table comprises of stress detection databases, lower part of FACS databases.

Reference	Sample Size (Participants)	Data Modality	Resolution	Frames per Second	Ground truth measurements
McDuff et al. [33]	10	RGB	960x720	30	Contact PPG, EDR
Bousefsaf et al. [15]	12	RGB	HD	30	EDR
Bevilacqua et al. [20]	20	RGB	1920x1080	50	Heart Rate
Liao et al. [31]	5	RGB	N/A	N/A	Psychological measurements
Giannakakis et al. [1]	23	RGB	526x696	50	Self-reports
Aigrain et al. [17]	14	Depth, RGB	640x480, 1440x1080	N/A	Self-reports
Sharma et al. [23]	35	RGB, Thermal	640x480	30	Self-reports
Cho et al. [50]	8	Thermal	120x120	N/A	Self-reports
Puri et al. [29]	12	Thermal	320x256	31	EE
Cohn-Kanade [126]	182	RGB, Grayscale	640x480, 640x490	30	FACS coder
Casme [127]	35	RGB	1280x720, 640x480	60, 60	FACS coder
CASME II [128]	35	RGB	280x340	200	FACS coder
Sayette GFT [129]	96	RGB	720x480	29.97	FACS coder
MMI (Part I-III) [134]	19	RGB	720x576,	24	FACS coder
DISFA [135]	29	RGB	1024x768	20	FACS coder
SEMAINE [136]	150	RGB, Grayscale	780x500	49.97	FACS coder

3.2.2 Data Modality

For the detection of facial expressions different data modalities can be used. Besides RGB imaging, thermal and hyperspectral imaging data, three-dimensional images containing depth information are of interest for facial expression recognition [132]. Three-dimensional facial expression analysis is an open research field and is at an early stage [132]. The majority of visual affect recognizer still use 2D images as input [137]. However, the rapid progress in depth-based imaging technology is supporting 3D facial analysis [137]. Stress detection approaches with the help of 3D facial expression analysis through depth data are not found at the time of writing. Current 3D facial expression analysis literature is focusing on facial action coding system (FACS) ¹ analysis and basic emotion (e.g. anger, happiness) recognition.

Several databases, including 3D static faces and image sequences, are established due to the progress in 3D data acquisition. Distinct 3D data acquisition techniques such as single image reconstruction, structured light, photometric stereo and multi-view stereo are used to create these databases. Especially the advancements in structured light scanning, stereo photogrammetry and photometric stereo allow the acquisition of 3D facial structure and motion. Structured light scanning basically projects one or more encoded light patterns onto the scene (e.g. the face) and then the deformations of the patterns on the object's surface is measured to extract shape information. Stereo photogrammetry utilizes multiple cameras situated at multiple known viewpoints from the subject to apply 3D facial reconstruction. The photometric stereo technique acquires the facial 3D structure by capturing a set of images under different illuminations. [132]

The automatic detection of action units (AUs) can be performed on two and three-dimensional data. However, AU detection from 2D illuminance images may have hit a performance ceiling. Research suggests that 3D face data could help to overcome the limits of 2D. Three-dimensional data enables true facial surface measurements, which may help to differentiate better between subtle differences of AUs. Further, 3D data has shown to be immune to illumination and in an extent to pose variations. Overall, the 3D modality has significant advantages in AU detection and performs in general better than 2D when comparing the same feature extraction and classification algorithms. Especially with increased pitch and yaw rotations of the face, the performance of AU recognizer using 2D data drops compared to 3D data. This performance decrease is due to occlusion effects and substantial distortion from out-of-plane rotations. Moreover, 3D data allows better normalization and luminance data can be compensated for moderate out-of-plane rotations. [138]

All in all, three-dimensional imaging has benefits over 2D data, such as immunity to illumination, better performance under different head poses and occluded faces as well as easier detection of out-of-plane movements of facial features [139]. Despite these benefits over 2D data, experiments show that 3D data is not necessarily better than 2D RGB data for facial analysis. Especially for the upper face regions 2D data show better performance due to noise of 3D data acquisition [133]. The fusion of 2D and 3D data

¹FACS; definition of 32 atomic facial muscle actions named AUs [49]

improves the performance of facial analysis [133]. In addition to single 2D and 3D facial analysis, the fusion of 2D and 3D data further improves facial analysis performance [133].

3.2.3 Image Resolution

The image resolution describes the number of pixels defined by the width and height of an image [140]. To successfully extract features and detect expressions of the face a minimum resolution of the captured data is required [110]. Depending on the recording setup, aspects such as camera resolution, sensor size, focal length, distance to object as well as the resolution of the face in the image have to be considered. In the case of facial stress detection, an appropriate image resolution of the face is of importance. To calculate the minimum object that can be seen in an image, the technical specifications of a camera system are decisive. The following variables and formulas outline, how to calculate the minimum viewable object for an image in order to detect e.g. a face accurately. [141]

With the following variables we can calculate the minimum viewable object for an image:

- D ; the distance from the lens to the subject
- AoV ; the angle of view of the lens
- $PixelsWidth$, $PixelsHeight$; pixel dimension of the target image
- angle of view from camera specification

To calculate the angle of view of a lens and the sensor following formula 3.1 can be applied,

$$AoV = 2 * \arctan(SD/2 * FL) \quad (3.1)$$

whereas SD specifies the diagonal *sensor dimension* of the camera and FL describes the *focal length* of the lens. The minimum viewable object can be interpreted as the smallest size that will map onto a single pixel for a given field of view, or when stated as a formula 3.2:

$$Minimum\ Viewable\ Object = Field\ of\ View/PixelsWidth \quad (3.2)$$

For a given distance the total *field of view* 3.3 is:

$$Field\ of\ View = 2 * D * \tan(AoV/2) \quad (3.3)$$

The minimum viewable object can be calculated for a specific camera in a test setup with these formulas. Taking McDuff et al. [33] experiment as an example with a distance of 3 meters from the lens to the subject, a resolution of 960×720 pixels and a focal length of 50 mm. For the missing parameters as the sensor dimension we can take an estimation. In this example the APS-C image sensor format of 25.1×16.7 mm is used. The calculation of the angle of view results in $AoV = 2 * \arctan(30.14/2 * 50)$ or 33.55 degrees. With the






$Field\ of\ View = 2 * 3 * \tan(0.58/2)$ or 1.79 m the *Minimum Viewable Object* for the 960 pixel wide image can be calculated. The result is *Minimum Viewable Object* = $1.79/960$ or 0.00186 m (0.186 cm) small. This value of 0.186 cm can be interpreted as the distance measured in the real world mapped onto 1 pixel in the image. To set this in the context of facial expressions, we are interested how many pixels the human face occupies in an image. The mean breadth and length of the German male face in Farkas et al. study [142] is 13.3×18.2 cm. Assuming these dimensions for our example, a person's face covers approximately 72×98 pixels in a captured image.

In order to assess the calculated face dimension of our example, we can look at Tian et al. [110] evaluation of face resolution for expression analysis. Their work evaluates the accuracy of expression analysis in regard to facial images with different resolutions. Table 3.2 illustrates results of their evaluation. In the left table description the following face processes are stated: face acquisition, feature extraction, FACS AUs and basic expressions. Each of these processes is specified in regard to the applied methods: face detector (FD), head pose estimation (HPE), geometric feature extraction by feature tracking (G1), geometric feature extraction by feature detection (G2) as well as appearance based feature extraction (AP). In Table 3.2, values range from yes/no values to percentage values. For the feature extraction process the values are yes or no if feature extraction was possible for the specific resolution. The percentage values from the rest of the table content describe the accuracy results of the used techniques for each different facial image resolution. The empirical study of Tian et al. show that head detection and HPE is able to detect faces in lower resolution than FDs. Furthermore, there is no difference in expression recognition based on geometric or appearance based feature extraction when the facial image resolution is higher than 72×96 pixel. Lower resolutions such as 36×48 pixel show better recognition rates when using appearance based features. In addition, finer levels of facial expressions can not be reliably obtained with low resolution image as 36×48 pixel. Comparing our example with the facial image resolution of 72×98 pixels with values in Table 3.2 from Tian et al., comparable results as in the fourth column shall be achieved. In our example, the image resolution is sufficient to extract features and detect expressions from the facial image data.

In addition to the image resolution of 2D data, a high image resolution of 3D image data is preferred to capture movements in very small parts of the face [132]. The acquisition of high resolution 3D data can depend on the used imaging technique [132]. Structured light imaging is able to record images in real-time or high speed and with low costs as in most cases only a projector is needed [132]. A drawback of this high temporal resolution is a lower image resolution [132]. Popular structured light capturing devices such as the Microsoft Kinect ² show low image quality and noise for facial analysis, which has to be compensated with e.g. multiple scans and data fusion [143]. Besides structured light, stereo cameras are utilized to capture 3D images or sequences [132]. These allow to capture high resolution 3D facial databases [139, 144]. However, the amount of captured data and the lower temporal resolution may make it unfeasible for applications such as micro-expression analysis or real-time systems [132].

²Microsoft Kinect; motion sensing input device

Table 3.2: Resulting effects of faces at different resolution. FD and HPE for face acquisition describes face detector and head pose estimation. G1 for feature extraction states geometric features extracted by feature tracking. G2 for feature extraction states geometric features extracted by feature detection. AP for feature extraction indicates appearance features extracted by Gabor wavelets. [110]

Face Process						
		288x384 (Original)	144x192	72x96	36x48	18x24
Face Acquisition	FD	100%	100%	100%	100%	0%
	HPE	98.5%	98%	98.2%	97.8%	98%
Feature Extraction	G1	Yes	Yes	Yes	No	No
	G2	Yes	Yes	Yes	Yes	No
	AP	Yes	Yes	Yes	Yes	Yes
FACS AUs	G1	90%	90.2%	89.9%	N/A	N/A
	G2	71%	70.8%	72%	54.3%	N/A
	AP	90.7%	90.2%	89.6%	72.6%	58.2%
	G1+AP	92.8%	93%	92.2%	N/A	N/A
	G2+AP	91.2%	90.8%	90%	87.7%	N/A
Basic Expressions	G1	92.5%	91.8%	91.6%	N/A	N/A
	G2	74%	73.8%	72.9%	61.3%	N/A
	AP	91.7%	92.2%	91.6%	77.6%	68.2%
	G1+AP	93.8%	94%	93.5%	N/A	N/A
	G2+AP	93.2%	93%	92.8%	89%	N/A

3.2.4 Robustness

Under the term *robustness*, the ability of a software to keep an "acceptable" behavior in spite of exceptional or unforeseen execution conditions (such as invalid or stressful inputs, unavailability of system resource etc.) is understood [145, 146]. The robustness of facial analysis algorithms and computational methods is crucial and may impact the outcomes of these solutions [51, 21]. The detection of faces in images is a fundamental task for vision-based computer interaction [147]. Robust and efficient ³ algorithms are required to enable face detection for distinct application and in environments with a variety of lighting conditions [148]. Facial expression analysis and facial stress detection approaches utilize face detection as one of the first steps of the processing pipeline [49, 1]. In computer vision, face detection is a well studied topic and modern face detectors achieve high performance in the detection of near frontal faces [149]. However, the detection of not near frontal faces is difficult for widespread algorithms such as the Viola-Jones [90] or histograms of

³efficient; operating with a minimum consumption of resources [146]

oriented gradients (HoG) [72] algorithm [149, 150, 151]. Especially when detecting faces from distinct angles or partly occluded faces these algorithms fail [152]. Recent research focuses on the detection of faces in unconstrained scenarios with factors such as extreme pose, exaggerated expressions and large portions of occlusions [153]. Convolution neural networks (CNNs) are used to address these problems and outperform algorithms as the previous mentioned Viola-Jones and HoG [154, 155]. Further, facial analysis from 3D data shows better performance under varying poses and illumination [138]. Despite the good results of CNNs for face detection, research in facial stress detection and facial expression analysis utilizes primarily appearance based methods such as the Viola-Jones or the HoG algorithm [49, 1]. The popularity of these algorithms is due to public availability of pre-trained models (e.g. in Matlab ⁴, OpenCV ⁵ or dlib ⁶), the computational simplicity and the reliability for frontal face detection [49].

In addition to the detection of faces, the facial landmark detection is important for facial stress detection and expression analysis [20, 99]. The goal of facial landmark detection is to automatically detect the location of facial landmark points on video or image data. Facial expression analysis or head pose estimation may rely on the landmark detection algorithms. Due to that outcomes of these approaches can depend on an accurate detection of landmarks. The performance of facial landmark detection is affected by distinct factors. Facial appearance changes from distinct facial expression and head poses can be significant. Further, environmental conditions such as illumination, affects the appearance of the face on images. Facial occlusions by other objects or self-occlusions due to extreme head poses can also lead to poor performance. The major cause of failure for landmark detection from these error sources shows to be extreme head poses (e.g. profile face). Significant head poses may lead to self-occlusion and missing landmark points. Moreover, the limited training data with head poses may result in bad performance for the detection of landmarks. Overall, face detection and facial landmark detection achieve good performance but are still an unsolved problem. [99]

Besides facial detection and facial landmark detection, the robustness of photoplethysmography (PPG) based and thermal imaging (TI) based stress detection approaches is affected by environmental factors. Varying lighting conditions, movements of the face, low sampling rate as well as sudden ambient and body temperature changes may influence the outcomes of these solutions [19, 21, 51].

3.3 Databases

Stress data is a crucial component in the detection of stress [46, 105]. Aspects such as data quality and acquisition can impact outcomes of stress detection solutions [106, 19]. Despite the central role of these data in research, only one open and accessible stress database [27] can be found at the time of writing. The dataset of Healey and Picard [27] describes stress condition, while real-world driving tasks. However, the database from Healey and

⁴Matlab, <https://www.mathworks.com/products/matlab.html>, last accessed 09.08.18

⁵OpenCV, <https://opencv.org/>, last accessed 09.08.18

⁶dlib, <http://dlib.net/>, last accessed 09.08.18

Picard comprises of contact sensor measurements and does not contain subject's facial images. In facial expression analysis, open and accessible databases describing FACS AUs [134] and basic emotions [126, 156], i.e. anger, disgust, fear, happiness, sadness and surprise, can be found. Further, accessible databases include the Sayette GFT [129], KDEF [157], DISFA [135], Bosphorus [158] and more⁷. Databases of stress conditions including facial images are collected through stress experiment protocols [33, 41, 17]. Open and accessible facial stress databases however are not found at the time of writing.

⁷Comprehensive collection of public facial expression databases, <https://www.behance.net/gallery/10675283/Facial-Expression-Public-Databases>, last accessed 03.10.18

Methodology

In order to detect a person's stress state facial analysis can be used [20]. Visual computing approaches allow to detect stress based on physiological signals or expressions derived from facial images or videos [18]. Due to non-available open stress datasets at the time of writing an own stress dataset is created. In the following experiment protocol chapter, the setup of stress data acquisition is outlined. The implemented stress detection solution is based on the research of Giannakakis et al. [1]. Figure 4.1 illustrates the developed solution in more detail. The image preprocessing steps include contrast enhancement, face detection and ROI determination. Afterwards, features from eyes, mouth, head and heart activity are extracted with the help of Euclidian distances, the optical flow algorithm and facial PPG. In the last machine learning step preprocessing and classification is performed. Subjects' facial data is identified as *stressed* or *non-stressed* based on the features extracted from the previous extraction step.

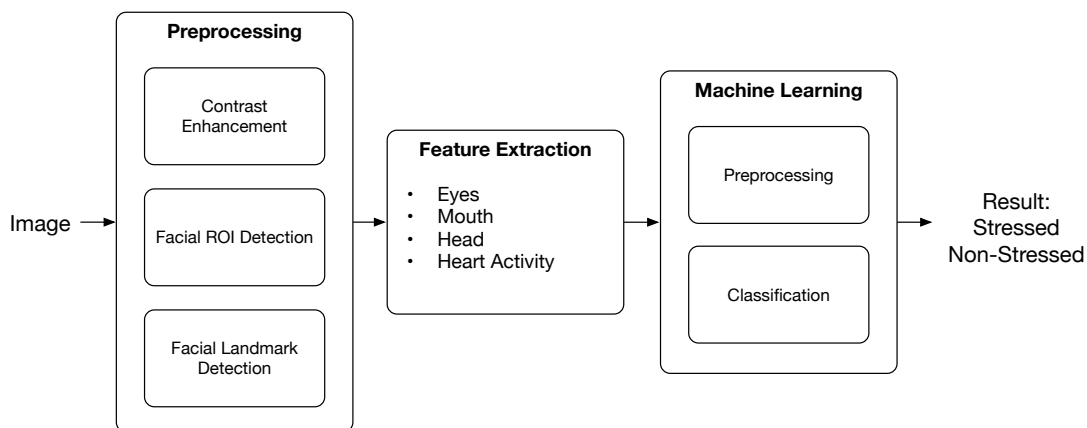


Figure 4.1: Overview of the stress detection framework from facial videos.

4.1 Experiment Protocol

In order to evaluate a person's stress conditions by machine analysis it is necessary to gather data of subjects in *stressed* and *non-stressed* states [46]. The main objective of this experiment protocol is to record facial stress data to perform stress detection by ML algorithms. In the conducted experiment, facial data from participants is recorded, while they are in a neutral and stressed emotional state. The acquired data consists of video sequences, heart rate measurements from a body contact sensor and self-reports regarding stress conditions. For the inducement of stress the *cold-pressor test* [159] is conducted. Overall, the experiment is designed functional and is not the main research goal of this thesis. Limitations regarding psychological procedures have to be taken into consideration.

4.1.1 Experiment Setup

Study participants are seated in front of a camera, which is placed at a distance of approximately two meters. The camera is situated with its field of view (FOV) to capture the frontal face of the subject. Next to the subject an ice bath is placed to perform the cold-pressor test (CPT). The overall setup of the experiment is illustrated in Figure 4.2. Ground truth measurements are captured with the help of a wristband, monitoring the heart rate. The technical equipment includes a DSLR RGB camera ¹ on a tripod, a 50mm lens ², a heart rate monitoring wristband ³ capable of exporting data as well as equipment to conduct the stress inducing test. Ambient lighting and natural lighting was used depending on the time of data acquisition. The video data is captured with a resolution of 1280×720 pixels and frame rate of 50 FPS.

