UNIVERSITY OF THE WEST *of* SCOTLAND

UWS

Tailored and Enhanced Automated Facial Expression Analysis to Tackle Practical Applications in Affective Computing

by

Arne Bernin

Thesis submitted in partial fulfilment of the requirements
of the University of the West of Scotland
for the award of Doctor of Philosophy

April, 2019

Table of Contents

Abstract

This thesis introduces an application-centric approach to affective computing, with a focus on adapting and enhancing automated facial expression recognition. Using a foundation of emotion models, algorithms for the recognition of facial expressions and their fields of application, this study presents an analysis of the potential to enhance existing automated facial expression recognition (AFER) algorithms, and a resulting strategy.

The EmotionBike – a variant of a cockpit scenario in the form of an affective bicycle exercise game – is introduced as an exemplary application context. Furthermore, the time-sensitive design of the EmotionBike framework and the evaluation results of the frameworks performance during the experiments are presented.

Four state-of-the-art algorithms for facial expression recognition are compared by utilising different annotated databases and three possible metrics for post-processing of algorithms' output to provide an optimal solution for the EmotionBike. To increase robustness of the detection, two novel approaches are presented. The first approach enhances the robustness of facial expression recognition by grouping expressions based on the post-processing of algorithms' output. The second approach of emotional shift analysis enables to automatically classify users' reactions to events, based on rapid changes in facial expressions.

Declaration

The research presented in this thesis was carried out by the undersigned. No part of the research has been submitted in support of an application for another degree or qualification at this or another university.

January 23, 2020	
Date	Signature

Acknowledgements

I am very grateful to everyone who helped me during my PhD studies. Without them, writing this thesis would have been very difficult or even impossible.

In particular, I would like to thank Florian Vogt for his contributions to the EmotionBike project and for investing so much time to help me in bringing this thesis to its final state.

Many thanks to my other supervisors Qi Wang, Christos Grecos and Kai von Luck for supporting this work throughout the years. Many thanks to Gunter Klemke, who supported me with the introduction of my doctoral thesis.

Special thanks to the dear people (Julia, Nadja, Zita, Katrin, Gavin, Qi, Kai) who supported me in the final phase of my work. With a very special thank to Sobin Ghose for all his help on the project and this thesis!

Many thanks to Ralf Jettke for his useful input on statistics, the EmotionBike system, experiments and all the rest, especially 'the kitchen'.

Thanks to the fine Scottish folk musicians (especially Kris Drever and Graeme E. Pearson) that have kept me motivated over the years. I've always enjoyed listening to or meeting you.

I would especially like to thank Erika Bernin her unwavering support.

In addition, many thanks to Julia Stock, Anja Schäfer, Nils Flagge, Lea Rahn and Jan-Christoph Winter for their support throughout the years.

The EmotionBike was a joint project by many contributors from the HAW Hamburg. Many thanks to the team, especially to Sobin Ghose and Larissa Müller for the fruitful discussions and the collective development of the project. Details on the contributions are described in Table A.1.

List of Publications

List of peer-reviewed publications related to this thesis, the assignment to the individual chapters is depicted in Figure 1.2.

P1: Inproceedings, **Arne Bernin**. *A Framework Concept for Emotion Enriched Interfaces*. International Conference on Entertainment Computing (ICEC) 2012, page 482-485

P2: Inproceedings, **Arne Bernin**, Larissa Mueller, Sobin Ghose, Kai von Luck, Christos Grecos, Qi Wang, Florian Vogt. *Towards More Robust Automatic Facial Expression Recognition in Smart Environments*. Pervasive Technologies Related to Assistive Environments (PETRA) 2017, page 37-44

P3: Inproceedings, **Arne Bernin**, Larissa Mueller, Sobin Ghose, Kai von Luck, Christos Grecos, Qi Wang, Ralf Jettke, Florian Vogt. *Automatic Segmentation and Shift Detection of Facial Expressions in Emotion-Provoking Environments*. PETRA 2018, page 194-201

P4: Inproceedings, Larissa Müller, **Arne Bernin**, Andreas Kamenz, Sobin Ghose, Kai von Luck, Qi Wang, Christos Grecos, Florian Vogt. *Emotional Journey for an Emotion Provoking Cycling Exergame*. IEEE 4th International Conference on Soft Computing Machine Intelligence (ISCMI) 2017, pages 104-108.

List of additional peer-reviewed publications related to the EmotionBike project.

Inproceedings, Larissa Müller, Sebastian Zagaria, **Arne Bernin**, Abbes Amira, Naeem Ramzan, Christos Grecos, Florian Vogt: *EmotionBike: A Study of Provoking Emotions in Cycling Exergames*. International Conference on Entertainment Computing (ICEC) 2015, pages 155-168

Inproceedings, Larissa Müller, **Arne Bernin**, Sobin Ghose, Wojtek Gozdzielewski, Qi Wang, Christos Grecos, Kai von Luck, Florian Vogt. *Physiological data analysis for an emotional provoking exergame*. IEEE Symposium Series on Computational Intelligence (SSCI) 2016, pages 1-8

Inproceedings, Larissa Müller, **Arne Bernin**, Kai von Luck, Andreas Kamenz, Sobin Ghose, Qi Wang, Christos Grecos, Florian Vogt: *Enhancing exercise experience with individual multi-emotion provoking game elements*. SSCI 2017, pages 1-8.

Inproceedings, Andreas Kamenz, Victoria Bibaeva, **Arne Bernin**, Sobin Ghose, Kai von Luck, Florian Vogt, Larissa Müller. *Classification of Physiological Data in Affective Exergames*. SSCI 2018, pages 2076-2081.

List of Abbreviations

6BE	Six Basic Emotions
AFER	Automated Facial Expression Recognition
API	Application Interface
AU	Action Unit
FACS	Facial Action Coding System
FER	Facial Expression Recognition
fps	frames per second
JSON	Javascript object notation
ML	Machine Learning
SVM	Support Vector Machine

List of Figures

List of Tables

1. Introduction

> To be able to describe and measure expression (and our ability to read it) as an element of a language in which a certain facial expression is assigned an intersubjective meaning, is to obtain another mysterious area. It is true, the face is the mirror of the soul, but its language is no longer ’unspeakable’.
>
> (Umberto Eco, *the face is the mirror of the soul* (Eco 1976))

1.1. Motivation

The expression and perception of emotions are essential in human communication. Emotions are also aroused when people interact with machines, but these emotions are widely ignored by the machines. In the ability to detect and respond adequately to emotions lies great potential to create truly smart and intelligent devices. This potential is at the core of future collaborative human-machine interfaces in which their joint capabilities hold the key to the next computer revolution of the 21st century.

The current state of computer science research offers exciting solutions to automatically detect human expressions and emotions. These solutions cannot be found in commercial applications, but rather only in the affective and social computing research. There is no agreed upon method for integrating emotion recognition into applications, but an application-specific approach may be feasible. To bridge the gap, context-specific validation is required.

Application areas with high potential to enhance user experience through emotion recognition include, firstly, learning platforms that adjust to the individual speed and difficulty to enhance learning effects; secondly, entertainment systems, adapting the storytelling to emotional responses to content and thereby increasing player motivation and satisfaction; thirdly, cockpit scenarios providing adaptive assistance to drivers, pilots or machine operators thereby increasing passenger safety and pilot performance by adjusting to situations, stress levels and emotional reactions.

For automated recognition of human emotions, different approaches have been demonstrated, often based on the analysis of physiological or visual information, especially camera-based facial expression recognition (FER) as a well-established method. The important connection between facial expressions and emotions has been emphasized in many publications over the last 150 years. It is exciting to observe the quality and quantity of promising algorithms for automated facial expression recognition (AFER), developed in the past decade. In these algorithms lies the true potential for revolutionising user interfaces towards human-orientated interaction in the foreseeable future.

However, the challenge of transferring emotional human-computer interaction from research environments to everyday application contexts remains unmet. The main gap is the missing consideration of specific application characteristics and the appropriate interpretation of the algorithm's output for an adequate system response. This work provides new approaches to effectively exploit existing state-of-the-art AFER systems in applications for smart environments[1] and to pragmatically close the gap between what is possible and what is already feasible by the example of an affective exercise game.

1.2. The EmotionBike as Platform for Affective Experiments

The EmotionBike was founded as a research project at the HAW Hamburg[2] to conduct experiments in the areas of affective gaming, exercise gaming and cockpit-scenarios. As an exemplary application, an affective exercise game scenario was developed, combining physical training on an ergometer and game-based emotion provocation (depicted in Figure 1.1).

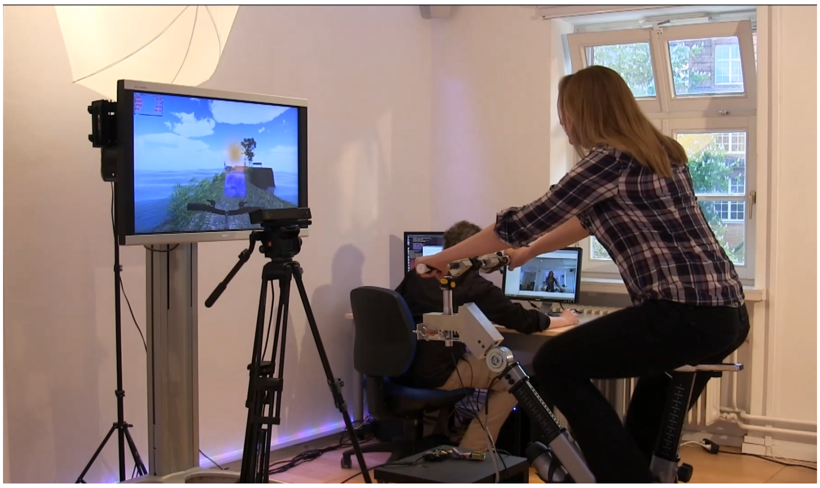


Figure 1.1: The EmotionBike, a cockpit-scenario as combination of an enhanced ergometer with a rotatable handlebar, a frontal camera and a video game.

[1] Smart environments are capable to acquire knowledge about the environment and its inhabitants and apply this knowledge to improve the experience in that environment (Cook and Das 2004).

[2] Hamburg University of Applied Sciences

In this affective exercise game, participants cycle through different game worlds with emotions provoked by game events (e.g. suddenly occurring monsters). Since this setup was novel, individual components (such as the game worlds) and the software framework, which interconnects all individual parts, had to be developed first. Two factors are important here: firstly, ensuring sufficient responsiveness of the interactive components to avoid artefacts in the data, and secondly, time-synchronous recording of all relevant events and sensor data. To achieve these goals, a protocol and cross-platform software framework was developed and successfully utilised and evaluated in several experiments.

This time-synchronous recording enables the success of the emotion provocation to be measured and allows the subsequent automated evaluation of the experiments without time-consuming manual annotation of the data set. To evaluate the emotion provocations, an event-based method was developed which considers the time windows around emotion provoking game events.

Within the EmotionBike project several aspects were developed by different contributors, as listed in Table A.1. Although in the overall research project, other approaches for detecting emotions (e.g. physiological sensors) have been evaluated, this work focuses on the interpretation of facial expressions by AFER which is considered a dominant method for detecting emotions.

1.3. Focus and Delimitation of This Work

Many works in the context of AFER have been proposed in recent decades[3] and approached the field almost exclusively from an algorithm developer's perspective, addressing the chal-

[3] The ACM digital library lists nearly 190.000 publications for 'facial expression recognition algorithm'.

lenge of how to improve recognition rates in comparison to the current state-of-the-art. Unfortunately, from an application developers perspective, the challenge of how to actually integrate these algorithms into a specific application remains unsolved, especially for non-desktop scenarios. This integration requires an understanding of the characteristics and boundaries of a specific environment, application and AFER solution to develop an adequate interpretation based on the AFER algorithms' output.

This thesis focuses on the development of methods for evaluation and integration of state-of-the-art AFER algorithmic solutions into affective applications by the example of the EmotionBike and on improving the results by adapting the algorithms output taking advantage of the provided dimensions and timing. The time-synchronization of the recorded data by the developed framework is hereby a key factor to allow post-experiment analysis of the data without time-consuming manual annotation of the data set.

1.4. Objectives

The following objectives define the scope of the thesis. In addition, they provide the cornerstones of this research as a result of strategic planning and the review of related work. They are related to the corresponding Chapters 3-6 of this thesis.

OBJ1: to develop the technical design and experimental evaluation of a multimodal, affective framework for the EmotionBike with a focus on unimodal, time-sensitive recording and processing of facial video data with AFER.

OBJ2: to develop a method for metric-based benchmarking and accuracy evaluation of algorithms for state-of-the-art AFER solutions related to the EmotionBike as application-specific context.

OBJ3: to enhance the accuracy of evaluated AFER algorithms with application-specific tailoring the algorithm's output by exploiting the dimensionality with grouping of expressions.

OBJ4: to develop a method of recognising subject-independent changes in facial expressions by exploiting the event-based analysis and timing in the EmotionBike data set.

1.5. Structure of the Thesis and Corresponding Research Questions

To provide orientation for this work, a brief description for each chapter and its content is presented. Corresponding research questions (RQ) are posed based on the objectives and to guide the contributions and narrative.

Chapter 2 focuses on the fundamental literature relevant to this thesis and gives a background of this work. The presented background includes emotion theories, facial expression analysis and application areas. Out of this literature review the following questions arise:

RQ 2.1[4]: What is an emotion and how is it modelled for human-computer interaction?

RQ 2.2: What is the relationship between emotion and facial expression?

RQ 2.3: What are the current application domains for facial expression based human-computer interaction?

[4] The research questions were numbered according to the chapters.

Chapter 3 presents the time-sensitive technical design of the EmotionBike framework based on the general literature review. This framework provides the foundation of the EmotionBike experimental system. The requirements for this type of interactive system are developed with a focus on the synchronised processing and recording of AFER-related data. Evaluation of the design was performed by a logging mechanism during the main experiments. Chapter 3 poses the following research questions:

> **RQ 3.1:** What are the requirements for building a cross-platform framework for the EmotionBike?
>
> **RQ 3.2:** In order for a system to satisfy the requirements from RQ 3.1, how would it have to be designed?
>
> **RQ 3.3:** What are appropriate metrics to measure whether the defined framework design fulfills the requirements and was the evaluation successful?

Chapter 4 provides a methodology for designing application-related benchmarks of AFER algorithms by applying metrics to the algorithm's output with a focus on the game-event-based approach of the EmotionBike described in Chapter 3. It also presents an exemplary evaluation of four state-of-the-art AFER algorithms based on three strategically selected facial expression databases related to the video data characteristics provided by the EmotionBike. Based on the findings in the literature in Section 2.6.7, this chapter primarily addresses these two questions:

> **RQ 4.1:** How effectively do existing systems for AFER perform? What is an appropriate method to generate an AFER algorithm benchmark that reflects characteristics of the EmotionBike?

RQ 4.2: What are the challenges and limitations in state-of-the-art AFER algorithms?

Chapter 5 describes one approach to overcome limitations identified in Chapter 4 by the method of *application-specific grouping* of AFER algorithm output dimensions. This method was evaluated based on data from the previous benchmark and from the EmotionBike experiments.

RQ 5.1: What is a suitable method to improve the performance of state-of-the-art AFER in a specific application context by exploiting the dimensionality of the algorithms output?

Chapter 6 describes a second approach to overcome limitations identified in Chapter 4 by the developed *emotional shift* analysis method for facial expression algorithms' output, which provides a subject- and expression-independent methodology in event-based environments by exploiting the dynamics of facial reactions.

RQ 6.1: What is a suitable method of exploiting the event-based provocation in the EmotionBike to advance the subject-independent, time-sensitive analysis of AFER?

Chapter 7 concludes the thesis with a summary, the discussion of the results and research contributions together with an outline for future work including promising application areas for AFER.

Figure 1.2 displays a graphical representation of the thesis structure linked to the corresponding peer-reviewed publications from the publication list described on page 13.

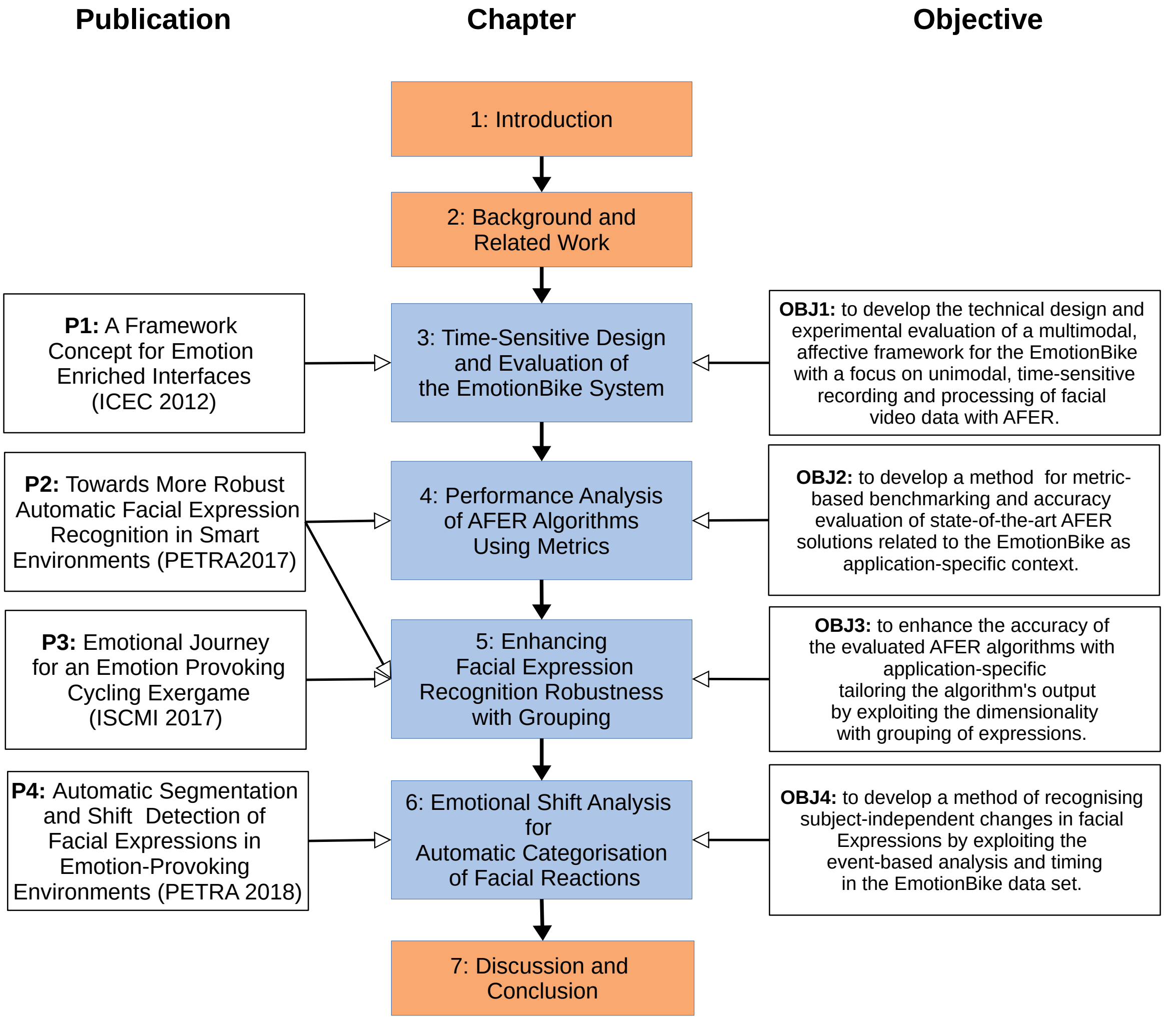

Publication
Chapter
Objective
1: Introduction
2: Background and Related Work
3: Time-Sensitive Design and Evaluation of the EmotionBike System
4: Performance Analysis of AFER Algorithms Using Metrics
5: Enhancing Facial Expression Recognition Robustness with Grouping
6: Emotional Shift Analysis for Automatic Categorisation of Facial Reactions
7: Discussion and Conclusion
P1: A Framework Concept for Emotion Enriched Interfaces (ICEC 2012)
P2: Towards More Robust Automatic Facial Expression Recognition in Smart Environments (PETRA2017)
P3: Emotional Journey for an Emotion Provoking Cycling Exergame (ISCMI 2017)
P4: Automatic Segmentation and Shift Detection of Facial Expressions in Emotion-Provoking Environments (PETRA 2018)
OBJ1: to develop the technical design and experimental evaluation of a multimodal, affective framework for the EmotionBike with a focus on unimodal, time-sensitive recording and processing of facial video data with AFER.
OBJ2: to develop a method for metric-based benchmarking and accuracy evaluation of state-of-the-art AFER solutions related to the EmotionBike as application-specific context.
OBJ3: to enhance the accuracy of the evaluated AFER algorithms with application-specific tailoring the algorithm's output by exploiting the dimensionality with grouping of expressions.
OBJ4: to develop a method of recognising subject-independent changes in facial Expressions by exploiting the event-based analysis and timing in the EmotionBike data set.

Figure 1.2: Graphical representation of the thesis structure. Chapter 3 describes the time-sensitive system design as a prerequisite for the post-experiment analysis of the recorded data set, while Chapters 4-6 focus on the contributions to the applicability of automated facial expression recognition (AFER) algorithms for the EmotionBike. The publications (P1-4, described on page 13) and objectives (Section 1.4) are linked to the corresponding chapters.

2. Background and Related Work

This chapter describes the works that are fundamentally to this thesis, with a focus on emotion theories, facial expressions, facial expression recognition algorithms, affective applications and smart environments. Additional, chapter-specific related work is presented in each subsequent chapter.

It has been shown that emotions actively influence cognitive processes such as planning and evaluating goals, recalling or saving of memories and reasoning (Calvo and D'Mello 2010; Poria et al. 2017). Since emotions and their expression are a fundamental component of human communication, they are highly important for human-computer interaction (HCI) (Zeng et al. 2009; Calvo and D'Mello 2010).

2.1. Affective Computing

The term **affective computing** was originally created by Rosalind Picard (Picard 1995). Picard defined four categories for affective systems Starting from class I as a normal computer with only a factual level of HCI and with no affect expression or recognition capabilities, the abilities are increased until class IV with bi-directional emotion recognition and expression. Calvo and D'Mello (2010) excluded Picard's class I (no affective capabilities) and proposed therefore three types of affective systems: '1. systems that detect the emotions of the user, 2. systems that express what a human would perceive as an emotion (e.g., an avatar, robot, and animated conversational agent), and 3. systems that actually 'feel' an emotion.' (Calvo and D'Mello 2010).

The applicability of affective computing was demonstrated for different areas. Some applications are directly linked to the environment (e.g. training on a exercise machine (Müller et al. 2015)), while other environments provide room for different applications; for example, learning (e.g. Graesser et al. (2016)), entertainment(e.g. Zagaria (2017)) or serious gaming (e.g Robinson (2014)) in a desktop environment.

2.2. Brief History of Emotion Theories and Facial Expression Analysis

A general overview, in the form of a brief history of facial expressions analysis and emotion theories is presented as an introduction.

The study of facial expression analysis (FEA) and emotions in science and philosophy reaches as far back as ancient Greece. Russell (1997) referred to the words of Aristotle (384–322 BC) as early evidence of the view of facial expression as 'glimpses into the heart of the other': 'there are characteristic facial expressions which are observed to accompany anger, fear, erotic excitement and all other passions'. Contemporaries of Aristotle, such as Socrates also contributed their perspective of facial expressions and their meanings, attributed to the 'science' of **physiognomy** – believing that facial geometry reveals the human character (Berland 1993). This was a common opinion until the middle of the 20th century, and was sometimes even used as justification for racism (Eliav-Feldon, Isaac, and Ziegler 2009).

When considering the literature on FEA in modern science, Charles Darwin (Darwin 1872) is often the first to be mentioned, although there are earlier works, such as the 'The mechanism of human facial expression' by G.-B. Duchenne de Boulogne, first published in 1865 (translated and reprinted in Boulogne (1990)). Duchene himself mentioned earlier works, such as the 'Dissertation on the Natural Varieties Which Characterize the Human Physiognomy' by P. Damper, published in 1792 (Hartley 2005) and 'The Anatomy and Philosophy of Expression' by Charles Bell (Bell 1844) as sources and inspiration for his research. Duchene and his predecessors focused on muscle activity and mechanics rather than the underlying emotions.

In 1855, the philosopher-psychologist Herbert Spencer articulated two fundamental principles that were to become the **psychological constructivist** approach to emotions. (Spencer 1855). Spencer argued that the class of mental states that people refer to as *emotion* is not of a different kind from the class of states that people refer to as *cognition*, even though people experience them as such. Instead, emotion and cognition differ in their emphasis on certain mental contents (Gendron and Barrett 2009).

Tomkins (1962) established the concept of **discrete** or **basic emotions**, which he attributed to Darwin's work. This concept was later adopted by Izard (1971) and Ekman and Friesen (1978). Basic emotion theories assume a fixed set of stimuli-response patterns (e.g. facial expressions) to a common set of non-overlapping, exclusive categories (basic emotions). According to Ekman (2009), Darwin (1872) treated emotions – derived from the facial expressions – as discrete entities or categories and as universal characteristic among all mammals. Gendron and Barrett (2009) disagreed, since Darwin thought that 'emotional expressions are vestiges of the past which are no longer functional in their present social context' (Gendron and Barrett 2009), and that 'although the basic emotion tradition of examining facial expressions of emotion is typically traced back to Darwin, Allport may be a more appropriate point of reference for this tradition' (Gendron and Barrett 2009).

As a contrasting concept to discrete emotional states, Wundt (1896) proposed a continuous space with three dimensions named *pleasure-displeasure*, *excitement-inhibition* and *tension-relaxation* (Reisenzein 1992). This concept of multidimensionality was later adapted by Schlosberg (1941) and Russell (1980).

Another concept of emotions is called **appraisal theory**. This was introduced by Lazarus (1966) and Arnold (1960) and was summarised by Moors et al. (2013) as 'an essence reflecting the ideas of Aristotle (384–322 BC), David Hume (1711 – 1776), Baruch de Spinoza (1632 – 1677), and Jean-Paul Sartre (1905 – 1980)'. The basic idea of appraisal theories is, that 'emotion is a reaction to some event after its implication for the self has been assessed by an individual' (Becker-Asano 2008), and is in reaction to a cognitive process rather than automatically triggered.

Panksepp (2007) argued against the conclusion of the 'attributional-dimensional constructivist view of human emotions that positive and negative core affects are the basic feelings – the primary processes – from which emotional concepts are cognitively and socially constructed'. From Panksepp's perspective, constructivism is speculation and is not supported by facts obtained from neuroscientific data. Instead, he argued for a concept of discrete emotions and the co-existence of basic emotions and appraisal (Panksepp 2007).