4.1.2 Experiment Procedure

The objective of the experiment is to induce affective states through a stressor. To this end participants are asked to take part in a stress phase and neutral phase. Subjects are invited to join the stress inducing experiment without the exact knowledge of the experiment procedures. Further, the subjects are invited individually to prevent information exchange about the experiment. These precautions are chosen to prohibit mental or physical preparations and increase stress conditions through the uncertainty. The experiments take place late in the morning and early afternoon to record the subjects while they were alert. Sleepiness in the early morning and late afternoon may impact facial expressions. At the beginning of the procedure, the subjects are asked to take place on a chair and the heart rate monitoring wristband is applied. Afterwards, instructions

¹Canon EOS 60D; <https://www.usa.canon.com/internet/portal/us/home/products/details/cameras/eos-dslr-and-mirrorless-cameras/dslr/eos-60d>, last accessed 18.08.18

²Canon EF 50mm f/1.8 II; <https://www.usa.canon.com/internet/portal/us/home/products/details/lenses/ef/standard-medium-telephoto/ef-50mm-f-1-8-ii>, last accessed 18.08.18

³Xiaomi Mi Band 2; <https://www.mi.com/en/miband2/>, last accessed 18.08.18



Figure 4.2: Cold-pressor test experiment setup.

are given to the participants to not speak and look into the camera while completing the stress phase (the CPT). During the experiment the subjects are alone in a room due to unintended social interaction (e.g. speaking, noises) when the instructor is present. After completion of the stress phase and time for resting, the second neutral phase is conducted. In the neutral phase, the subject are asked to look into the camera without performing any tasks or interaction. To get feedback from stress conditions, the subjects watch the previously recorded stress videos of themselves and commented stressful moments. Feedback and time of these stressful moments are noted for afterward annotation of the data. At the end of the procedure the participants are informed about the general aim of the experiment and their contribution.

4.1.3 Stress Phase

The stress inducing phase of the conducted experiment is performed with the help of the so called *cold-pressor test (CPT)*. This test is used in medicine [160], neuroscience [159] and psychology [161]. Further, the CPT is one of the most frequently used stress protocol in humans [161]. The CPT works as a physiological stressor and activates the major stress systems of the body, the sympathetic nervous system and the hypothalamus-pituitary-adrenal axis [159, 161]. Subjects performing the CPT are asked to put their left arm in ice-cold water (0° to 1°C) to elicit stress conditions. The participants are given the instructions to submerge the arm for as long as possible, while looking into the camera. Moreover, they are told to emerge their arm when reaching their pain threshold.

4.2 Datasets

The first dataset is gathered in a conducted experiment with the help of 22 adults (10 women, 12 men). Participants are between the age of 18 and 52 (27.3 ± 8.3 years). The previously outlined experiment protocol is applied, which leads to videos in *stressed* and *non-stressed* emotional states. Duration of video sequences range from 30 seconds to 11 minutes with an average video duration of 3:05 minutes. Videos are recorded at 50 FPS with a resolution of 1280×720 pixels. In addition to the videos, heart rate (HR) measurements with the use of a wristband device and feedback of participants are acquired. The annotations of the videos are applied frame-wise by assigning the labels *stressed* or *non-stressed*, which leads to 210901 frames labeled as non-stressed and 41070 frames labeled as stressed. Indicators for data annotations are the acquired HR measurements and feedback of the subjects. In Figure 4.3 example images of faces from three female and three male participants under stress conditions are illustrated. Mouth movement, eye movements and an overall stressed impression characterize these images.

A study conducted with 23 participants (7 women, 16 men) at the age of 45.1 ± 10.6 years characterizes the second dataset [1]. Recorded videos in the dataset include subjects in *neutral*, *stress/anxiety* and *relaxed* emotional states. These states are induced by experimental phases including social exposure, emotional recall, stressful videos and stressful image/mental task. Videos are captured with a resolution of 526×696 pixels and at 50 FPS. The duration of these videos are between 0.5, 1 and 2 minutes. In contrast to the first dataset, annotations in this dataset are applied for each video. Furthermore, this dataset consists of facial landmarks due to privacy protection.

4.3 Data Preprocessing

Data preprocessing is performed on the captured video sequences and was done image-wise. Video preprocessing includes histogram equalization for contrast enhancement, color space conversion, tilting correction and ROI detection. The color conversion from RGB to gray values and histogram equalization is applied to increase face and facial landmark detection accuracy and speed. After these first preprocessing steps, face and landmarks are detected. With the extracted landmarks, the tilting correction can be calculated. This correction is performed in order to optimize subsequent ROI extraction. For the tilted faces a rotation correcting the tilt is performed. To this end the bridge of the nose is assumed to be a straight line. The angular offset between the nose bridge line and a vertical line on the frontal plane, starting at the nose tip, are used for correction of the tilted faces. Figure 4.4 illustrates the tilting correction in an example. Another preprocessing step is the ROIs detection, which is required for afterward mouth and heart activity feature extraction. Extracted ROIs comprise of mouth, upper and lower mouth as well as a part of the forehead region. The mouth ROIs enable extraction of mouth motions. The forehead ROI allows the extraction of heart rate activity as McDuff et al. [33] describes. All ROI extraction is based on facial landmarks.



Figure 4.3: Facial image examples of participants in stressed conditions from the video sequences of the conducted experiment protocol.

4.4 Face Detection

In facial analysis face detection from image/video recordings is a fundamental step. For automated facial expression analysis (AFE) face detection is an essential preprocessing step [49]. Face detection algorithms can be categorized into knowledge based, feature invariant, template matching and appearance based. These distinct algorithms utilize features, such as edges, colors, size, shape information as well as modelling and classification methods as Gaussians, ASMs, eigenvectors, shape templates, HMMs, support vector machines (SVMs) and more [147]. Recent face detection is based on deep learning approaches such as CNNs [154, 155]. For the detection of faces in this work, the histogram of oriented gradients (HoG) algorithm, a popular appearance-based face detector is chosen. The choice falls on the HoG algorithm due to the availability in common frameworks, the computational simplicity, the performance and the frontal and near frontal faces in the datasets [49, 72]. The face detection used in the developed solution implements a HoG feature descriptors with linear classifiers and image pyramids as Felzenszwalb et al. [162] describe.

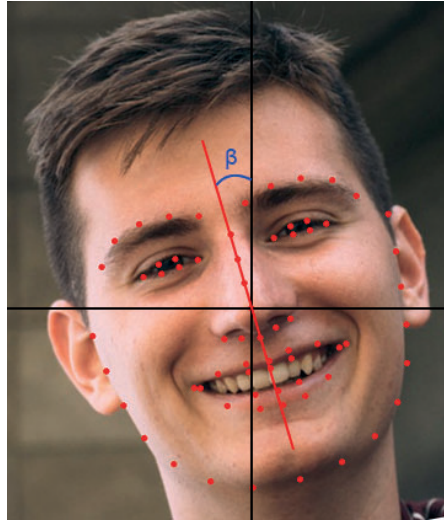


Figure 4.4: Tilting correction example image. The image is rotated clockwise by the angular offset β with the assumption that the nose bridge is always vertical in relation to the frontal plane.

4.4.1 Histogram of oriented gradients

The histogram of oriented gradients (HoG) descriptor is based, as the name suggests, on local histograms of oriented gradients and shows good performance for object detection. These histograms are normalized and computed over a dense grid of the image. The general idea of the HoG descriptors is that local object appearance and shape can be characterized by the distribution of local intensity gradients or edge direction. This can even work without accurate knowledge of edge positions or corresponding gradient. The implementation of the HoG method is done by dividing the image window into small spatial regions ("cells"). For each cell then a local one-dimensional histogram of gradient direction or edge orientation over the pixels of the cells will be calculated. The histograms of the cells are accumulated, normalized and as combined feature vector used for classification. [72]

Suleinman and Sze [163] outline the extraction of the HoG descriptor in an example. In Figure 4.5 the steps of feature extraction are illustrated. The input image is divided into 8×8 pixels patches (cells). On each of these pixels the horizontal and vertical gradient is calculated with the gradient filter $[-1 \ 0 \ 1]$. Further, the orientation and magnitude of the gradient are calculated from the horizontal and vertical gradients. From the orientation and magnitude then a cell histogram consisting of 9 bins is generated. A cell histogram represents the gradient orientations. As a next step, histogram normalization is performed by the values of the neighboring cells. The normalization is applied over a block of 2×2 cells and increases the robustness to lighting variations and to texture. After accumulation of the blocks to a HoG feature vector (e.g. 36 values), a linear SVM classifier can be trained to identify as object/non-object. Variations of the HoG

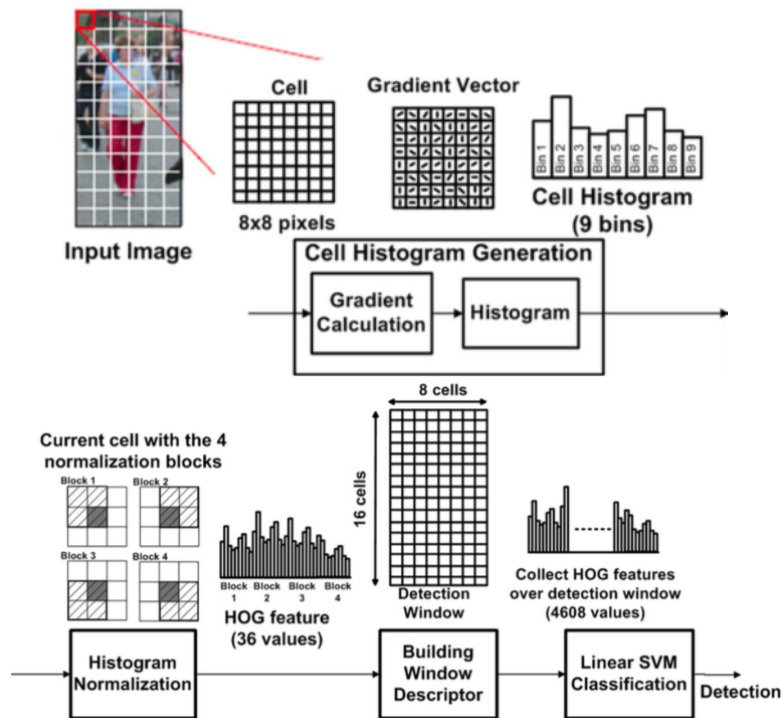


Figure 4.5: Object detection algorithm based on HoG features [163]

descriptor extraction method can be in the cell size, overlapping cells, block size for gradient normalization and other parameters. [163, 72]

4.4.2 Deformable Part Models

Felzenszwalb et al. [162] implement deformable part models for object detection given HoG features with a linear classifier and image pyramids. Characteristics of the deformable part models (DPMs) are its dense feature representation, it is a discriminative non-probabilistic model, uses templates, has explicit structures and performs translation and scale. In their approach a deformable part-based model was defined by a "root" filter, additional part filters and deformation models (see Figure 4.6). The so called root filter is analogous to Dalal and Triggs filter, which uses a sliding window and is applied at all positions and scales of an image [72]. In addition to the root filter, part filters allow to extract features at distinct spatial resolution relative to the features captured by the root filter. The part filter feature extraction is applied on a standard image pyramid and allows to model visual appearance at different levels and resolution. The deformation models represent the costs, which penalizes parts that are far away from where they are supposed to be.

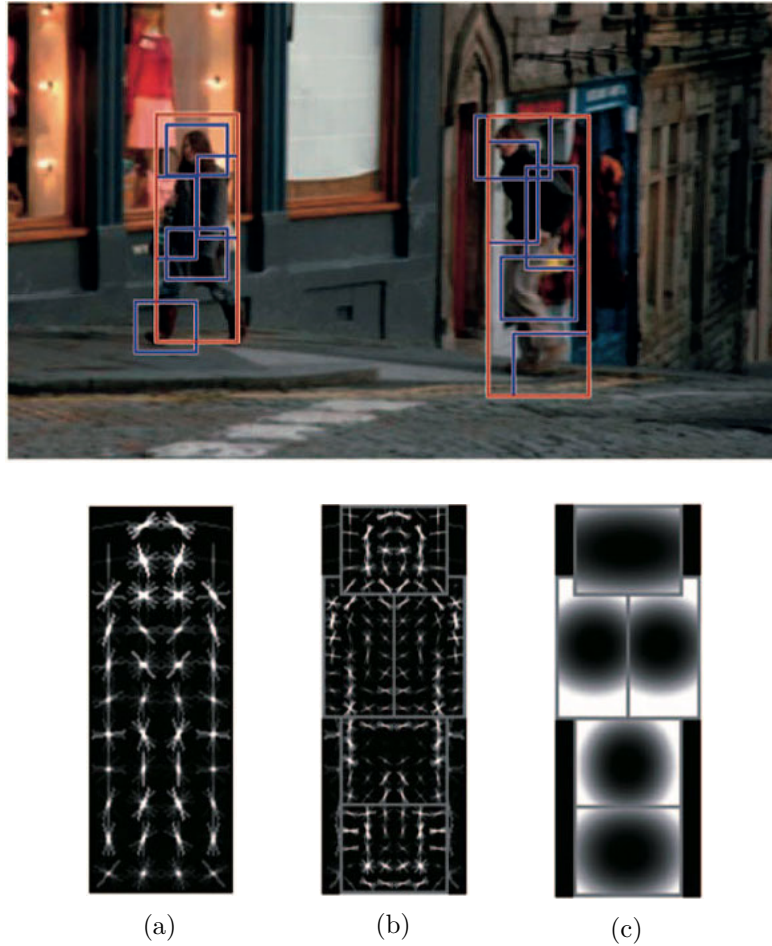


Figure 4.6: Detection with the help of a person model. The model is characterized by (a) a coarse root filter, (b) part filters in several resolutions and (c) the spatial deformation model. Weights are defined by the filters for the HoG features. The visualization of the deformation model represents the costs of placing the center of a part at distinct locations relative to the root. [162]

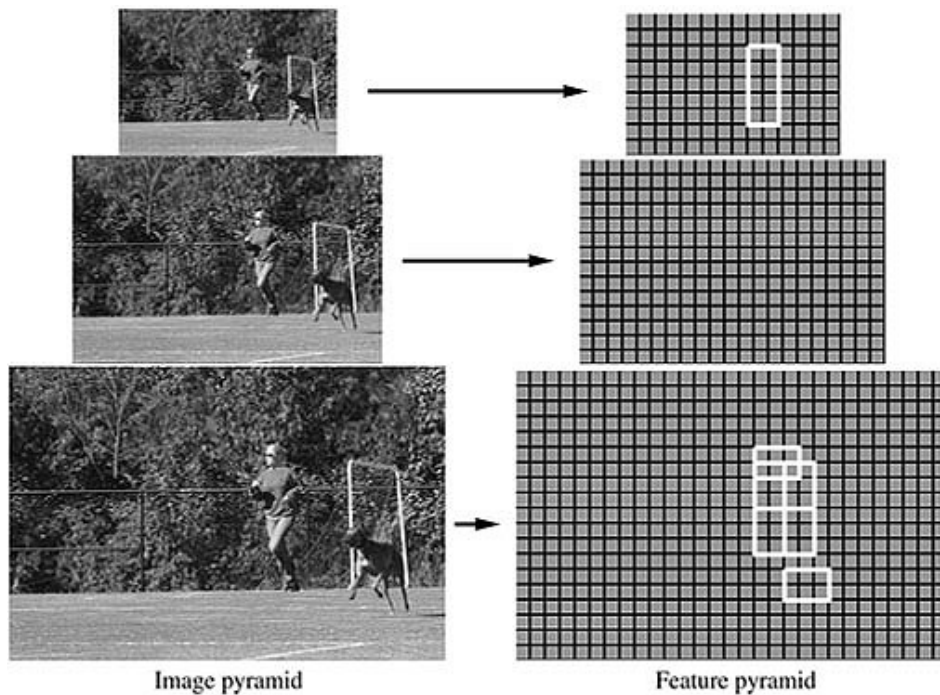


Figure 4.7: A feature pyramid constructed from different resolutions of the image pyramid. A person is instantiated within the feature pyramid. At the top of the pyramid the root filter is placed, the part filters are located at twice the spatial resolution of the placement of the root. [162]

The root filter and the part filters characterize the deformable parts model. A deformable parts model represents an object which can have multiple appearances due to movable (deformable) parts. With deformable the configuration of the parts and how they are positioned is meant, the parts are not deforming or changing their appearance. For the detection of an entire object it is covered by the root filter. The part filters then cover smaller parts of the object at higher resolution. As illustrated in Figure 4.7, the root filter defines a detection window in which the part filters try to cover the smaller parts at lower levels of the image pyramid.

As an example, we can consider to build a model of the face. The root filter can model the face boundaries by coarse resolution edges and the part filters could characterize details such as nose, mouth and eyes. A deformable parts model for an object with n parts is formally defined by $(F_0, P_1, \dots, P_n, b)$. F_0 is a root filter with n part models where P_i is a model for the i -th part and b is a real-valued bias. To detect objects with the help of the deformable parts model a score of the location of the root filter and the part filters is calculated. A higher score represents a hypothesis that at the given location an object is present. The calculation of the score is done given the following formula 4.1:

$$score(p_0, \dots, p_n) = \sum_{i=0}^n F_i * \phi(H, p_i) - \sum_{i=0}^n d_i * \phi_d(dx_i, dy_i) + b \quad (4.1)$$

A score is the sum of the filter scores, minus the deformation costs. In the formula, the score is calculated over all locations, where p_0 is the location of the root filter and p_1 to p_n are the part filters locations. F_i are the filters, $\phi(H, p_i)$ are the HoG features at the location and pyramid level p_i . The second term of the formula describes the deformation costs. The deformation parameters d_i is a four-dimensional vector and $\phi_d(dx_i, dy_i)$ is the displacement of the part i relative to its anchor position. The bias b is characteristic for all SVMs.

Object detection is performed on the overall score of a root location defined by the root filter. The best possible placements of the parts is computed according to 4.2:

$$score(p_0) = \max_{p_1, \dots, p_n} score(p_0, \dots, p_n) \quad (4.2)$$

As a result, the high-scoring root locations by computing also the scores of the parts define the detections. These locations are found by a sliding-window approach calculating the scores. Further, a distance transform step is performed for detection. With this step, the contribution of the parts inside of the root detection window are included in the score. The response of the distinct parts are calculated for each part filter i in the l -th level of the feature pyramid.

$$R_{i,l}(x, y) = F_i * \phi(H, (x, y, l)) \quad (4.3)$$

The response $R_{i,l}$ 4.3 of a part filter is calculated with the formula above. The first part F_i is the part filter. The second part of the formula $\phi(H, (x, y, l))$ is the HoG at the place x, y at level l in the pyramid. These part responses are then transformed given the following formula 4.4, where root is at (x, y) .

$$D_{i,l}(x, y) = \max_{dx, dy} (R_{i,l}(x + dx, y + dy) - (d_i * \phi_d(dx, dy))) \quad (4.4)$$

Given the root location (x, y) all possible displacements (dx, dy) of the part responses and the possible deformation costs are calculated. For all the possible displacements from the anchor position (x, y) the maximum score is selected. This calculation is done for every part in parallel.

With the training of the model, the unknown model parameters $\beta = (F_0, \dots, F_n, d_1, \dots, d_n, b)$ including root filter, part filters, deformation vectors and bias are learned. Positive training example with labeled bounding boxes are used for learning. Latent (unknown) variables such as the locations of the parts of the object, are not labeled in the training example and have to be learned. The classifier scores an example x by 4.5:

$$f_\beta(x) = \max_{z \in Z(x)} \beta * \phi(x, z) \quad (4.5)$$

The $z \in Z(x)$ are the possible places the parts can be, where z are the latent values and $Z(x)$ are possible latent values for example x . The term $\phi(x, z)$ is the HoG for the latent values z and an example x . Aim of the training is to identify the unknown variables β . To this end an objective function is minimized, which is comparable to SVM learning. The objective function is the following 4.6:

$$L_D(\beta) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i f_\beta(x_i)) \quad (4.6)$$

The β value is similar to the w in SVMs, with the exception that β is a filter (vector), consisting of root filter, part filters etc.. The last part of the formula $\max(0, 1 - y_i f_\beta(x_i))$ is the standard hinge loss with a constant regularization term C . The training examples $D = (\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle)$ are labeled as $y_i \in -1, 1$ with -1 for a negative example with no object and 1 positive example containing an object. At the beginning of training, a latent SVM the β values are initialized by assuming that all parts are at a fixed location z . The iterative optimization tries then to find the new best location z while β is fixed. This step is similar to the detection step mentioned before. In the next iterative step then z is fixed and β is optimized. This procedure is done until the iterative optimization terminates. [162]

4.5 Facial Landmark Detection

Facial landmark detection is developed to automatically detect the locations of facial landmark points. Distinct facial analysis task rely on facial landmark detection, hence landmark detection is of importance. Algorithms for facial landmark detection can be categorized into holistic methods, constrained local model (CLM) methods and regression-based methods. Holistic methods explicitly build models which represent the global facial appearance and shape information. CLMs utilize the global shape model but create local appearance models. The regression-based methods implicitly describe facial shape and appearance. Further, these methods do not explicitly build a global face shape model. Besides these three categories, more recent techniques perform landmark detection with deep learning and global 3D shapes. The best performing algorithms for facial landmark detection are shown to be regression-based and deep learning-based regression methods. Overall, there is a trend to use deep learning approaches for face detection and landmark detection instead of traditional methods. [99]

The facial landmark detector utilized in the developed solution is based on deformable part models with a pose estimator based on an ensemble of regression trees and a sparse subset of pixel intensities [162, 164]. This facial landmark detector is chosen, as in the case of the face detector, due to the availability in common frameworks and pre-trained models [49]. The training of the landmark detector is performed on the iBUG 300-W face landmark dataset [165]. This dataset consists of 300 Indoor and 300 Outdoor facial images captured under totally unconstrained conditions. Large variations of the images include identity, expression, illumination conditions, pose, occlusion, and face size. Further, a large percentage of the images are partially-occluded and faces show a

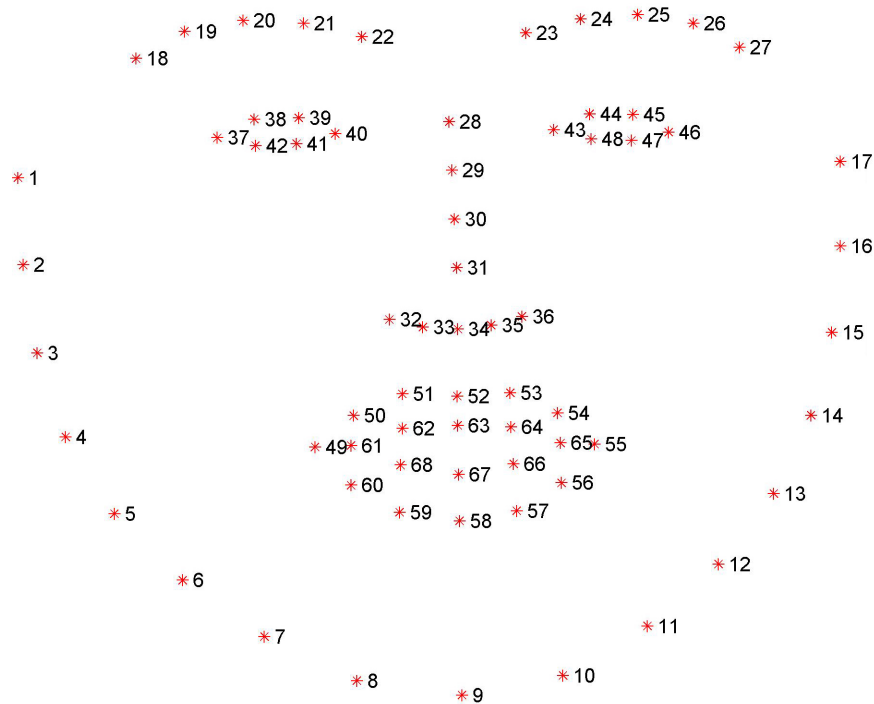


Figure 4.8: The 68 points mark-up annotation of the iBUG 300-W dataset [165]. Jaw line, eyebrows, eyes, nose and mouth points are annotated. [166]

large variety of expressions besides more common ones, such as *neutral*, *smile*, *surprise* or *scream*. Image annotations of the iBUG 300-W database are applied semi-automatic and used the 68-points mark-up, illustrated in Figure 4.8. The detection of facial landmarks in the implemented solution is performed frame-wise on image sequences and on frontal or nearly frontal faces. Extracted landmark points are used for facial feature extraction.

4.6 Eye Related Features

Extracted features of the eye region comprise of eye aperture as well as the blinking rate. Blinking of a person can occur due to reasons such as an internal or external reflex or voluntarily [32]. Blinking also increases with emotional arousal including anxiety and stress levels [39]. Further states affecting the blink rate can be psychological health issues as well as environmental conditions (e.g. lighting, temperature) [167, 168]. Another eye related feature is the eyelid closure, which is shown to be more significant in anxious persons as compared with non-anxious individuals [38].

The extraction of the eye aperture and blinking rate is performed with the help of facial landmarks. Especially the eyes surrounding landmarks 37 to 48 are of interest for feature extraction. The eyeball is segmented using the six discrete landmark points for each eye. To determine the eye aperture, two-dimensional coordinates (x_i, y_i) of

the six landmarks are used. After calculation the aperture of both eyes, the mean is calculated. Computation of the eyelid aperture is done with the area formula of the simple polygon 4.7:

$$A = \frac{1}{2} \sum_{i=1}^N |x_i y_{i+1} - y_i x_{i+1}| \quad (4.7)$$

The area A is computed with $N = 6$ and the convention that $x_{N+1} = x_1$, $y_{N+1} = y_1$. For machine analysis the eye aperture is also averaged as a total over all frames of a video sequence as well as over a sliding window.

The second eye related feature is the blinking rate. Eye blinks can be seen as a sudden aperture decrease in the eye aperture plot (see Figure 4.9). Blinking is detected with the eye aspect ratio (EAR) as Soukupová and Čech [169] outline. In their method the aspect ratio between height and width of the eye is computed. The following formula 4.8 outlines the calculation of the EAR

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}, \quad (4.8)$$

where p_1, \dots, p_6 are the 2D landmark locations around the eyes. Figure 4.9 depicts the landmarks and EAR for several frames of a video sequence. When an eye is open, the EAR is constant and while closing an eye the EAR is getting near to zero. Further characteristics of the EAR include the head pose insensitivity, fully invariant to uniform scaling of an image and in-plane rotations. The EAR is computed for both eyes and averaged. This is done due to that blinking is performed by both eyes synchronously. Computation of the EAR is performed for each frame. A blink is then determined by a threshold, if the EAR is below a specified value a blink is detected. The choice of the threshold is chosen on manual inspection of the EAR signal. Similar to the described blink detection from the EAR value, a threshold-based approach with the eye aperture as input value is implemented. The threshold is chosen as the 15th percentile of the eye aperture. Further, to prevent false positives of blink detection, a custom filter function based on the average blinking duration of 100-400 ms is implemented [131, 1]. Closing the eye for a longer period resulting in multiple detected blinks shall be filtered with this function. The number of blinks per minute is additionally used as a feature for classification.

4.7 Mouth Related Features

Mouth related features such as particular lip movements and deformations as well as mouth opening, are linked to stress/anxiety conditions [32]. Asymmetric lip deformations are shown to be more present under high stress levels [34]. Further, the frequency of mouth openings can be associated with stress levels and higher cognitive workload [170].

To extract mouth related features the optical flow algorithm is used. With this algorithm it is possible to extract movements from image objects between two consecutive

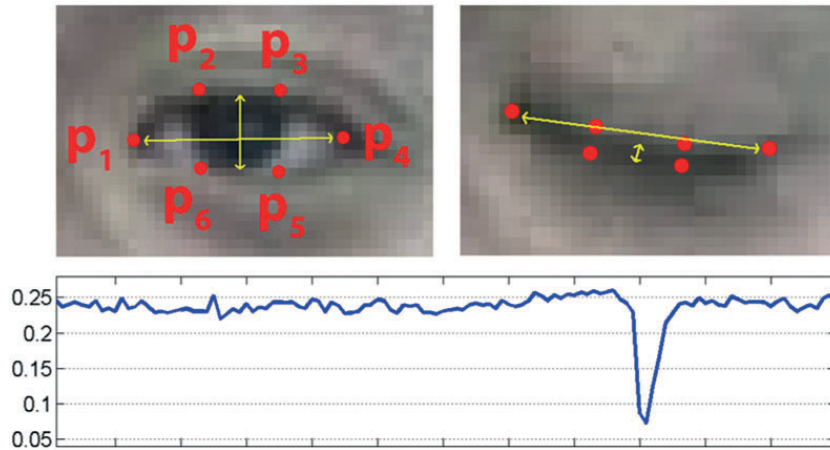


Figure 4.9: Calculation of the EAR from eye landmarks. EAR is defined as a ratio of height to width. In the picture below a graph of the EAR values for multiple frames is illustrated, a blink is shown as a value close to zero. [169]

frames. As a feature the general mouth motion is extracted. The feature extraction is applied after mouth ROI determination. For each frame then the maximum magnitude or maximum velocity is calculated. The calculation of the mouth motion is performed by using dense optical flow as Farneback [171] describes. Further features calculated from the mouth movement include the mean, median and variance of mouth movement over a sliding window, variance of time intervals (VTI) between adjacent motion peaks, mean, median and variance of variance of time intervals (VTI) as well as the total mouth movement of a video sequence. Moreover, these features were also calculated for the upper and lower mouth. VTI between mouth motion peaks shall describe the reduced rhythmicity of lip movements during increased levels of stress [1].

4.7.1 Optical Flow

The optical flow method is developed by Horn and Schunck [172]. In general, the aim of the optical flow algorithm is to describe motions in images. For two subsequent frames, the two-dimensional vector or optical flow for each pixel is calculated. These vectors describe the motions or displacements in the images from one frame to the next frame. Figure 4.10 illustrates this in an example. The computation of the optical flow is done pixel-wise and can be described as in 4.9:

$$f(x, y, t) = f(x + dx, y + dy, t + dt) \quad (4.9)$$

The intensity of each pixel $f(x, y, t)$ is calculated for two frames, in which the object in the subsequent image at the time $t + dt$ moves by distance dx and dy . For a higher temporal resolution it is assumed that the intensity or brightness of the pixels is constant and the motion is smooth. With this brightness constancy assumption, we can describe the small brightness with a Taylor series, which results in the optical flow equation 4.10.

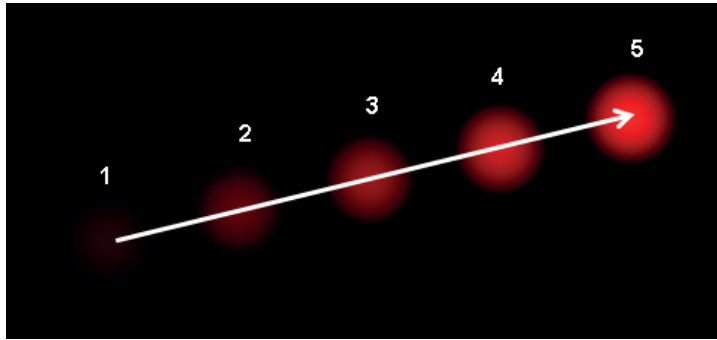


Figure 4.10: This example illustrates the optical flow for a ball moving in five consecutive frames. The optical flow or displacement vector is shown by the arrow.⁴

$$f_x u + f_y v + f_t = 0 \quad (4.10)$$

In the optical flow equation u and v are the change dx and dy over the time dt . Further, f_x and f_y are image gradients and f_t is the gradient along the time. These gradients can be found, however to solve the optical flow equation with the two unknown variable u and v several methods can be applied. The assumption of brightness constancy and smoothness constraint are crucial. These constraints are included into the optical flow formula 4.11

$$\int \int \{(f_x u + f_y v + f_t)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2)\} dx dy \quad (4.11)$$

The double integral is used to calculate the optical flow for each pixel. In the first part of the formula the brightness constancy is described. Ideally the brightness constancy will be zero due to the previously outlined assumption. The second smoothness constraint states that (nearly) all pixels are moving in a similar motion. With these constraints it is possible to compute the optical flow by formulating it as a minimization problem. This minimization problem can be solved by using variational calculus. Further different methods to find the unknown variables and compute the optical flow can be used, e.g. the method described by Lucas-Kanade [173]. [172]

The dense optical flow method is a variation of the optical flow and produces a dense motion field able to capture spatially varying motions (e.g. lip movements). In the approach Farneback [171] describes the velocity vector (optical flow) of each pixel is computed with the help of a two-frame motion estimation algorithm. In order to calculate the dense optical flow the neighborhood of each pixel is approximated with a polynomial. This polynomial expansion is applied on two-frames. The neighborhood of both frames is described by quadratic polynomials outlined in 4.12.

⁴Image from https://en.wikipedia.org/wiki/File:Optical_flow_example_v2.png, last accessed 04.09.18

$$f_1(x) = x^T A_1 x + b_1^T x + c_1 \quad f_2(x) = x^T A_2 x + b_2^T x + c_2 \quad (4.12)$$

These quadratic polynomials, where A is a symmetric matrix, b a vector and c a scalar, is estimated from a weighted least square fit to the signal values in the neighborhood. The two-frames characterized by quadratic polynomials are related by a global displacement d . This relationship can be described with the the following equation 4.13

$$f_2(x) = f_1(x - d) \quad (4.13)$$

and solving for displacement d results in 4.14

$$d = -\frac{1}{2}A_1^{-1}(b_2 - b_1) \quad (4.14)$$

The displacement d can be calculated from the coefficients A_1 , b_1 and b_2 , if A_1 is singular. When implementing the dense optical flow, the polynomial expansion coefficients ($A_1(x)$, $b_1(x)$, $c_1(x)$, $A_2(x)$, $b_2(x)$, $c_2(x)$) are first calculated for the two frames and the approximation for $A(x)$ is made from 4.15.

$$A(x) = \frac{A_1(x) + A_2(x)}{2} \quad (4.15)$$

With the equations above and with 4.16:

$$\Delta b(x) = -\frac{1}{2}(b_2(x) - b_1(x)) \quad (4.16)$$

the primary constraint 4.17 can be obtained, where $d(x)$ is a spatially varying displacement field. This equation 4.17 is solved over a neighborhood of each pixel.

$$A(x)d(x) = \Delta b(x) \quad (4.17)$$

4.8 Head Related Features

Head movements as indicator for stress are used in current literature [20]. During stressful conditions reports, it is shown that head movements are more frequent and rapid [31, 32]. Overall greater head motions are associated with stress [40]. In order to extract head movements stable portions of the face such as nose landmark points are tracked. These specific facial landmarks comprise of the points 28, 31 and 33-35 of the 68 points mark-up illustrated in Figure 4.8. With the selected landmarks it is possible to determine translations, rotations and head movements. To extract the head related features, the Euclidian distance is calculated for each frame. The first frame of the analyzed video sequence is chosen as reference to compute the distances of the head. Movement of the head is defined as in 4.18

$$movm = \frac{1}{5} \sum_{k=1}^5 \|p_k - p_k^{ref}\| \quad (4.18)$$

with the landmark points $p_k, k = 1, \dots, 5$ and the Euclidean norm $\|\cdot\|$. In addition to the head movement, the speed/velocity of the head motion is calculated as in the following 4.19

$$velocity = \frac{1}{5} \sum_{k=1}^5 \|p_k(t) - p_k(t-1)\| \quad (4.19)$$

where $p_k(t)$ is the landmark point at the time t . As in the case of the head movement, the velocity is calculated frame-wise over a video sequence. Furthermore, head related features such as the total mean head movement, total mean head velocity for sliding windows are computed. These include also the head movement and velocity for the x and y axis.