Gendron and Barrett (2009) identified three fundamental directions of emotion theories in the last 150 years: **basic emotions, appraisal theory and psychological constructivism**. Calvo and D'Mello (2010) added a fourth, **emotions as embodiment**, to this list and referred to James (1884) as the originator of this idea, while Gendron and Barrett (2009) stated, that James had an approach of psychological constructivism, and that it was Watson (1919) who reduced emotions to a physical state. The thesis that bodily responses induce emotions, also called the **facial feedback hypothesis**, was presented by Floyd Allport in the 1920s (Gendron and Barrett 2009). According to Calvo and D'Mello (2010), there are two more directions of emotion theories as products of neural circuitry: **neurobiology and core affect**.Additionally, Clore and Ortony (2013) proposed the classification of emotions by describing the situations in which they occur.

In the early 20th century, experimental psychologists aimed to prove that faces express emotions. They attempted to verify this under controlled experimental conditions using emotional stimuli and recorded the facial expressions that occurred (Landis 1924; Buzby 1924). They discovered a difference between the recognition rate of posed and spontaneous emotions by human observers and started to examine the role of the context in which facial expression occurred (Russell 1997).

In 1970, the Swedish anatomist Carl-Herman Hjortsjó published his book 'Man's Face and Mimic Language' (Hjortsjö (1970), describing a numbering scheme for facial muscles and muscle activity for emotional expressions such as disgust or surprise. Unfortunately, this work was ignored by the research community (Russell 1997). Ekman and Friesen (1978) published a similar work on coding facial expressions for basic emotions, called the facial action coding system (FACS), which had a significant influence on research and in particular on the later computational interpretation of facial expressions. FACS defines action units (AUs) for quantifying human facial movements by their facial appearance.

One of the main conflicts between the advocates of discrete (e.g. Ekman, Izard) and multidimensional emotion theories (e.g. Wundt, Schlosberg, Russell) was over whether the interpretation of facial expressions of emotions is universal for all humans since birth, or a social construct that is learned through social interaction and based on simpler non-emotion-specific (cognitive) processes. Russell (1994) added that facial expression not only includes seven basic emotions, but ‘laughs, pouts, yawns, winces, grimaces, and all manner of actions difficult to describe’.

In recent years, approaches have attempted to combine different directions of emotion theories. The ‘Lövheim cube of emotion’ is a combination of neurobiology and discrete emotions with a shared 3D space (Lövheim 2012). Cambria, Livingstone, and Hussain (2012) reinterpreted Plutchik’s ’wheel of emotion’ (Plutchik 1980) by organising primary emotion labels around four independent but simultaneous dimensions, for which their different levels of activation make up the total emotional state. This four-dimensional ‘hourglass of emotions’ might potentially be used to describe the full range of emotions discovered by humans (Poria et al. 2017).

Apart from discrete emotion theories, different multidimensional or hierarchical models for both, appraisal and constructivism have been proposed.

2.3. Selected Emotion Theories

The following section describes selected emotion theories that are relevant to computational emotional modelling and their relationship to facial expressions.

2.3.1. Terminology of Affective States

The meaning of the terms ‘emotion’ and ‘affect’ differs in literature. According to Gross (2010), affect refers to a superordinate category, while attitudes, moods and emotions refer to subordinate categories. However, others have suggested that affect refers to the experience of an emotion (or the emotion itself) or to the behavioural aspects of emotion (Gross 2010).

The term emotion entails a degree of uncertainty that is also apparent in the relevant literature. According to Gross (2010), the term emotion was directly adopted from common language, causing the problem of indefiniteness, as it inherits its meaning from folk theories on emotions (Russell 2003), or as Scherer (2005a) cited Averill and Frijda: ‘emotions are what people say they are’. The meaning of ‘emotion’ also differs among emotional models. In the context of the OCC model, it refers to a state of mind (Anderson et al. 2013), while in appraisal theories it refers to a process (Gratch and Marsella 2004) involving bodily changes and cognition (Scherer 2005a). According to Picard (1995), another term for an emotional state is a ‘sentic state’, referring to the Latin word sentire, ‘the root of the words sentiment and sensation’ (Picard 1995).

For this work, the author proposes the method of organising key terms in affective science as published by Gross (2010) and illustrated in Figure 2.1. In this concept, the term **affect** or affective state refers to a valence-based state (e.g. good and bad). **Attitudes** refer to the general positive or negative opinion with regard to somebody or something. **Moods** are not as stable as attitudes, and are often not specifically directed towards objects or persons. **Emotions** are the shortest in duration of these three processes. **Feelings**, **Behaviour** and **Physiology** are components within emotions.

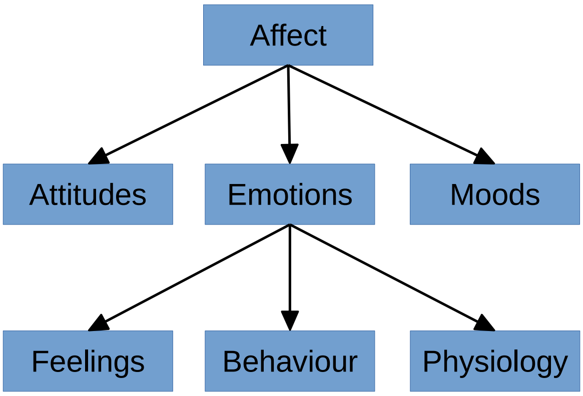


Figure 2.1: Graphical hierarchy of affect and emotions according to Gross (2010)

In the further course of this thesis, facial expressions that are associated with emotions are enclosed in quotation marks: e.g. ‘fear’ refers to the facial expression of the emotion fear.

2.3.2. Classifying Emotion Theories

Different classifications have been proposed to group emotion theories. Gendron and Barrett (2009) proposed a structure based on the underlying psychological concept of cognitive appraisal, basic emotion or psychological constructivism. Calvo and D’Mello (2010) separate emotion theories into the categories of ‘expressions, embodiments, outcomes of cognitive appraisals, social constructs, and products of neural circuitry’. Barrett and Lindquist (2008) grouped emotion theories according to the relationship between body, mind and emotional response. In this terminology, emotions are either caused by bodily changes resulting or an emotion felt in the mind causes bodily changes, or both are complementary parts of an emotion.

From a computer scientists' and application developers' perspective, it seems useful to group these theories according to the model structure (e.g. discrete, multidimensional, hierarchical), instead of the psychological or neurophysiological concepts that they describe, as the focus of this study is on their applicability for computer programs.

2.3.3. Discrete Emotion Models

Discrete emotion models propose separate, independent classes of emotions, either as mutually exclusive (e.g. basic emotions), categorical (e.g. positive-negative) or combined states (e.g. compound emotions) and are often closely related to the corresponding facial expressions.

Basic Emotions

One of the most influential theories on emotion models – especially in the field of computer science – is the concept of basic emotions (Izard 2007), sometimes referred to as primary emotions (Gendron and Barrett 2009) or prototypical emotions (Russell 1997).

Basic emotion theories assume that emotions, as a state of mind, automatically trigger biological programmes that share a common cause (Gendron and Barrett 2009). As these programmes are believed to be **universal**, every emotion triggers the same patterns of bodily responses, behaviours and **facial expressions** throughout all humans and can this be easily recognised (Gendron and Barrett 2009). In addition, basic emotions are defined as discrete, **mutually exclusive** states (Russell and Barrett 1999).

Author(s)	Basic Emotions (BEs)	# BEs
James (1884)	Fear, grief, love, rage.	4
McDougall (1908)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder.	7
Watson (1930)	Fear, love, rage.	3
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness.	11
Mowrer (1960)	Pain, pleasure.	2
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise.	10
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise.	8
Ekman, Friesen, and Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise.	6
Panksepp (1982)	Expectancy, fear, rage, panic.	4
Gray (1982)	Rage, terror, anxiety, joy.	4
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise.	9
Weiner and Graham (1984)	Happiness, sadness.	2
Ekman and Friesen (1986)	Anger, disgust, fear, joy, sadness, surprise , contempt.	7
Frijda (1986)	Desire, happiness, interest, surprise, wonder, sorrow.	6
Oatley and Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness.	5
Shaver, Schwartz, Kirson, and O'Connor (2001)*	Love, joy, anger, sadness, fear, surprise.	6
Jack, Garrod, and Schyns (2014)*	Happyness, sadness, fear-surprise, disgust-anger.	4

Table 2.1: List of proposed basic emotions, original from Ortony and Turner (1990); extended entries are marked *.

Although the number of proposed basic emotions ranges from two to 20 (Picard 1995), the best-known basic emotion model is the one developed by Ekman and Friesen (Ekman, Friesen, and Ellsworth 1982), typically referred to as the **six basic emotions**. This model originally contained the emotions of anger, disgust, fear, joy, sadness and surprise, but was later extended to seven basic emotions by including contempt (Ekman and Friesen 1986). Table 2.1 presents different numbers of basic emotions (6BE) as proposed in the literature.

Universality Debate

The debate about the universality of basic emotions, sometimes referred to as ‘basicality’, is still unfinished in psychology. The theory of universal basic emotions postulates that facial expressions that signal basic emotions are universal for all humans, regardless of their cultural background or spoken language, and are a fundamental concept of basic emotion theories (Nelson and Russell 2013).

Several studies have been conducted to find evidence for the universality of basic emotions. The outcome and especially the methodology of theses studies is highly controversial. According to Russell (1994) and Nelson and Russell (2013), the use of a within-subject study design combined with a forced-choice method for the emotion labels are the main reasons why the results of the studies on universality succeeded in finding evidence. Studies that utilised free labels could not produce evidence of the claimed universality of basic emotions, except for joy (Russell 1994). The recognition rate of distinct facial expressions from still images seemed to be related to the degree of knowledge of ‘Western culture’, which might be related to the tradition of using acted facial expressions in western societies (Russell 1994).

Jack et al. (2012) stated, that facial expressions of emotion are not culturally universal, but depend on cultural influence. Russell (1994) argued that basic emotions are not ideal for describing spontaneous communication, but that ‘it could be either an encoding or a decoding problem’ (Russell 1994). D’Mello and Kory (2015) mentioned that the six basic emotions rarely occur in real-life scenarios, but that other discrete expressions, such as boredom and frustration, are more common.

Secondary and Non-basic Emotions

Based on the theory that there are primary or basic emotions, some theorists postulate that there is a second group, called secondary emotions. Panksepp (2007) distinguished between primary ('feelings'), secondary ('related to learning and thinking') and tertiary ('thoughts about thought') emotions. Becker-Asano et al. (2008) defined secondary emotions as those that arise from higher cognitive processes that are based on the evaluation of outcomes and expectations, and named frustration and hope as examples. Zhang et al. (2018) named depression, agreement, distress and disappointment as examples of secondary emotions.

There is an overlap of definitions for secondary and non-basic emotions: Sariyanidi, Gunes, and Cavallaro (2015) simply defined non-basic emotions as all emotions not included in the (six or seven) basic emotions, thus including the secondary emotions. D'Mello and Kory (2015) stated, that 'boredom, confusion, frustration, engagement, and curiosity share some, but not all, of the features commonly attributed to basic emotions. Consequently, these are labeled as non-basic states'.

Categorical and Compound Emotion Models

Du, Tao, and Martinez (2014) proposed the model of compound facial expressions of emotion, combining the facial expressions of basic emotions into 15 new combined categories (e.g. 'happily surprised' and 'angrily surprised'). They also added 'neutral' and the standard six basic emotions to form 22 expressions overall. Du, Tao, and Martinez (2014) reported, that humans can reliably distinguish between these 22 expressions. Although no underlying emotion model was proposed, their findings supported the hypothesis that there are more complex emotional states than the six basic emotions (Yarkoni and Westfall 2017).

Scherer (2005b) suggested that free-response labels of subjects with regard to their emotional state should be sorted into 36 affective categories based on a synonym list[1], while Zeng et al. (2006) suggested a simple positive-negative model for their analysis of facial expressions 'to improve the quality of interface in HCI' (Zeng et al. 2006).

2.3.4. Hierarchical Models

Ortony, Clore, and Collins (1990) developed the OCC model, which mainly defines emotions as a hierarchical and logical structure, in the tradition of appraisal theories (Gendron and Barrett 2009). Clore and Ortony (2013) themselves mentioned that the OCC model is 'consistent with a constructivist approach', although they postulate that in the OCC-model, emotion is the response to an event, an action or an object, which could be considered to be an appraisal model.

Different enhancements to the OCC model were later proposed, as computer scientists argued that the logical structure was not consistent. These scientists published an updated version of the OCC tree (Steunebrink, Dastani, and Meyer (2009); see Figure 2.2) and a logically formalised version (Adam, Herzig, and Longin 2009).

[1] The categories were later included in his 2D Genevieve emotion wheel.

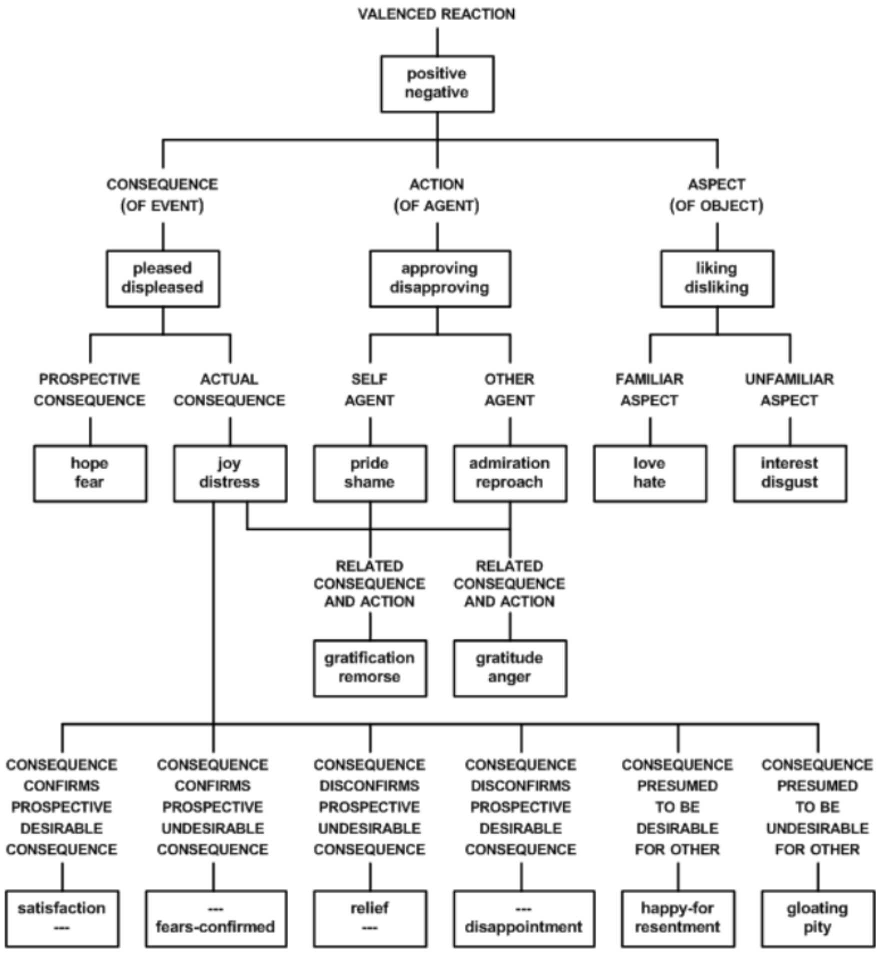


Figure 2.2: Updated version of OCC (Steunebrink, Dastani, and Meyer 2009)

2.3.5. Dimensional Models

Multidimensional models represent a set of continuous models for emotions (Picard 1995), typically with two, three, and recently, four dimensions.

Wundt

William Wundt (Wundt 1896) proposed the first continuous emotion model: a 3D affective space with the axes of *calm-excitement*, *strain-relaxation* and *pleasantness-unpleasantness*, which can be transferred to *intensity*, *arousal* and *valence* (Gendron and Barrett 2009). According to Russell (1980), Wundt developed his model on the basis of introspection rather than experiments. Wundt's approach is considered to follow a psychological constructivist approach (Gendron and Barrett 2009) and opposes the categorisation of basic or discrete emotions Becker-Asano (2008).

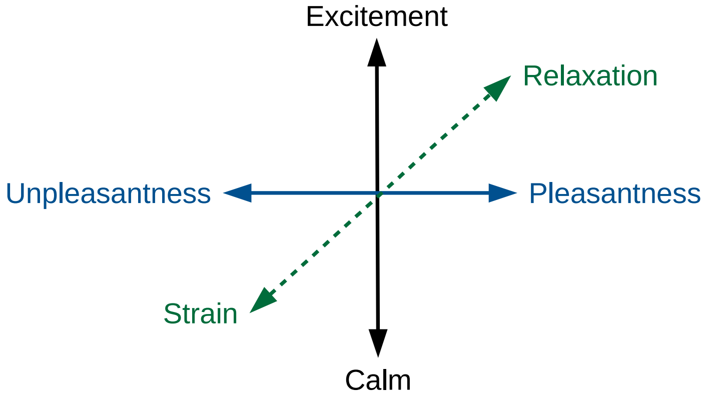


Figure 2.3: Wundt's model of emotion with the dimensions of pleasure (blue), arousal (black) and strain (green) (Wundt 1896).

PAD Space

PAD space was developed as an emotion model by Mehrabian and Russell (1974). It is named after its three numerical dimensions:*pleasure, arousal* and *dominance*. Pleasure is defined as a degree of valence, arousal as an indication of the level of emotional activation and dominance as the degree of situational control (Marsella, Gratch, and Petta 2010).

The PAD space model is similar to the 3D model by Wundt, with which it shares the first two dimensions. In contrast to Wundt's model, which classified 'strain' as a third dimension, in PAD space the third axis indicates the feeling of situational control (dominance) that is experienced. (Marsella, Gratch, and Petta 2010). The mapping of basic emotions within PAD space is common (Becker, Kopp, and Wachsmuth 2004; Zhang et al. 2010; Gebhard 2005), and Figure 2.4 depicts an example of eight basic emotions mapped into the 3D space.

PAD has been used in computer science to build believable affective agents (Becker-Asano 2008) and emotional avatars (Zhang et al. 2010). It has also been applied in organisational studies (Ashkanasy 2008) and consumer marketing studies (Ratneshwar, Mick, and Huffman 2003).

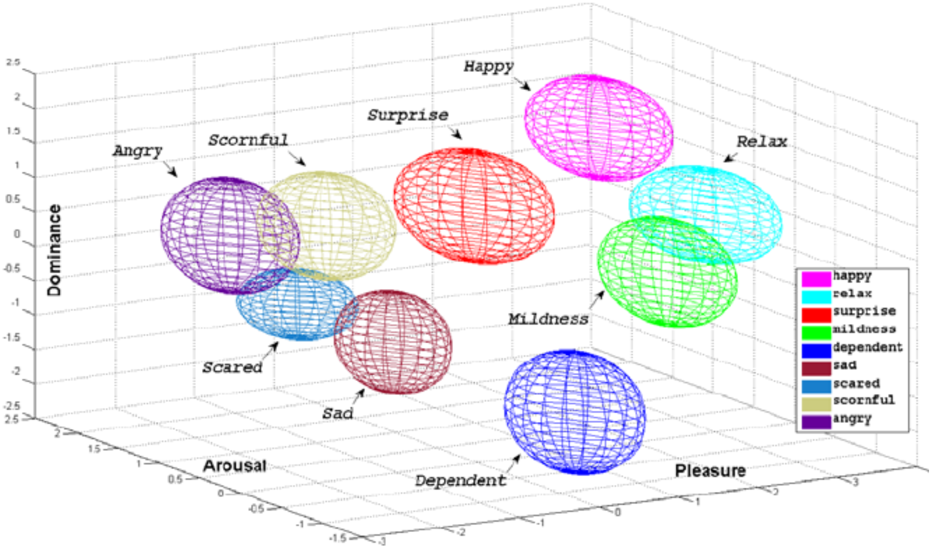


Figure 2.4: PAD space with mapped basic emotions (Zhang et al. 2010).

Russell and Barrett (1999) criticised the inclusion of dominance, since dimensions other than pleasure and arousal were founded on subsets of emotion-related words, and reflected the causes and consequences of the emotion, rather than being a third independent dimension.

Core Affect

In his framework in ‘Core Affect and the Psychological Construction of Emotion’, Russell (2003) proposed two primitives to describe the concepts of emotions: core affect and affec-

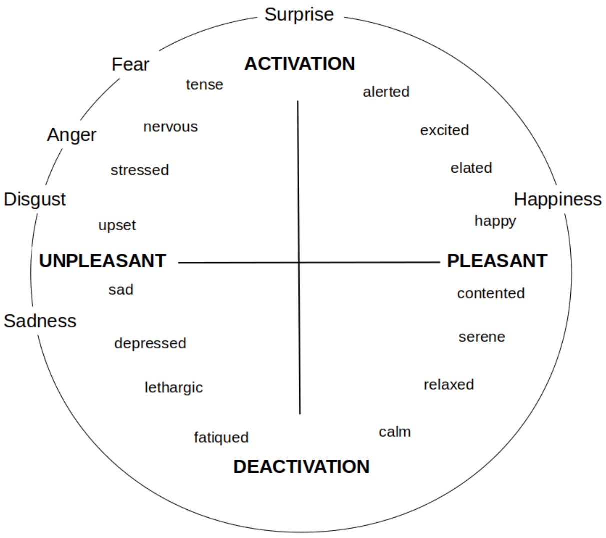


Figure 2.5: Core affect according to Russell, is formed by the *pleasant-unpleasant* and *activation-deactivation* dimensions. The outer circle describes the mapping of core affect to the six basic emotions as rare prototypical emotional episodes (Russell and Barrett 1999).

tive quality. Core affect is basically a 2D model of values for *pleasantness* and *activation* (depicted in Figure 2.5) that defines the current state of subjective feeling in a similar way to Wundt's model (Becker-Asano 2008), and was first presented by Russell and Barrett (1999). Russell (2003) described core affect as 'a neurophysiological state that is consciously accessible as a simple, non-reflective feeling'. The ability to change this state is called affective quality (Russell 2003), which is a similar concept to the experience that is commonly referred to as appraisal (Plass and Kaplan 2016). Basic emotions (which Russell calls 'prototypical emotional episodes') can be mapped on core affect, as depicted in Figure 2.5, but according to Russell (2003), these emotions seldom occur.

Panksepp (2007) criticised the reduction to only one core system, as 'cross-species neuroscience strongly supports the existence of many core emotional systems'.

Lövheim Cube of Emotion

The ‘cube of emotion’ was proposed by Lövheim (2012) and is a fusion of a 3D continuous model, basic emotions and findings from neurobiology. Lövheim assumed, that the monoamine neurotransmitters (serotonin, dopamine and noradrenaline) play an essential role in the control of behaviours and emotions and deliver ‘emotional information to large and dispersed areas of the brain’(Lövheim 2012). The levels of these monoamines form the three axes of his model. Lövheim assigns the eight basic emotions from Tomkins (1984) that have a high intensity (interest, enjoyment, surprise, distress, fear, shame, contempt, and anger) to each corner of the formed cube, representing the extremes of these emotions. ‘All emotions, including everyday tepid emotions, lie within the bounds of these basic emotions’ (Lövheim 2012). According to Lövheim, the main advantage of his model compared to the 3D models that were published before it is ‘its neurobiological correlate’ (Lövheim 2012). The Lövheim cube of emotion was mainly utilised in the field of artificial and computational intelligence to build cognitive-affective architectures that were able to simulate emotions in computational systems (e.g. Kugurakova, Talanov, and Ivanov (2016) and Vallverdú et al. (2016)).

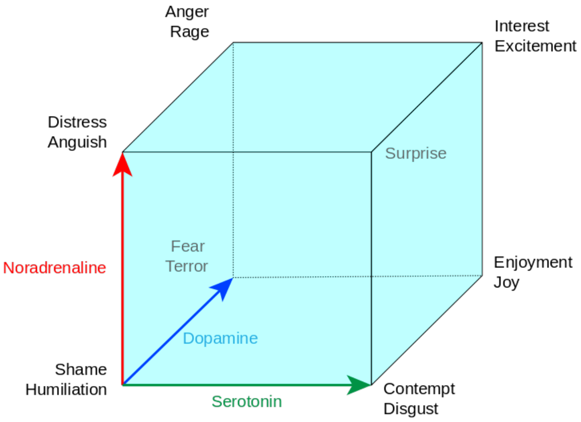


Figure 2.6: Lövheim cube of emotion (Lövheim 2012) combining basic emotions and monoamine neurotransmitters. Figure by Fred the Oyster[2].License: CC BY 3.0[3].

The Hourglass of Emotions (HGE)

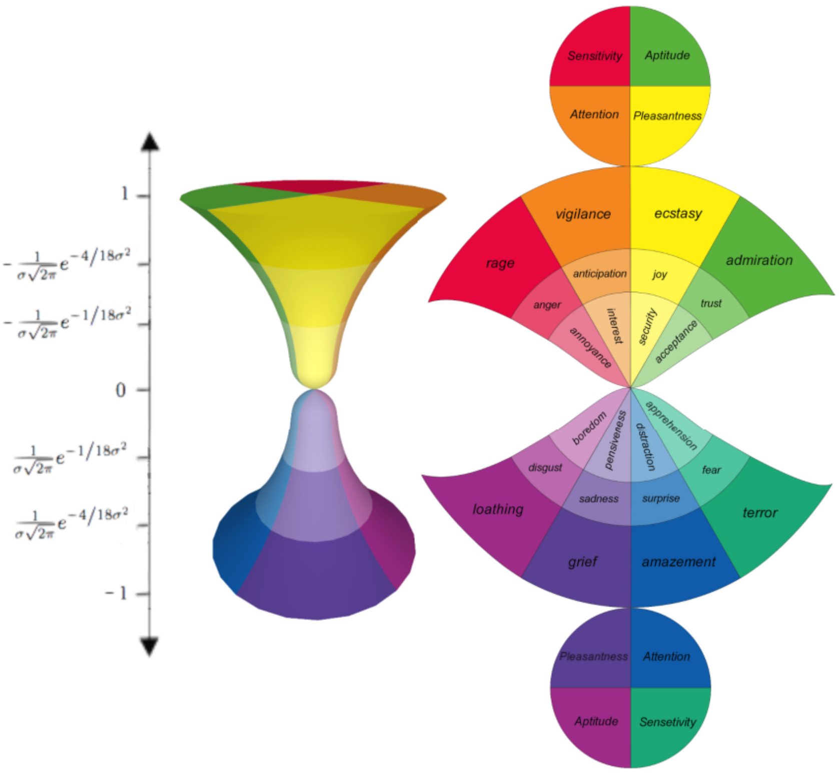


Figure 2.7: The 3D model and the net of the hourglass of emotions by Cambria, Livingstone, and Hussain (2012).

The HGE is a model proposed by Cambria, Livingstone, and Hussain (2012) and depicted in Figure 2.7. In contrast to other approaches, it utilises four dimensions: valence, potency, arousal and unpredictability. The HGE was inspired by Plutchik's studies on human emotions (Cambria, Livingstone, and Hussain 2012) and was mainly based on a study by Fontaine et al. (2008): 144 emotion features were rated according to 24 prototypical emotion terms and principal component analysis was utilised to reduce the number of dimensions to four. This model was applied to studies in the field of sentic computing, 'a

[3] https://commons.wikimedia.org/wiki/File:Lövheim_cube_of_emotion.svg
[3] https://creativecommons.org/licenses/by/3.0/

multi-disciplinary approach to opinion mining and sentiment analysis, to semantically and affectively analyse text and encode results in a semantic aware format according to different web ontologies' (Cambria et al. 2012).