4.9 Heart Rate from Facial Video

The heart rate from facial video is extracted from camera-based photoplethysmography (PPG) as Poh et al. [21] describe. Camera-based PPG measures variations in the reflected light from the skin. These variations can be used to establish the underlying blood volume pulse (BVP) for computation of the heart rate [21]. In the different color channels of an image a mixture of the reflected plethysmographic signal is contained. Each color sensor captures a mixture of the original signal source with slightly different weights. These three sensor signals can be formulated as $y_1(t)$, $y_2(t)$ and $y_3(t)$, which describe the amplitudes of the signal at the time t . The underlying source signals from each channel are denoted by $x_1(t)$, $x_2(t)$ and $x_3(t)$. To compute the source signals an independent component analysis (ICA) is employed. The ICA model assumes that the observed signals are linear mixtures of the sources, i.e.,

$$y(t) = Ax(t) \quad (4.20)$$

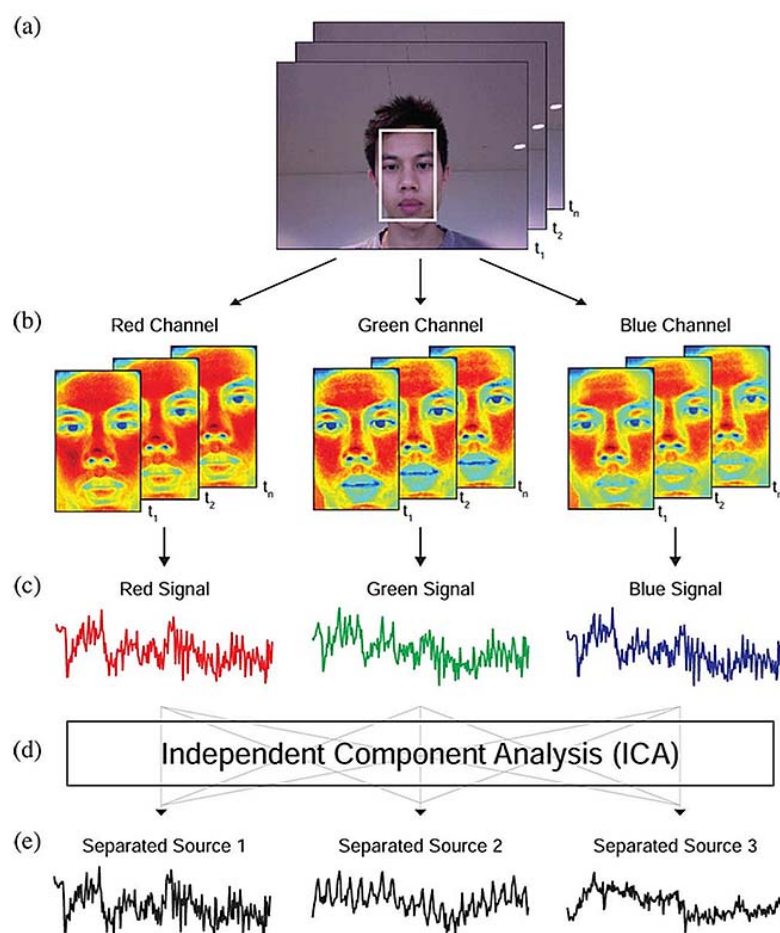
These linear mixtures are outlined in 4.20, where $y(t) = [y_1(t), y_2(t), y_3(t)]^T$, $x(t) = [x_1(t), x_2(t), x_3(t)]^T$ and the mixture coefficients a_{ij} are contained in the 3×3 matrix A . With the ICA a demixing matrix W is approximated, which is the inverse of the original mixture matrix A . The output of the ICA can be described as the following 4.21

$$\hat{x}(t) = Wy(t) \quad (4.21)$$

where an estimate of the vector $x(t)$ contains the underlying source signal. ICA assumes that the sources are independent and maximizes W , the non-Gaussianity of each source. The maximization of W is performed by minimizing a given cost function.

The implementation of facial PPG is illustrated in Figure 4.11. As a first step of extraction, the facial ROI is extracted for each frame of a video sequence. For each pixel of the ROI the red, green and blue values is averaged separately resulting in a raw signal $y_i(t)$, where $i = 1, 2, 3$ for each color channel. These average values are then normalized with mean μ_i and standard deviation σ_i as following 4.22:

Figure 4.11: Implementation of the facial PPG extraction. (a) Facial ROI is detected. (b) For each frame the ROI image is split into the red, green and blue channels and spatially averaged. (c) The constructed raw signals of the previous step is normalized. (d) ICA is applied on the raw signals. (e) Results of the ICA are three independent and separated source signals. In this example the BVP is visible in the second source signal. [21]



$$y'_i(t) = \frac{y_i(t) - \mu_i}{\sigma_i} \quad (4.22)$$

The normalized signal are then used as input for the ICA based on the joint approximate diagonalization of eigenmatrices (JADE) algorithm [174]. The separated source signals of the ICA are then transformed and filtered. The frequency with the greatest power is then assumed to be the frequency of the heart rate. In order to smooth the heart rate signal, an average of the previous five calculated heartbeats is computed.

4.10 Machine Learning

In the last machine learning (ML) step of the presented processing pipeline 4.1, data preprocessing and classification of image data is performed. Preprocessing methods are applied on the extracted features. These methods include calculation of additional features, normalization, smoothing, filtering and imputation. Further, features are averaged due to the different annotations of the datasets and to provide additional information for machine analysis. Computation of these "*meta-features*" are based on the facial features described in the previous chapters. Created meta-features from these facial features are the total calculation of a feature defined as the mean feature value over all frames of a video. These total feature calculation are necessary to allow classification of the dataset [1] due to the video-wise annotation. In the case of blinking, a mean value of blinks per sliding window is calculated. Further calculated meta-features include mean, VTI between adjacent motion peaks and the variance of features. These meta-features are also calculated over sliding windows. A detailed overview of created meta-features is given in Table 4.1.

In addition to the calculation of meta-features, facial features are filtered, smoothed and imputed. Filtering is performed by simple thresholding and imputing the filtered value with the mean value of previous feature values. Smoothing is done with the Savitzky and Golay filter [175] on features showing high fluctuations and noise of landmark tracking. Further, instances of the dataset are deleted, when the landmark detector is not able to track the face and no features were extracted. Missing values are replaced by the respective mean of the feature. Normalization of the features is done for each video to allow comparison of the different values across subjects. As normalization method min-max scaling, normalizing the feature values between the interval [0;1], is applied. To balance imbalanced datasets for subsequent classification the more frequent *non-stressed* instances are undersampled. For the classification of the data distinct ML algorithms are employed and evaluated.

4.11 Implementation

The implementation of the developed stress detection solution is based on the processing pipeline outlined in Figure 4.1. Core parts of the solution are preprocessing, feature

Table 4.1: Meta-features calculated from eye, mouth, head and heart rate features. Feature computed over a sliding window of 5**, 15* sec.

Eye	Mouth	Head	Heart Rate
mean, median aperture*	mean, median, variance of movement*	mean movement (x, y-axis)*	median**
	VTI between adjacent motion peaks	mean velocity (x, y-axis)*	
	mean, median, variance of VTI*		
	mean, median, variance upper/lower mouth VTI*		

extraction and machine analysis. Image acquisition is not considered in the implementation. However, requirements on image data consist of RGB video files, frontal or near frontal facial images as well as certain image quality standards. The different processing steps are modularized and configurable over separate files. This leads to a more flexible solution allowing to perform distinct data operations, such as filtering, smoothing etc.. Further, extracted facial landmarks and features can be exported to files and statistical measurements (e.g. mean, variance) of features can be calculated. This file-based approach is chosen to provide flexibility in the use of different frameworks (e.g. for ML). Moreover, extracted landmarks or features enable faster processing, as these steps in the pipeline can be skipped. The specific solution was implemented in *python* and used the popular frameworks *OpenCV*⁵ and *dlib*⁶ for image processing. Face and facial landmark detection is performed with the provided algorithms in *dlib*. Image input, contrast enhancement and extraction of features of the mouth is handled with the help of *OpenCV* methods. Data preprocessing and analysis is performed with the *scikit-learn*⁷ ML framework. In addition to the stress detection implementation, a video player is developed to enable frame-wise data annotation of the captured stress dataset.

⁵OpenCV, <https://opencv.org/>, last accessed 01.09.18

⁶dlib, <http://dlib.net/>, last accessed 01.09.18

⁷scikit-learn, <http://scikit-learn.org>, last accessed 01.09.18

Results & Discussion

The evaluation of the developed stress detection solution is conducted on two datasets. These datasets comprise of stress data gathered in a conducted experiment as well as a dataset from [1]. Extracted features of the implemented solution such as eye, mouth, head and heart rate related features are discussed. Statistical analysis with the help of t-tests and box plots describe extracted features in more detail. Further, classification performance of distinct machine learning algorithms are presented. Performance scores and plots illustrate outcomes of ML algorithms. Moreover, a comparison of outcomes with results of stress detection solutions in literature is conducted.

5.1 Eye Related Features

The eye related features are calculated on both datasets. Most important features of the eyes comprise of eye aperture and blinking rate. To assess these features the calculations of both datasets are presented. For the conducted stress experiments a comparison between eye related features from videos of the *stress* and the *non-stress* phase is performed. The total eye aperture as mean feature value over all frames for each video is computed. In Figure 5.1 the box plot illustrates the distribution of the eye aperture for stress and non-stress phase. In the case of videos from the stress phase the mean eye aperture (\pm s.d.) is 274.9 ± 74.9 pixels² and for the non-stress phase 317.3 ± 84.5 pixels². The inter-quartile range of the two box plots overlaps, ranges are [196.4; 323.1] for the stress phase and [243.3; 385.9] for the non-stress phase. Differences can be seen for the non-stress data, where the box plot is situated higher than the box plot for the stress phase. This suggests overall more closed eyes in the stress-phase compared to the non-stress phase. A statistical significant difference of a decreased mean eye aperture in the stress phase can be found ($p < 0.05$).

Evaluation of the eye aperture of the second dataset [1] is performed based on the different stress inducing tasks in relation to the respective reference task. These distinct elicitation tasks are outlined in the following:

1. Social Exposure
 - 1.1 Neutral
 - 1.2 Self-description speech
 - 1.3 Text reading
2. Emotional Recall
 - 2.1 Neutral
 - 2.2 Recall anxious event
 - 2.3 Recall stressful event
3. Stressful images/Mental Task
 - 3.1 Neutral/stressful images
 - 3.2 Mental Task (Stroop)
4. Stressful videos
 - 4.1 Neutral
 - 4.3 Adventure/heights video
 - 4.4 Home invasion video

In Table 5.1 the mean eye aperture for each task is presented. Further, the box plot in 5.3 visualizes the mean eye aperture for each task with blue boxes representing the neutral reference task and brown boxes the stress inducing tasks. A statistical significant difference of the eye aperture between stress inducing tasks and corresponding neutral task can not be found. However, a decrease in the eye apertures for the task 1.2 vs 1.2, 4.1 vs. 3.2 is illustrated in Figure 5.3.

The second eye related feature, the number of blinks per minute (BPM), is extracted from the EAR signal. For blink detection a EAR threshold of 0.3 is chosen for the conducted experiment data. This threshold is selected after manual inspection and visualization of the EAR signal from the video data. The mean blink rate of the stress phase from the conducted experiment videos is 39.9 ± 20.1 BPM and of the non-stress phase 22.5 ± 24.0 BPM. The increased blink rate in the stress phase is statistically significant ($p < 0.05$) compared to the non-stress phase. In Figure 5.2 the differences of the blink rate between the two phase is shown. The inter-quartile range of the two box plots are [19; 57] for the stress phase and [4; 56] for the non-stress phase. The box plot for the stress phase is situated higher than the plot for the non-stress phase, which illustrates the differences in blinking behavior.

In the case of the second dataset, blink detection is performed based on the eye aperture and the 15th percentile as threshold. This extraction method is chosen due to the insufficient detection from the EAR signal. A statistically significant increased blink rate is found for the tasks 1.2, 3.1, 3.2, 4.3 and 4.4 in comparison to the corresponding

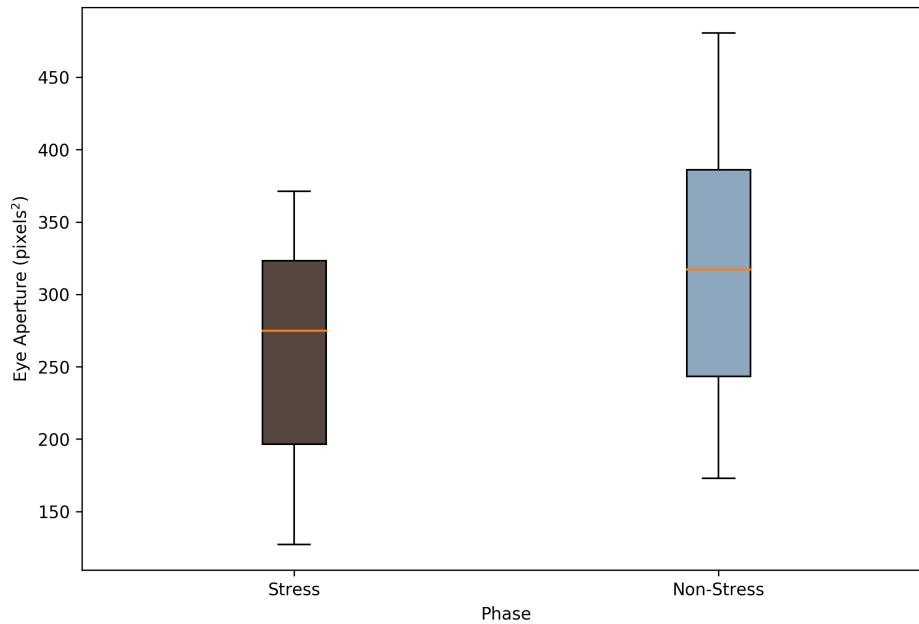


Figure 5.1: Box plot of the mean eye aperture (pixels²) for the stress and non-stress phase.

reference task ($p < 0.05$). An illustration of the mean blink rate for each task can be found in Figure 5.4. The box plots situated higher compared to the reference task show statistically measurable differences. Overall, blinking detection by a simple threshold-based approach has its limitations such as the strong requirements on the setup in regard to face-camera pose (head rotation), motion and more [169]. An approach based on a trained classifier for blink detection as outlined in [169] is preferable.

During the stress phase for the conducted experiment data a statistically decreased eye aperture is found. In outcomes from Bevilacqua et al. [20] also a statistically decreased eye aperture in the stress phase compared to the non-stress phase is present. The percentage of the mean eye aperture change between stress and non-stress phase is -13.3% for the conducted experiment data. Bevilacqua et al. states a percentage change of the eye aperture between -2.6% and -8.9% for the work-load stress phases. Furthermore, an increased eye blinking during the stress phase can be found, which is also suggested in outcomes from [1].

5.2 Mouth Related Features

The mouth related features are extracted with the use of the optical flow as well as with the calculation of the mouth area. Mouth motion is defined as the maximum magnitude of the optical flow for the conducted experiment data. In the case of the second dataset, mouth related features can not be extracted. The second dataset consists

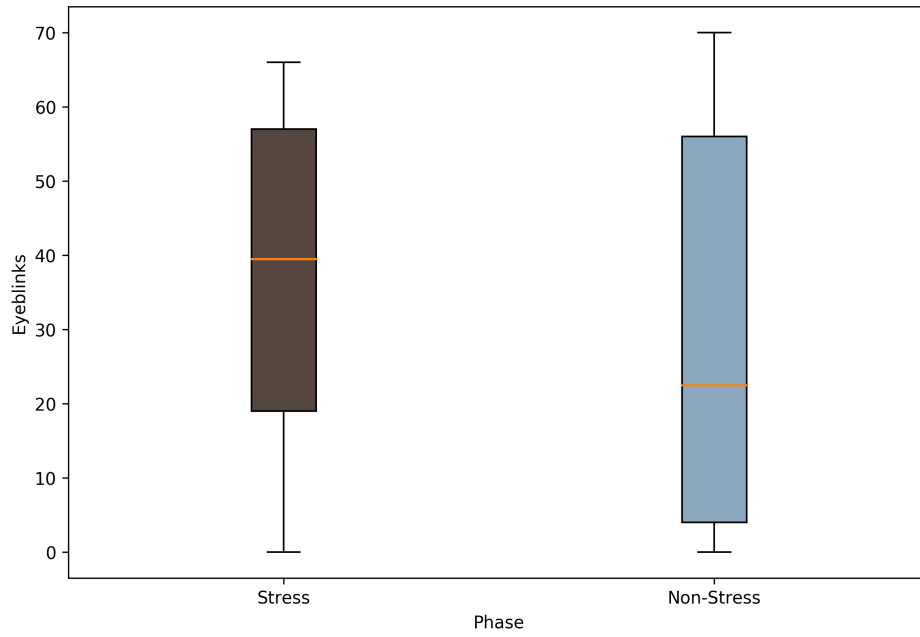


Figure 5.2: Box plot of the mean eyeblinks (in blinks per minute (BPM)) for the stress and non-stress phase.

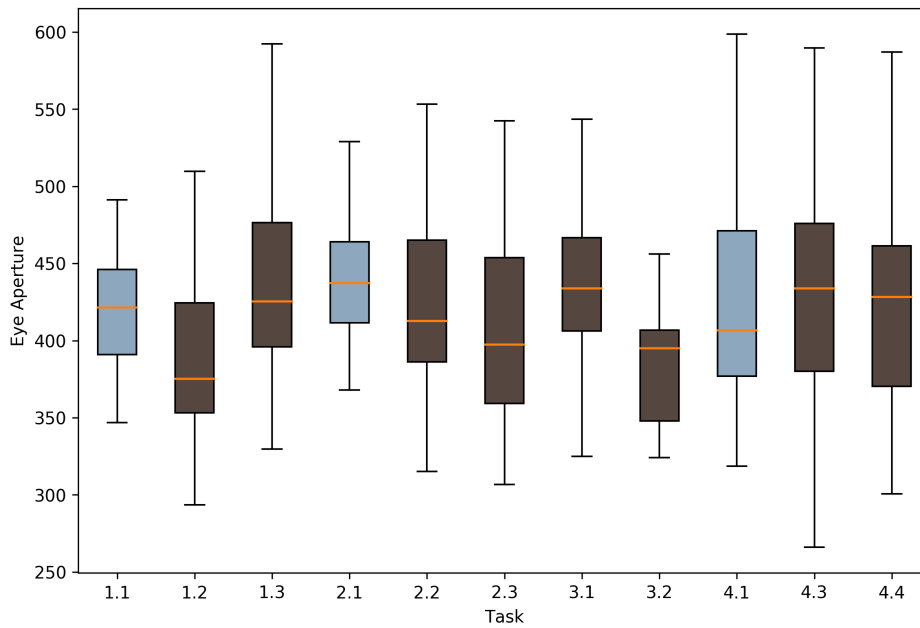


Figure 5.3: Box plot of the mean eye aperture for each emotional elicitation task (see 5.1).

Table 5.1: Mean (\pm s.d.) of eye aperture and blink rate for each emotion inducement task with corresponding reference (neutral) baseline.