2.3.6. Embodiment in Emotion Theories

The relationship between physicality and emotion is particularly important for facial expression recognition, as facial muscles are always involved in displaying an expression.

Picard (1995) referred to the **body response theories**, proposing that emotions are induced by sensory feedback from (facial) muscles. Schlosberg (1954) attributed this theory to James (1884), but partly disagreed, as he stated that the feedback from skeletal and visceral responses is a required – but not sufficient component – of emotions. Ignoring visceral responses, this theory is termed the **facial feedback hypothesis** and its origin has been attributed to Darwin (LaFrance 2000) and later to Allport and Tomkins (Gendron and Barrett 2009).

Body response theories (especially James (1884)) are attributed to constructivism (Gendron and Barrett 2009), and are considered independent of the resulting classification model of emotion (e.g. discrete or dimensional).

The underlying debate of body response theories seems to be a variant of the *chicken or the egg* paradox: are emotions responses to bodily changes, are bodily changes responses to emotions, or are they independent of each other?

Reisenzein and Stephan (2014) answered that bodily feedback is not a necessary component for emotions, as emotions still occur even if facial muscles are completely blocked by curare or botulinum toxin injections. Facial muscles are also naturally blocked during rapid eye movement sleep, yet emotions still often occur during dreams in this sleep phase (Reisenzein and Stephan 2014). Physical states can, however, trigger emotions. Sadness, fear or panic can be induced by stimulating the subcortical deep-brain (Panksepp 2007); fear or panic may be induced by unusual physical states, such as in the case of panic disorders, even leading to a positive feedback loop between felt fear and bodily sensations (Roth 2010).

2.3.7. Attributed and Situational Emotion Theories

Russell (2003) mentioned that emotions are always directed towards an event or object, e.g. 'Alice was afraid of the bear, or more precisely, of the bear attacking and harming her' (Russell 2003). This contradicts to the concept of emotions as psychological primitives, as proposed by the basic emotion theories. A similar approach is to classify emotions by describing the situations in which they occur (Clore and Ortony 2013), but as people react individually to situations, their resulting emotions can be different (Picard 1995).

While numerous emotion theories have been proposed, discussed and evaluated in the research on emotions, their applicability for practical solutions must be demonstrated in reality.

2.4. Current State of Emotion Theories and Models

From the author's perspective, there seems to be an ongoing and unfinished debate rather than a clear definition in psychology and its related disciplines with regard to what a facial expression means, what an emotion is, and how they are related to each other. Or, to quote Reisenzein,

> ‘From the perspective of an applied affective computing researcher, psychology would ideally be at an advanced stage where the ultimate correct theory of emotion (UCTE) or even better, the ultimate unified theory of the mind (UUTM)had been attained. Furthermore, that theory would ideally be formulated as a computational model, or would at least be available in a format that lends itself readily to implementation as a computer program. Unfortunately, psychology has not yet arrived at this stage.’ (Reisenzein et al. 2013)

This opinion is not uncommon: ‘In the earliest days of psychology, William James famously asked ’What is an emotion?’ (James, 1884). He still hasn’t received a satisfactory answer’ (Gross 2010). As Izard (2007) described it, ‘there is no consensus on a definition of the term emotion, and theorists and researchers use it in ways that imply different processes, meanings, and functions’. Reisenzein et al. (2013) referred to the work of Strongman (2003), who listed approximately 150 psychological and philosophical emotion theories that had been proposed throughout history until 2003. This number has since increased (e.g. Cambria, Livingstone, and Hussain (2012); Lövheim (2012)).

The debate between supporters of the different directions of emotion theories seems unfinished, but Izard (2007) argued that these alternative theories might be considered to be complementary yet necessary constructs. From a practical perspective, different types of emotion theories have been applied to affective systems. Multidimensional approaches have been mostly applied to virtual agents (e.g. Becker, Kopp, and Wachsmuth (2004), Gebhard (2005)) or when utilising physiological sensors (e.g. Valstar et al. (2016)), while **discrete emotion models have been mostly applied to AFER systems** (e.g. Littlewort et al. (2011) and Stöckli et al. (2017)).

Although the type of emotion theory that is closest to reality is as yet uncertain, the assumption that there is a connection between emotions and facial expressions is common to many emotion theories (Clore and Ortony 2013; Gendron and Barrett 2009; Russell 1994).

While a precise definition of emotion is still unclear, emotion recognition can still be applied in computer science applications, as ‘this problem of not being able to precisely define categories occurs all the time in pattern recognition and fuzzy classification’ (Picard 1995).

2.5. Modalities for Emotion Recognition

Several approaches to recognise affective states that utilise or combine different modalities have been demonstrated in the field of computer science. The comprehensive list by D’Mello and Kory (2015) included AFER, voice and speech analysis, text analysis, body postures and gestures, eye gaze, electroencephalography (EEG) and physiological signals (e.g. electrodermal activity (EDR), respiration, electromyography (EMG) and electrocardiography (ECG)).

Most approaches (even for physiological signals) utilised discrete emotion modelling (basic emotions and non-basic emotions), and few dimensional – especially one-dimensional (arousal) and 2D (e.g. valence-arousal) – modelling Calvo and D’Mello (2010).

Thermographic cameras provide another modality for AFER. Thermal imaging is limited due to dull areas in the images; therefore, an appearance-based approach that utilises the complete face may be applied (Corneanu et al. 2016). An additional method is the registration of normal camera and thermal images.

To achieve 3D facial expression recognition, a 3D facial model is commonly registered on a 3D point cloud, resulting in a 3D geometrical model with already known landmarks (Corneanu et al. 2016).

According to a survey by Noroozi et al. (2018), the main sources of bodily emotion recognition are hand, head and torso position (body posture) and movement (body gesture). Several approaches have been demonstrated to recognise the valance-arousal and mostly discrete emotional signals from these body gestures (Noroozi et al. 2018).

Various approaches have been proposed and demonstrated for combining different modalities by utilising multimodal fusion, most of which have focused on visual (facial expressions, body posture) and acoustic (voice, speech) modalities (Poria et al. 2017). Although the multimodal fusion offers the possibility to increase the robustness of the whole detection, Harley (2016) argued that body gestures are often largely redundant to other modalities (e.g. facial expressions and context), and that multiple modalities may use different emotion models, which makes fusion challenging. An exception to this redundancy is AFER in communication scenarios that involve speech, as utilising a fusion with aural emotion cues can compensate for the distortion of facial expressions caused by speaking (Shah et al. 2013). Despite this increase in robustness, only moderate increase in overall recognition rates are to be expected (D'Mello and Kory 2015).

Despite the availability of additional modalities, facial expressions are still considered to be the dominant modality for understanding emotions (Poria et al. 2017; D'Mello and Kory 2015; Harley 2016) and were therefore chosen as primary source of emotion recognition for the EmotionBike. In addition, efficient unimodal emotion detection provides the fundament for effective multimodal recognition and is thus a requirement for building well-performing multimodal affect recognition (Poria et al. 2017).

2.6. Automated Facial Expression Recognition

This section provides a detailed description of the nature of facial expressions, detection algorithms and open challenges.

2.6.1. Facial Expressions

A common view on defining facial expressions is to divide them into three different phases, named **onset** (muscular contractions start and increase), **apex** (maximum muscular contraction) and **offset** (muscular relaxation) (Sariyanidi, Gunes, and Cavallaro 2015), as depicted in Figure 2.8. According to Sariyanidi, Gunes, and Cavallaro (2015), a neutral, expressionless phase can occur before the onset and after the offset.

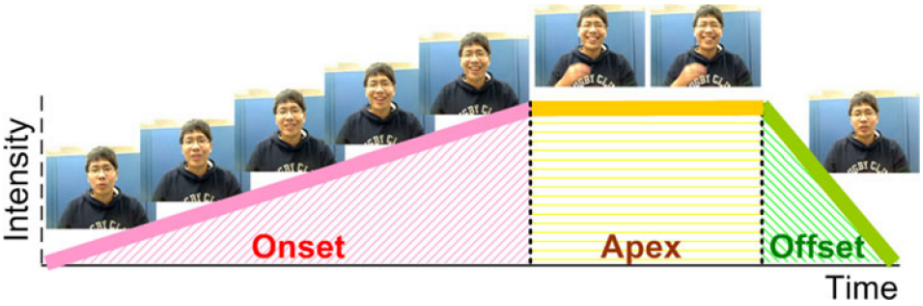


Figure 2.8: Three phases of a prototypical facial expression in their normal order: onset, apex and offset. The duration of the phases is variable, but onset and offset are usually shorter than the apex. The image was taken from the survey by Chung-Hsien Wu, Lin, and Wei (2014) (License: CC BY 3.0[4].)

Although there is often a great variance in the duration of all three phases, onset and offset are typically short in length, while apex is usually the longest phase. Combinations of

[4] https://creativecommons.org/licenses/by/3.0/

the phases are often displayed in spontaneous facial expressions with multiple expression apexes (Koelstra, Pantic, and Patras 2010), while datasets often contain acted expressions (e.g. Kanade, Tian, and Cohn (2000)).

Facial expressions can be divided into *normal* and *micro-expressions*; the latter are sometimes called *leaking expressions* (Ekman and Friesen 1978; Yan et al. 2013). Although discussion continues about the relevance of duration as a criterion of differentiation (Yan et al. 2013), micro-expressions appear to last less than 0.5 s (less than 0.3 s according to Corneanu et al. (2016)), while normal expressions typically last longer, often exceeding 1 s (Yan et al. 2013).

2.6.2. Spontaneous and Posed Expressions

Facial expressions are often distinguished in terms of 'conscious' or 'unconscious'[5] movements of facial muscles. This differentiation in *spontaneous/posed* (Russell 1997), *naturalistic/acted* (Bosch et al. 2015), *actual/simulated* (Russell 1997) or *involuntary/voluntary* (Ekman 2009) expressions seems to be grounded in neurophysiological studies and literature (Borod, Haywood, and Koff 1997) and the suggested different pathways for spontaneous and posed expressions in the human brain. However, the review by Borod, Haywood, and Koff (1997) of 49 studies that focused mainly on the asymmetry of expressions found no support for this thesis. Instead, the neurophysiological pathways for posed and spontaneous expressions seem to be identical; 'this review strongly supports the notion that the left hemiface is more involved than the right hemiface in the expression of facial emotion, regardless of valence, face part, elicitation condition, or the operation of social display rules' (Borod, Haywood, and Koff 1997).

[5] Although this distinction is controversial.

Although the (neuro-) physiological cause for this differentiation in naturalistic and acted expressions may be unclear, the distinction between spontaneous and posed expressions has been proposed in psychology since the 19th century. An example of this is the differentiation in Duchenne (spontaneous) and non-Duchenne (posed) smiles (Boulogne 1990; Schmidt and Cohn 2001). Besides the suggestion of a difference in psychology, significant differences between posed and spontaneous facial expressions have been reported in experiments. For example, Bosch et al. (2015) referred to the study of Hoque, McDuff, and Picard (2012) who classified video recordings of frustrated and delighted smiles for frustration. Smiles were present in 90% of the cases that demonstrated spontaneous (provoked) frustration, but only in 10% of the posed variants.

Naturalistic and acted expressions also often differ in the displayed intensity. Naturalistic expressions are typically of low to moderate intensity and may have multiple peaks in intensity (Corneanu et al. 2016), while datasets that contain posed expressions typically have a fixed climax, progressing directly from a neutral state to an expression (e.g. Kanade, Tian, and Cohn (2000)). The way in which spontaneous expressions are displayed may be altered by the social context or (culture dependent) displaying rules (see Section 2.6.3 for details).

To **obtain spontaneous expressions, a provoke-response pattern is applied to induce an emotional state**. Typically, the expressions are provoked with an emotional stimulus or obtained in a communication setup between at least two protagonists (Weber, Soladie, and Seguier 2018). As emotional reactions to stimuli are subject-dependent, a certain level of uncertainty is unavoidable as ‘it is impossible to know objectively what emotion is felt by the subject, how it is perceived by a third party and how much the facial expression reflects it’ (Weber, Soladie, and Seguier 2018). Posed expressions are easier to acquire and reproduce. In addition, posed expression data provides the benefit of eliminating disturbances that are widespread in daily emotional communication, such as social display rules or personal emotion control strategies (Krumhuber et al. 2017). Choosing the best detection approach may also vary for posed and spontaneous data (Sariyanidi et al. 2013).

The differentiation between the two types of expressions is based on the assumption, that (spontaneous) facial expressions are universally natural and are not learned from social interaction. In the latter case, all expressions would be ’posed’ to some extent, and thus the differentiation would be unnecessary.

2.6.3. Social Context of Facial Expressions

Fridlund (1991) discovered that watching a film with or without company had an effect on the intensity of the viewer’s smile; the smiling facial expression was less intense if the film was watched alone, and increased when a co-viewer was physically or virtually present. A repetition of the experiment only confirmed the effect in case of a physical presence and that the intensity of the smile had no effect on the intensity of the underlying emotion: happiness (Fischer, Manstead, and Zaalberg 2003).

Fischer, Manstead, and Zaalberg (2003) derived three conclusions from their experiments. Firstly, with the physical presence of another person, a more intense smile was expressed. Secondly, the intensity and frequency of smiling was directly connected to the type of film (e.g. moderately or highly amusing). Thirdly, the relationship of the co-viewer to the viewer was important; more smiles of enjoyment were displayed in the presence of friends than strangers.

Schützwohl and Reisenzein (2012) confirmed the social influence on smiling in the results of their experimental setting, but did not find a significant influence with regard to surprise; only 33% of their subjects displayed the facial expression ‘surprise’, regardless of the presence of another person. Unfortunately, they did not mention whether a change to other facial expressions occurred.

Although the normal ‘surprise’ expression seems to be difficult to detect by itself, if the surprise is combined with a frightening aspect (a jump-scare), reactions are much different. The resulting facial expression may be termed ‘surprise’ or ‘fear’ (Cheng et al. 2017).

In another experiment, Jakobs, Manstead, and Fischer (2001) provoked sadness by using a film as a stimulus. They noticed significant differences in the expressiveness of the displayed ‘sad’ facial expressions. However, in contrast to the experiments with amusing stimuli, the effects were reversed: viewers displayed less expressiveness in the presence of others (regardless of whether they were friends or strangers). Fischer, Manstead, and Zaalberg (2003) suggested that displaying sadness in the presence of others might be considered as inappropriate and was thus avoided.

Zaalberg, Manstead, and Fischer (2004) investigated the relationships between emotions, display rules, social motives, and facial behaviour. Laughing or smiling was again the object of investigation, provoked by a joke. As a result, they could distinguish between honest laughter and something they termed ‘social or polite smile’.

There may also be a significant sex or gender difference in emotion expressions (Timmers, Fischer, and Manstead 1998) that result in different intensities of expressions and whether a certain facial expression is displayed in a social context.

Although the question of whether facial expressions are universal and thus culturally independent (see Section 2.3.3) has not been answered, cultural differences in displaying emotional facial expressions (display rules) exist. As an example, the Japanese often mask their negative expressions (sadness, fear, disgust, anger) with a smile (Matsumoto 1991).

To summarise, social context can have a significant effect on the occurrence of facial expressions. This is especially important for non-laboratory scenarios (refer to Section 2.6.7)

2.6.4. AFER Algorithms

Numerous algorithms for recognising facial expressions in processing videos and images have been demonstrated over the years since the first publication on AFER Suwa (1978). The surveys of Zeng et al. (2009), D'Mello and Kory (2015), Sariyanidi, Gunes, and Cavallaro (2015), Martinez and Valstar (2016) and Corneanu et al. (2016) provide a comprehensive overview. In addition, although the primary focus of Martinez et al. (2017) is the automated coding of facial actions (FACS-based), they provide a recent survey on methods for facial detection, registration, filtering and feature extraction, which are the same technologies that are applied to direct expression detection.

Many state-of-the-art AFER algorithms utilise a discrete emotion scheme (described in Section 2.3.3), whether it is the classical six basic emotions or the extended models with additional expressions referred to in this work as **extended basic emotions** As examples for this additional expressions, Bartlett et al. (2008) added 'contempt', McDuff et al. (2016) 'contempt' and 'smirk', Mone (2015) 'contempt', 'confusion' and 'frustration'.

Although different machine learning techniques (e.g. support vector machines (SVM), neuronal networks (NN)) have been applied to AFER algorithms, there is little difference in the accuracy of the different approaches in practical applications (Martinez and Valstar 2016).

General Approach

A common approach for a processing pipeline of an AFER algorithm is depicted in Figure 2.9. Many AFER solutions consist of the following components:

1. **Face detection:** detecting the face is the first stage and is essential for further processing, as this depends on the correct localisation of faces in the image.

2. **Filtering and registration:** the recognised face serves as the input for different stages of (optional) pre-processing filters. Typically, this includes the normalisation of size and alignment and the compensation of illumination variations. The face is registered to a geometric or appearance-based model.

3. **Feature extraction:** the next stage reduces dimensions of the data through feature extraction based on the registered geometric or appearance-based model.

4. **Classification:** based on trained machine learning or a statistical model the features are classified to generate the algorithm's output.

Algorithms can be trained to recognise AUs (e.g. McDuff et al. (2016)) or facial landmarks (e.g. Michel and El Kaliouby (2003)) as an intermediate step, or to directly recognise the facial expression from the features of the image (e.g. Valstar et al. (2011a)), which has been referred to as message-based (facial expression) or sign-based (AU or landmark) (Pantic 2009).

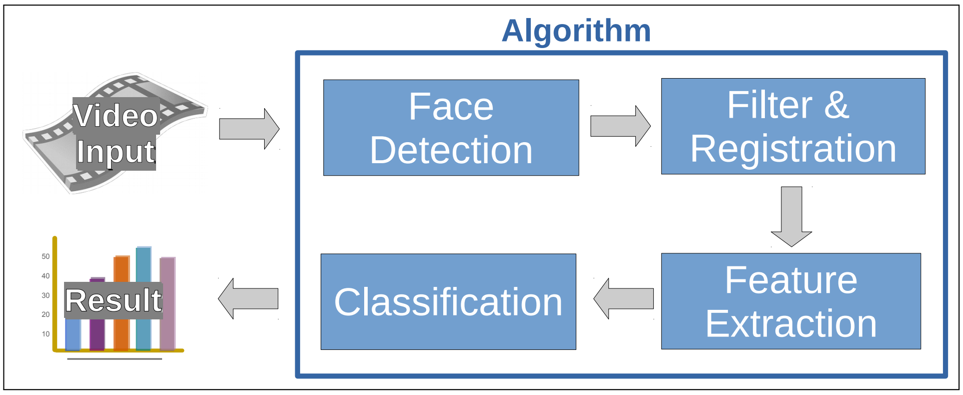


Figure 2.9: Example of a pipeline structure for a facial expression recognition algorithm derived from Corneanu et al. (2016). Inputs from video or camera are analysed utilising face detection, filtering and registration, feature extraction and classification that are mostly based on machine learning models.

Face Detection

The first step of AFER, as depicted in Figure 2.9, is detecting the face with algorithms are on machine-learning and pre-trained models approaches.

A standard approach for face detection is the algorithm developed by Viola and Jones (2004) based on Haar-like feature models trained with AdaBoost. Other face detectors utilise deformable parts models (DPMs) (e.g. Mathias et al. (2014)) or a combination of histogram of oriented gradients (HOG) and support vector machines (SVMs) (e.g. Zhu and Ramanan (2012)) providing a better detection rate for not near-frontal images but at higher computational costs (Martinez et al. 2017). According to Mathias et al. (2014), 'it turns out that for face detection the children of two classic detection approaches, Viola & Jones and HOG+SVN, are the best performing methods'.

The reasons why Viola and Jones (2004) is still being widely used is the usually sufficient performance ,especially for near-frontal images, and the ease of availability due to the freely available implementation[6] (including trained models) as part of OpenCV[7], a widely-used open source library for computer vision (Martinez et al. 2017).

While these algorithms usually work accurately for frontal or near-frontal views, face detection in larger areas (e.g. smart homes) remains challenging. Brauer, Grecos, and Luck (2014) proposed a face detection algorithm that was specially designed for 180-degree fisheye cameras to cover a wide area from the ceiling. Maglogiannis (2014) detected the silhouette as a pre-stage to face detection with this type of camera.

Ding and Tao (2016) mentioned, that although great advances have been accomplished in facial recognition, there is still much room for improvement, and performance testing of existing approaches with real-life databases is still needed for further development.

A detailed overview of face detection algorithms may be found in the surveys of Zafeiriou, Zhang, and Zhang (2015) and Ding and Tao (2016).

Filtering and Face Registration

Depending on the model used for the registration of the face, some prior filtering may be applied, including illumination compensation, alignment or cropping. These steps are a common case when utilising low-level models such as Gabor representations. Facial expression models are commonly categorised as geometric (e.g. a 2D mesh model) or appearance-based (e.g. a Gabor filter) (Zeng et al. 2009; Corneanu et al. 2016).

Registration has a critical effect on the robustness of the algorithm against pose variations, which often occur in non-frontal scenarios. While simple systems that are designed for frontal

[6] even for commercial applications.

[7] https://opencv.org/

data use two to four points, more sophisticated systems that are more robust to head rotations use whole faces, parts of them, or point registration (Sariyanidi, Gunes, and Cavallaro 2015)

Feature Extraction

The main purpose of feature extraction is to reduce dimensionality (Sariyanidi, Gunes, and Cavallaro 2015) before classifying the data. Image difference, edge detection and Gabor wavelets have been applied to extract features in appearance-based models (Poria et al. 2017). Typical features of geometric models include facial point locations (Poria et al. 2017; Sariyanidi, Gunes, and Cavallaro 2015), active contours that describe parts of the face (e.g. eyes and mouth) or deformable models (e.g. AAM) (Corneanu et al. 2016).

Classification of Facial Expressions

Many AFER systems have applied SVMs as the primary approach for the classification of facial expressions (Sariyanidi, Gunes, and Cavallaro 2015). In addition, numerous classifiers have been applied, such as nearest neighbour, Bayesian networks, AdaBoost and neural networks (Poria et al. 2017; Corneanu et al. 2016).

2.6.5. Dynamics of Facial Expressions

Spatial representations that process the video frame-by-frame are still the standard approach for AFER algorithms. Spatio-temporal representations of facial expressions that consider the neighbourhood of frames (Sariyanidi, Gunes, and Cavallaro 2015) for AFER algorithms have been proposed since 2005 (Pantic 2009), but until now, the dynamics of facial expressions have only been exploited to ensure the consistency of labelling over time (Martinez and Valstar 2016).

This is surprising, since the dynamic nature of facial expressions is one of their main characteristics and the information provided by the dynamic motion enhances the accuracy of the recognition of facial expressions by human observers (Krumhuber, Kappas, and Manstead 2013). The advantages of dynamic information are most evident when static information is limited, such as with occlusions or unfavourable viewing angles (Krumhuber, Kappas, and Manstead 2013). While dynamic video sequences are not superior to human observers in recognising **intense** emotional facial expressions, such as those that are posed, they provide an additional source of recognition if more subtle expressions are displayed, as in spontaneous scenarios (Krumhuber, Kappas, and Manstead 2013). According to Martinez et al. (2017), posed and spontaneous facial expressions are significantly different in terms of their temporal dynamics, which was previously reported by Zeng et al. (2009).

There are multiple reasons why the temporal dynamics have not been widely included in state-of-the-art AFER systems. Firstly, the form and function of facial expression temporal dynamics are not fully understood (Jack, Garrod, and Schyns 2014; Martinez et al. 2017). Secondly, as Krumhuber, Kappas, and Manstead (2013) stated, the advantages emerge only if there are expressions with low intensity (spontaneous expressions) or restricted conditions (occlusions and view angles). This is true for real-life or in-the-wild scenarios, which only recently became a focus for AFER. Thirdly, a way in which the dynamics can be modelled or included into existing representations has not yet been identified (Martinez and Valstar 2016).

2.6.6. Intensity and Probability of Facial Expressions

It is common for AFER algorithms to output the probability (likelihood) of all possible facial expressions (e.g., six basic emotions) that they are able to detect (Bosch et al. 2015). In an ideal situation, AFER algorithms would additionally provide the intensity of these expressions. However, the intensity of a facial expression is often closely related to the probability of detecting this expression; the more expressive a face is, the easier the expression is to detect (Wells, Gillespie, and Rotshtein 2016; Martinez et al. 2017).

Differences in intensity and timing can allow for differentiation between acted and spontaneous smiles, and polite and embarrassing smiles (Corneanu et al. 2016). AU intensities can also be used to detect pain levels from facial expressions (Kaltwang, Rudovic, and Pantic 2012). Although there is no general scale for the intensity of an expression, Dhall et al. (2012b) proposed *neutral* ,*small smile*, *large smile*, *small laugh*, *large laugh* and *thrilled* for the intensity labels in their HAPPEI dataset. As stated by Corneanu et al. (2016), most video (RGB) datasets do not have labelling for the intensity of expressions.

The AU intensities are evaluated on a five-level ordinal scale, A-B-C-D-E, where A indicates the lowest and E the highest intensity (Martinez et al. 2017). Littlewort et al. (2011) proposed utilisation of the distance to the hyperplane for their SVM-based detection approach as an indication of the intensity of the expression. However, this was questioned by later research (Martinez et al. 2017). Martinez et al. (2017) considered the estimation of AU intensities to be an unsolved problem, especially because the task of annotating AU intensities in datasets is challenging due to the small variations between the levels (Corneanu et al. 2016). Furthermore, inter-rater reliability for AU intensity labelling is lower than for occurrence labelling (Martinez et al. 2017). In recent years, the intensity of facial expressions and AUs has become a trend in the research field.

According to Hess, Adams Jr, and Kleck (2004), there is also a gender aspect to the intensity of facial expressions: displays of happiness among women were reported to be more intense than among men. With disgust and anger, men's expressions were reported to be more intense. The same was true for the self-perception of the emotions for participants.

2.6.7. Challenges of Automated Facial Expression Recognition

Although many AFER approaches have been developed in the last two decades and some have become available as commercial products, fundamental problems remain unsolved, especially in demanding, real-life environments (Pantic 2009; Martinez et al. 2017).

> In reality, felt emotions are not always visibly manifest because the experience of them is subjective, nor do they map cleanly onto Ekman's six categories. Another limitation is that expressive facial signals are highly context-dependent and are involved in communication besides emotions, such as cognitive load, back-channelling, and turn-taking (Gunes and Hung (2016)).

Outside the AFER research community, the common belief is that facial expression recognition is a solved problem. This seems especially true for the media, as they focus on success stories of the application of AFER (Gunes and Hung 2016). Inside the AFER research community, the view is different, as the analysis of facial expressions from common, real-life interactions is still limited (Gunes and Hung 2016).

In their comprehensive survey on automatic analysis of facial affect, Sariyanidi, Gunes, and Cavallaro (2015) named two major obstacles that hinder further development of AFER algorithms. Firstly, the validation of AFER systems on posed data, which are 'often insufficient for everyday life settings' (Sariyanidi, Gunes, and Cavallaro 2015). Secondly, the application of sophisticated statistical models, such as SVMs, even for rather trivial problems, instead of resolving them with simpler methods of classification.