Experimental Phase	Eye Aperture (pixels²)	Blink Rate (blinks per minute)
Social Exposure		
1.1 Neutral	435.1 \pm 67.0	55.5 \pm 27.54
1.2 Self-description speech	393.5 \pm 68.64	140.3 \pm 43.31*
1.3 Text reading	444.2 \pm 71.1	46.69 \pm 17.0
Emotional recall		
2.1 Neutral	443.5 \pm 64.3	49.82 \pm 30.41
2.2 Recall anxious event	430.4 \pm 78.5	56.60 \pm 26.48
2.3 Recall stressful event	407.8 \pm 80.7	52.0 \pm 23.72
Stressful images/Mental task		
3.1 Neutral/Stressful images**	435.2 \pm 60.5	118.75 \pm 46.14*
3.2 Mental Task (Stroop)**	407.8 \pm 80.7	135.92 \pm 40.18*
Stressful videos		
4.1 Neutral	430.1 \pm 77.1	58.78 \pm 31.54
4.3 Adventure/heights video	433.6 \pm 73.7	88.02 \pm 34.35*
4.4 Home invasion video	425.5 \pm 73.6	129.9 \pm 56.79*
**Reference task 4.1		* $p < 0.05$

of facial landmarks and does not contain image data due to privacy protection. Hence, the optical flow can not be extracted from this dataset. Extracted data from videos of stress and non-stress phase are averaged for each video and phase. Results of the conducted experiment data show a mean mouth movement of 8.17 ± 2.34 for the stress phase and 3.30 ± 2.22 for the non-stress phase. Statistically significant increased mouth movement is found between stress and non-stress phase ($p < 0.05$). In Figure 5.5 the differences of mouth movements in the two phases is depicted. The inter-quartile range of the two box plots are [6.68; 10.12] for the stress phase and [2.27; 5.40] for the non-stress phase. The box plot of the stress phase is situated higher than the plot for the non-stress phase emphasizing the discrepancy in mouth motion between the two phases. Further, mouth related feature values are presented in Table 5.2. These mean values comprise of mouth movement of upper and lower mouth, VTI between mouth motion peaks as well as mean variance of (upper/lower) mouth movement. Due to the correlation of upper and lower mouth movement as well as the overall more frequently occurring mouth movements these values show statistically significant differences between the two phases.

Overall, results indicate an increased mouth movement in the stress phase compared to the non-stress phase. Outcomes of Liao et al. [31] also state frequent mouth motion in the form of mouth openings in stress conditions. Significant differences between mouth

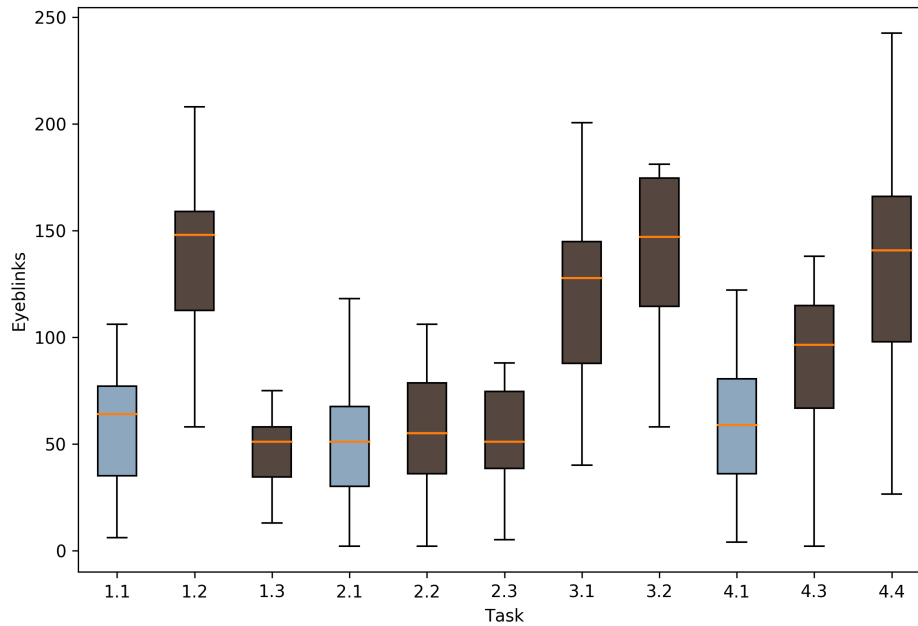


Figure 5.4: Box plot of the mean blink rate (in BPM) for each emotional elicitation task (see 5.1).

related features in stressful and non-stressful events are found in [20]. In their work a decrease of mouth movement during work-load (stressful state) is found. These differences of mouth motion between the presented results and [20] may be due to the distinct stress elicitation tasks (work-load vs. cold-pressor test). McDuff et al. [33] study shows distinct outcomes for different stress inducing tasks. Furthermore, the higher VTI of mouth motion peaks between stress and non-stress phase implies a reduced rhythmicity of lip movements associated with stress as outlined in [1].

5.3 Head Related Features

The head related features comprise of head movement and head velocity. Head movement and head velocity are calculated from Euclidian distances. Movement and velocity are measured in pixels and pixels/frame between facial landmarks from video frames. Evaluation of the head movement and head velocity is based on the mean movement/velocity over all frames of videos from stress and non-stress phase. The extracted head movement of the conducted experiment dataset demonstrates increased head motion of participants in the stress phase. In comparison to the head movement in the non-stress phase, increased head motion is statistically significant ($p < 0.05$). Mean head movement of videos from the stress phase is 35.4 ± 32.7 and 7.73 ± 22.6 pixels for videos from the non-stress phase. In Figure 5.8 the mean head movement is displayed in a box plot. The data of head movement is illustrated for each phase. The inter-quartile range of the data in the

Table 5.2: Mean (\pm s.d.) of mouth related features for stress and non-stress phase. All of the values show statistically significant differences between the phases ($p < 0.05$)

Mouth Related Feature	Stress Phase	Non-Stress Phase
Mean Mouth Movement	8.17 \pm 2.34	3.3 \pm 2.22
Mean Variance Mouth Movement	6.53 \pm 8.26	0.92 \pm 3.23
Mean VTI Mouth Movement	16.35 \pm 1.99	13.33 \pm 1.2
Mean Upper Mouth Movement	5.22 \pm 2.11	3.10 \pm 1.95
Mean Variance Upper Mouth Movement	2.99 \pm 8.81	0.77 \pm 2.34
Mean VTI Upper Mouth Movement	15.15 \pm 2.11	13.08 \pm 1.14
Mean Lower Mouth Movement	4.71 \pm 1.57	2.55 \pm 1.65
Mean Variance Lower Mouth Movement	3.68 \pm 5.27	0.85 \pm 2.26
Mean VTI Lower Mouth Movement	16.89 \pm 2.3	12.84 \pm 1.12

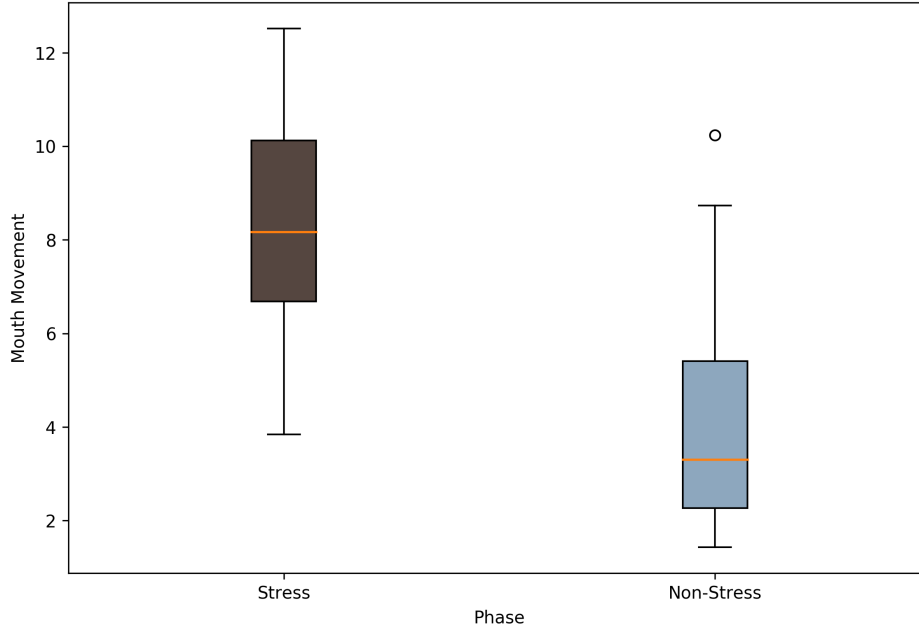


Figure 5.5: Box plot of the mean mouth movement for the stress and non-stress phase.

stress phase is [27.3; 64.0] and of the data in the non-stress phase is [2.90; 20.1]. The box plot of the mean head movement of the non-stress phase is shorter in comparison to the mean head movement of the stress phase. This suggests an agreement of head movements in the non-stress phase. Further, the plot in the stress phase is situated higher in comparison to the non-stress phase, which indicates more head movement in the stress phase.

The head velocity has also statistically significant differences between videos from stress and non-stress phases, where the velocity is increased in the stress phase. Mean head velocity of videos from the stress phase is 0.99 ± 1.44 and 0.61 ± 0.56 pixels/frame of videos from the non-stress phase. An illustration of the mean head velocity is displayed in Figure 5.7. In the shown box plot the characteristics of the head velocity in the individual phases is shown. The inter-quartile range of the head velocity in the stress phase is [0.62; 1.62] and of the non-stress phase is [0.51; 0.89]. The box plot of the head velocity of the non-stress phase is more compressed and overlaps with the head velocity of the stress phase in the range of [0.62; 0.89]. This indicates a same head velocity for the overlapping areas in both phases, which may results from non-movement periods of the head. However, the distribution of the head velocity of the stress phase stretches over a wider range when also comparing the whiskers of the data.

Results from the second dataset are outlined in Table 5.3. Statistically significant differences between emotion elicitation task and reference task have been found for the tasks 1.2, 1.3, 3.1, 3.2, 4.3 and 4.4 for the head movement in comparison the respective reference task. Figure 5.8 illustrate the head movement for each task. For the head velocity a statistically significant difference has been found for the 3.1 in comparison to the reference task 4.1. The head velocity for each task is displayed as a box plot in Figure 5.9.

The head movement of the stress phase are more prominent compared to the head movement in the non-stress phase. Frequent head movements in stress conditions are also reported in [31]. Furthermore, head movements tend to be increased in stress elicitation tasks of [1]. Bevilacqua et al. [20] also evaluated the head movements between a stressful (work-load) and non-stressful task. In their work, a decrease of head movement was found in the stress inducing task. The distinct stress experiment of [20] compared to the conducted stress experiment may lead to the different behavior of head movements. In addition to the head movement, the head velocity is increased during the stress phase compared to the non-stress phase. An elevated head velocity can be found for tasks by Giannakakis et al. [1].

5.4 Heart Rate from Facial Video

The extracted heart rate is evaluated for the conducted experiment dataset. Heart rate measurements for the second dataset are not established. The second datasets consists of facial landmarks, image information required for facial PPG is not provided by this dataset due to privacy protection. To compare the heart rate between the stress and non-stress phase, a mean heart rate over all videos in the respective phase is computed.

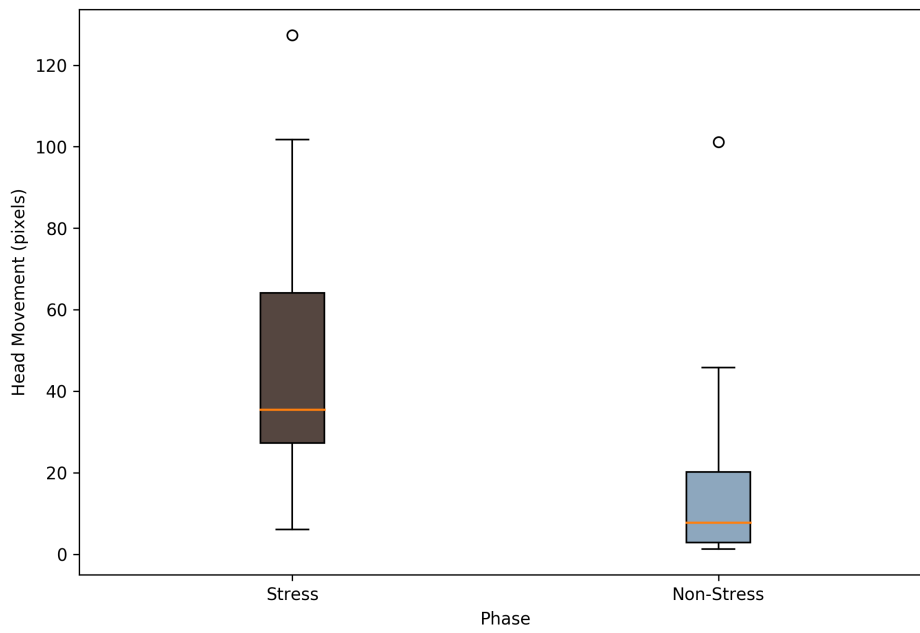


Figure 5.6: Box plot of the mean head movement (in pixels) for the stress and non-stress phase.

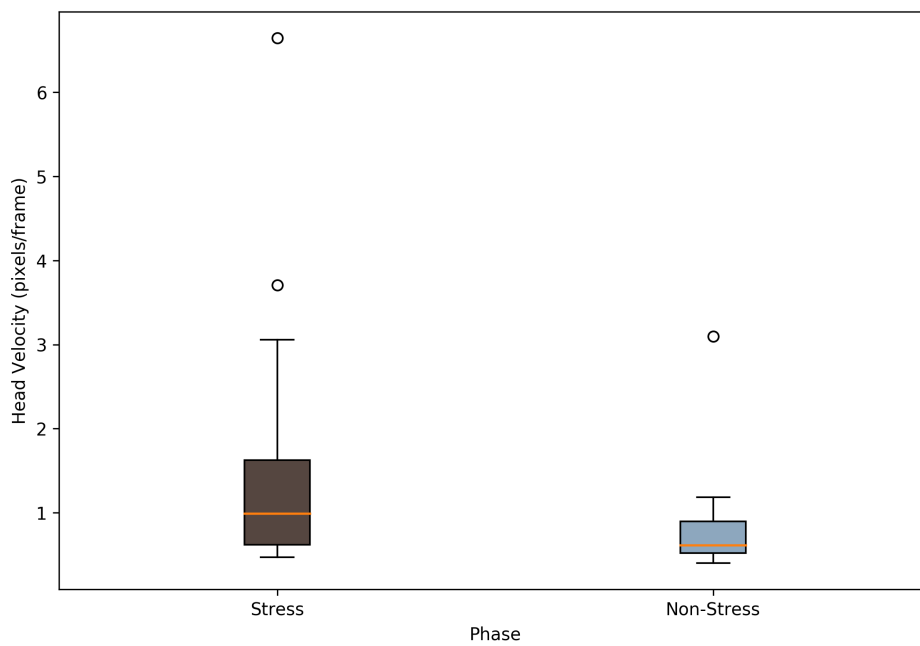


Figure 5.7: Box plot of the mean head velocity (in pixels/frame) for the stress and non-stress phase.

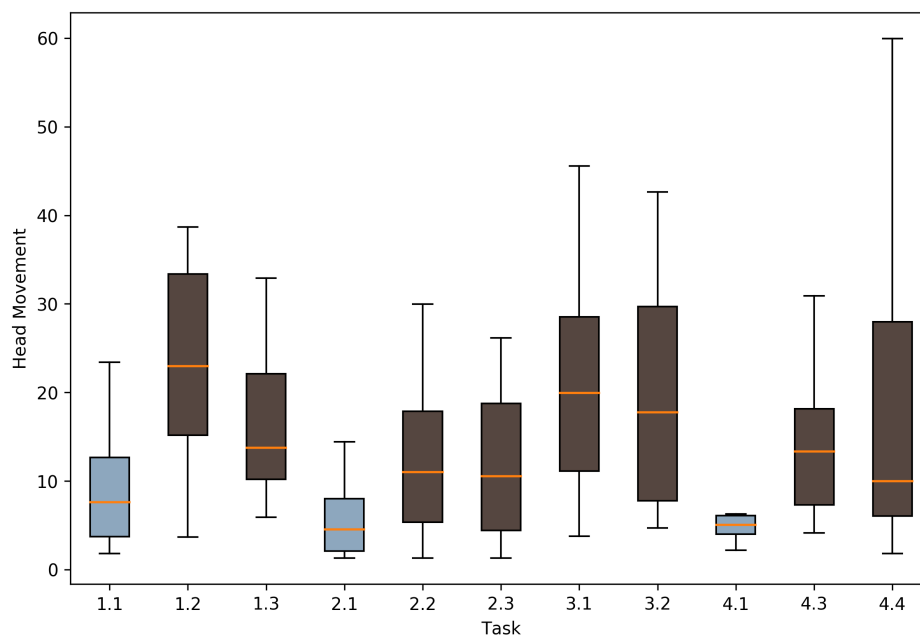


Figure 5.8: Box plot of the mean head movement for each emotional elicitation task (see 5.3).

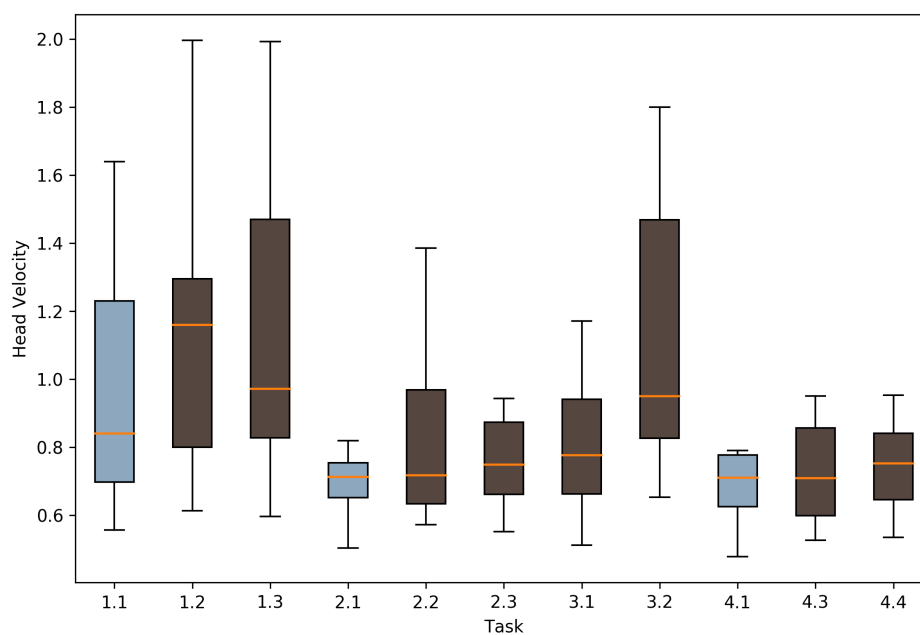


Figure 5.9: Box plot of the mean head velocity for each emotional elicitation task (see 5.3).