Quality of Training Data

Martinez and Valstar (2016) summarised the effects of training data quality on the performance of supervised machine learning-based AFER algorithms:

> Any learning problem is primarily determined by the data available. The dependence of performance on the quality and quantity of data can hardly be overestimated. An inferior method trained with more abundant or higher quality data will most often result in better performance than a superior method trained with lower-quality or less abundant data. This is particularly dramatic in the case of facial expression recognition. (Martinez and Valstar (2016))

The quality and quantity of the data depends on the way it was recorded and annotated, the complexity of the process and how suitable the training data is for the target environment. (e.g. posed versus spontaneous expressions) (Sariyanidi, Gunes, and Cavallaro 2015).

Labelling datasets with emotional facial expressions is a laborious and challenging task, especially when the expressions are spontaneous rather than posed (Pantic 2009). The subjective perception of the manual annotators of emotions causes flaws in the labelling of spontaneous expressions more often than in posed ones (Sariyanidi, Gunes, and Cavallaro

2015). In general, **for posed expressions, subjectivity lies with the actor who interprets the instructions, and with the (manual) annotators**. **For spontaneous expressions**, the actors subjectivity is replace by the **subject-dependent nature of responses to emotional stimuli** (e.g. Schaefer et al. (2010), Rumpa et al. (2015)).

The manual annotator's subjective criteria often have a significant effect when assigning labels to videos (Martinez and Valstar 2016). One way to reduce this problem is to have the data labelled by multiple annotators and then use the average or majority decision for reliable labelling. Unfortunately, this increases the number of manual annotators required for consistent labelling, and therefore the required resources (Martinez and Valstar 2016). This leads to the problem of the availability of sufficient (high-quality) training data, especially as modern methods for machine learning (e.g., CNNs) require a greater amount of training data than those currently available (Martinez et al. 2017). To address this, automatic (e.g. Fabian Benitez-Quiroz, Srinivasan, and Martinez (2016)) and semi-automatic (e.g. Dhall et al. (2012a)) labelling has been proposed, to reduce the resources of annotating large datasets. However, the use of automatic systems for labelling large datasets only transposes the problem, because these systems also need to be trained and validated. In addition, the labelling of databases does not necessarily reflect the genuine feelings of the subjects who displayed the expressions; the labels are based on the human annotator's intention, on the basis of previous research or emotion theory (Krumhuber, Kappas, and Manstead 2013).

The selection of data for training can also have undesired side effects if the data is not balanced in reflecting a general basic population. In addition to performance problems, this may cause ethical problems (see Section 2.8.3).

To summarise the current level of development, although the situation for high-quality datasets has improved over the years, finding appropriate datasets still remains an open issue (Martinez et al. 2017).

Comparing Different Approaches

Comparing different approaches for AFER algorithms is a challenging task, mainly because there is no standard experimental configuration or procedure: ‘studies are often different in terms of validation procedures, the number of test images/videos, subjects or labels’ (Sariyanidi, Gunes, and Cavallaro 2015).

The absence of standardised procedures and datasets has been repeatedly noted over the years, both for direct expression detection and AU-based approaches (Zeng et al. 2009; Valstar et al. 2011b; Sariyanidi, Gunes, and Cavallaro 2015; Martinez et al. 2017).

Nevertheless, in the absence of a standardised procedure, most comparisons use publicly available datasets (e.g. CK+ (Lucey et al. 2010)) for comparison. This procedure is described in more detail in Section 4.2.1. The **major risk** in utilising these datasets for comparison is that developers will optimise their AFER systems to **boost results in these benchmarking datasets**. While this leads to a high performance in the benchmarks, it does not solve the general problem.

Special attention should be directed towards the environment when comparing algorithms, as the results for posed or naturalistic expressions may differ, because the AFER system ‘that attains the highest performance in a posed validation scheme may be attaining the lowest in a spontaneous scheme, or vice versa’ (Sariyanidi, Gunes, and Cavallaro 2015).

Applicability for Real-life Scenarios

There are three different scenario types for applying AFER solutions: **lab** situations with a fixed task and a controlled environment; **real-life** situations with a fixed task and a non-controlled environment; and as an extension to real-life, **in-the-wild** scenarios where the task is not fixed nor is the environment controlled.

When applying AFER to real-life applications, the systems must be able to withstand the applications' requirements, which can differ significantly from conditions in the lab. In order to evaluate the applicability of AFER systems for use in real-life scenarios, different factors must be considered and can be grouped into different classes, from an AFER systems perspective:

External factors, such as illumination variations, colour variations, view-angle and distance of the camera to the subject's head (Stöckli et al. 2017) and low camera resolution or fps (Weber, Soladie, and Seguier 2018). Additional challenges are errors in registration, occlusions and identity bias[8] (Sariyanidi, Gunes, and Cavallaro 2015).

According to Zhang et al. (2018), the effect of occlusions on AFER has 'received relatively less investigations previously', although this is a common case in real-life and especially in in-the-wild scenarios.

Internal factors include the fact that in the real world facial expressions are often spontaneous, typically of low intensity, and may contain multiple apexes of different expressions (Sariyanidi, Gunes, and Cavallaro 2015). Additionally, the facial expression may

[8] Deviations of the subject's face from an imaginary average face.

change in conversational situations, as a result of speaking (Corneanu et al. 2016). Spontaneous facial expressions are often associated with fluctuations in the posture of the head, which must be considered and modelled before the facial expressions are measured (Sariyanidi, Gunes, and Cavallaro 2015; Corneanu et al. 2016).

Data for training and evaluation is a requirement, in order to be prepared for the two aforementioned classes. AFER systems, therefore, must be trained with data from realistic scenarios. However, collecting authentic data for evaluation can be difficult, as 'they are relatively rare, short lived, and filled with subtle context-based changes' (Ringeval et al. 2013). Another challenge is labelling the data. Especially for in-the-wild scenarios, this still seems to be unsolved:

> The challenge is addressing, how we can link the spontaneous behaviour that we exhibit as we navigate through our everyday lives and how this relates to real emotions and feelings. How do we label these? Can we rely on clean labels? Probably not. We will end up with a multitude of noisy labels that could be associated with all sorts of activities, embedded in a whole load of short-term and long-term contexts. This is an extremely challenging problem but one that is interestingly fundamental to computer science, and yet, not sufficiently tackled. (Gunes and Hung (2016))

Adequate Emotion Modelling is a requirement for real-life scenarios. Apart from the fundamental discussion of which emotion model best reflects reality, many modern AFER systems are based on discrete emotions (e.g. the six basic emotions). Gunes and Hung (2016) criticised this, as they argued that this is a simplification and not suitable for the 'majority of everyday applications'. Instead, they proposed the use of continuous dimensional modelling (Gunes and Hung 2016). Although the fundamental issue is still unsolved, Chapter 5 discusses an approach to improve AFER for real-life applications based on existing, discrete emotion theory algorithmic solutions.

Another important factor for the adequate modelling of emotions is to include the **real-life context** (Gunes and Hung 2016), as 'affective expressions can never be divorced from context' (Calvo and D'Mello 2010). This context has two perspectives: recognising the external situational context (e.g. in-the-wild), and possible internal context (e.g. events), if the situation is fixed (e.g. a gaming scenario in a real-life application).

2.6.8. Summary of AFER Algorithms

Various algorithmic approaches for AFER have been demonstrated in the last two decades and some have reached the status of professional products. However, the greatest performance improvement was achieved in the first step of the processing pipeline (see Section 2.6.4), face detection (Martinez and Valstar 2016).

This is not surprising, as the knowledge of what a face is and how it can be modelled is well defined and high-quality data for training and evaluation exist, resulting in several high performance approaches. In addition, the commercial interest in facial recognition in general application areas such as security was extensive.

In contrast, at present, the emotion recognition of facial expressions has few commercial application areas (although great potential) and is significantly more challenging. Especially the absence of an agreed general model for emotions, the underlying neurophysiological principles and the corresponding facial expressions have resulted in a number of major issues:

limited computational models, difficulties in comparing algorithms and an insufficient understanding of facial expression dynamics and intensities.

In addition, the labelling of ground truth for training and evaluation data remains challenging, as interpretation of facial expressions is subjective.

Although major challenges remain, AFER is applicable to certain environments and applications, as described in the next section.

2.7. Applications and Environments of AFER

The following section describes important applications and environments for AFER algorithms from the related literature as displayed in Table 2.2. Applications and environments are considered separately because an application may be used in different environments and therefore with different environmental conditions (e.g. lighting constraints).

Application	Environment
Medical	Desktop
Ability Training	Cockpit
Learning	Smart Environment
Serious Gaming	Exhibition
Entertainment	Virtual Reality
Exercise gaming	Augmented Reality
Cockpit	Mobile
Artwork	Robotics
...	...

Table 2.2: Selection of applications and environments for AFER. The highlighted applications and environments are closely related to the EmotionBike setup.

2.7.1. AFER Application Areas

The following section presents a selection of important application areas for AFER as described in Table 2.2.

Medical Applications

Medical applications have always been a driving force behind affective computing (Valstar 2014). Automatic analysis of affective signals can be utilised to support the diagnosis, monitoring and treatment of various diseases. According to Valstar (2014), the main medical focus of affective computing is on mood and anxiety disorders, neuro-developmental disorders and the recognition of pain. Pain is often detected from the facial expressions of patients by utilising an AU-based approach (Martinez and Valstar 2016). Another proposed application is the development of interactive toys for therapy (Scherer, Banziger, and Roesch 2010).

Ability Training

Ability training includes the social behavioural training of individuals to optimise their performance under certain social conditions (Martinez and Valstar 2016). Examples for this are the improvement of public speaking abilities (Martinez and Valstar 2016) or social coaching for job interviews by utilising a virtual agent (Anderson et al. 2013). In contrast to serious gaming, ability training does not necessarily involve a playful approach, although game engines are often applied to display avatars or scenarios.

Learning

Emotions play an **important role in learning** and education as they may help or hinder the acquisition of knowledge (Vela, Vela, and Jensen 2016). With a mixture of negative and positive affective states experienced during learning, computer learning systems must respond adequately to these affective states to optimise learning and motivation (Graesser et al. 2016). One approach to address emotions in learning is to build an intelligent tutoring system (ITS), in an attempt to recreate a one-to-one tutoring situation, as this has been shown to be the most effective learning strategy (Gordon et al. 2016). This has, for example, been demonstrated by the AutoTutor system (D'Mello, Picard, and Graesser 2007). Typically, in modern ITS, virtual agents guide the students through the learning process to increase their interest in learning and address the absence of self-regulation utilising cognitive and affective scaffolding, which would normally be provided by a tutor or other students in a class (Park 2016). Gordon et al. (2016) investigated their game-based learning system in combination with the AFFDEX AFER system and a social robot instead of a virtual agent generating feedback for the student.

The classical six basic emotions have only minor importance in learning scenarios, as non-basic learning-centric emotions typically include boredom, frustration, confusion, curiosity, flow and anxiety (Calvo and D'Mello 2010).

Serious gaming

Serious gaming, as a special kind of computer based learning, utilises game scenarios to facilitate teaching through the gamification of the learning task. Conati (2002) proposed the probabilistic modelling of users' emotions and engagement during the interaction with an educational game that focused on students' understanding of primes and factorisation. Robinson (2014) applied AFER, gesture and speech analysis to a serious game platform, enabling young patients with conditions on the autism spectrum to train for their daily lives.

Entertainment Gaming

On one hand, computer games can influence players' emotions through game mechanics and narrative elements (Wilkinson 2013). On the other hand, players can influence the storyline and parameters such as the game's difficulty when their emotions are observed. Zagaria (2017) created an affective stealth game that featured a virtual companion reacting to the facial expressions of the player, resulting in an increase (positive expressions) or decrease (negative expressions) in the companion's collaboration. Moniaga et al. (2018) applied AFER to adjust the difficulty of their real-time combat game, resulting in an increase in positive player experience.

Exercise Gaming and Physical Training

Exercise gaming, often abbreviated to **exergaming** (Hoda, Alattas, and Saddik 2013) describes the scenario of physically exercising while playing a computer game. Warburton et al. (2007) discovered, that exergaming had a positive effect on health-related physical fitness when compared to traditional exercises. Süssenbach et al. (2014) developed an indoor cycling companion based on a Nao 3 robot to increase training motivation. Hoda, Alattas, and Saddik (2013) demonstrated that the combination of indoor cycling and computer games increases exercise parameters such as virtual speed and rpm. While all of the aforementioned approaches utilised exergaming, none applied AFER.

Additional Application Areas

The application areas of **cockpit** and **artwork** are closely related to their corresponding environment and are therefore described in Section 2.7.2.

Challenges of Applications

There are several challenges for affective applications, especially with regard to application-specific expressions, emotion modelling and their correspondent interpretation, because in real-life applications, emotional reactions are typically complex. In addition, the question of how to include the context (task, environment and subject) into the emotion model remains unsolved (Zeng et al. 2009).

2.7.2. Environments for Affective AFER Applications

Environments for affective applications that utilise AFER range from constrained settings such as the desktop (one person in front of a screen), smart rooms or homes (multiple users in a larger area), to entire districts or cities.

Desktop

This laboratory setup has one person with a frontal view of a screen, normally with optimal lighting and camera placement (a frontal view of the face). It is a standard one subject scenario. Typical applications for such a setup are market research (Stöckli et al. 2017), learning (D'Mello, Graesser, and Picard 2007) and ability training (Anderson et al. 2013).

Cockpit

A cockpit scenario is an enhanced desktop scenario, typically with the addition of a physical or motor challenge (e.g., exercising or driving). In real-life setups, such as in driving a real car outside, difficult lighting conditions may occur. In addition, turning the driver's face results in an increased view angle. An example for a cockpit scenario is the 'Driver Behaviour and Situation Aware Brake Assistance for Intelligent Vehicles' by McCall and Trivedi (2007), which adapted the system's reaction based on situational severity and driver attentiveness and intent by utilising a camera that was pointed at the driver's head. Doshi and Trivedi (2011) provided an overview of systems for driver behaviour prediction and intent inference.

Smart Environments

Cook and Das (2004) defined a smart environment as 'one that is able to acquire and apply knowledge about the environment and its inhabitants in order to improve their experience in that environment'. Sensors in smart environments may be directly integrated into the environment in the form of ubiquitous computing (Weiser 1995).

In contrast to cockpits and desktops, intelligent or 'ubiquitous environments' (Cook and Das 2007) are more difficult: the person can be moving indoors or outdoors, and multiple cameras, sensors, and even people can be involved. In addition, smart environments often have a variety of possible facial positions other than frontal poses, occlusions and irregular illumination, making AFER approaches difficult (Bosch et al. 2015).

Another example of a mobile smart environment is a transportable measurement setup for the analysis of balance performance in human movement science studies (Bernin, Jettke, and Vogt 2016).

Kanjo, Al-Husain, and Chamberlain (2015) provided their 'platform for understanding the growing field of pervasive affective sensing' as an introduction for ' designers, computer scientists, and researchers from other related disciplines' (Kanjo, Al-Husain, and Chamberlain 2015) which is a valuable introduction and presentation of the different approaches and modalities of emotion recognition in ubiquitous or pervasive environments.

Smart Home

A smart home is a special manifestation of the smart environment, with a stationary setting that is usually a flat or house. Ellenberg et al. (2011) developed their environment for context-aware applications in smart homes which was evaluated in the ’Living Place Hamburg’, a smart home lab at the HAW Hamburg, providing a 140 sqm loft-style apartment suitable for conducting experiments under real-life conditions (Luck et al. 2010).

A special form of smart homes is their residents’ assistive environments, for example, within the context of ambient assisted living (Lundström et al. 2016). For example, Maglogiannis (2014) developed methodologies and human-centred systems for the integration in assistive environments, such as smart homes, focusing on the understanding of human emotions and behaviour.

Exhibition

An exhibition is also a special form of intelligent environment, in which emotion recognition is built into artefacts, for example, to trigger a reaction. In the case of utilising artefacts, the main benefit for AFER lies in attracting the observer’s view and face towards the object. For example, Müller et al. (2012) created an interactive surface that responded to the measured facial expressions of observers. Di Mauro et al. (2017) considered that similar concept could be transferred to an entire museum, a – ‘smart-museum’ – as a possible way to integrate virtual or augmented reality and affective installations. Using an urban district as an exhibition area, Public Face[9] was an installation in HnafenCity, Hamburg, that publicly displayed a

[9] https://kunstundkulturhafencity.de/en/event/public-face

mean of captured facial expressions (‘sad’, ‘angry’, ‘surprised’ or ‘happy’) by exploiting CCTV cameras in the district.

Virtual, Augmented and Mixed Reality

Virtual reality comes in multiple forms, such as in cave virtual environments and virtual reality headsets (e.g. HTV Vive or Occulus Rift). The main challenge for headset-based approaches lies in the occlusion of important parts of the face, which makes the detection of facial expressions by camera-based approaches difficult (Zhang et al. 2018). A possible solution could be the integration of electromyography sensors (EMG) in the headsets, as EMG has been demonstrated to be an alternative method for detecting facial expressions (Calvo and D’Mello 2010). In a cave scenario, AFER is possible with a multi-camera setup.

Mobile Devices

Mobile devices, such as smartphones and tablets, are becoming the new standard computing devices for daily HCI, leading to a shift away from desktop computers and notebooks (Gunes and Hung 2016). Several approaches have already been evaluated to examine the effectiveness of AFER in analysing marketing campaigns (Martinez and Valstar 2016; Stöckli et al. 2017). With the acquisition of the AFER manufacturer Emotient by Apple in 2015, AFER algorithms reached the consumer market (Gunes and Hung 2016), although it is currently unclear, whether mobile devices (such as smartphones) provide enough computing power for sophisticated, machine learning-based approaches (Martinez and Valstar 2016). Other challenges for AFER are uncontrolled lighting conditions and changing view angles of the face, due to movements while holding the mobile device.

Robotics

The utilisation of social robots for disabled patients began in 1993 with research that led to the development of PARO, an artificial seal (Inoue, Wada, and Ito 2008). Advanced robots with camera sensing capabilities (e.g. Sony's AIBO and ATR's Robovie; Corneanu et al. (2016)) evolved and can be understood as a mixture of a mobile device and a mobile environment – or sensing sphere – around the current position of the robot. AFER for human-robot interaction (HRI) has high potential to increase the bond between humans and robots (Martinez and Valstar 2016). Robots are also known to improve e-learning experiences (Corneanu et al. 2016). Affect recognition is seen as a fundamental component of personal robotics (Sariyanidi et al. 2013) and is also fundamental to the perception of robots as partners, rather than sophisticated tools (Kim, Smith, and Thayne 2016). Robots can have a collection of sensors at their disposal, including multiple cameras, resulting in a more optimal view angle than in mobile devices. Although they also suffer from unconstrained lighting conditions and unfavourable view angles of the user's face, they can rely on considerable computing power[10] to conduct sophisticated AFER analysis.

Summary of Environments

Table 2.3 describes the main characteristics of environments for affective applications: the typical size of the camera's **covered area**, constrained or unconstrained **lighting conditions**, available **computing power**, **camera field of view**, fixed or changing **facial position**, the camera **view angle related to the face** and the **number of users** usually found in this

[10] For example, NVIDIA offers Jetson Xavier, a high performance ML platform designed specifically for autonomous robots, see https://www.theverge.com/2018/6/4/17424118/nvidia-ai-chip-jetson-xavier-robot-platform-isaac.

environment. The environments for affective applications range from the most constrained desktop environments to the least restricted exhibitions that cover entire urban districts.

The main challenges of non-desktop environments for affective applications relate to external factors, such as facial position, camera view angles, camera coverage, lighting conditions and multiple users making most environments challenging. For the EmotionBike, the characteristics of the cockpit scenario apply but with a less constrained facial position, as a result of movement of the head during pedalling and steering.

Environment	Covered area	Lighting conditions	Computing power	Camera field of view	Face position	Face view angle	Number of users
Desktop	Small	Constrained	High	Small	Constrained	Constrained	One
Cockpit	**Small**	**Unconstrained**	**High**	**Small**	**Constrained**	**Unconstrained**	**One**
Smart Environment	Medium	Unconstrained	High	Wide	Unconstrained	Unconstrained	One – Multiple
Smart Home	Medium	Constrained	High	Wide	Unconstrained	Unconstrained	One – Multiple
Mobile Device	Small	Unconstrained	Low	Wide	Constrained	Unconstrained	One
Robotics	Medium	Unconstrained	Medium	Wide	Unconstrained	Unconstrained	One – Multiple
Virtual or Augmented Reality	Small	Constrained	High	Wide	Unconstrained	Unconstrained	One
Exhibition	Small	Unconstrained	High	Small	Constrained	Unconstrained	One – Multiple
Urban District Exhibition	Large	Unconstrained	High	Wide	Unconstrained	Unconstrained	One – Multiple

Table 2.3: Main characteristics of environments for affective applications. The EmotionBike setup is close to the Cockpit Environment. The Desktop Environment has the most constrains and is therefore already included in the Cockpit scenario.

2.8. Ethical Implications of Automated Facial Expression Recognition and Affective Systems

The application of AFER and multimodal affective systems provide opportunities and challenges that must be considered when designing, applying and utilising these systems. How can technology and citizens cooperate to take advantage of opportunities and avoid threats (such as in an Orwellian world) to the individual and society?

2.8.1. Potential Effects on Society

> A nervous tic, an unconscious look of anxiety, a habit of muttering to yourself – anything that carried with it the suggestion of abnormality, of having something to hide. In any case, to wear an improper expression on your face (to look incredulous when a victory was announced, for example) was itself a punishable offence. There was even a word for it in Newspeak: FACECRIME, it was called.
>
> (George Orwell, Nineteen Eighty-Four (Orwell 1949))

Besides the extreme of the dystopia[11] described by George Orwell in his novel Nineteen Eighty-Four (Orwell 1949), systems that are capable of analysing expressions or emotions can have significant social effects, even if unintended by the developers. Mittelstadt et al. (2016) summarised this as follows: 'gaps between the design and operation of algorithms and our understanding of their ethical implications can have severe consequences affecting individuals as well as groups and whole societies'.

There is growing scepticism about the side effects of algorithmic decisions in society and the academic community (Chander 2016), which are discussed in the following sections.

2.8.2. Black Boxing and Biased Decisions

The black box problem is a well-known problem of machine learning (Mittelstadt et al. 2016; Matthias 2004), and describes the fact that even for the human trainer of a complex system based on machine learning (e.g. neural networks) it is impossible to understand exactly how and why a system makes its decisions internally. While the developer of such a system

[11] Of which facial expressions are only a small part.

can at least compare an input with a desired output, for those affected by the widespread application of such a system, this possibility does not exist. Burrell (2016) named this the ‘opacity problem’.

Mittelstadt et al. (2016) argued, that ‘determining whether a particular problematic decision is merely a one-off ’bug’ or evidence of a systemic failure or bias may be impossible (or at least highly difficult) with poorly interpretable and predictable learning algorithms’. Unjustified actions at the disadvantage of users could be the result of this inaccurate evidence (Mittelstadt et al. 2016).

2.8.3. Algorithmic Discrimination

Algorithmic discrimination – or discrimination by algorithms – are part of the academic debate about the side effects of algorithms. Kitchin (2017) stated, that ‘far from being neutral in nature, algorithms construct and implement regimes of power and knowledge and their use has normative implications.’ Chander (2016) argued, that it is not the implementation, but the re-implementation of existing societal structures that causes discrimination, because ‘algorithms trained or operated on a real-world data set that necessarily reflects existing discrimination may well replicate that discrimination.’ Chander terms this ‘viral discrimination’.

Mittelstadt et al. (2016) supported this opinion:

> To give a formally precise account of this fact, the informal ’garbage in, garbage out’ principle clearly illustrates what is at stake here, namely that conclusions can only be as reliable (but also as neutral) as the data they are based on. Evaluations of the neutrality of the process, and by connection whether the evidence produced is misguided, are of course observer-dependent.

Ignoring the outward appearance of certain groups of people in the process of recognition (e.g., by only detecting Caucasians) and ignoring cultural differences in facial expressions (with regard to specific social display rules) discriminates these groups of people. As an example of the first type of discrimination, face detection algorithms used in cameras sometimes failed to detect dark skin tones in faces due to poor training data, which contained only a few African Americans (Narayan Soni, Datar, and Datar 2017).

Chander (2016) argued that algorithms are often more neutral than human decision makers, but that they also operate in ‘a world permeated with the legacy of discrimination’ (Chander 2016).

2.8.4. Personal Effects of Multimodal Affective Systems

Although the focus of this work is mainly in the area of facial expression analysis, the (possible) ethical implications should be understood in the greater context of camera-based (Bernin 2011) and multimodal affective systems. Scherer, Banziger, and Roesch (2010) mentioned three important ethical aspects for users of multimodal systems that are capable of recognising human emotions:

1. A negative effect, as the user’s privacy will be potentially violated if a computing entity analyses the user’s emotional reactions without prior notice and permission.

2. A positive effect, as this technology can help in the diagnosis of emotional dysfunction and enable the success of treatments to be monitored.

3. A neutral effect, as media and entertainment industries could utilise it to offer a new level of interaction.

When applying AFER, at least these three possible effects should be considered, which are discussed in more detail below.

Negative Effects of Surveillance

The importance of the negative effect described by Scherer, Banziger, and Roesch (2010) is supported by Turk (2014):

> There are potentially quite significant privacy issues associated with multimodal systems that must be considered early on in order to provide potential users with the assurance and confidence that such systems will not violate expectations of security and privacy. (Turk (2014))

To further support the relevance of the privacy issue, the following should be considered: a user of an affective enabled system could be aware of the fact that this system *might* at any point analyse his or her reactions. The result could be a kind of *emotional panopticon*, which could lead to a premature change in the user's behaviour (Koskela 2000). The term 'panopticon' has been described by Bentham (1791) as a concept for designing buildings (mainly prisons and factories) with constant potential[12] surveillance of the inhabitants and was perceived by Foucault as a manifestation of surveillance in modern societies (Foucault 2012).

Analysing the emotions of people could potentially result in emotional manipulation. Nevertheless, one might argue, that simpler methods of the algorithmic manipulation of people[13]

[12] The buildings were designed in such a way that the inmates could not see whether the supervisor was looking in their direction or not.

[13] Such as by manipulating the elections in the US by utilising data analysis from Facebook accounts (Cadwalladr 2018).

are already in place, and are also less computationally expensive than analysing people's facial expressions.

Neutral and Positive Effects

As it is sometimes difficult to distinguish between 'positive' and 'neutral' effects, especially because both often use game-based settings, they are discussed together.

Numerous examples of the neutral effect (enhancing interaction in media and entertainment) in the field of gaming – especially educational gaming – have been published. For instance, Wilkinson (2013) published an overview of affective educational games (serious gaming). Moniaga et al. (2018) evaluated applying AFER as input for dynamic game balancing, developing a serious game to adjust to the players' experience. Grafsgaard et al. (2013) presented their approach to detect engagement and frustration in an online tutoring environment, and Anderson et al. (2013) developed a virtual agent designed to train young unemployed people for job interviews.