Table 5.3: Mean (\pm s.d.) of head movement and head velocity for each emotion inducement task with corresponding reference (neutral) baseline.

Experimental Phase	Head Movement (pixels)	Head Velocity (pixels/frame)
Social Exposure		
1.1 Neutral	10.53 \pm 10.1	1.49 \pm 2.13
1.2 Self-description speech	25.4 \pm 14.6*	1.20 \pm 0.60
1.3 Text reading	18.6 \pm 11.6*	1.33 \pm 0.32
Emotional recall		
2.1 Neutral	8.27 \pm 11.03	0.76 \pm 0.33
2.2 Recall anxious event	17.15 \pm 20.0	0.94 \pm 0.59
2.3 Recall stressful event	17.3 \pm 19.93	0.95 \pm 0.92
Stressful images/Mental task		
3.1 Neutral/Stressful images**	20.96 \pm 11.6*	0.83 \pm 0.28
3.2 Mental Task (Stroop)**	19.35 \pm 12.43*	1.12 \pm 0.36*
Stressful videos		
4.1 Neutral	8.41 \pm 9.0	0.78 \pm 0.33
4.3 Adventure/heights video	16.81 \pm 14.7*	1.12 \pm 1.39
4.4 Home invasion video	18.5 \pm 16.6*	1.34 \pm 2.66
**Reference task 4.1		* $p < 0.05$

Measurements of the heart were filtered and smoothed to reduce noise and fluctuations from facial PPG. In Figure 5.10 the distribution of the mean heart rate values from videos of the two phases is illustrated. In the stress phase an elevated heart rate distribution can be found with an inter-quartile range of [80.1; 94.2] versus [68.2; 85.5] for the non-stress phase. Mean heart rate values for the stress and non-stress phase are 88.25 \pm 11.7 and 79.3 \pm 11.3 beats per minute (BPM). A statistically significant difference between the mean heart rate in the stress and non-stress phase could not be found ($p > 0.05$). Overall, the heart rate of participants ranged from 61.6 to 109.8 BPM.

Measurements, illustrated in Figure 5.10, indicate an elevated heart rate for the stress phase. However, in comparison to the non-stress phase these are not statistically significant. This agrees with outcomes are presented by McDuff et al. [19]. In their work, the heart rate is also not significantly different. This suggests that the heart rate alone may not be a discriminative indicator of stress.

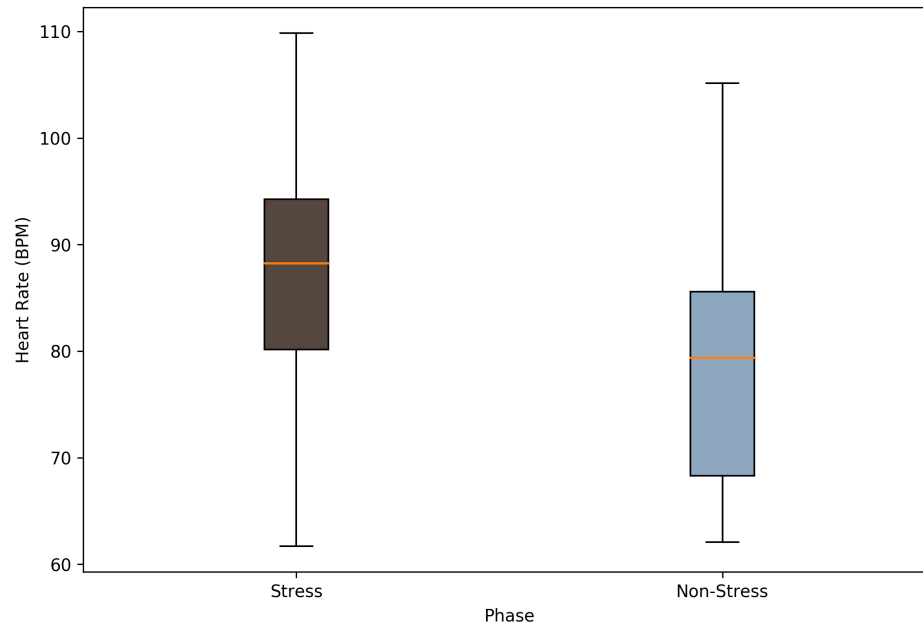


Figure 5.10: Box plot of the heart rate from facial PPG for the stress and non-stress phase.

5.5 Classification Performance

Classification of stress detection categorizes data into stressed and non-stressed emotional states. Features extracted from facial video data determine the outcomes of ML algorithms. For the evaluation of datasets performance measurement such as accuracy, precision, recall and F_1 score are chosen. The datasets are assessed with the use of cross-validation (CV) and train-test splits.

Achieved scores of CV for the conducted experiment dataset are presented in Table 5.4. The highest accuracy reaches the Naive Bayes classifier with 77%, followed by the SVM, Random Forest (RF), AdaBoost and k-NN. The F_1 score, the harmonic average of precision and recall, is best for Naive Bayes and SVM with 0.76. The poorest performance is reached by the k-NN with an accuracy of 70% and a F_1 score of 0.70. Further evaluation is done by a split of 70% training and 30% test data. The ROC curve in Figure 5.11 depicts the outcomes of the distinct ML algorithms. In the ROC curve the relation between true positive and false positive rate is plotted. The dotted line dividing the ROC plot in the diagonal represents equal amount of true positives and false negatives and an area under the ROC curve (AUC) of 0.5. A AUC of 0.5 characterizes a predictor making random guesses. Hence, the best results are located in the upper left corner with a AUC of ideally one. Best performing classifier of the ROC plot in 5.11 is the SVM with an AUC of 0.861 followed by RF and the AdaBoost classifier. A detailed view on outcomes of the SVM is shown in Figure 5.12. In the confusion matrix the ratio

of samples identified as false positives, false negatives, true positives and true negatives are displayed. This enables a more detailed analysis of classification outcomes. In the case of the SVM classification 25% of non-stressed instances are false positives and 14% of stressed instances are false negatives. The falsely identification of non-stressed as stressed instances leads to the major performance decrease. A perfect classification will ideally reach 100% true positives and true negatives.

Table 5.4: Results of classification from the conducted experiment dataset for different scores, evaluated with 10-fold CV.

Score	Naive Bayes	k-NN	AdaBoost	RF	SVM
Accuracy	0.77	0.70	0.74	0.74	0.75
Precision	0.83	0.74	0.79	0.78	0.79
Recall	0.74	0.67	0.71	0.75	0.76
F ₁	0.76	0.70	0.71	0.75	0.76

In addition to classification results, a feature selection method is applied. Feature selection allows to find features contributing most to the prediction outcomes. In this work, the feature importance of ensemble ML algorithms i.e. Trees are used for feature selection. The feature importance of tree classifiers is established by a relative rank corresponding to depth of a feature used as decision node [176]. The calculation of feature importance from an ensemble of trees leads to the top five feature importance ranking in Table 5.5. Best ranked features comprise of mouth and head related features as well as of heart rate from facial PPG.

Table 5.5: Feature importance ranking computed from an ensemble of tree classifier.

Feature Importance Ranking
1. Mean Lower Mouth Movement
2. Mean Mouth Movement
3. Mean Head Movement
4. Median Heart Rate
5. Mean Head Velocity (x-axis)

The evaluation of the second dataset [1] is performed on the distinct emotional elicitation tasks. Classifiers such as k-NN, Naive Bayes, AdaBoost and SVM, distinguished the data from stress inducing tasks and corresponding neutral reference task. To create comparable results with outcomes of [1], the mean accuracy from 10-fold CV is chosen as classification score. Outcomes of the developed solution and the values achieved from Giannakakis et al. [1] are presented in Table 5.6. Accuracy values of the developed solution differ depending on the emotional elicitation task. The accuracy values of the first social exposure task are values between 72.5% (k-NN) and 77% (SVM) with a mean difference of 11.5% to the results from [1]. The worst performance is achieved for the

second emotional recall task with an accuracy ranging from 40% to 53.5%. These low accuracies may be due to the missing mouth and heart rate features, which can not be extracted from the facial landmark data. Feature selection conducted by Giannakakis et al. shows that heart rate and mouth related features contribute most to prediction outcomes of the emotional recall task. For the tasks 3 and 4 classifiers achieve accuracy values between 71% and 82% with a respectively mean difference of 6.87% and 7.92%. Overall, performance of task 1, 3 and 4 is better than those of task 2. This may be due to the fact that feature selection determines extracted head velocity and head movement as main contributing feature for these tasks [1].

Table 5.6: Average classification accuracy results for each task. The first accuracy value is the performance achieved by the developed solution. The second value is the presented value from [1].

Task	k-NN (%)	Naive Bayes (%)	AdaBoost (%)	SVM (%)
1. Social Exposure	74.0/85.5	76.5/86.0	72.5/91.6	77.0/82.9
2. Emotional Recall	40.0/88.7	53.5/73.2	46.5/81.0	40.0/65.8
3. Stressful images/Mental task	76.0/88.3	78.0/83.4	78.5/87.2	82.0/83.1
4. Stressful Videos	71.0/88.3	71.0/71.6	71.5/83.8	77.5/76.0

A comparison of outcomes of the developed solution with results in current literature allows to assess overall performance. The most comparable approach for stress detection, besides Giannakakis et al. [1] work, is the solution of Bevilacqua et al. [20]. Their model-based solution also employs facial landmark tracking with features based on Euclidian distance. However, a classification by ML algorithms is not employed in their approach. In comparison to the model-based approaches by Aigrain et al. [17] and Dinges et al. [32] with an accuracy of 77% and 75% to 88% respectively, the classification accuracy of the developed solution reaches similar levels. Higher accuracy values ranging from 80% to 91.68% are achieved by [1]. Stress detection utilizing only facial PPG reaches performance with a classification accuracy of 80% to 86%. Thermal and hyperspectral imaging solutions classify stress and non-stress emotional states with an accuracy ranging from 56.52% to 88.9%. A more comprehensive comparison of the distinct outcomes of research in stress detection is outlined in Table 2.3.

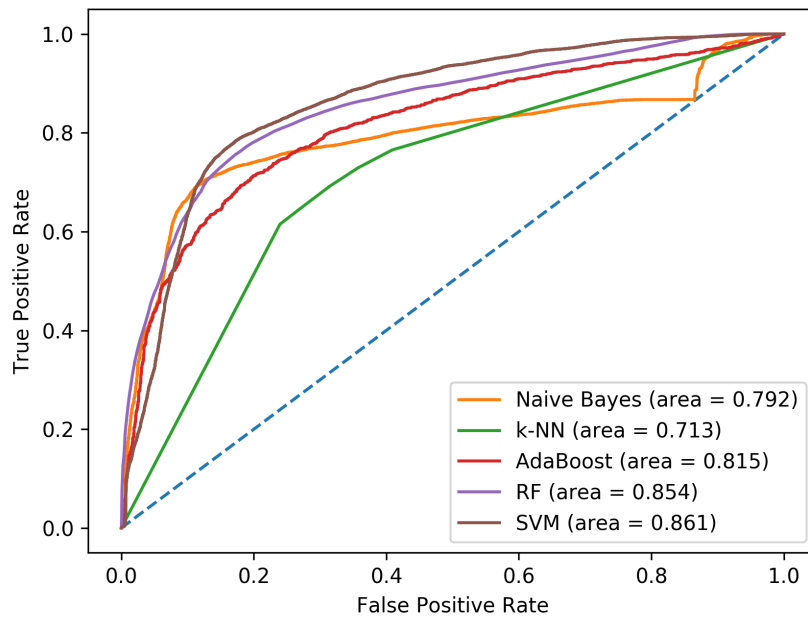


Figure 5.11: Results of classifiers from a 70/30 train-test split presented in a ROC plot.

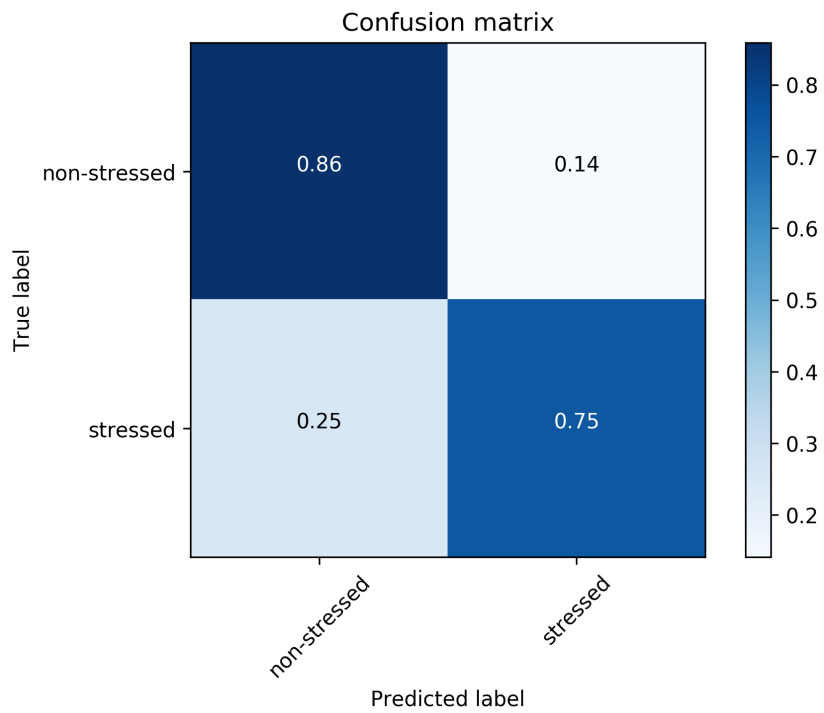


Figure 5.12: Results of a SVM classifier from a 70/30 train-test split presented in a confusion matrix.

Conclusion

In this thesis, a stress detection solution based on the analysis of facial signs from video sequences was introduced. Stress related features from eyes, head and mouth as well as the heart rate were assessed. Extraction of these features was based on computational methods such as Euclidian distances, optical flow and facial PPG. A face and landmark detection built the basis of feature extraction. Head and eye related features such as head movement, head velocity, eye aperture and blinking were computed with the help of facial landmarks. Moreover, selected ROIs allowed to determine the optical flow and facial PPG to extract mouth movements and heart rate respectively.

Evaluation of the presented solution was based on two distinct datasets. Both of these sets comprise of facial data from subjects in stressed and non-stressed emotional states. The first dataset was gathered through an experiment protocol employing the stress inducing cold-pressor test. The second dataset comprised of facial landmark data from distinct stress elicitation and reference tasks. For each dataset facial features of subjects in stressed and non-stressed/neutral emotional states were computed. Statistical analysis allowed to compare extracted feature values between these states. Significant differences, in the first dataset, could be found for eye, head and mouth related features. The analysis of the second dataset resulted in statistically significant differences depending on the stress elicitation task.

Further evaluation of the developed solution was focused on classification performance of employed ML algorithms. The classifiers categorized data as stressed or non-stressed. Prediction results have shown a classification accuracy of up to 77% in the analysis of the first dataset. Moreover, a feature selection identified specific facial features, which impact classification outcomes the most. In the case of the second dataset, outcomes of classification differed depending on the stress elicitation task. Overall, classification performance of both datasets is comparable with results in literature [17, 32].

Despite the ability to detect stress in this setting, limitations include the conducted stress elicitation experiment as well as the nature of stress. The elicitation of stress took place in a controlled laboratory environment, which impacts the generalizability of

6. CONCLUSION

results. Real-life situation may differ in case of the stressor as well as in the intensity of stress. Furthermore, environmental, physical or psychological conditions can impact features utilized for stress detection [168, 167, 177].

List of Figures

2.1	Automatic facial expression analysis steps of image/video data. Pre-Processing (Face acquisition) step, Feature Extraction based on distinct approaches, Machine Analysis of Facial Actions by Action Unit (AU, atomic facial muscle actions) processing. [49]	6
2.2	Typical physiological and physical measures utilized to detect stress. The usual measurement sources, electroencephalography (EEG), electromyography (EMG), heart rate variability (HRV), blood volume pulse (BVP), galvanic skin reponse (GSR), are shown in the figure. [18]	7
2.3	Illustration of an automated method used to recover a HRV spectrogram from video sequences of a human face. 1) Detection of facial landmarks and segmentation of face region of interest (ROI) (excluding the region around the eyes), 2) Spatial averaging over time of each color channel from the ROI, 3) Calculating source signal via Independent Component Analysis, 4) Selection of strongest blood volume pulse (BVP) signal and inverting if necessary, 5) Interpolation of BVP signal, BVP peak detection, inter-beat interval (IBI) calculation, 6) HRV spectrogram calculation. [33]	10
2.4	a) Experiment set-up. Participants are completing tasks on a laptop while sitting in front of a camera recording them. b) Screenshots of cognitive stress inducing tasks performed on a laptop. [19]	12
2.5	Posterior probability prediction of cognitive stress. i) For each task and participant. ii) Average probability (with 95% confidence intervals) for each task across all participants. During the ball task the predicted stress is significant higher compared to the rest period. Rest task, N=20; Ball task, N=30; Card task, N=30. [19]	12
2.6	Framework overview. (a) Face detection and pan, tilt, zoom (PTZ) computation is performed. (b) Skin detection isolating pixels containing PPG information. (c) Conversion of color space RGB to LUV. (d) U component calculation combined (AND) with skin detection information. (e) Spatial averaging over U pixel intensities of a set of frames, resulting into a single raw signal. [15]	13
2.7	Results of stress detection for two participants. The dashed lined plot corresponds to the low-cost camera derived stress signal. The solid-line plot represents the EDR trace. [15]	14
		85

2.8	Illustration of the face model with left and right mouth corner movements [34]	17
2.9	Development of the deformable mask from generation 1 to 3 (A-C) used for tracking facial features [32]	19
2.10	Components of the stress monitoring system [31].	20
2.11	Facial landmarks and features. (a) The 68 detected facial landmarks. (b) Facial features F_1 to F_6 visually represented. [20]	22
2.12	A facial region segmented into 3x3 blocks. (a) Block of the frame in the VS. (b) Blocks of the corresponding frame in the TS [23]	24
2.13	(a) Subject's forehead and ROI. (b) The frontal vessels (~10% hottest pixels in ROI) marked in pink. [29]	27
2.14	Cho et al. [50] CNN architecture comprising of two convolution layers, two pooling layer and one fully connected layer.	28
2.15	Results of stress detection by blood oxygenation measurements from HSI. The experiment consists of task (a) sitting on a chair for baseline measurements (b) deep and slow breaths for HSI sensitivity reference (c) subject ran for five minutes. The illustrated HSI maps are shown in false colors with high (red) and low (blue) concentrations of blood oxygenation. [88]	29
2.16	ROC curve of a binary classifier [51].	30
4.1	Overview of the stress detection framework from facial videos.	47
4.2	Cold-pressor test experiment setup.	49
4.3	Facial image examples of participants in stressed conditions from the video sequences of the conducted experiment protocol.	51
4.4	Tilting correction example image. The image is rotated clockwise by the angular offset β with the assumption that the nose bridge is always vertical in relation to the frontal plane.	52
4.5	Object detection algorithm based on HoG features [163]	53
4.6	Detection with the help of a person model. The model is characterized by (a) a coarse root filter, (b) part filters in several resolutions and (c) the spatial deformation model. Weights are defined by the filters for the HoG features. The visualization of the deformation model represents the costs of placing the center of a part at distinct locations relative to the root. [162]	54
4.7	A feature pyramid constructed from different resolutions of the image pyramid. A person is instantiated within the feature pyramid. At the top of the pyramid the root filter is placed, the part filters are located at twice the spatial resolution of the placement of the root. [162]	55
4.8	The 68 points mark-up annotation of the iBUG 300-W dataset [165]. Jaw line, eyebrows, eyes, nose and mouth points are annotated. [166]	58
4.9	Calculation of the EAR from eye landmarks. EAR is defined as a ratio of height to width. In the picture below a graph of the EAR values for multiple frames is illustrated, a blink is shown as a value close to zero. [169]	60
4.10	This example illustrates the optical flow for a ball moving in five consecutive frames. The optical flow or displacement vector is shown by the arrow.	61

4.11 Implementation of the facial PPG extraction. (a) Facial ROI is detected. (b) For each frame the ROI image is split into the red, green and blue channels and spatially averaged. (c) The constructed raw signals of the previous step is normalized. (d) ICA is applied on the raw signals. (e) Results of the ICA are three independent and separated source signals. In this example the BVP is visible in the second source signal. [21] 64

5.1 Box plot of the mean eye aperture (pixels²) for the stress and non-stress phase. 69

5.2 Box plot of the mean eyeblinks (in blinks per minute (BPM)) for the stress and non-stress phase. 70

5.3 Box plot of the mean eye aperture for each emotional elicitation task (see 5.1). 70

5.4 Box plot of the mean blink rate (in BPM) for each emotional elicitation task (see 5.1). 72

5.5 Box plot of the mean mouth movement for the stress and non-stress phase. 73

5.6 Box plot of the mean head movement (in pixels) for the stress and non-stress phase. 75

5.7 Box plot of the mean head velocity (in pixels/frame) for the stress and non-stress phase. 75

5.8 Box plot of the mean head movement for each emotional elicitation task (see 5.3). 76

5.9 Box plot of the mean head velocity for each emotional elicitation task (see 5.3). 76

5.10 Box plot of the heart rate from facial PPG for the stress and non-stress phase. 78

5.11 Results of classifiers from a 70/30 train-test split presented in a ROC plot. 81

5.12 Results of a SVM classifier from a 70/30 train-test split presented in a confusion matrix. 81

List of Tables

1.1	Workers in the EU-27 reporting each individual symptom in percentage [26]	2
2.1	Facial features associated with stress/anxiety categorized [1].	8
2.2	Overview of the classification accuracy for Naive Bayes Model and SVM with different health input features [33].	10
2.3	Summary of the outcomes for each assessed approach. The approaches are based on photoplethysmography (PPG), models, thermal imaging (TI) and hyperspectral imaging (HI). Abbreviations: Electrodermal response (EDR), energy expenditure (EE), tissue oxygen saturation (StO2).	33
3.1	Facial expression databases and corresponding characteristics of image/video recordings. Upper part of the table comprises of stress detection databases, lower part of FACS databases.	39
3.2	Resulting effects of faces at different resolution. FD and HPE for face acquisition describes face detector and head pose estimation. G1 for feature extraction states geometric features extracted by feature tracking. G2 for feature extraction states geometric features extracted by feature detection. AP for feature extraction indicates appearance features extracted by Gabor wavelets. [110].	43
4.1	Meta-features calculated from eye, mouth, head and heart rate features. Feature computed over a sliding window of 5**, 15* sec.	66
5.1	Mean (\pm s.d.) of eye aperture and blink rate for each emotion inducement task with corresponding reference (neutral) baseline.	71
5.2	Mean (\pm s.d.) of mouth related features for stress and non-stress phase. All of the values show statistically significant differences between the phases ($p < 0.05$)	73
5.3	Mean (\pm s.d.) of head movement and head velocity for each emotion inducement task with corresponding reference (neutral) baseline.	77
5.4	Results of classification from the conducted experiment dataset for different scores, evaluated with 10-fold CV.	79
5.5	Feature importance ranking computed from an ensemble of tree classifier.	79
		89

5.6	Average classification accuracy results for each task. The first accuracy value is the performance achieved by the developed solution. The second value is the presented value from [1].	80
-----	--	----

Acronyms

- AAM** active appearance model. 15, 31, 36
- AFEA** automated facial expression analysis. 5, 51
- AP** appearance based feature extraction. 42, 43, 89
- ASM** active shape model. 19, 51
- AU** action unit. 7, 16, 31, 40, 42, 45
- AUC** area under the ROC curve. 78
- BPM** blinks per minute. 68, 70, 72, 87
- BPM** beats per minute. 77
- BR** breathing rate. 9, 11
- BVP** blood volume pulse. 8, 9, 32, 35, 36, 38, 63, 64, 87
- CLM** constrained local model. 57
- CLNF** constrained local neural field. 21
- CNN** convolution neural network. 26–28, 33, 44, 51, 86
- CPT** cold-pressor test. 48, 49
- CV** cross-validation. 78, 79, 89
- DBN** dynamic bayesian network. 19, 20
- DPM** deformable part model. 53
- EAR** eye aspect ratio. 59, 60, 68, 86
- ECG** electrocardiogram. 3, 35, 36
- EDA** electrodermal activity. 35, 36

EDR electrodermal response. 13, 33, 38, 39, 89

EE energy expenditure. 3, 26, 33, 36, 39, 89

EEG electroencephalography. 3, 35

FACS facial action coding system. 7, 16, 38–40, 42, 45, 89

FD face detector. 42, 43, 89

FOV field of view. 48

FPS frames per second. 38, 48, 50

G1 geometric feature extraction by feature tracking. 42, 43, 89

G2 geometric feature extraction by feature detection. 42, 43, 89

GA genetic algorithm. 23, 24

GSR galvanic skin response. 3, 38

HDTP histogram of the dynamical thermal pattern. 23, 24

HF high frequency. 11

HI hyperspectral imaging. 4, 21, 22, 31–33, 89

HMM hidden Markov model. 17, 18, 33, 51

HoG histogram of oriented gradients. 16, 44, 51–54, 56, 57, 86

HPE head pose estimation. 42, 43, 89

HR heart rate. 9, 11, 13, 15, 18, 30, 36, 50

HRV heart rate variability. 9, 11, 13

HSI hyperspectral imaging. 28–30

ICA independent component analysis. 9, 63–65, 87

JADE joint approximate diagonalization of eigenmatrices. 65

k-NN k-nearest neighbors. 15, 33, 78, 79

LBP local binary pattern. 23

LBP-TOP local binary patterns top. 23, 24

LF low frequency. 11

ML machine learning. 1, 48, 65–67, 78–80, 83

NN neural network. 27

OCR optical computer recognition. 17

PPG photoplethysmography. 4, 5, 8, 9, 11, 12, 15, 21, 31–33, 36, 38, 39, 44, 47, 63, 64, 74, 77–80, 83, 87, 89

PSD power spectral density. 26

RF Random Forest. 78

ROC receiver operating characteristic. 30, 78, 81, 86, 87

ROI region of interest. 24–27, 30, 47, 50, 60, 63, 64, 83, 86, 87

RVS respiration variability spectrogram. 26, 27

SIFT scale invariant feature transform. 25

StO₂ tissue oxygen saturation. 28–30, 33, 89

SVM support vector machine. 9, 15, 16, 23–25, 33, 51, 52, 56, 57, 78, 79, 81, 87

TI thermal imaging. 4, 21, 22, 25, 26, 30–33, 44, 89

TS thermal spectrum. 23, 24

TSST Trier Social Stress Test. 29

VS visible spectrum. 23, 24

VTI variance of time intervals. 60, 65, 66, 71–73

Bibliography

- [1] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. Simos, K. Marias, and M. Tsiknakis, “Stress and anxiety detection using facial cues from videos,” *Biomedical Signal Processing and Control*, vol. 31, pp. 89 – 101, 2017.
- [2] H. Selye, “The stress syndrome,” *The American Journal of Nursing*, pp. 97–99, 1965.
- [3] N. Schneiderman, G. Ironson, and S. D. Siegel, “Stress and health: psychological, behavioral, and biological determinants,” *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 607–628, 2005.
- [4] S. Folkman and J. T. Moskowitz, “Stress, positive emotion, and coping,” *Current directions in psychological science*, vol. 9, no. 4, pp. 115–118, 2000.
- [5] L. Vitetta, B. Anton, F. Cortizo, and A. Sali, “Mind-body medicine: Stress and its impact on overall health and longevity,” *Annals of the New York Academy of Sciences*, vol. 1057, no. 1, pp. 492–505, 2005.
- [6] G. E. Miller, S. Cohen, and A. K. Ritchey, “Chronic psychological stress and the regulation of pro-inflammatory cytokines: a glucocorticoid-resistance model,” *Health psychology*, vol. 21, no. 6, p. 531, 2002.
- [7] R. S. Surwit, M. S. Schneider, and M. N. Feinglos, “Stress and diabetes mellitus,” *Diabetes care*, vol. 15, no. 10, pp. 1413–1422, 1992.
- [8] J. A. Seltzer and D. Kalmuss, “Socialization and stress explanations for spouse abuse,” *Social Forces*, vol. 67, no. 2, pp. 473–491, 1988.
- [9] P. R. Johnson and J. Indvik, “Stress and violence in the workplace,” *Employee Councelling Today*, vol. 8, no. 1, pp. 19–24, 1996.
- [10] M. Kalia, “Assessing the economic impact of stress - The modern day hidden epidemic,” *Metabolism*, vol. 51, no. 6, pp. 49–53, 2002.
- [11] C. Cherniss, *Staff burnout: Job stress in the human services*. Sage Publications Beverly Hills, CA, 1980.

- [12] C. Lloyd, R. King, and L. Chenoweth, “Social work, stress and burnout: A review,” *Journal of mental health*, vol. 11, no. 3, pp. 255–265, 2002.
- [13] A. Iacovides, K. Fountoulakis, S. Kaprinis, and G. Kaprinis, “The relationship between job stress, burnout and clinical depression,” *Journal of affective disorders*, vol. 75, no. 3, pp. 209–221, 2003.
- [14] G. S. Everly Jr and J. M. Lating, *A clinical guide to the treatment of the human stress response*. Springer Science & Business Media, 2012.
- [15] F. Bousefsaf, C. Maaoui, and A. Pruski, “Remote assessment of the heart rate variability to detect mental stress,” in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 348–351, IEEE, 2013.
- [16] D. McDuff, S. Gontarek, and R. W. Picard, “Improvements in remote cardiopulmonary measurement using a five band digital camera,” *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, 2014.
- [17] J. Aigrain, S. Dubuisson, M. Detyniecki, and M. Chetouani, “Person-specific behavioural features for automatic stress detection,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 03, pp. 1–6, May 2015.
- [18] N. Sharma and T. Gedeon, “Objective measures, sensors and computational techniques for stress recognition and classification: A survey,” *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1287 – 1301, 2012.
- [19] D. J. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard, “Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4000–4004, ACM, 2016.
- [20] F. Bevilacqua, H. Engström, and P. Backlund, “Automated analysis of facial cues from videos as a potential method for differentiating stress and boredom of players in games,” *International Journal of Computer Games Technology*, vol. 2018, 2018.
- [21] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in noncontact, multi-parameter physiological measurements using a webcam,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.
- [22] M. A. Kompier, C. L. Cooper, and S. A. Geurts, “A multiple case study approach to work stress prevention in europe,” *European Journal of Work and Organizational Psychology*, vol. 9, no. 3, pp. 371–400, 2000.
- [23] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, “Thermal spatio-temporal data for stress recognition,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 28, 2014.

- [24] M. H. Bekker, A. Nijssen, and G. Hens, “Stress prevention training: Sex differences in types of stressors, coping, and training effects,” *Stress and Health*, vol. 17, no. 4, pp. 207–218, 2001.
- [25] Arbeiterkammer Oberösterreich, “The Austrian Employee Health Monitor.” “https://media.arbeiterkammer.at/ooe/presseunterlagen/arbeitsgesundheitsmonitor/PKU_Gesundheitsmonitor_12122012_engl.pdf”, 2012. [Online; accessed 12. October 2017].
- [26] European Foundation for the Improvement of Living and Working Conditions, “Fourth European Working Conditions Survey.” “<https://www.eurofound.europa.eu/publications/report/2007/working-conditions/fourth-european-working-conditions-survey>”, 2007. [Online; accessed 13. October 2017].
- [27] J. A. Healey and R. W. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.
- [28] J. Pretty, J. Peacock, M. Sellens, and M. Griffin, “The mental and physical health outcomes of green exercise,” *International Journal of Environmental Health Research*, vol. 15, no. 5, pp. 319–337, 2005.
- [29] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren, “Stresscam: non-contact measurement of users’ emotional states through thermal imaging,” in *CHI’05 extended abstracts on Human factors in computing systems*, pp. 1725–1728, ACM, 2005.
- [30] S. Goel, G. Kau, and P. Toma, “A novel technique for stress recognition using eeg signal pattern,” *Current Pediatric Research*, vol. 21, no. 4, pp. 674–679, 2018.
- [31] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, “A real-time human stress monitoring system using dynamic bayesian network,” in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pp. 70–70, 2005.
- [32] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, “Optical computer recognition of facial expressions associated with stress induced by performance demands,” *Aviation, Space, and Environmental Medicine*, vol. 76, no. 6, pp. B172–B182, 2005.
- [33] D. McDuff, S. Gontarek, and R. Picard, “Remote measurement of cognitive stress via heart rate variability,” in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 2957–2960, 2014.

- [34] D. Metaxas, S. Venkataraman, and C. Vogler, “Image-based stress recognition using a model-based dynamic face tracking system,” in *International Conference on Computational Science*, pp. 813–821, Springer, 2004.
- [35] N. Sharma and T. Gedeon, “Modeling observer stress for typical real environments,” *Expert Systems with Applications*, vol. 41, no. 5, pp. 2231 – 2238, 2014.
- [36] E. Kaniusas *et al.*, *Biomedical Signals and Sensors II*. Springer, 2015.
- [37] J. A. Levine, “Measurement of energy expenditure,” *Public Health Nutrition*, vol. 8, no. 7a, p. 1123–1132, 2005.
- [38] J. A. Taylor, “The relationship of anxiety to the conditioned eyelid response.,” *Journal of Experimental Psychology*, vol. 41, no. 2, p. 81, 1951.
- [39] C. S. Harris, R. I. Thackray, and R. W. Shoenberger, “Blink rate as a function of induced muscular tension and manifest anxiety,” *Perceptual and Motor Skills*, vol. 22, no. 1, pp. 155–160, 1966.
- [40] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose, “Head movement correlates of juncture and stress at sentence level,” *Language and Speech*, vol. 26, no. 2, pp. 117–129, 1983.
- [41] M. N. Mohd, M. Kashima, K. Sato, and M. Watanabe, “Internal state measurement from facial stereo thermal and visible sensors through svm classification,” 2015.
- [42] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
- [43] S. Cohen, R. C. Kessler, and L. U. Gordon, *Measuring stress: A guide for health and social scientists*. Oxford University Press on Demand, 1997.
- [44] W. A. Freiwald, D. Y. Tsao, and M. S. Livingstone, “A face feature space in the macaque temporal lobe,” *Nature neuroscience*, vol. 12, no. 9, p. 1187, 2009.
- [45] S. Setterlind and G. Larsson, “The stress profile: a psychosocial approach to measuring stress,” *Stress Medicine*, vol. 11, no. 1, pp. 85–92, 1995.
- [46] A. Alberdi, A. Aztiria, and A. Basarab, “Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review,” *Journal of Biomedical Informatics*, vol. 59, pp. 49 – 75, 2016.
- [47] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, “Activity-aware mental stress detection using physiological sensors,” in *Mobile Computing, Applications, and Services* (M. Gris and G. Yang, eds.), (Berlin, Heidelberg), pp. 211–230, Springer Berlin Heidelberg, 2012.

- [48] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, “A psychometric evaluation of the facial action coding system for assessing spontaneous expression,” *Journal of Nonverbal Behavior*, vol. 25, pp. 167–185, Sep 2001.
- [49] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, “Automatic analysis of facial actions: A survey,” *IEEE Transactions on Affective Computing*, 2017.
- [50] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, “Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings,” *CoRR*, vol. abs/1708.06026, 2017.
- [51] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She, “Detection of psychological stress using a hyperspectral imaging technique,” *IEEE Transactions on Affective Computing*, vol. 5, pp. 391–405, Oct 2014.
- [52] T. H. Holmes and R. H. Rahe, “The social readjustment rating scale,” *Journal of psychosomatic research*, vol. 11, no. 2, pp. 213–218, 1967.
- [53] B. S. McEwen, “The neurobiology of stress: from serendipity to clinical relevance11published on the world wide web on 22 november 2000.,” *Brain Research*, vol. 886, no. 1, pp. 172 – 189, 2000. Towards 2010, A brain Odyssey, The 3rd Brain Research Interactive.
- [54] K. Hong, P. Yuen, T. Chen, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, W. Oxford, J. Piper, *et al.*, “Detection and classification of stress using thermal imaging technique,” in *Optics and Photonics for Counterterrorism and Crime Fighting V*, vol. 7486, p. 74860I, International Society for Optics and Photonics, 2009.
- [55] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, “Wearable physiological sensors reflect mental stress state in office-like situations,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 600–605, Sept 2013.
- [56] M. Bickford, “Stress in the workplace: A general overview of the causes, the effects, and the solutions,” *Canadian Mental Health Association Newfoundland and Labrador Division*, pp. 1–3, 2005.
- [57] Y. Okada, T. Y. Yoto, T. Suzuki, S. Sakuragawa, and T. Sugiura, “Wearable ecg recorder with acceleration sensors for monitoring daily stress: Office work simulation study,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4718–4721, July 2013.
- [58] J. A. Harrigan and D. M. O’Connell, “How do you look when feeling anxious? facial displays of anxiety,” *Personality and Individual Differences*, vol. 21, no. 2, pp. 205 – 212, 1996.
- [59] E. Fox, A. Mathews, A. J. Calder, and J. Yiend, “Anxiety and sensitivity to gaze direction in emotionally expressive faces.,” *Emotion*, vol. 7, no. 3, p. 478, 2007.