For the group of positive effects, Valstar (2014) distinguishes between three groups, focusing on medical applications: applications for mood and anxiety disorders, neuro-developmental disorders, and pain estimation (Martinez and Valstar 2016).

Systems in the field of research have been published for all three application areas, such as systems designed to detect depression and PTSD, combined with a digital interviewer (Stratou and Morency 2017), interactive emotional games for children with autism (Schuller et al. 2013) or pain detection based on facial expressions (Ryan et al. 2009).

2.8.5. Summary of Ethical Implications

Affective systems offer a range of potential benefits and dangers. Positive effects have been described in the literature, such as helping people to experience a (more) normal life, teaching new skills or monitoring the progress of their therapy.

The negative consequences of (possible) continuous surveillance should always be considered, but so should the possible neutral and positive consequences as proposed by Scherer, Banziger, and Roesch (2010), as they have the potential to improve the everyday lives of users of such systems.

Independent, critical studies of ‘researching algorithms’, as proposed by Kitchin (2017), provide a valuable source of reflection. In addition, the critical engagement of civil society can help to assess the effects of technology. This is also necessary to create and maintain trust in ML-based technologies and a base from which to implement the necessary ethical and formal/legal contexts, as proposed by the European Commission (Craglia et al. 2018).

2.9. Discussion of Related Work

The automatic recognition of (emotional) facial expressions is part of an extensive, active field of interdisciplinary research, especially in computer science. In this chapter, the relevant literature was subsumed in order to describe the state-of-the-art.

2.9.1. Summary of Related Work

Various emotion theories and models have been proposed in the last 150 years, mostly in a psychological context, yet it is still unclear, which of these is the most accurate to model human emotions. Apart from this fundamental issue, different emotion models have been applied to computing contexts, discrete models have been applied to recognise emotions (e.g., facial expressions) and multidimensional or hierarchical models have been applied to represent internal affective states (e.g., of virtual avatars).

Various modalities have been proposed and evaluated for the detection of emotions, but the recognition of facial expressions utilising AFER is considered the most dominant modality. Multimodal affect recognition is a promising approach, but requires reliable unimodal robustness, such as robust facial expression recognition.

Many AFER solutions follow a general approach: a pipeline of face detection, filtering, registration, feature extraction and classification by applying machine learning (depicted in Figure 2.9). Although various solutions have been published, many open challenges remain, such as the modelling of timing, dynamics and intensities of facial expressions. Missing standard protocols for a general comparison of approaches and the data quality for training and evaluation – especially for real-life or in-the-wild scenarios – remain open issues. Although different applications have been demonstrated to benefit from AFER, the integration of AFER into new scenarios remains challenging due to the demands of environments and application-specific interaction schemes.

In addition to the foundation and technical possibilities of emotion detection and AFER discussed in this chapter, ethical consequences have also been discussed, and challenges and opportunities were highlighted. It should be noted that any new technology should be applied with care and reflection on the social consequences.

The strategy resulting for the general objective of this work, integrating AFER in the EmotionBike, is explained as follows.

2.9.2. Strategy for this Thesis

Table 2.4 summarises the relevant key characteristics of affective systems based on the presented literature review and the surveys of Zeng et al. (2009), Calvo and D'Mello (2010), D'Mello and Kory (2015), Sariyanidi, Gunes, and Cavallaro (2015), Martinez and Valstar (2016), Corneanu et al. (2016) and Poria et al. (2017). These key characteristics provide the basis to define the strategy and focus of this thesis.

Application	Environment	Modality	Affective Model	
			Type	Concrete Model
Medical	Desktop	Audio (Speech)	**Discrete emotions**	**6 Basic Emotions (BE)**
Ability Training	**Cockpit**	Physiological	**Discrete emotions**	**Extended BE**
Learning	Smart Environment	Thermographic	Facial Movement	FACS
Serious Gaming	Robot	**2D Facial expression**	Facial Movement	FAP
Entertainment	Virtual Reality	3D Facial Expressions	Multidimensional	PAD-Space
Exercise Gaming	Augmented Reality	Multimodal	Multidimensional	HGE
Cockpit	Mobile	Gestures	Multidimensional	Core Affect
Artwork	Exhibition	Body Pose	Hierarchical	OCC
...	...	...	...	...

Table 2.4: A compact table summarising key characteristics for affective computing systems: applications, environments, modalities and emotional models for automated human emotion recognition. The focus of this work is on the highlighted areas as they reflect the AFER-related characteristics of the EmotionBike.

The reasons for selecting characteristics presented in Table 2.4 as the focus of this work are founded in their relevance for the goal of applying AFER to the EmotionBike. The EmotionBike has a main application focus on cockpit-scenario and exercise gaming (refer to Section 1.2). These applications are within the cockpit-like environment, which also includes aspects of a Desktop environment. As facial expressions are considered the main modality for emotion recognition (refer to Section 2.5), this thesis focuses on AFER. As many state-of-the-art AFER systems utilise 2D camera-based and discrete emotion approaches with basic emotions or extended basic emotions, these type of affective model was chosen for this work. The EmotionBike project itself also investigated physiological and thermographic approaches, but as they are not AFER-related, they are not discussed in this thesis. The following chapters of this thesis refer to the selected key characteristics of affective systems and focus on two areas.

Firstly the novel EmotionBike software framework as foundation for interaction and especially the synchronised recording of data avoiding manual annotation of the recorded data set, as manual annotation is a challenging, laborious and time-consuming task and should therefore be avoided (Section 2.6.7). The EmotionBike provides the ability of automatic annotation by its event-based emotion provocation approach. A requirement for this is time-synchronous processing and storage of events and data as described in Chapter 3. This recording poses a precondition for the integration of AFER into the EmotionBike as it enables post-experiment processing and evaluation of the data.

Secondly the selection of an appropriate state-of-the-art AFER solution. For this, the challenge is to select a robust and mature AFER solution in order to achieve an adequate response of the EmotionBike to the provoked emotion of the probands. The focus of this thesis hereby is on the evaluation of the integration of *existing* systems for AFER into a *novel* application such as the EmotionBike. Hereby, the challenges of the comparability of approaches and the applicability in non-lab applications (Section 2.6.7), such as the cockpit-scenario of the EmotionBike, are of particular importance. In addition, Table 2.3 describes the important characteristics of the environment needed to be considered for the selection of an appropriate AFER solution. This selection process is described in Chapter 4.

The following chapter describes the software framework for the EmotionBike system.

3. Time-Sensitive Design and Evaluation of the EmotionBike System

Based on the chosen strategy for this thesis explained in Section 2.9.2, with a focus on 2D camera-based AFER and to ensure evaluability of the experimental data in relation to AFER, an appropriate software framework for the EmotionBike system is required. This system framework demands high flexibility to integrate system components, different AFER approaches and other sensors. A time-sensitive design is required to ensure appropriate responses to user actions, provoked emotions and the recording of data with common time stamps to allow automatic annotation of game events and user reactions, thus enabling experiments with the EmotionBike.

3.1. Introduction

Although this work focuses on one modality, 2D-based AFER, the EmotionBike system is in general multimodal, since it integrates game controllers, various sensor modalities and a game-engine (refer to Table A.1). Although these additional modalities are not in the focus of this thesis, they are part of the overall system and were therefore considered in the framework design.

> Multimodal interfaces [...] represent a new direction for computer interfaces and a paradigm shift away from conventional graphical interfaces because they involve recognition-based technologies designed to handle continuous and simultaneous input from parallel input streams, [...] distributed processing and time-sensitive architectures. (Ruiz, Chen, and Oviatt (2010))

The EmotionBike is a multimodal, affective application enabling users to cycle through an interactive game environment on a stationary bike trainer with steering capabilities. The EmotionBike framework is a distributed system, providing support for live multimodal interaction and analysis as well as extensive capabilities for data recording and logging of the system status. The framework is capable of handling the interaction during runtime and storing data with synchronized timestamps, which allows post-experiment processing without the need for manual annotation of the data set.

The distribution of the system components is depicted in Figure 3.1, the components are described in Table 3.1. This thesis focuses on the AFER-related components marked in both, Figure and Table.

The exergame provided by the EmotionBike is a variant of a cockpit scenario also suitable for research on the topic of games for physical therapy and orthopaedic rehabilitation.

Different game scenarios were created to challenge the user and provoke different emotions[1]. Since these scenarios and the game environment were developed especially for the

[1] For examples, please refer to Section 6.4.2.

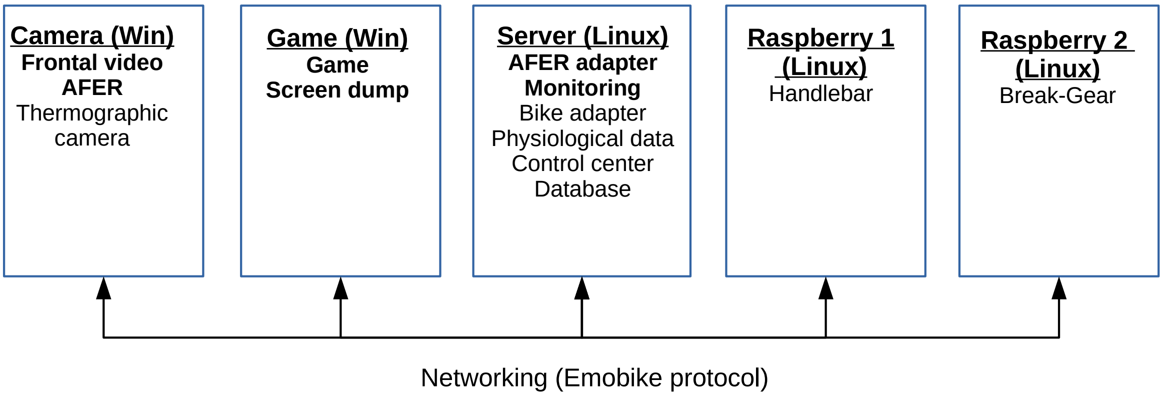


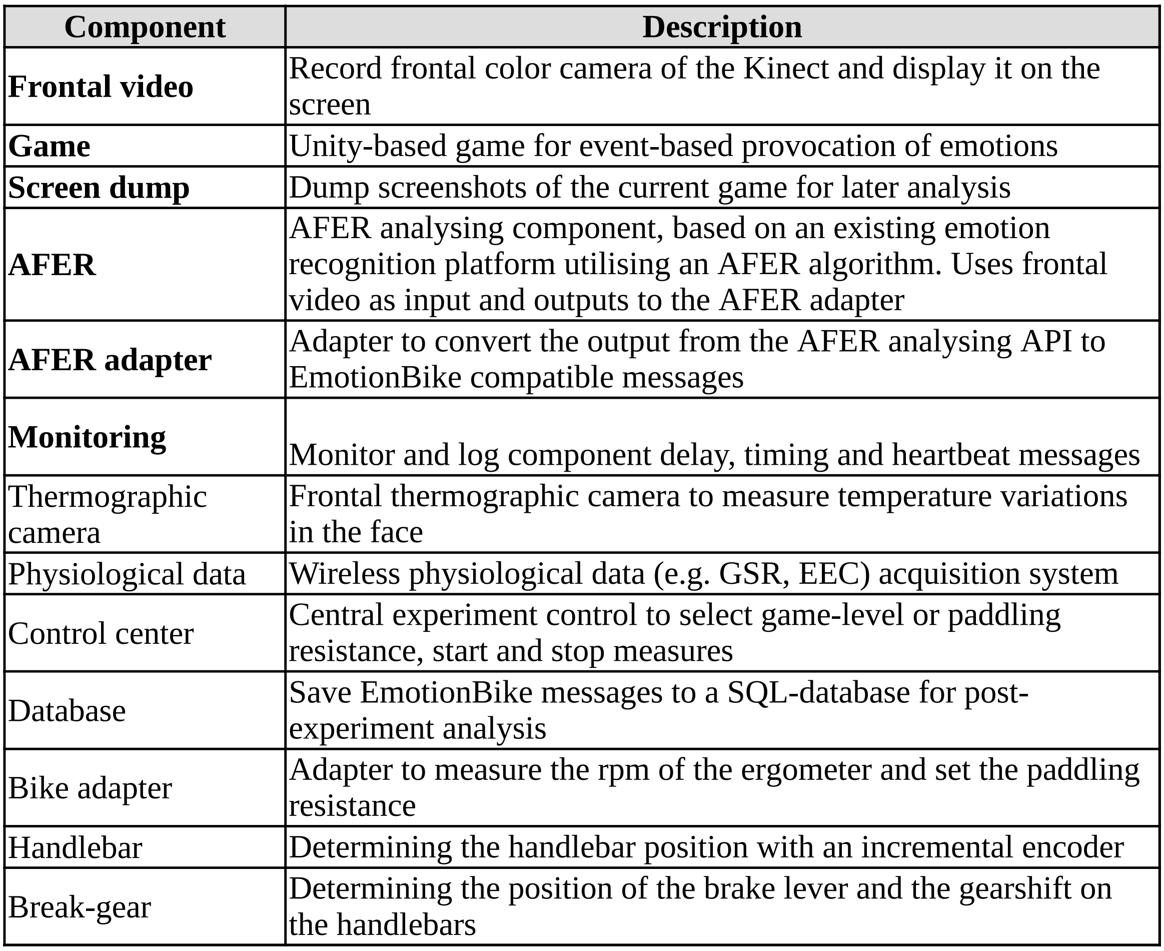
Figure 3.1: Sytem component distribution on computing nodes with Windows (Win) or Linux operating systems. The AFER-related components covered in this thesis are **marked**. All listed component are described in Table 3.1. All components communicate over the network by utilising the EmotionBike protocol, based on AMQP and JSON as described in Section 3.3.2.

EmotionBike, automatic logging of all relevant events within the game was integrated to allow automatic annotation of video data.

To sense the user, camera-based recognition of facial expression and physiological sensors were applied. Details on the sensors and the physical setup are provided in Section 3.3.4.

While previous publications (Müller et al. 2015; Müller et al. 2017; Müller 2018) focussed on the application and experimental viewpoints, this chapter explores a technical perspective. Time management plays a central role in the human-machine interaction as well as the recording of sensor data, since inadequate handling of the timing aspects in the most unfortunate case leads to disruption of the interaction and uselessness of the recorded data.

This chapter offers novel insights and evaluation results on the aspects of time-dependent behaviour, software patterns and the architecture of the EmotionBike system.



Component	Description
Frontal video	Record frontal color camera of the Kinect and display it on the screen
Game	Unity-based game for event-based provocation of emotions
Screen dump	Dump screenshots of the current game for later analysis
AFER	AFER analysing component, based on an existing emotion recognition platform utilising an AFER algorithm. Uses frontal video as input and outputs to the AFER adapter
AFER adapter	Adapter to convert the output from the AFER analysing API to EmotionBike compatible messages
Monitoring	Monitor and log component delay, timing and heartbeat messages
Thermographic camera	Frontal thermographic camera to measure temperature variations in the face
Physiological data	Wireless physiological data (e.g. GSR, EEC) acquisition system
Control center	Central experiment control to select game-level or paddling resistance, start and stop measures
Database	Save EmotionBike messages to a SQL-database for post-experiment analysis
Bike adapter	Adapter to measure the rpm of the ergometer and set the paddling resistance
Handlebar	Determining the handlebar position with an incremental encoder
Break-gear	Determining the position of the brake lever and the gearshift on the handlebars

Table 3.1: Components of the EmotionBike system. The **marked** components are AFER-related and are therefore discussed in this chapter. The other components are explained for the completeness of the EmotionBike system and are the reason why the framework has to handle multimodal data.

3.2. System Requirements and Related Approaches

The development process of the EmotionBike framework started in 2012 with analysis of the initial requirements of an affective, multimodal computing framework as described in Bernin (2012). The initial general design is depicted in Fureigure 3.2. It has been refined during the development process, as presented later in this chapter.

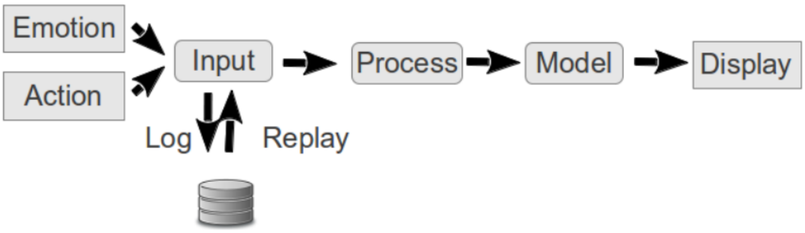


Figure 3.2: The general design of an affective computing framework from Bernin (2012).The emotions and actions of the user generate input to the system which is recorded and processed and an adequate reaction is generated and displayed utilising an interaction model.

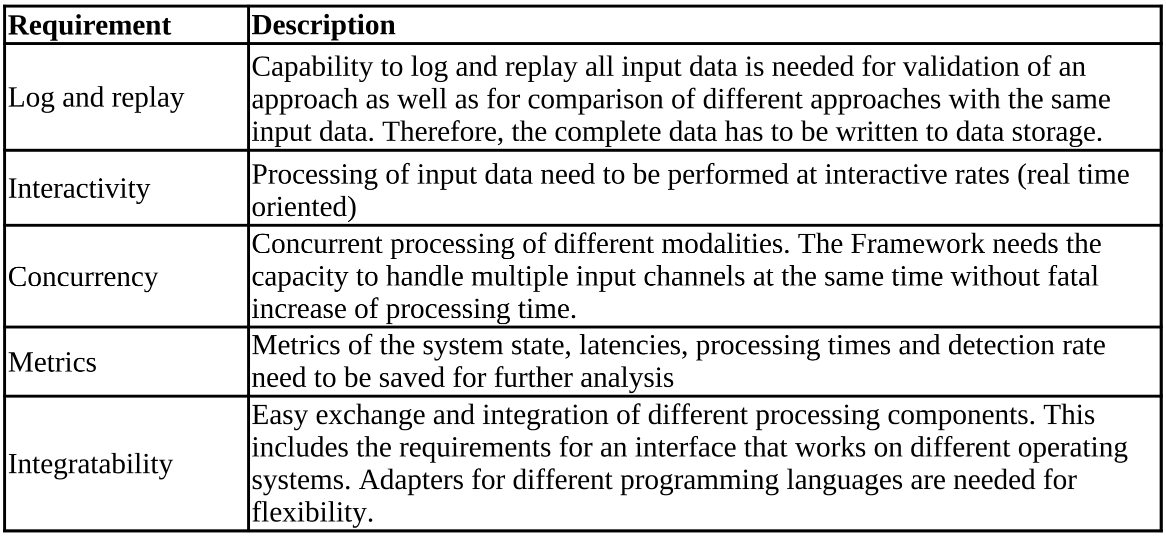
Afterwards, the idea arose to use a game-based scenario to evaluate the recognition of emotions in an application scenario. This approach has also been proposed in the literature: ‘What are the issues in using game-like scenarios as complex stimuli to evaluate models and elicit emotions?’ (Broekens, Bosse, and Marsella 2013).

Table 3.2 presents the requirements for an emotion-enriched (multimodal) framework proposed in Bernin (2012), which build the foundation of the following comparison of related approaches for interactive and multimodal frameworks and the technical design of the EmotionBike framework.

3.2.1. Refined Key Requirements for Multimodal Interaction and Recording

The key requirements for the design of multimodal interaction and recording frameworks were adapted and extended based on Table 3.2.

1. **Interactivity**: The maximum time for the input-response loop has to be lower than 100 ms (see Section 3.2.2 for details) to ensure interactive responsiveness with soft real-



Requirement	Description
Log and replay	Capability to log and replay all input data is needed for validation of an approach as well as for comparison of different approaches with the same input data. Therefore, the complete data has to be written to data storage.
Interactivity	Processing of input data need to be performed at interactive rates (real time oriented)
Concurrency	Concurrent processing of different modalities. The Framework needs the capacity to handle multiple input channels at the same time without fatal increase of processing time.
Metrics	Metrics of the system state, latencies, processing times and detection rate need to be saved for further analysis
Integratability	Easy exchange and integration of different processing components. This includes the requirements for an interface that works on different operating systems. Adapters for different programming languages are needed for flexibility.

Table 3.2: Initial requirements for emotion enriched interfaces from Bernin (2012). A short name for the requirement was added for simpler reference.

time (Dick, Wellnitz, and Wolf 2005; Bernin 2012). The architecture has to consider time-sensitive requirements.

2. **Concurrency**: Concurrent processing of different modalities has to be possible. The architecture needs to be distributed (Dumas, Lalanne, and Oviatt 2009).

3. **Recording**: All sensors and system data need reliable, time-sensitive recording (e.g. with timestamps). The system needs to support the preparation and conducting of measures during experiments with the ability to provide a prepare, start and stop measure cycle.

4. **Logging**: For identifying artefacts created by the system, intensive traceability of components is necessary, in addition to the recording of sensor data (Bernin 2012).

5. **Metrics**: To complete the logging of data and components, additional metrics for the

framework components (e.g. latency and delay, clock drift, frame-rate for videos) have to be integrated (Bernin 2012).

6. **Integrability**: A distributed architecture with component- or agent-based design, extensibility, pluggability and easily reused components (Dumas, Lalanne, and Oviatt 2009) is needed, providing interfaces to common programming languages (APIs of applied software such as gaming or sensor platforms) and operating systems.

7. **Communication**: For enabling an interactive system, different components need channels to exchange relevant data (e.g. a handlebar with steering capabilities needs to communicate with the game engine).

3.2.2. Input Response Time in Multimodal Interfaces

Dumas, Lalanne, and Oviatt (2009) postulated, that ‘the grand challenge of multimodal interface creation is to build reliable processing systems able to analyse and understand multiple communication means in real-time’. However, Dumas, Lalanne, and Oviatt (2009) did not define the term ’real-time’. In principle, this term can be interpreted in different ways in a computer science context: for example with a machine-centric point of view as describing the deadline characteristics of the overall system (’hard’, ’firm’, ’soft’, see Shin and Ramanathan (1994)), or from a user perspective focussing on the ‘interactive real-time responsiveness’ (Dick, Wellnitz, and Wolf 2005), input-response loop, which is most important for the user experience.

The maximum time tolerated by users is task and modality dependent. Miller (1968) postulated the maximum reaction time for interaction with physical pointing devices (light-pen) to be less than 100ms, and 100–200 ms for keyboard interaction. The maximum tolerable response times of mouse–window interaction was experimentally evaluated to be between 150 ms and 200 ms (Dabrowski and Munson 2001; Dick, Wellnitz, and Wolf 2005). One key point for the tolerable system latency in interactive contexts is, whether it degrades user performance or experience.

For gaming scenarios, especially first-person shooter and car-racing simulations with a required high sensitivity, Claypool and Claypool (2006) evaluated, that a threshold above 100 ms should not be exceeded. For third-person games, the maximum latency of 500 ms is much higher. As the EmotionBike, as cockpit scenario, is similar to a car racing simulation, a **threshold of 100 ms for interactive delays seems acceptable**. As the speed of the virtual bike in the game is rather low compared to a car race, it might be even above.

3.2.3. Related Approaches for Interactive Multimodal System Frameworks

Although different frameworks for multimodal interfaces and especially multimodal affect recognition have been developed in recent decades, a universal affect detector for HCI has not been implemented according to the literature:

> some of the other key outstanding challenges in this exciting field include: [...] synchronization of frames, voice and utterance, reduction of multi-modal Big Data dimensionality to meet real-time performance needs, etc. These challenges suggest we are still far from producing a real-time multimodal affect detector which can effectively and affectively communicate with humans, and feel our emotions. (Poria et al. (2017))

As Dumas, Lalanne, and Oviatt (2009) stated, ‘distributed and time-sensitive architectures’ are key requirements for multimodal user interfaces, which constitutes a good link to the second aspect of the EmotionBike framework: an adapter for the integration of different components (requirement *integratability*, refer to Table 3.2). Dumas, Lalanne, and Oviatt (2009) describe typical characteristics of frameworks in this multimodal context as ‘component- or agent-based design, sensor fusion capabilities, API, extensibility, pluggability, the ease of reusing components and availability (e.g. open-source software) ’.

Different frameworks have been developed and applied to multimodal HCI. Nine relevant systems are described briefly, and compared to the EmotionBike Framework. Table 3.3 presents the comparison summary. All framework are presented in the order of their publication, although in some cases they have been referred to earlier or hinted at in the literature without sufficient detail.

In addition to the key requirements described in Section 3.2.1, several detailed characteristics were evaluated for the summary in Table 3.3: **Component-based**, **agent-based** and **distributed** are differentiated characteristics of the **integrability** requirement, while **centralized storage**, **decentralised storage** and **time synchronisation** are part of the **recording** requirement.

As the developed application for the EmotionBike system is an exercise game, **game-engine integration** is a key characteristic, since it enables automatic event-based analysis.

The planned hardware for sensors and the enhanced ergometer required Windows and Linux **platforms** to be supported. The same is true for the **programming APIs**. Details on the frameworks are described in the following.

Characteristic	Key Requirements							Detailed Characteristics										
Framework	Interactivity	Concurrency	Recording	Logging	Metrics	Integrability	Communication	Centralized Storage	Decentralised storage	Time synchronisation	Game-engine integ.	Component-based	Agent-based	Heartbeat	Distributed	Year (published)	Platform	Programming APIs
CRN Toolbox	X	X	X	X	O	X	X	X	X	X	O	X	O	O	X	2006 (2010)	Android, iPhone, PC	JAVA, MATLAB
SSI	X	X	X	X	O	X	O	X	X	X	X	X	O	O	X	2013	Windows	C++
IMI2S	X	X	X	X	X	X	X	X	X	X	O	O	X	O	X	2014	Windows, Linux	C++, Java, Python
EmotionBike	X	X	X	X	X	X	X	X	X	X	X	X	O	X	X	2015	Windows, Linux	C++, C#, Python
FILTWAM	X	O	X	O	O	O	O	X	O	O	O	O	O	O	O	2016	MacOS	C/C++
PSI	X	X	X	O	O	X	X	X	O	X	O	X	O	O	O	2017	Windows	C#, .NET
MultiSense	X	X	X	X	O	X	X	X	X	X	O	X	O	O	X	2017	Windows	C++
PHASER	X	O	O	O	O	O	O	X	X	O	O	O	X	O	X	2016	MacOS	unknown
FARMI	X	X	X	O	O	X	X	X	X	X	O	X	O	O	X	2018	Windows	C++, Python
SSJ	X	X	X	O	O	O	O	O	X	O	O	X	O	O	O	2018	Android	Java

Table 3.3: Characteristics of frameworks for multimodal interaction and recording. In addition to the requirements described in Section 3.2.1 (marked in green), characteristics describing special handling of data have been added (marked in orange). 'x' donates, that the framework provides the capability, while 'o' marks the fact, that the requirement was not met or is unknown from the publications.

Context Recognition Network Toolbox

Bannach et al. (2006) developed the Context Recognition Network (CRN) Toolbox, specially focussing on activity monitoring and recording in the area of wearable computing and sensors. The CRN provides an event-based synchronisation mechanism for multiple sensors and streams. First developed in Java for PC-based systems (Bannach et al. 2006), it was later extended for the Android and iPhone platforms (Bannach et al. 2010).

Social Signal Interpretation

Wagner et al. (2013) published their work on Social Signal Interpretation (SSI), a C++ framework mainly for live recognition of multimodal signals (Wagner et al. 2013). Social Signal Interpretation concentrates on recording, processing and classifying multimodal signals on the same computing node, only forming a partly distributed system (Mauro and Cutugno 2016; Scherer et al. 2013). It has a plugin-based architecture with a C++ API and is available as open-source software[2].

Interaction, Multimodal Integration and Social Signals

The Interaction, Multimodal Integration and Social Signals framework (IMI2S) is a lightweight framework proposed by Anzalone et al. (2014). The IMI2S framework utilises small computational modules (similar to agents), that can be interconnected using messages in XML format transported over zeroMQ streams with a public-subscribe pattern. Publisher and subscriber are configured into the modules. For synchronising the system clock of computing nodes, network time protocol (NTP) is recommended, and all messages contain a time-stamp. Binary streams are not provided, all binary data need base64 encoding and decoding. Large binary data (e.g. Video or Kinect raw data) is recorded locally to decentralised storage.

EmotionBike

Details of the design and architecture of the EmotionBike system framework are given in Section 3.3. While preparing the EmotionBike framework for the second set of experiments

[2] https://github.com/hcmlab/ssi

(ES2), utilising the robot operating system version 2 (ROS2)[3] based on data-distributed services (DDS)[4] was considered as alternative middleware. The ROS2 also implements the publish-subscribe pattern utilised in the EmotionBike. At that time, early 2016, there was no ROS2 compatible open-source implementation of DDS available. For this reason, the approach was discontinued.

Framework for Improving Learning Through Webcams And Microphones

The Framework for Improving Learning Through Webcams and Microphones (FILTWAM) was published by Bahreini, Nadolski, and Westera (2016) and is oriented towards enhancing e-learning with emotion recognition. It mainly provides a fusion of sensors and detected emotion recognition by applying speech analysis and facial expression analysis. The FILTWAM is neither distributed nor component based; it appears as a monolithic solution integrating existing software for emotion recognition into a single application. Nevertheless, the applied techniques and algorithm are described in detail, providing interesting insight into how to build such systems.

Platform for Situated Intelligence

Bohus, Andrist, and Jalobeanu (2017) developed the Platform for Situated Intelligence (PSI) which provides a .NET runtime for time-sensitive analysis of multimodal signals. Component adapters for different sensors ((e.g., camera, microphone, Kinect) and processing components (e.g. speech recognition or face tracking) are available. The built PSI application runs on a single Windows-based machine.

[3] http://www.ros.org/

[4] https://www.omg.org/spec/DDS/1.4

MultiSense

MultiSense enhances SSI with inter-component communication by applying VHMsg[5] with ActiveMQ (Stratou and Morency 2017), and integrates modules for 3D head position-orientation, facial tracking, facial expression analysis and face detection. Although MultiSense was already mentioned by Wagner et al. (2013) and Scherer et al. (2013), details about the framework were not revealed until four years later by Stratou and Morency (2017).

Pervasive Human-centred Architecture for Smart Environmental Responsiveness

Mauro and Cutugno (2016) published the Pervasive Human-centred Architecture for Smart Environmental Responsiveness (PHASER), a decentralised, distributed approach with dynamic configuration. The PHASER focusses on interaction in smart homes and with internet-of-things (IoT) devices.

Framework for Recording Multi-Modal Interactions

Jonell et al. (2018) developed the Framework for Recording Multi-Modal Interactions (FARMI) on the basis of a component-based, distributed architecture. Components exchange messages using a central message broker (RabbitMQ[6]) with a public-subscribe pattern and high-bandwidth demanding streams directly by utilising zeroMQ[7]. The FARMI provides an interesting way to handle time-offset in the distributed system: Instead of synchronising clocks or components or operating systems, only the time-offset is transferred and appended to the timestamps of sent messages.

[5] http://vhmsg.sourceforge.net/

[6] https://www.rabbitmq.com/

[7] http://zeromq.org/

SSJ

Damian, Dietz, and Andre (2018) introduced SSJ as a framework to analyse social signals of users on Android devices in real-time and provide multimodal feedback while in social interaction. The SSJ is based on SSI and provides an Android app, which allows the creation of a social signal processing pipeline and feedback system with a graphical user interface without the need to write code. Data can be recorded directly to the device for offline analysis. Although modern smartphones provide comprehensive computing power, 'certain tasks such as training neural networks or processing videos with high frame rates and resolutions are currently still limited by the processing capabilities of the underlying hardware ' (Damian, Dietz, and Andre 2018). SSJ is available as open-source software[8].

Results of Comparison of Approaches

In conclusion, different and interesting approaches exist for scientific frameworks for multimodal interaction and recording of data on a variate of platforms.

EmotionBike framework development started in early 2014, based on previous experiences with JavaScript Object Notation (JSON) and message broker (ActiveMQ) (Müller et al. 2012).

The IMI2S framework provides similar concepts and characteristics to that of the EmotionBike; it seems to be the closest approach, although missing a heartbeat capability. The also missing integration of the game engine support, which is a prerequisite for the event-based approach of EmotionBike, is essential. In addition, the IMI2S paper was published at the end of 2014, after development of the EmotionBike framework had already begun.

[8] https://hcmlab.github.io/ssj/

The EmotionBike system is also quite similar to MultiSense, which offers many interesting possibilities, but also lacks heartbeats and game engine integration. In addition, MultiSense was not yet available at the time of the development of our framework since details on the capabilities were revealed in May 2017 when the EmotionBike framework was already fully developed, and the experiments had already ended.

EmotionBike is also the only framework providing a builtin experiment cycle which enables the preparation of sensors before actually starting and stopping a measure to ensure, that the sensor is already online and prepared for actually starting to record data.

The next section provides insights into the technical design of the EmotionBike platform.

3.3. Aspects of the EmotionBike Platform Technical Design

3.3.1. EmotionBike System Software Design

The EmotionBike framework was designed with two different main applications in mind: firstly, as an interactive platform for generating multimodal user interfaces; and secondly, as a scientific recording system for multimodal data for analysis after the experiment and (post) processing. As from the beginning, inter-application and inter-operating system were required, the framework was designed as a time-sensitive distributed system as proposed by Dumas, Lalanne, and Oviatt (2009).

Stage	Property	Description
Design	Considered	Include requirements for adequate temporal behaviour
Development	Testable	Test temporal behaviour of components
Operation	Measurable	Reliably save data and system timing and status information from components
Analysis	Verifiable	Recorded system information and logs enable verification of recorded data

Table 3.4: Model time-dependent adequate behaviour while developing a time-sensitive system.

Timing behaviour is essential for interactive experiments and recording; therefore, time dependent modelling is a core functionality and was addressed throughout all developmental stages as described in Table 3.4.

Timing can be considered from an application view (e.g. interactive timing), technical view (e.g. time synchronization of distributed systems) and implementation view (e.g. implementation in components). For instance, for an affective response based on facial expressions, a time frame of 300 ms poses no problem as emotions usually manifest rather slowly in the face (see Section 2.6.1), for the user's physical interaction the target timing should be significantly less than 100ms. An important objective in data acquisition is that any artefacts that occur are not caused by a failure of time synchronization.

In contrast to frameworks such as MultiSense, which facilitate central recording of streams, a more lightweight approach was chosen for the EmotionBike: raw image data[9] is mostly saved and processed locally by the components, and only aggregated data is sent using the EmotionBike protocol, which was also suggested by Poria et al. (2017). This prevents the classical bottleneck of network-based raw data streams.

[9] For example, Kinect raw data is nearly 200 MByte/s.

3.3.2. EmotionBike Architecture and Patterns

The design of the EmotionBike system was inspired by the concept of a service-oriented architecture (SOA) (Buschmann et al. 1996; Lalanda 1997), using independent components in different programming languages and hardware architectures. Stal (2006) describes the forces that drive an SOA as ‘distribution, heterogeneity, dynamics, transparency and process-orientation’. Heterogeneous applications or components are loosely coupled by applying a proxy design pattern (Buschmann et al. 1996; Schmidt et al. 2000).

Although SOAs are often associated with web services (’Web 2.0’), web services are only one possible application. In general, SOAs bundle different existing components by utilising a network protocol to form new services or applications (Stal 2006).

Service-oriented architecture has been applied to different multimodal frameworks. Heinzl et al. (2009) presented their scalable SOA for multimedia analysis, synthesis and consumption. Grifoni et al. (2017) proposed the usage of SOA for creating server-side multimodal interaction for integrating with cloud-services. Gonzalez-Sanchez et al. (2011) published their work on an agent-based design for multimodal emotion recognition framework with a service-oriented approach towards external communication with third-party systems such as games or learning environments.

For the EmotionBike framework, the flexibility in the SOA design enables one to choose the optimal application software for each type of sensor, such as camera APIs, AFER software or physiological sensors due to the addition of a system-wide control layer based on components.

In contrast to a traditional SOA containing a dynamic service repository for automatic discovery and registration, the services were statically configured utilising context-dependent pipelines defined during system start-up. Although a fully dynamic discovery and registration could easily be implemented and was originally planned, it was postponed due to the limited number of dynamic changes required for the EmotionBike setup and the overhead of implementation in all components. Newly developed components interested in specific data (e.g. the output of AFER) start to consume data by subscribing to the corresponding channel without changes to the producer.

Figure 3.3 displays a pipeline for the complex service that provides game reactions from the camera recording to the game engine as AFER output is interpreted to determine the current facial expression. The Game engine may then react on this expression, enabling emotional context-aware game responses. The occurring latency of the AFER-related detection is later explained in Section 3.5.2.

For communication between the EmotionBike components, a public-subscribe pattern (Birman and Joseph 1987) was applied with a JSON-based protocol (Crockford 2006) on top of the application layer Advanced Message Queuing Protocol (AMQP)[10]. As AMQP is an official standard published by the Organization for the Advancement of Structured Information Standards (OASIS[11]), client libraries for multiple programming languages for the different components (in C/C++, Java, C#) are available as open-source software. The analysis of alternative frameworks for multimodal data acquisition revealed that many of the open source approaches also utilise ActiveMQ as middleware for communication between components.

The JSON standard was chosen as the message standard to develop an application-specific

[10] https://www.amqp.org/

[11] https://www.oasis-open.org/

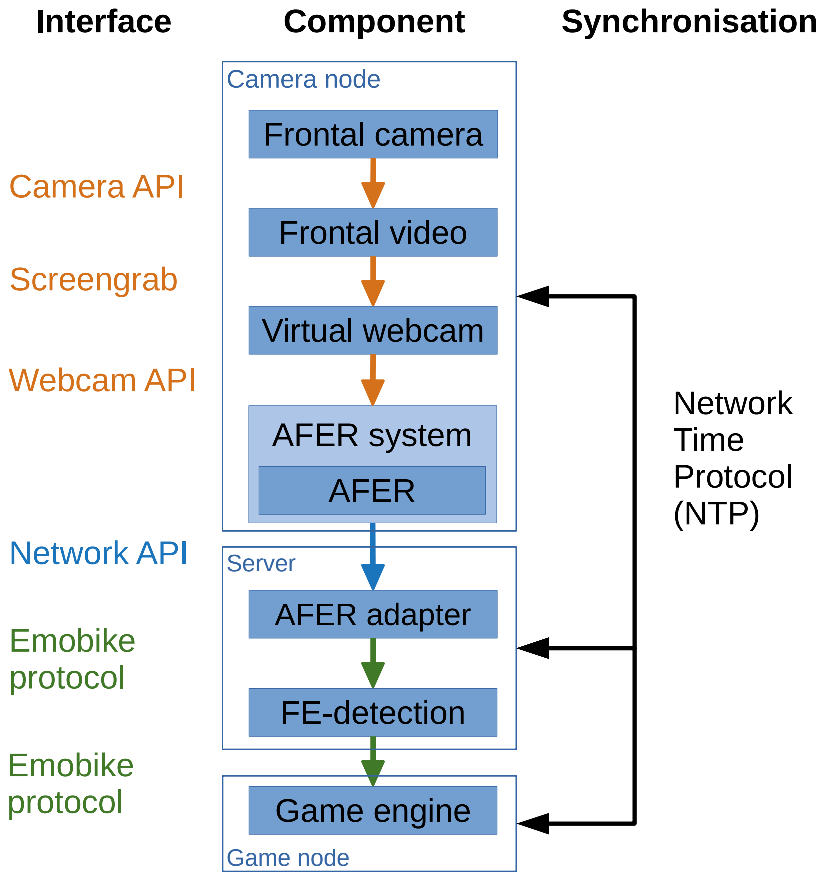


Figure 3.3: Example of a complex, distributed service built out of different components that are interconnected with APIs (orange), Network protocols (blue) and the Emobike protocol (green). This complex service describes the pipeline for reactions on facial expressions from recording to the game. The camera image is grabbed by the frontal video component, and transferred to AFER via a virtual camera driver. After processing, the output of AFER is send via an networked API to the AFER adapter, which converts it to a EmotionBike protocol message. This message is sent to a facial expression detection, determining the current expression and sends this to the game engine.

protocol and provide an easy facility for integrating new components and debugging features in the development phase. Santos, Saleme, and Andrade (2015) also proposed JSON in their systematic review of multimedia data exchange format to ensure correct encapsulation.

All messages were handled by an Apache ActiveMQ[12] message dispatcher on the server side and recorded to a SQL database for later analysis.

In order to facilitate the implementation of new applications as services, a platform-independent layer (called Emobike-lib) was designed and implemented with Qt/C++[13]. The different communication layers of the components are depicted in Figure 3.4.

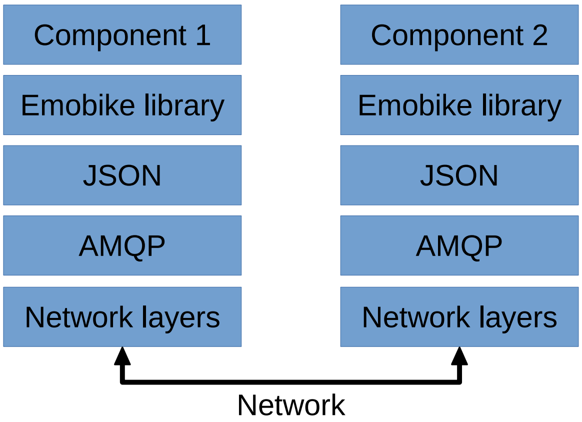


Figure 3.4: Layers of the Emobike-lib design. Components utilise the Emobike-lib which uses the levels below to ensure communication to other components.

Since the exchange of high-bandwidth raw data can be a critical limitation to networking and system latency, only the required, mostly aggregated data (e.g. analysed facial expressions instead of raw video data) was exchanged via the messaging system. Raw data was first processed, saved locally and later copied to the server to enable additional post-processing analysis of the experiments.

[12] http://activemq.apache.org/

[13] https://blog.qt.io

3.3.3. System-wide Time Synchronisation

Timing and synchronisation is a core factor in distributed systems, including latency and reaction delay of components. The EmotionBike system is distributed across several computing nodes as depicted in Figure 3.1.

Time synchronisation was implemented on the operating system level by adding a NTP server and client scheme. Timestamps for the components were generated by requesting the current time from the operating system, the temporal resolution was one microsecond.

3.3.4. Physical Setup of the EmotionBike

The EmotionBike framework provides a multi-machine distributed system that is independent of operating system and of programming language. For exercising, an ergometer (Daum premium 8i) with a networking API was extended with steering (Inkrementalgeber Kübler 2400) and break–gear (potentiometer Inc 2013) capabilities for controlling the game. The user cycles through a game environment displayed on a frontal display and the face of the user is illuminated with LED lamps. The user's face is recorded with a Kinect camera and a thermographic imaging system (Infratec VarioCam HD 875). All these components are displayed in Figure 3.5.

Physiological data were recorded with a BIOPAC MP36 (for the first set of experiments) and a biosignalsplux (for the second series). For computational processing, a PC-based camera system, a PC-based game system, a data server and two raspberry Pis were integrated.

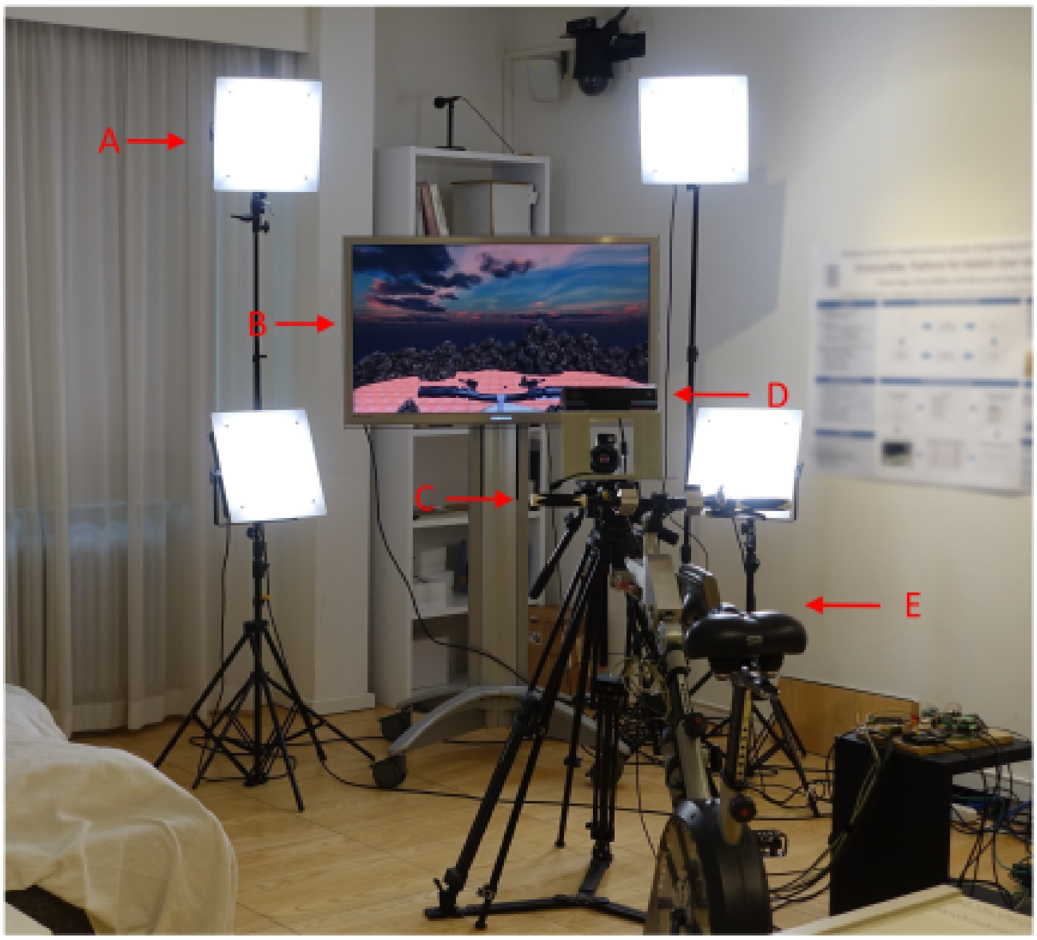


Figure 3.5: Physical setup for ES2: (A) illumination lamps; (B) frontal display; (C) rotatable handlebar, gear and brake; (D) Kinect Camera; (E) physical exergame controller (ergometer).

3.4. Experiments with the EmotionBike System

Two main sets of experiments have been conducted with the EmotionBike system.

3.4.1. Experimental Series with the EmotionBike System

A pretest with eight participants was performed before the first experimental series (ES1) at the beginning of 2015 to verify the general availability and accuracy of the overall system. After these tests, the first set of experiments was conducted.

The first set of experiments was conducted in March and April 2015 with 11 participants. Eight were male and three female with age ranging from 19 to 41 (with an average of 27). The focus in ES1 was to generate a first data set for evaluation and post-experiment processing of the data without an emotional response of the system, thus not requiring a real-time AFER solution.

Twenty-five participants took part in ES2 in February and March 2017. Fifteen were female and 10 male with ages ranging from 18 to 51 (with an average of 29). Different measures were recorded per user and are later referenced as Mnnnn (e.g. M2036 for measure labelled 2036). The focus on this set of experiments was to provide an *emotional journey*, where the order of the game levels was controlled by emotions occurring in the previous level. This required the integration of a real-time AFER solution as depicted in Figure 3.3[14].

[14] In ES2, the analysis of emotional responses was semi-automatic to ensure a correct interpretation during the experiments.

3.4.2. General Structure of the Game Levels

In general, all game levels required the subjects to drive from the start of the level to an end sign signalling the end of the level. Different events and tasks were implemented depending on the level. The general structure of these levels is depicted in Figure 3.6.

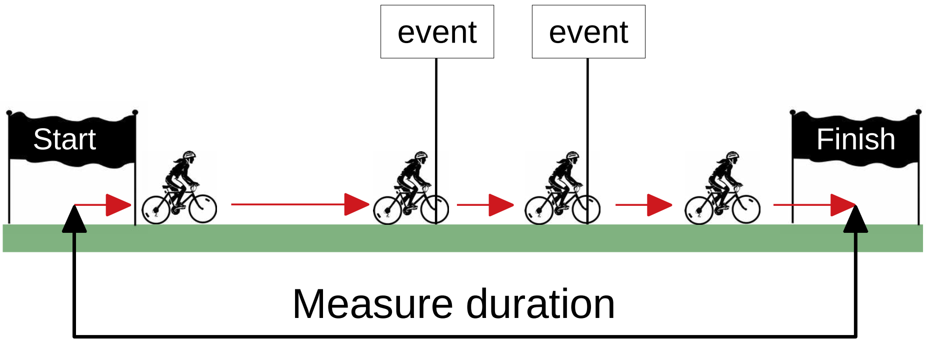


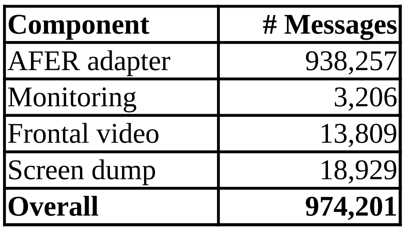
Figure 3.6: Fundamental structure of an EmotionBike level. A measurement was conducted over the complete time when the user is in the level, marked by start and stop. Events occurring during the level (e.g. a jump-scare) are logged to analyse the users' reaction afterwards.

Examples of two levels are depicted in Figure 6.11 and 6.10 as data from both levels is evaluated in later chapters of this thesis. The detailed explanation of all levels is beyond the scope of this work. They are described in (Müller et al. 2015; Müller et al. 2017; Müller 2018).

3.5. Performance Results of the EmotionBike Framework

As described in Section 3.4, two different series of experiments were conducted in the EmotionBike project. As a result of ES1, additional metrics for evaluating the performance (see Sections 3.5.3, 3.5.4, 3.5.5) were added to the framework.

Table 3.5 presents the number of all messages sent and recorded during the 258 measurements (with 518 min total recording time) of ES2 for the components responsible for the time-sensitive behaviour related to AFER of the EmotionBike platform.



Component	**# Messages**
AFER adapter	938,257
Monitoring	3,206
Frontal video	13,809
Screen dump	18,929
Overall	**974,201**

Table 3.5: Number of all messages sent by the AFER-related components during the 8.6 h of recording during experiment ES2.

3.5.1. Comparison of Image Recording Frame Rates

The frame rates for screen dump and the frontal image recording were calculated with the number of recorded frames in comparison to the length of the measure as logged with the start-measure and stop-measure commands. For ES1, 70 measures were evaluated, but only 67 are considered due to failures in recording in three cases. For ES2, two hundred and fifty-three measures were analysed. Both measures showed the same tendency for the **screen recording**: the longer the recording time (from 21s to 286s for ES1 and 10s to 505s for ES2), the closer the recorded frame-rate to the optimal rate of 30 fps, ranging from 26.8 fps to 29.8 fps for ES1 and 29.6 fps to 30.0 fps for ES2. The mean increased from 29.5 fps for ES1 to 29.8 fps for ES2.

For the **frontal videos**, the results differ. The resulting frame-rate is independent of the duration of the measurement. The fps fluctuated between 26.8 fps and 30.0 fps for ES1 and 28.8 fps and 30.0 fps for ES2. The mean fps increased from 28.84 for ES1 to 29.58 for ES2.

One main reason identified for the fluctuating fps was lighting conditions for ES1 which decreased the fps from the Kinect camera as the source of the frontal imaging. Therefore, when the data was later analysed (e.g. Chapter 6), the interval was spread over the actual length of the data in order to achieve correspondence to the time of important events.

3.5.2. Latency of the Complex AFER Component

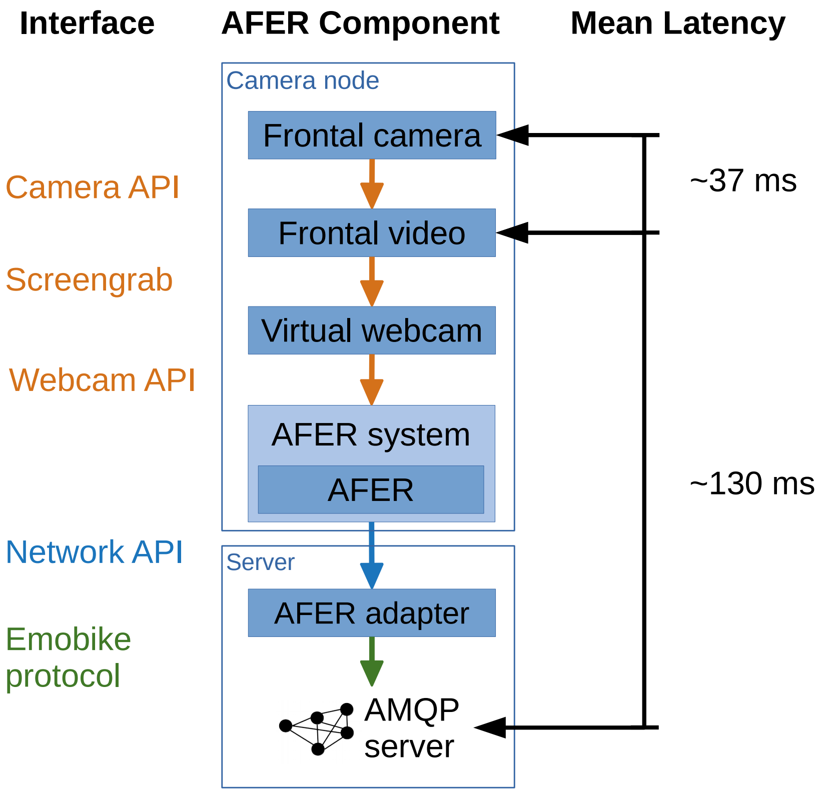


Figure 3.7: Latency of the complex AFER component. The mean latency between Kinect camera recording and the final message containing the detected probabilities of emotions from the AFER adapter send to the AMQP messaging server is 167ms, with a minimum of 120ms and a maximum of 213ms.

The complete pipeline for AFER processing (as depicted in Figure 3.7) within the EmotionBike starts with the video frame recording of the frontal camera by the frontal video component. This video is transferred to the AFER processing system, utilising a virtual camera adapter. After processing with AFER the data is transferred AFER API to the AFER adapter which converts it to a compatible message for the EmotionBike Protocol.