- [60] M. Honma, “Hyper-volume of eye-contact perception and social anxiety traits,” *Consciousness and Cognition*, vol. 22, no. 1, pp. 167 – 173, 2013.
- [61] J. Allen, “Photoplethysmography and its application in clinical physiological measurement,” *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.
- [62] W. Verkruyse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light.,” *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.
- [63] M. Tkalcic and J. F. Tasic, *Colour spaces: perceptual, historical and applicational background*, vol. 1. IEEE, 2003.
- [64] A. Garde, B. Laursen, A. Jørgensen, and B. Jensen, “Effects of mental and physical demands on heart rate variability during computer work,” *European journal of applied physiology*, vol. 87, no. 4-5, pp. 456–461, 2002.
- [65] G. J. Edwards, T. F. Cootes, and C. J. Taylor, “Face recognition using active appearance models,” in *European Conference on Computer Vision*, pp. 581–595, Springer, 1998.
- [66] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [67] B. Fasel and J. Luetttin, “Automatic facial expression analysis: a survey,” *Pattern Recognition*, vol. 36, no. 1, pp. 259 – 275, 2003.
- [68] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *ACM Computing Surveys (CSUR)*, vol. 35, pp. 399–458, Dec. 2003.
- [69] H. van Kuilenburg, M. Wiering, and M. den Uyl, “A model based method for automatic facial expression recognition,” in *Machine Learning: ECML 2005* (J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, eds.), (Berlin, Heidelberg), pp. 194–205, Springer Berlin Heidelberg, 2005.
- [70] M. Padiaditis, G. Giannakakis, F. Chiarugi, D. Manousos, A. Pampouchidou, E. Christinaki, G. Iatraki, E. Kazantzaki, P. G. Simos, K. Marias, and M. Tsiknakis, “Extraction of facial features as indicators of stress and anxiety,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3711–3714, Aug 2015.
- [71] J. Nicolle, K. Bailly, and M. Chetouani, “Facial action unit intensity prediction via hard multi-task metric learning for kernel regression,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, pp. 1–6, May 2015.

- [72] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893 vol. 1, June 2005.
- [73] S. K. Goldenstein, C. Vogler, and D. Metaxas, “Statistical cue integration in dag deformable models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 801–813, 2003.
- [74] Q. Ji, Z. Zhu, and P. Lan, “Real-time nonintrusive monitoring and prediction of driver fatigue,” *IEEE Transactions on Vehicular Technology*, vol. 53, no. 4, pp. 1052–1068, 2004.
- [75] Z. Zhu and Q. Ji, “A head motion free gaze tracker with one-time calibration,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA*, 2005.
- [76] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 200–205, 1998.
- [77] A. Lanitis, C. J. Taylor, and T. F. Cootes, “Automatic interpretation and coding of face images using flexible models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 743–756, Jul 1997.
- [78] S. L. Lauritzen, “The em algorithm for graphical association models with missing data,” *Computational Statistics & Data Analysis*, vol. 19, no. 2, pp. 191–201, 1995.
- [79] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “Constrained local neural fields for robust facial landmark detection in the wild,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.
- [80] J. Stewart, *Calculus*. Cengage Learning, 2011.
- [81] M. Vollmer, M. Klaus-Peter, *et al.*, *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2017.
- [82] P. M. Mehl, Y.-R. Chen, M. S. Kim, and D. E. Chan, “Development of hyperspectral imaging technique for the detection of apple surface defects and contaminations,” *Journal of Food Engineering*, vol. 61, no. 1, pp. 67 – 81, 2004. Applications of computer vision in the food industry.
- [83] A. Gowen, C. O’Donnell, P. Cullen, G. Downey, and J. Frias, “Hyperspectral imaging – an emerging process analytical tool for food quality and safety control,” *Trends in Food Science & Technology*, vol. 18, no. 12, pp. 590 – 598, 2007.
- [84] G. Lu and B. Fei, “Medical hyperspectral imaging: a review,” *Journal of Biomedical Optics*, vol. 19, no. 1, p. 010901, 2014.

- [85] S. Ioannou, V. Gallese, and A. Merla, “Thermal infrared imaging in psychophysiology: potentialities and limits,” *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.
- [86] Y. Yoshitomi, S.-I. Kim, T. Kawano, and T. Kilazoe, “Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face,” in *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No.00TH8499)*, pp. 178–183, 2000.
- [87] M. M. Khan, R. D. Ward, and M. Ingleby, “Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature,” *ACM Transactions on Applied Perception (TAP)*, vol. 6, pp. 6:1–6:22, Feb. 2009.
- [88] P. Yuen, K. Hong, T. Chen, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, L. Williams, W. Oxford, *et al.*, “Emotional & physical stress detection and classification using thermal imaging technique,” 2009.
- [89] E. Ring, “The historical development of temperature measurement in medicine,” *Infrared physics & technology*, vol. 49, no. 3, pp. 297–301, 2007.
- [90] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [91] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [92] Y. Masaoka and I. Homma, “Anxiety and respiratory patterns: their relationship during mental stress and physical load,” *International Journal of Psychophysiology*, vol. 27, no. 2, pp. 153 – 159, 1997.
- [93] Y. Cho, S. J. Julier, N. Marquardt, and N. Bianchi-Berthouze, “Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging,” *Biomedical Optics Express*, vol. 8, pp. 4480–4503, Oct 2017.
- [94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.
- [95] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, “The ‘trier social stress test’—a tool for investigating psychobiological stress responses in a laboratory setting,” *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.
- [96] N. Sharma and T. Gedeon, “Modeling observer stress for typical real environments,” *Expert Systems with Applications*, vol. 41, no. 5, pp. 2231 – 2238, 2014.

- [97] N. Tack, A. Lambrechts, P. Soussan, and L. Haspeslagh, “A compact, high-speed, and low-cost hyperspectral imager,” in *Silicon Photonics VII*, vol. 8266, p. 82660Q, International Society for Optics and Photonics, 2012.
- [98] K. Uto, H. Seki, G. Saito, Y. Kosugi, and T. Komatsu, “Development of a low-cost, lightweight hyperspectral imaging system based on a polygon mirror and compact spectrometers,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, pp. 861–875, Feb 2016.
- [99] Y. Wu and Q. Ji, “Facial landmark detection: A literature survey,” *International Journal of Computer Vision*, May 2018.
- [100] M. Haak, S. Bos, S. Panic, and L. Rothkrantz, “Detecting stress using eye blinks and brain activity from eeg signals,” *Proceeding of the 1st Driver Car Interaction and Interface (DCII 2008)*, pp. 35–60, 2009.
- [101] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, “Stress detection from speech and galvanic skin response signals,” in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pp. 209–214, IEEE, 2013.
- [102] J. Zhai and A. Barreto, “Stress recognition using non-invasive technology,” in *FLAIRS Conference*, pp. 395–401, 2006.
- [103] T. G. M. Vrijkotte, L. J. P. van Doornen, and E. J. C. de Geus, “Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability,” *Hypertension*, vol. 35, no. 4, pp. 880–886, 2000.
- [104] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, “Cas (me)²: a database for spontaneous macro-expression and micro-expression spotting and recognition,” *IEEE Transactions on Affective Computing*, 2017.
- [105] J. Zhai and A. Barreto, “Stress detection in computer users based on digital signal processing of noninvasive physiological variables,” in *Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE*, pp. 1355–1358, 2006.
- [106] G. A. Liebchen and M. Shepperd, “Data sets and data quality in software engineering,” in *Proceedings of the 4th International Workshop on Predictor Models in Software Engineering*, PROMISE ’08, (New York, NY, USA), pp. 39–44, ACM, 2008.
- [107] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.
- [108] L. L. Pipino, Y. W. Lee, and R. Y. Wang, “Data quality assessment,” *Commun. ACM*, vol. 45, pp. 211–218, Apr. 2002.

- [109] J. C. Wyatt and J. L. Y. Liu, “Basic concepts in medical informatics,” *Journal of Epidemiology & Community Health*, vol. 56, no. 11, pp. 808–812, 2002.
- [110] Y.-l. Tian, “Evaluation of face resolution for expression analysis,” in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW’04. Conference on*, pp. 82–82, IEEE, 2004.
- [111] R. L. Winkler and A. H. Murphy, “Experiments in the laboratory and the real world,” *Organizational Behavior and Human Performance*, vol. 10, no. 2, pp. 252 – 270, 1973.
- [112] G. Dedeoğlu, S. Baker, and T. Kanade, “Resolution-aware fitting of active appearance models to low resolution images,” in *European Conference on Computer Vision*, pp. 83–97, Springer, 2006.
- [113] F. El-Amrawy and M. I. Nounou, “Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?,” *Healthcare Informatics Research*, vol. 21, no. 4, pp. 315–320, 2015.
- [114] A. Shcherbina, C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler, and E. A. Ashley, “Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort,” *Journal of Personalized Medicine*, vol. 7, no. 2, p. 3, 2017.
- [115] N. L. Elwess and F. D. Vogt, “Heart rate and stress in a college setting.,” *Bioscene: Journal of College Biology Teaching*, vol. 31, no. 4, pp. 20–23, 2005.
- [116] N. K. Singh and D. O. Rieke, “Towards an open data framework for body sensor networks supporting bluetooth low energy,” in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 396–401, June 2016.
- [117] E. C. Nelson, T. Verhagen, and M. L. Noordzij, “Health empowerment through activity trackers: An empirical smart wristband study,” *Computers in Human Behavior*, vol. 62, pp. 364–374, 2016.
- [118] N. Y. Oei, W. T. Everaerd, B. M. Elzinga, S. van Well, and B. Bermond, “Psychosocial stress impairs working memory at high loads: an association with cortisol levels and memory retrieval,” *Stress*, vol. 9, no. 3, pp. 133–141, 2006.
- [119] J. D. Payne, L. Nadel, J. J. Allen, K. G. Thomas, and W. J. Jacobs, “The effects of experimentally induced stress on false recognition,” *Memory*, vol. 10, no. 1, pp. 1–6, 2002.
- [120] W. E. Winkler, “Methods for evaluating and creating data quality,” *Information Systems*, vol. 29, no. 7, pp. 531 – 550, 2004. Data Quality in Cooperative Information Systems.

- [121] S. Singhal and M. Jena, “A study on weka tool for data preprocessing, classification and clustering,” *International Journal of Innovative technology and exploring engineering (IJITEE)*, vol. 2, no. 6, pp. 250–253, 2013.
- [122] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Data preprocessing for supervised learning,” *International Journal of Computer Science*, vol. 1, no. 2, pp. 111–117, 2006.
- [123] R. Engels and C. Theusinger, “Using a data metric for preprocessing advice for data mining applications,” in *ECAI*, pp. 430–434, 1998.
- [124] N. M. Nasrabadi, “Pattern Recognition and Machine Learning,” *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.
- [125] R. Singh, M. Vatsa, and A. Noore, “Face recognition with disguise and single gallery images,” *Image and Vision Computing*, vol. 27, no. 3, pp. 245 – 257, 2009. Special Issue on Multimodal Biometrics.
- [126] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 46–53, 2000.
- [127] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, “Casm database: a dataset of spontaneous micro-expressions collected from neutralized faces,” in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–7, 2013.
- [128] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, “Casm II: An improved spontaneous micro-expression database and the baseline evaluation,” *PLoS one*, vol. 9, no. 1, p. e86041, 2014.
- [129] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, “Sayette Group Formation Task (GFT) Spontaneous Facial Expression Database,” in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pp. 581–588, IEEE, 2017.
- [130] E. J. Candes and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, March 2008.
- [131] I. Fatt and B. A. Weissman, *Physiology of the eye: an introduction to the vegetative functions*. Butterworth-Heinemann, 2013.
- [132] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, “Static and dynamic 3d facial expression recognition: A comprehensive survey,” *Image and Vision Computing*, vol. 30, no. 10, pp. 683 – 697, 2012. 3D Facial Behaviour Analysis and Understanding.

- [133] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, “Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1548–1568, Aug 2016.
- [134] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 5–pp, 2005.
- [135] S. M. Mavadati, M. H. Mahoor, K. Bartlett, and P. Trinh, “Automatic detection of non-posed facial action units,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 1817–1820, IEEE, 2012.
- [136] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.
- [137] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1113–1133, June 2015.
- [138] A. Savran, B. Sankur, and M. T. Bilge, “Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units,” *Pattern Recognition*, vol. 45, no. 2, pp. 767 – 782, 2012.
- [139] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, “Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database,” *Image and Vision Computing*, vol. 32, no. 10, pp. 692 – 706, 2014. Best of Automatic Face and Gesture Recognition 2013.
- [140] T. Waldraff, *Digitale Bildauflösung: Grundlagen, Auflösungsbestimmung, Anwendungsbeispiele*. Springer-Verlag, 2013.
- [141] B. Ballard, *Designing the Mobile User Experience*. John Wiley & Sons, 2007.
- [142] L. G. Farkas, M. J. Katic, and C. R. Forrest, “International anthropometric study of facial morphology in various ethnic groups/races,” *Journal of Craniofacial Surgery*, vol. 16, no. 4, pp. 615–646, 2005.
- [143] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, “Facewarehouse: A 3d facial expression database for visual computing,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, pp. 413–425, March 2014.
- [144] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, “A high-resolution 3d dynamic facial expression database,” in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6, Sept 2008.

- [145] J.-C. Fernandez, L. Mounier, and C. Pachon, “A model-based approach for robustness testing,” in *Testing of Communicating Systems* (F. Khendek and R. Dssouli, eds.), (Berlin, Heidelberg), pp. 333–348, Springer Berlin Heidelberg, 2005.
- [146] J. Radatz, A. Geraci, and F. Katki, “IEEE standard glossary of software engineering terminology,” *IEEE Std*, vol. 610121990, no. 121990, p. 3, 1990.
- [147] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.
- [148] T. Mita, T. Kaneko, and O. Hori, “Joint haar-like features for face detection,” in *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, vol. 2, pp. 1619–1626 Vol. 2, Oct 2005.
- [149] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [150] C. Zhang and Z. Zhang, “A survey of recent advances in face detection,” 2010.
- [151] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.
- [152] S. S. Farfadi, M. J. Saberian, and L.-J. Li, “Multi-view face detection using deep convolutional neural networks,” in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR ’15*, (New York, NY, USA), pp. 643–650, ACM, 2015.
- [153] S. Yang, P. Luo, C.-C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [154] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, “A convolutional neural network cascade for face detection,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [155] H. Jiang and E. Learned-Miller, “Face detection with the faster r-cnn,” in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 650–657, May 2017.
- [156] N. Aifanti, C. Papachristou, and A. Delopoulos, “The mug facial expression database,” in *Image analysis for multimedia interactive services (WIAMIS), 2010 11th international workshop on*, pp. 1–4, IEEE, 2010.

- [157] D. Lundqvist, A. Flykt, and A. Öhman, “The karolinska directed emotional faces (kdef). cd rom from department of clinical neuroscience. psychology section, karolinska institutet; 1998,” tech. rep., ISBN 91-630-7164-9.
- [158] N. Alyüz, B. Gökberk, H. Dibeklioglu, A. Savran, A. A. Salah, L. Akarun, and B. Sankur, “3d face recognition benchmarks on the bosphorus database with focus on facial expressions,” in *Biometrics and Identity Management* (B. Schouten, N. C. Juul, A. Drygajlo, and M. Tistarelli, eds.), (Berlin, Heidelberg), pp. 57–66, Springer Berlin Heidelberg, 2008.
- [159] D. Schoofs, O. T. Wolf, and T. Smeets, “Cold pressor stress impairs performance on working memory tasks requiring executive functions in healthy young men.,” *Behavioral Neuroscience*, vol. 123, no. 5, p. 1066, 2009.
- [160] D. L. Wood, S. G. Sheps, L. R. Elveback, and A. Schirger, “Cold pressor test as a predictor of hypertension.,” *Hypertension*, vol. 6, no. 3, pp. 301–306, 1984.
- [161] L. Schwabe, L. Haddad, and H. Schachinger, “Hpa axis activation by a socially evaluated cold-pressor test,” *Psychoneuroendocrinology*, vol. 33, no. 6, pp. 890 – 895, 2008.
- [162] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, Sept 2010.
- [163] A. Suleiman and V. Sze, “Energy-efficient hog-based object detection at 1080hd 60 fps with multi-scale support,” in *2014 IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 1–6, Oct 2014.
- [164] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [165] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: database and results,” *Image and Vision Computing*, vol. 47, pp. 3 – 18, 2016. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.
- [166] iBUG 300-W, “Facial Points Annotations.” “<https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/>.” [Online; accessed 24. August 2018].
- [167] J. Mackintosh, R. Kumar, and T. Kitamura, “Blink rate in psychiatric illness,” *The British Journal of Psychiatry*, vol. 143, no. 1, pp. 55–57, 1983.
- [168] G. W. Ousler III, K. W. Hagberg, M. Schindelar, D. Welch, and M. B. Abelson, “The ocular protection index,” *Cornea*, vol. 27, no. 5, pp. 509–513, 2008.

- [169] T. Soukupová and J. Cech, “Real-Time Eye Blink Detection using Facial Landmarks,” *21st Computer Vision Winter Workshop*, vol. February 3, pp. 1–8, 2016.
- [170] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, “A decision theoretic model for stress recognition and user assistance,” in *Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 2005, pp. 529–534, 2005.
- [171] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Image Analysis* (J. Bigun and T. Gustavsson, eds.), (Berlin, Heidelberg), pp. 363–370, Springer Berlin Heidelberg, 2003.
- [172] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1, pp. 185 – 203, 1981.
- [173] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” 1981.
- [174] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.
- [175] A. Savitzky and M. J. Golay, “Smoothing and differentiation of data by simplified least squares procedures.,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [176] G. Louppe, “Understanding random forests: From theory to practice,” *arXiv preprint arXiv:1407.7502*, 2014.
- [177] R. C. Chan and E. Y. Chen, “Blink rate does matter: a study of blink rate, sustained attention, and neurological signs in schizophrenia,” *The Journal of Nervous and Mental Disease*, vol. 192, no. 11, pp. 781–783, 2004.