The applied Kinect camera sensor has a known minimal latency of 20ms for the camera (Sell and O'Connor 2014; Fankhauser et al. 2015), resulting in a maximum latency of 53ms when running at 30Hz, so the mean latency is around 37ms. The latency from the frontal video to the output of the AFER adapter was calculated from the recorded data to be between 100ms - 160ms with an average of 130ms, resulting in an overall maximum latency of 215ms.

As the the dynamic of facial expressions is relatively slow, with normal facial expression often exceeding a duration of 1s (refer to Section 2.6.1), an additional latency of 215ms is acceptable, because the reaction to changing facial reactions in the game are less demanding than for interactive controllers. Abrupt changes in the game flow in reaction of facial expressions should be avoided, since they could compromise the consistency of the game world from a users perspective. This approach posed no problem, as the response of the EmotionBike in ES2 was to select the following game level based on the emotions that occurred, not to trigger in-level changes. **The 100ms threshold mentioned in Section 3.2.2 refers to direct interaction (e.g. turning the handlebar) only**.

3.5.3. System Latency and Component Delay

Measurement of General Component Latency

The latency of the system components was measured by collecting the delay. To measure this metric, the monitoring sent a time request to the corresponding channel. Each component capable of answering this request (some older components had not implemented this feature) sent its own time as depicted in Figure 3.8. Delay and clock drift was calculated utilising the values *t1*, *t2* and *t3*. The latency is measured from monitoring to monitoring and therefore not the normal runtime of the signal, but normally twice the runtime.

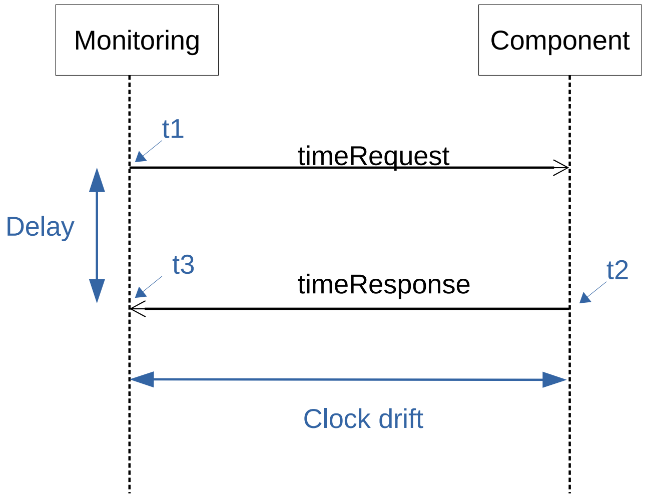


Figure 3.8: Sequence of measuring component delay and clock drift. Delay of the component is the difference between *t1* and *t3*, and clock drift is calculated as *t3* - *t2*.

Delay for Frontal Video and Screen Dump Components

Figures 3.9 and 3.10 depict two representative example delay behaviours for the measure M2036. The delay for frontal video often starts with an increased value due to a synchronous API-call to start the Kinect camera recording. The mean delay for this measure is 4.53 ms for frontal video and 4.67 ms for screen dump. Both components show smaller aberrations, but all together are reasonably stable during measurement.

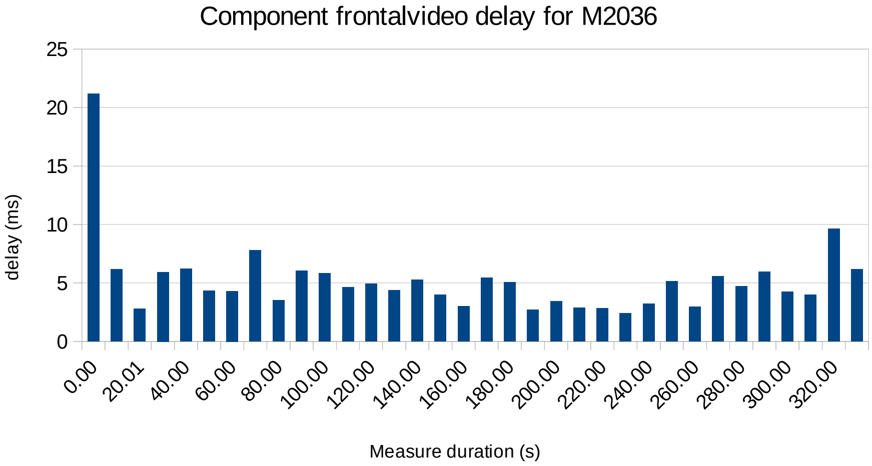


Figure 3.9: Frontal component video delay for measure M2036 of ES2 (338.5s duration, mean of 6.4ms). At the start of a measurement, the process is busy setting up the Kinect video recording which increases the first delay. After this delay, the process is as responsive as other components.

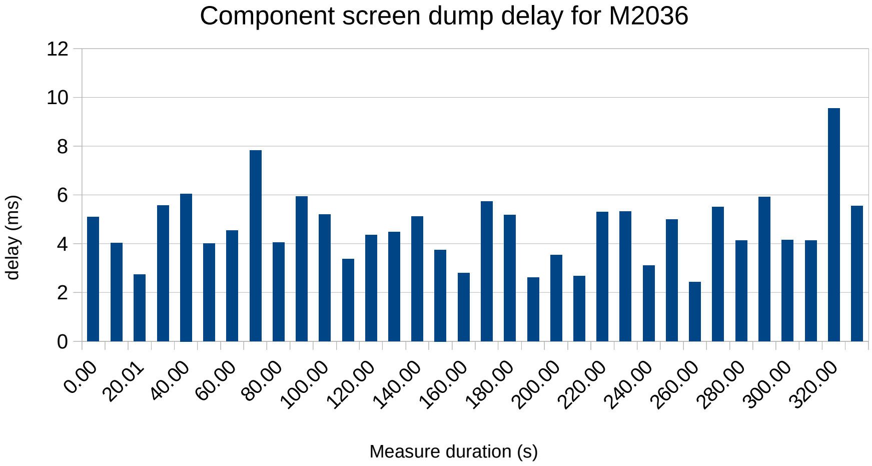


Figure 3.10: Delay of screen dump component for measure M2036 of ES2 (338.5s duration, mean of 6.4ms).

Component Delays for ES2

Table 3.6 displays the measured component delays for all 258 valid measures from ES2. With the exceptions of rare conditions, all AFER related components provided a low delay and the system was therefore capable of an appropriate interactive response.

Component	Meanmaxdelay (ms)	Meandelay (ms)
AFER adapter	5.19	2.711
Screen dump	5.128	2.68
Game	18.256	8.557
Frontal video	5.225	2.584

Table 3.6: Delay for AFER-related components of the EmotionBike system. Component delay metric for the second set of experiments (ca. 6000 samples for each component). Meandelay was calculated over all samples, meanmaxdelay is the mean of all single maximum delays for a given component and measure. A low meanmaxdelay indicates, that outliers only occurred in rare conditions, with no impact of the overall performance of the component.

Interaction Delay from Handlebar to Display

A latency measure was conducted between the handlebar adapter and the video game display as described in Hornschuh (2015) to measure the visible delay for the user. The average results were around 80 ms which is lower than the maximum interaction threshold of 100 ms described in Section 3.2.2 and therefore provides acceptable performance.

3.5.4. Component Clock Drift for Synchronisation

Synchronising the time in distributed systems is a key requirement if the data is recorded for later analysis as drifting of clocks can lead to artefacts and the unusability of the data in the worst case.

The clock for the different system components is synchronised using the NTP protocol. During a measurement, the logging component sends a time request to each registered component to log the clock drift during measurement. An example result is depicted for measure M2036 in Figure 3.11: The maximum drift, in this case, is under 6 ms after a measure duration of 330s which has no significant negative effect on the recorded data. For longer

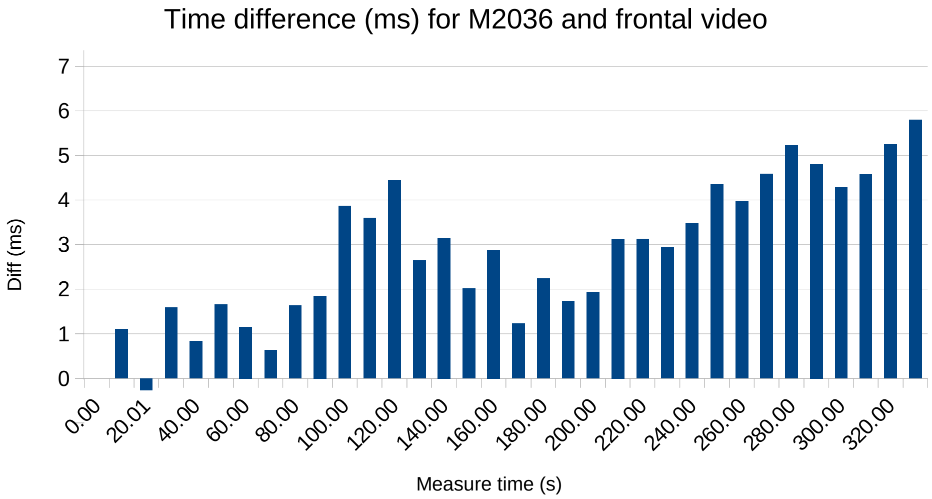


Figure 3.11: Example clock drift for component frontal video and measure M2036 (Downhill level, 338.5 s duration). The maximum drift is 5.8 ms at the end of the measure.

measurements it is important, that the clock drift is low; otherwise, the recorded timestamps for data of the components would need to be adjusted to ensure system-wide consistency.

There were two occurrences of significant jumps (maximum 1.3s) in time for the frontal video component during the 258 measurements for ES2 due to the synchronisation of the system clock by NTP, which could be corrected afterwards in the recorded data since the drift was logged. During all other measures, this effect did not occur. This effect occurred only on the camera server which utilises a Windows-operating system. The Linux-based machines performed significantly better on the issue of synchronising the clocks. To avoid these rare cases, forced sync when preparing the measure should be implemented.

Table 3.7 displays the measured clock drift values for the AFER-related components during ES2.

Component	Maxmeandrift (ms)	Meandrift (ms)
AFER adapter	2.73	1.31
Screen dump	3.91	2.28
Game	16.66	7.91
Frontal video	22.66	12.10

Table 3.7: Table displaying the clock drift for different framework components. Maxmeandrift is the mean of maximum difference values for each measurement with a low value indicating, that few outliers occurred. Meandrift shows the mean of differences over all time messages. The maxmeandrift is substantially smaller than the time between two facial video frames (33ms), fulfilling the requirements for the setup.

3.5.5. Heartbeat Capabilities

As Hou et al. (2003) described, 'heartbeat mechanism is widely used in the high availability field of monitor network service and server nodes'. The heartbeat capabilities in the EmotionBike framework are designed in the following way: components send a heartbeat message to the central system monitoring component at a (component-dependent) predefined rate to indicate that they are still in an operational state. Table 3.8 displays the results for ES2: the number of heartbeat messages sent and the absolute and mean difference compared to the expected timing of heartbeats.

As the heartbeat is based only on an internal timer event, it gives insight into the accuracy

Component	# Heartbeats	Absolut (ms)	Mean (ms)
AFER adapter	9826	4	0.20
Screen dump	10421	83	9.20
Game	10392	1404	91.54
Frontal video	8054	11	3.16

Table 3.8: Results for heartbeats: number, absolute and mean difference from the expected time of sending the messages. The heartbeat timer for the Qt-based components (all except the game) is more precise due to a better implementation of the timer in Qt.

of the timer. The components (all in Table 3.8 except the game) provided a low absolute and mean value compared to the non Qt-based components.

3.6. Summary

This chapter has discussed the technical aspects of the EmotionBike framework, focusing especially on a time-sensitive design by refining the requirements for an experimental affective interaction system to ensure the verifiability of the experiments. Related approaches were discussed resulting in the decision, that the existing approaches does not fulfill all requirements, especially the needed integration of the game engine, metrics and providing an overall system state. In order to fulfill the requirements, a novel design of the EmotionBike framework was designed, developed and evaluated.

The refined EmotionBike system complies with all requirements for multimodal interaction and recording systems as described in Section 3.2.1. Compliance with the **interactivity** requirement was ensured as the measured input-response loop for direct interactions was below 100 ms. **Concurrency** and **integratability** requirements were fulfilled, as the EmotionBike utilises a distributed, component-based system architecture. The **recording** requirement was exemplarily investigated via the frame rate of screen-dump and frontal video recording. For ES2, frontal video (at 29.6 fps) and screen dump (at 29.8 fps) almost reached the optimal frame rate of 30 fps. Additionally, clock drift and local recording ensured low network load. Measuring component delay and heartbeat messages were introduced to apply with the **metrics** requirement and to monitor the status of the components. The results of these metrics were logged, and component-dependent logfiles completed the **logging**.

The presented results of performance evaluation showed insight into the time-dependent behaviour of the EmotionBike framework and the synchronised recording of data. Important parts of the EmotionBike system, especially with a focus on facial expressions, are the image recording and processing components which form the foundation for further processing, particularly for AFER. **For this, a proper synchronisation of these components was integrated in the framework synchronising game events and corresponding video data of facial expressions, enabling post-experiment analysis without the need of manual annotating recorded data sets**.

The EmotionBike system framework provides a solid basis for integrating different approaches to AFER for live and post-processing, as described in the next chapter.

4. Performance Analysis of AFER Algorithms Using Metrics

The integration of an established AFER solution in the EmotionBike opens the possibility of a system reaction based on detected emotions and the automatic annotation of the recorded data.

Comparing different AFER approaches is challenging, since there is no standard procedure or data set as described in Section 2.6.7. The EmotionBike utilises an event-based emotion provocation. Before processing the EmotionBike data, it is important to develop metrics suitable for this new event-based facial data processing as there is no agreed standard approach for determining the primary facial expression in video data. This determination is necessary as it is required for an adequate response.

The developed EmotionBike framework offers the advantage that the time of the game-based provocation is choreographed and thus offers the possibility to tackle the challenge of determining the primary emotion within a time window around the provocative event.

Since the EmotionBike data doesn't contain ground truth about facial expressions video data, the three metrics are evaluated with labelled databases that reflect the characteristics of the applications cockpit-like environment as described in Section 2.7.2. Four state-of-the-art algorithms are benchmarked with three different metrics and three databases to find the most appropriate solution for the EmotionBike.

4.1. Introduction

Research on AFER is at the intersection of psychology, human-computer interaction, image processing and machine learning and is therefore an interesting research area for all these disciplines, each of which often has a different perspective of the domain.

Initially, the development of AFER algorithms was mainly driven by research in the field of computer science and provided to other researchers (e.g. the computer expression recognition toolbox (CERT) by Littlewort et al. (2011)). These approaches have since been further developed into commercial products[1], with the advantage that they offer professional development and support. Unfortunately, these systems are black boxes in two ways: firstly, because they are commercial software systems, their internal mechanisms are not publicly known or customisable, as they are provided only as binary to the customer, thus concealing the internal mechanisms (Kitchin 2017). Secondly, and more important, **all pre-trained AFER algorithms are black bloxes**, as they involve machine learning and thus the internal parameters of trained systems are often too complex to understand, even for the human trainer (Mittelstadt et al. 2016). As **the focus of the EmotionBike and this thesis is to apply existing AFER solutions to a novel application**, rather than developing and train-

[1] and were utilised in numerous studies, refer to Section 4.3.1.

ing a new AFER algorithm, the fact that most of the algorithms investigated are commercial systems plays a secondary role.

Most AFER systems are used within the context of market research and the study of consumer behaviour, while some are used in affective science, educational research and the study of user experience (Lewinski, Uyl, and Butler 2014). Although AFER has been proposed for online learning systems (Tettegah and Gartmeier 2015) or personal assistants in smart environments (Castellano et al. 2010), the consumer market still lacks a real ’killer application’ (Gunes and Hung 2016).

Many applied research scientists and developers of applications are confronted with the problem, that there is no neutral, manufacturer independent and cross-dataset comparison of available AFER systems with a specific focus on (non-desktop) applications. There is also no common procedure to benchmark AFER algorithms for specific applications.

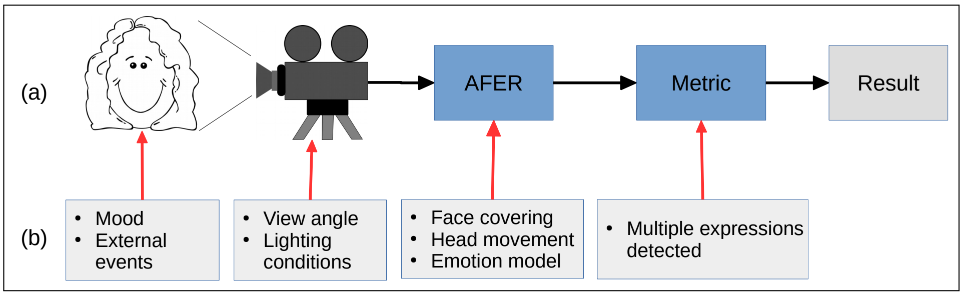


Figure 4.1: Interference factors for AFER processing: (a) depicts the pipeline with a user filmed by a camera. This video is analysed by an AFER algorithm (often including pre-processing), and metrics are applied to the output to determine a resulting categorisation. (b) lists factors that influence different stages of the pipeline with the potential to alter detection results.

Although emotions derived from facial expressions are a valuable source of information for event and reaction detection in affective applications (Dormann 2003), they are also difficult to interpret, to relate with user actions and profiles, and to reconcile other data. Consequently, the AFER analysis is quite complex and often not very reliable due to reaction fluctuations (Bernin et al. 2017). Possible factors that interfere with AFER analysis – especially in the context of labelled databases as references – are depicted in Figure 4.1. Each step of the pipeline is influenced by specific factors.

In this chapter, a benchmark protocol for testing AFER methods in application-specific conditions is presented. Four state-of-the-art AFER algorithms were selected and evaluated based on three independent databases and metrics for the algorithms output. The datasets were selected to cover the important characteristics of the cockpit environment as explained in Section 4.3.3. Three metrics are introduced and compared to find the most relevant (primary) expression in a time window of a displayed emotion. The results are later provided in a form similar to Table 4.1.

Algorithm →		Algorithm 1			Algorithm 2			…			Ground Truth
DB ↓	Metric → Emotion ↓	Metric 1	Metric 2	Metric 3	Metric 1	Metric 2	Metric 3	…	…	…	Number of labelled videos
DB1	expression A	??	??	??	??	??	??	…	…	…	??
	expression B	??	??	??	??	??	??	…	…	…	??
	…	…	…	…	…	…	…	…	…	…	…
	overall	??	??	??	??	??	??	…	…	…	…
…	expression A	??	??	??	??	??	??	…	…	…	??
	expression A	??	??	??	??	??	??	…	…	…	??
	…	…	…	…	…	…	…	…	…	…	…
	overall	??	??	??	??	??	??	…	…	…	…

Table 4.1: Example of a table that displays benchmark results. For each algorithm (white) and database (blue), the correct matches for every facial expression (red) in the classified videos are displayed based on one of the metrics (green) evaluated. The last column (violet) displays the videos labelled with this expression category.

4.2. Related Work

Details of the general principles of AFER algorithms are provided in Section 2.6.4 and in the comprehensive surveys published by Sariyanidi, Gunes, and Cavallaro (2015) and Zeng et al. (2009). This ’related work’ section details published work that concern the benchmarking performance analysis and metrics for the output.

4.2.1. Testing and Benchmarking of AFER Systems

A well-established approach to benchmark AFER algorithms is to evaluate their performance on labelled datasets (e.g. Valstar et al. (2011a), Littlewort et al. (2011), and Sariyanidi, Gunes, and Cavallaro (2015)) using similar comparison pipelines to the one depicted in Figure 4.2 with the difference, that metrics are usually not explicitly mentioned. This is unfortunate as the metric is essential for deciding for the primary expression and makes integration into an application difficult.

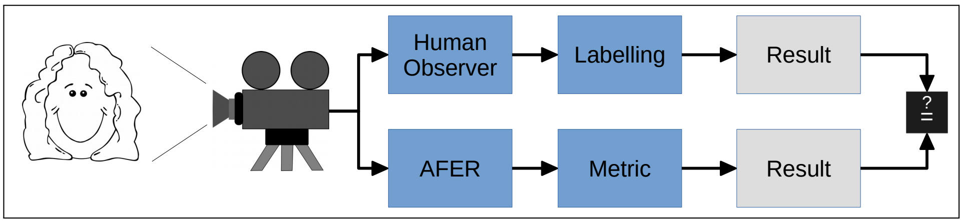


Figure 4.2: General pipeline for benchmarking AFER algorithms: the recorded videos are labelled by both the algorithm and a human observer, and their results are then compared.

A variety of annotated facial expressions databases is available to the scientific community, for details see Section 4.2.2.

For commercially available systems, few comparisons have been published. One example was published by Lewinski, Uyl, and Butler (2014), who utilised the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP) (Olszanowski et al. 2008) and the Amsterdam Dynamic Facial Expression Set (ADFES) (Van Der Schalk et al. 2011) for benchmarking the Noldus FaceReader[2] AFER system. In contrast to the benchmark applied in the present study, only the single frame that displayed the apex of expression intensity was compared to the label of the video for categorisation. Only comparing the apex is a common approach for benchmarking (Sariyanidi, Gunes, and Cavallaro 2015).

Lewinski, Uyl, and Butler (2014) defined matching score as ‘the percentage of observers who selected the predicted label’ and calculated it, based on the apex, for every basic emotion. Stöckli et al. (2017) applied the same accuracy measure for their study.

Stöckli et al. (2017) utilised WSEFEP and ADFES for their validation study on Affectiva (Affdex) and Emotient (FACET). In contrast to Lewinski, Uyl, and Butler (2014), Stöckli et al. (2017) generated videos instead of applying still images, as this is a requirement of Affectiva and Emotient. While ADFES already contains video sequences, Stöckli et al. (2017) converted still images from WSEFEP into five-second videos by repeating a single frame.

Performance evaluation with a database is often performed either to compare the results of (newly) developed algorithms (e.g. Littlewort et al. (2004)) or in the context of challenges that offer their own data sets (e.g. Valstar et al. (2011a)). In the first case, two or more databases (often with the extended Cohn-Kanade data set (CK+) (Lucey et al. 2010) as a reference) are usually selected, while in the second case only the challenge dataset is evaluated for comparison. The selection of databases is often heterogeneous, which complicates a general comparison of algorithms: ‘Unfortunately, the experimental results of different systems

[2] https://www.noldus.com/human-behaviour-research/products/facereader

can be seldom compared against each other directly, as the experimental configurations of different studies are often different in terms of validation procedures, the number of test images/videos, subjects or labels' (Sariyanidi, Gunes, and Cavallaro 2015).

For application developers, it appears beneficial to include data records that reflect important characteristics of the target environment. Therefore, three databases were selected for frontal, non-frontal and smart environments as relevant to the EmotionBike setup (described in Section 4.3.3).

4.2.2. Databases

Published databases provide the ability to reproduce the results of other researchers and to evaluate new – or more recent versions – of algorithms. Databases that display emotional facial expressions have, for 20 years, been applied to training, and to the evaluation of AFER algorithms.

Although the Cohn-Kanade database (Kanade, Tian, and Cohn 2000) is considered to be the first modern dataset (Corneanu et al. 2016), Lyons et al. (1998) published the Japanese Female Facial Expression (JAFFE) database two years earlier. Both databases offer a frontal camera perspective on the expressions of different subjects for the standard set of basic emotions (happiness, sadness, surprise, anger, disgust, fear) with the difference, that JAFFE only offers single images for each expression while CK provides image sequences ranging from neutral to the highest intensity.

Since 1998, over 60 databases have been published that provide posed or spontaneous expressions (Valstar and Pantic 2010), naturalistic expressions ('in-the-wild') (Bosch et al.

2015) or 3D point-clouds with corresponding textures (Yin et al. 2006), sometimes with additional modalities, such as audio (Valstar et al. 2013), body movement (Tcherkassof et al. 2013) or physiological signals (Ringeval et al. 2013).

To provide a survey on all available datasets for AFER is beyond the scope of this work. The most comprehensive and recent survey on facial expression databases was published by Weber, Soladie, and Seguier (2018) and covers the review of 61 databases (including multimodal datasets). Previous surveys have been published by Zeng et al. (2009) and D'Mello and Kory (2015).

4.2.3. Identifying the Primary Expression

Categorisation of the expression displayed is important for generating interactive affective applications as well as for labelling videos in databases.

In order to label the videos in databases, often only the video frame with the highest intensity of expression is selected to label the complete video (Lucey et al. 2010). The same methodology was applied by Lewinski, Uyl, and Butler (2014) in their evaluation of the FaceReader AFER system, which utilised the frames with the highest peak intensity. Stöckli et al. (2017) classified a match as true if 'the highest value (out of all generated values for all basic emotions) matched with the database's emotion label'. This is especially problematic, as some algorithmic approaches utilise an approach of fast boosting the probability of an expression to 100% as later described in Section 4.4.3.

While this 'highest intensity' evaluation works well in settings with one facial expression acted, in more natural environments, there may be multiple expressions in a period of time (Gunes and Hung 2016; D'Mello and Kory 2015). The challenge in such an environment is to recognize **the primary expression, the expression that is most dominant**. One approach is to use a fixed window of time around an occurring event, as described for the EmotionBike setup in Müller et al. (2015).

Bosch et al. (2015) utilised a combination of maximum, median and standard deviation in a fixed time window for their classification of facial expressions.

Grafsgaard et al. (2013) applied an empirically determined threshold of t=0.25 to classify the output (with a probability range of 0.0 to 1.0) of the computer expression recognition toolbox (CERT), in order to detect the dominant expression based on Action Units.

Sariyanidi, Gunes, and Cavallaro (2015) mentioned that the average recognition rate or average area under the curve are mostly utilised for the evaluation of AFER systems. Unfortunately, they did not report how the decision for the primary emotion was implemented.

Littlewort et al. (2004) applied SVMs with an one-against-all decision scheme for selecting one expression among all others and analysed the output with multi-class voting. To choose for the primary emotion, Softmax allocates a number between 0 and 1, based on division of the sum over all classes for an image (Littlewort et al. 2004).

Although basic emotions are by definition mutually exclusive (Russell and Barrett 1999), the output of AFER systems is usually a multi-channel output with signals for every possible expression, so the dominant/primary emotion must be determined (Bernin et al. 2017). While selecting the frame with the maximum intensity for a clean database video that displays only one expression works well, the primary expression in an application context is usually relevant for a certain period of time, and not only for a single frame. To identify this expression, a

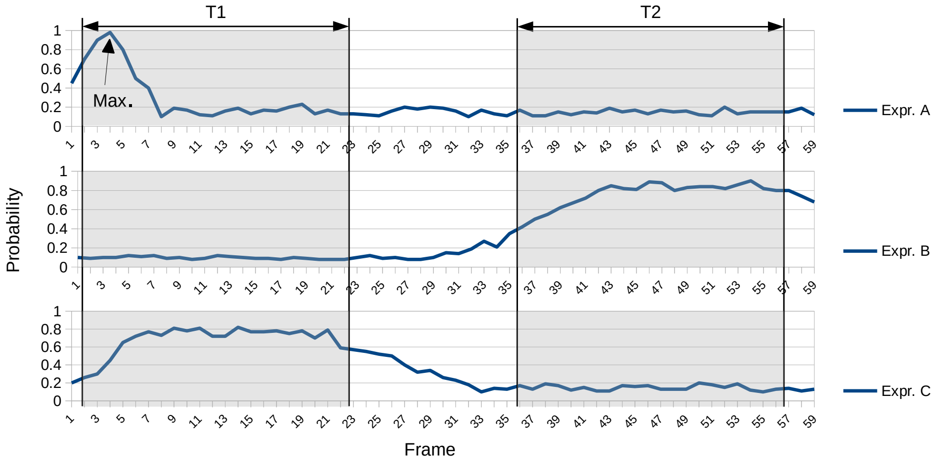


Figure 4.3: Example of a graph that displays three output channels for the probability of expressions (Expr. A-C) of an AFER system with two time periods (T1,T2) and different dominant expressions. In period T1, expression C is the most dominant over most of the time, but A is dominant for one frame, while in period T2, expression B is the primary one, as its outputs display the highest values.

window-based approach is applicable, especially for provocative-reaction-based interaction schemes that provide the event position in time (Müller et al. 2015). Figure 4.3 illustrates the case of conflicting expressions (A to C): with the maximum intensity approach, expression A would be selected for time period T1, while expression C might be chosen if using a mean or median based approach. For T2, both approaches would select expression B.

4.3. AFER Algorithms Evaluation Method

This section describes the evaluation method applied to the selected databases and AFER algorithms. To evaluate the proposed method for benchmarking, four state-of-the-art AFER systems were selected. These systems were selected based on their general availability for research with a focus on the practical approach and their widespread use in research.

As the algorithms are provided in binary form only, published information about internal

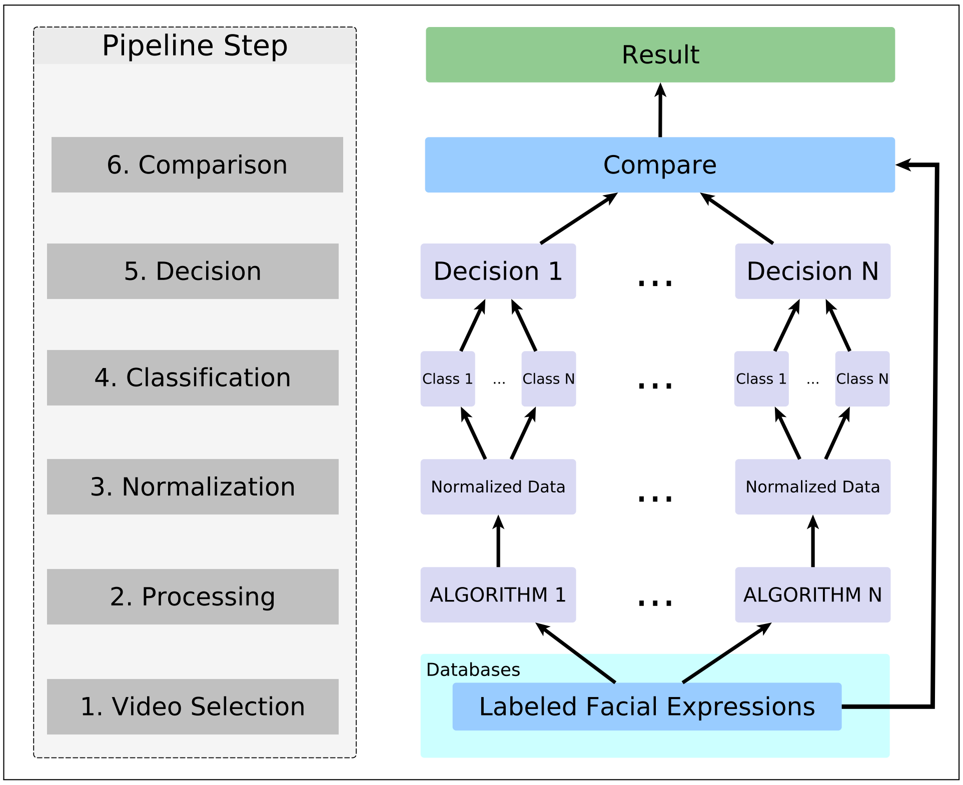


Figure 4.4: Steps of the evaluation pipeline (bottom-up): Labelled videos from databases are selected and processed with different AFER algorithms, and normalised output is analysed with metrics to identify the primary expression with the three evaluated metrics. The result is compared to the emotional label of the video.

methods is difficult to verify. In addition, the parameters inside the AFER system cannot be changed, therefore, a classical 'black box testing' (Patton 2005) was conducted with the systems default parameters.

The performance evaluation procedure is depicted in Figure 4.4, with a bottom-up diagram:

1. Select the videos labelled with one of the six basic emotions from the databases.

2. Process the videos with the different algorithms.

3. Normalise the output to probability values between 0.0 and 1.0.

4. Classify the output as described in Section 4.3.4 with the three described metrics.
5. Decide the primary expression detected in the video (see Section 4.3.4 for details).
6. Compare the result of previous step with the input-label from the database.

4.3.1. Utilised Algorithms for AFER Analysis

System	Platforms	Interface	Labelled Expressions	Output
Emotient (FACET)	Windows	IMOTIONS API/UDP	**Anger, Joy, Surprise, Sadness, Disgust, Fear**, Contempt, Confusion, Frustration, Neutral	Evidence ($\sim$ -10–10)
Affectiva (Affdex)	Linux, Windows, Mobile	C++, C#	**Anger, Joy, Surprise, Sadness, Disgust, Fear**, Contempt, Smirk, Smile	Probability (0–100)
InSight	Linux, Windows, Mac OS, Mobile	C++	**Anger, Joy, Surprise, Sadness, Disgust, Fear**, Neutral	Probability (0.0–1.0)
CERT	Mac OS	Manual csv output	**Anger, Joy, Surprise, Sadness, Disgust, Fear**, Contempt, Neutral, Smile	Probability (0.0–1.0)

Table 4.2: Overview of the AFER algorithms benchmarked. Ekman's six basic emotions are highlighted in the list of recognized labelled expressions. Except for Emotient, the output of all algorithms is a probability for the occurrence of a certain facial expression.

The four algorithms have been selected, because they have been and are still widely used in research and were utilised for AFER in a large number of studies, applying CERT (e.g. Grafsgaard et al. (2013), Rychlowska et al. (2014), Valstar et al. (2016), and Arguel et al. (2017)), Emotient (e.g. Liang et al. (2018), Izquierdo-Reyes et al. (2018), and Boucher et al. (2016)), Affectiva (e.g. Liu et al. (2017), Jaques et al. (2016), Tussyadiah and Park (2018), and Magdin and Prikler (2018)) and Insight (e.g. Blom et al. (2014), Boychuk et al. (2016), and Blom, Bakkes, and Spronck (2019)).

The four state-of-the-art AFER-systems evaluated are capable of processing video data near soft real time[3] (<1s), which is important to minimise response time when creating interactive applications. Details on the algorithms are described in Table 4.2 in addition to the following descriptions:

The designers of **InSight**[4], do not provide documentation to describe their methods[5]. Since they have added OpenCV credits to their licence, it is reasonable to assume that they have used the standard OpenCV facial recognition algorithm based on hair cascades[6] and Viola and Jones (2001).

CERT (Bartlett et al. 2008) is based on an enhanced version of Viola-Jones for face recognition together with Gabor filtering for feature extraction (Littlewort et al. 2011). The features are fit into independent linear SVMs, one for each AU probably using a one-against-all partitioning scheme as previously published by Littlewort et al. (2004). The output of this layer is fed into a multivariate logistic regression (MLR) classifier, which provides the posterior probability of each emotion as outputs (Littlewort et al. 2011).

In contrast to the three other algorithms, CERT also processes the data in almost real time, but does not provide an appropriate output interface for integrating in a live application, as only batch processing of the input is supported in the program version that was provided for research. CERT was applied for the post-experiment evaluation of ES1 and still has been applied in recent scientific publications (e.g. Valstar et al. (2016), Dufner et al. (2018), and Arguel et al. (2017)).

[3] With the exception of CERT which only supports batch processing of the input data.

[4] http://sightcorp.com/insight/

[5] To the knowledge of the author.

[6] https://docs.opencv.org/trunk/d7/d8b/tutorial_py_face_detection.html

Emotient/FACET, as the successor to CERT, utilises the same fundamental algorithms (Mone 2015). In contrast to CERT, categorisation of the expressions is not based on the AUs output but is separately trained on the raw feature set. Emotient was purchased by Apple in 2015[7], which terminated its availability as a product in its own right. However, it is still available as part of the iMotions platform[8].

Affectiva (also named Affdex SDK) utilises a pipeline for face detection through Viola-Jones and localization of important facial areas (McDuff et al. 2016). After an extraction of texture characteristics with a histogram of oriented gradients, the AUs are classified with the use of separate SVMs. The prototypical emotions are categorised based on the AUs, as described by the emotional facial coding system (EMFACS-7) (McDuff et al. 2016).

4.3.2. Normalising AFER Output

Some AFER algorithms such as Emotient (see Table 4.2) provide results in the form of 'evidence values', which must first be converted into probability values ranging from 0.0 to 1.0 to be comparable with the output of other AFER approaches. The evidence value represents the distance to the hyperplane of an SVM (as described in Littlewort et al. (2004)).

$$p(x) = \frac{1}{1 + 10^{-1*x}} \tag{4.1}$$

For the transformation of data in this study, Equation 4.1 is suggested in the iMotions documentation, with x as the evidence value in a data range from $\sim$ -10 to 10, and p(x) as the

[7] https://www.wired.co.uk/article/apple-emotient-ai-emotions

[8] https://imotions.com/facial-expressions

resulting probability. As this conversion results in a value of 0.5 for an evidence of 0 (representing no signal found), the original formula was the value was altered to return 0 instead of 0.5 for consistency with the three other algorithms (Equation 4.2).

$$p(x) = \begin{cases} 0.0 & \text{if } x = 0 \\ \frac{1}{1+10^{-1*x}} & \text{otherwise.} \end{cases} \tag{4.2}$$

4.3.3. Utilised Databases for Benchmarking

The approach of identifying for databases with relevant characteristics to the EmotionBike setup was based on the characteristics described in Table 4.3. This table describes the characteristics of the three environments relevant to the EmotionBike. In this table there is an increasing difficulty level from top to bottom, which causes cockpit to contain all limitations of cockpit, and go beyond them. The same is true for smart environment. This circumstance is essential for the evaluation of the AFER algorithms, because for cockpit and especially for an affective exergame, such as the EmotionBike, no database existed.

Environment	Covered area	Lighting conditions	Computing power	Camera field of view	Face position	Face view angle	Number of users
Desktop	Small	Constrained	High	Small	Constrained	Constrained	One
Cockpit	**Small**	**Unconstrained**	**High**	**Small**	**Constrained**	**Unconstrained**	**One**
Smart Environment	Medium	Unconstrained	High	Wide	Unconstrained	Unconstrained	One – Multiple

Table 4.3: Main characteristics of environments for AFER-based affective applications relevant to the EmotionBike (abbreviated Table 2.3) The highlighted environments are close to the EmotionBike setup. As no database specific to the cockpit-scenario was available, smart environment was chosen as more demanding environment, already including the characteristics of cockpit and desktop scenarios.

The challenging characteristics for a cockpit scenario are robustness against changing lighting conditions and the view angle from camera to face as described in Table 4.3.

The CK+, ‘Binghamton university 3D facial expression’ (BU-4DFE) and ‘affective facial expressions in the wild’ (AFEW) databases were selected to represent various aspects of benchmarking for smart environments. All three databases cover different aspects of facial expression such as head movements in the scene, temporal phases, and frontal or non-frontal scenes (see Table 4.4 for details).

Database	Datatype	Head Movement	Performer	Origin	Phases	Environment
CK+	2D frontal	No	Non-professional	Acted expression	Onset	Desktop
BU-4DFE	2D frontal / 3D and textures	No	Non-professional	Acted expression	Onset apex offset	Desktop
AFEW	2D frontal / non-frontal	Yes	Professional actors	Movies	mixed	Smart Environment

Table 4.4: Video characteristics of utilised databases. This table describes type of data (2D/3D), movement of head, type of performer, origin of expression, covered phases and the corresponding environment.

Only videos with a basic emotion facial expression label were chosen from the databases for the generated test bed. Videos not labelled or labelled with other expressions from the CK+ (18) and AFEW (194) databases were therefore ignored as a comparison with the expressions detected by other AFER algorithms is not possible. Table 4.5 lists the available videos in each database.

Database	Included Data			Labelled Expressions (6BE)							Additional	
Name	Subjects	Videos	Labelled	Anger	Joy	Surprise	Sadness	Disgust	Fear	All	Contempt	Neutral
CK+	123	593	327	45	69	83	28	59	25	**309**	18	0
BU-4DFE	41	605	605	101	100	101	101	101	101	**605**	0	0
AFEW	330	1645	1106	182	208	119	168	112	123	**912**	0	194

Table 4.5: This table displays the number of subjects and (labelled) videos per database. AFEW and CK+ contain additional labelled video sequences.

The three databases were selected because of the following reasons.

The **CK+** dataset (Kanade, Tian, and Cohn 2000) is the unofficial gold standard for AFER algorithm testing and training. Although the original CK database was created in the year 2000, with the CK+ update by Lucey et al. (2010), it has still been utilised in recent publications as a benchmark (Happy and Routray 2015; Lopes et al. 2017). The extension of CK by the CK+ dataset contains new subjects, spontaneous recordings, as well as associated annotations and labels, and expressions for 'contempt'. (Sariyanidi, Gunes, and Cavallaro 2015).

BU-4DFE (Yin et al. 2008) is a database that includes facial 2D (textures) and 3D point-clouds of the head as sequences (called 4D data). This combination of data enables to generate frontal and non-frontal (to approximately 30 degrees) images of the face. This type of facial view is very similar to the data provided by the Kinect camera setup of the EmotionBike (see Chapter 3), in which the depth camera creates point-clouds and texture information.

AFEW (Dhall et al. 2012a) provides short scenes (up to a maximum length of six seconds) from movies to display facial expressions by professional actors in a near real life setup, compared to lighting, (partial) covering of faces and changing view angles. This dataset has been included as the videos are the most challenging task for AFER algorithms in almost realistic application conditions.

Examples of video frames from the three databases are depicted in Figure 4.5.

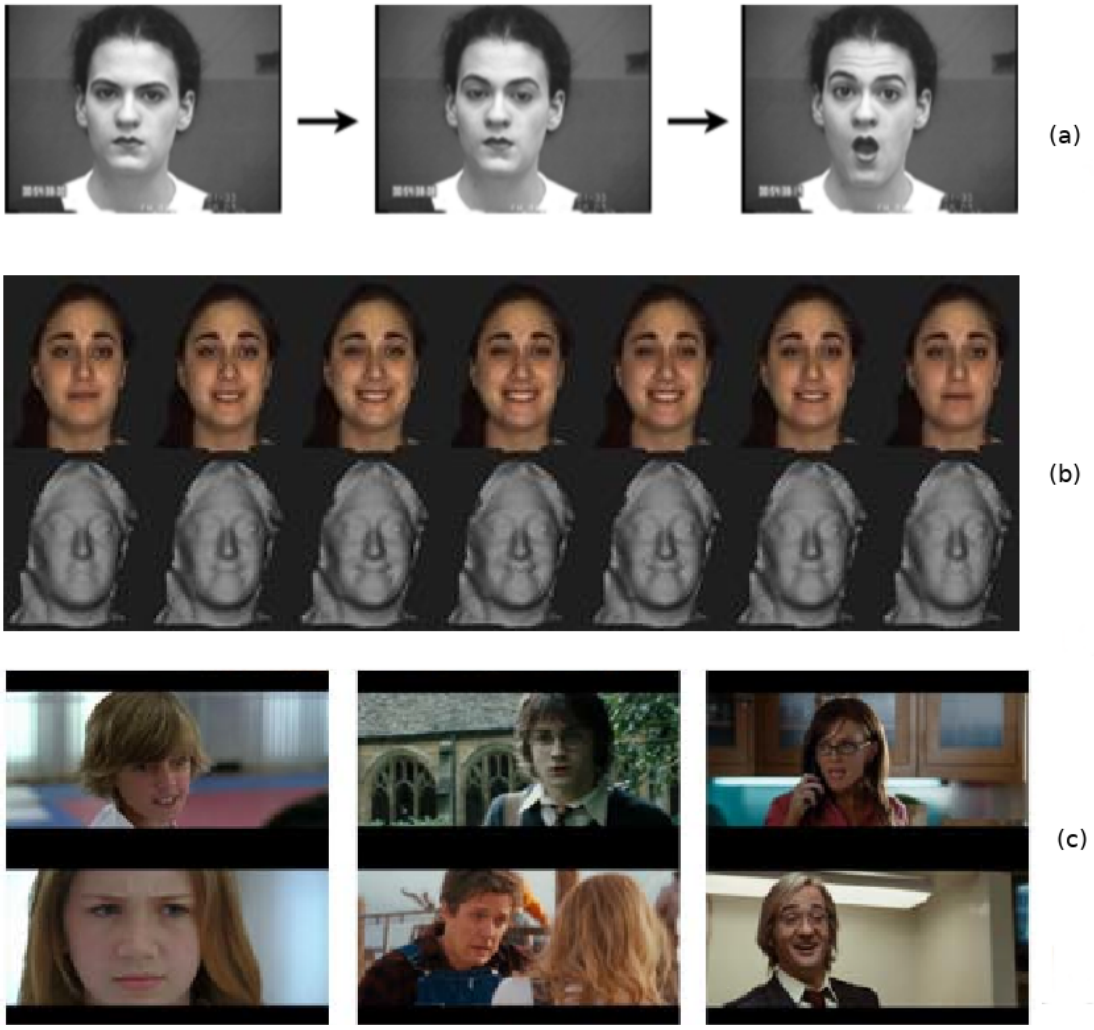


Figure 4.5: Example images from the utilised databases: Cohn-Kanade+ (a) provides 2D frontal videos with acted expressions by non-professional performers. BU-4FDE (b) contains 3D point cloud data sequences and textures that display frontal acted expressions by non-professional performers. (c) depicts examples of AFEW, which present short snippets from commercial films with changing view angles and lighting conditions.

4.3.4. Metrics for Classification and Decision of the Primary Expression

As mentioned in Section 4.2.3 there are different approaches to identify the most relevant expression in a video sequence. For early work in the EmotionBike project (Müller et al. 2015; Müller et al. 2016) a primarily manual based approach was applied. To enhance the window based method and fulfil this task automatically, three possible metrics to identify the primary expression (Equations 4.3, 4.4 and 4.5) have been evaluated.

For all equations, x is the normalised probability (range = 0.0-1.0) of an expression and n is the count of frames processed.

$$mean(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.3}$$

$$meanp5(x) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0 & \text{if } x_i <= t \\ x_i - t & \text{if } x_i > t \end{cases} \tag{4.4}$$

$$binaryp5(x) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0 & \text{if } x_i <= t \\ 1 & \text{if } x_i > t \end{cases} \tag{4.5}$$

In addition to a standard mean (Equation 4.3), the other two metrics (Equation 4.4 and 4.5) include a threshold for values lower than t=0.5. Using a binary filter and thresholding are common approaches in signal processing. The *binaryp5* approach was chosen to increase the divergence of data for expressions with lower values. The purpose of *meanp5* was to give preference to expressions with higher values. Lowering the threshold leads to an increases in false-positive detection and higher values lead to missing occurrences.

To choose the primary expression, the maximum rule, also described as 'winner takes it all', was then applied to the metrics outcomes for each group of expressions and for every algorithm. The highest value classifies the expression as primary, and this result is then compared to the database label.

4.4. Performance Results for the Benchmarked AFER Algorithms

4.4.1. Example of a Complex Result

Figure 4.6 displays a typical example of AFER output to illustrate the complexity that an application developer may encounter, even with the simple CK+ database. This Figure also depicts an example of a possible false interpretation; as the values for 'frustration' and 'confusion' increase earlier and have higher averages, this would be classified as 'confusion'. A continuous presence of 'sadness' also appears in the output; even in the beginning, where the expression should be 'neutral'[9]. This is an indication that a base-line to correct the expressions would be required for this subject. The occurrence of conflicting expressions is common in multi-channel approaches, an advice on handling this is explained in Chapter 5.

[9] CK+ videos always start with the neutral expression

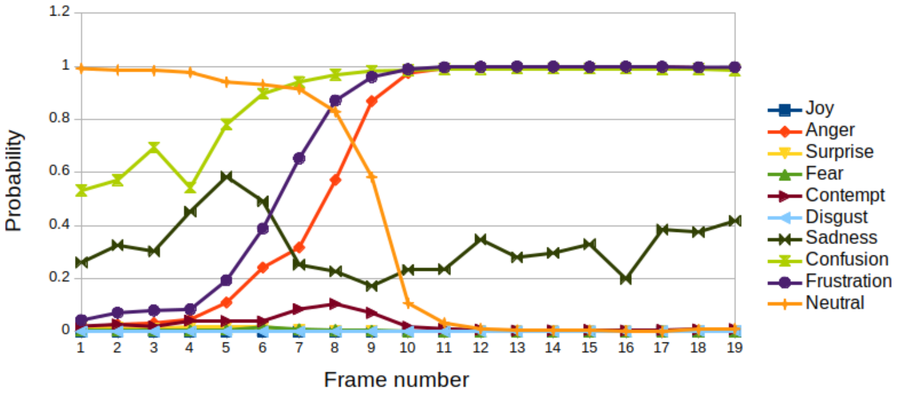


Figure 4.6: Sample output of subject S010 from the CK+ database labelled with 'anger'. The video has been analysed with AFER algorithm Emotient. In addition to the six basic emotions, the extra expressions provided by Emotient are presented: frustration, contempt, confusion and neutrality. 'Anger' is present, but 'confusion' and 'frustration' begin to increase earlier and therefore have higher average values.

Although the example displayed in Figure 4.6 refers to more channels than just the 6BE, conflicting expressions also occur frequently among these as well.

4.4.2. Evaluation of Metric Results

The mean, *meanp5* and *binary5* metrics (described in Section 4.3.4) were evaluated for their respective values for primary expression recognition. The details are presented in Table 4.6.

The best overall accuracy of all algorithms and metrics was achieved by CERT, with 886 videos (out of 1826) correctly classified using the mean metric, followed by Emotient, with 861 videos. Relative to the number of evaluated videos in the databases, the accuracy is only ∼49% for CERT and ∼47% for Emotient. The difficulty increases progressively with a decreasing recognition rate for the databases with the mean approach, starting with 96% for

CERT and CK+ to 31% with AFEW. Emotient dropped from 96% on CK+ data to $\sim 18\%$,only slightly higher than chance ($\sim$ 16.6%).

Meanp5 and *binaryp5* values were usually similar with some exceptions: e.g. 123 videos from AFEW that were labelled ’sadness’ were correctly detected by CERT using the mean metric and only 92 by applying *meanp5*. An explanation for this behaviour could be, that the intensity of the sad expressions is not so expressive and therefore the values forming the mean are too low to pass the threshold for p5.

There are a few occasions where *binaryp5* delivers better results than mean or *meanp5*: ’disgust’ and ’fear’ for BU-4DF and Insight, ’anger’, ’sadness’, ’fear’ and ’overall’ for AFEW and Emotient. But these differences are negligible in comparison to the total number of labelled videos.

The mean metric seems to be the principal approach for the best results in classification of the output data of the algorithms and is therefore the default metric.

4.4.3. Intensity Insensitive Boosted Output

One visible difference between the algorithms, especially for videos from AFEW, is the difference between true-positive detections of mean, *meanp5* and *binaryp5* in Table 4.6. If the count for mean is similar to the others, this can indicate a ’boosting the output’; regardless of the intensity, probability output of the algorithm increases quickly to the maximum of 1.0 indicated by similar results for mean, *meanp5* and *binaryp5*. This is displayed in the last two lines of Table 4.6; For CERT and Emotient the differences between all three metrics ($\sim$1.6% and $\sim$0.5%) are much smaller than for InSight ($\sim$2.6%) and especially Affectiva ($\sim$9%). This

Algorithm →		**Insight**			**Emotient**			**Affectiva**			**CERT**			**Ground Truth**
DB ↓	**Metric →** **Emotion ↓**	mean	meanp5	binaryp5	mean	meanp5	binaryp5	mean	meanp5	binaryp5	mean	meanp5	binaryp5	Number of labelled videos
CK+	Anger	26	24	26	42	41	40	16	8	8	43	43	43	45
	Joy	53	54	53	69	69	69	51	53	50	68	68	68	69
	Surprise	81	81	80	80	81	81	66	29	28	74	76	76	83
	Sadness	17	10	10	25	26	25	15	9	9	25	26	25	28
	Disgust	44	44	44	56	59	53	59	54	53	57	59	58	59
	Fear	21	20	18	20	22	21	2	2	2	23	24	24	25
	overall	**242**	**233**	**231**	**292**	**298**	**289**	**209**	**155**	**150**	**290**	**296**	**294**	**309**
BU-4DF	Anger	64	61	62	61	59	57	39	15	16	54	50	53	101
	Joy	49	48	49	97	97	96	88	89	88	76	75	75	100
	Surprise	72	72	71	86	89	86	57	35	34	33	33	34	101
	Sadness	21	8	8	67	61	62	25	14	14	77	67	67	101
	Disgust	9	9	10	73	73	72	72	67	67	48	48	49	101
	Fear	8	9	10	21	22	21	4	4	4	29	30	29	101
	overall	**223**	**207**	**210**	**405**	**401**	**394**	**285**	**224**	**223**	**317**	**303**	**307**	**605**
AFEW	Anger	34	30	34	14	11	17	8	4	5	26	29	27	182
	Joy	9	9	9	49	50	50	94	94	94	97	85	82	208
	Surprise	33	35	34	51	52	48	17	2	2	10	9	10	119
	Sadness	29	10	14	34	33	37	7	0	1	123	92	97	168
	Disgust	1	1	1	12	11	12	46	28	28	15	16	17	112
	Fear	6	5	5	4	4	5	0	0	0	8	8	8	123
	overall	**112**	**90**	**97**	**164**	**161**	**169**	**172**	**128**	**130**	**279**	**239**	**241**	**912**
All Dbs	**overall**	**577**	530	538	**861**	860	852	**666**	507	503	**886**	838	842	1826
%	**overall**	**31.60**	29.03	29.46	**47.15**	47.10	46.66	**36.47**	27.77	27.55	**48.52**	45.89	46.11	100.00

Table 4.6: Database benchmark with classification results for all four AFER algorithms, three metrics and three databases with the number of labelled videos for ground truth. The cells display the number of correctly classified videos, with a background colour between white and green that corresponds to the percentage of the total number of videos for each expression. The difficulty of detection increases with the databases from top to down. The best overall performance over all three databases is reached with the mean metric for all four AFER algorithms as highlighted in the last two lines. It should be noted, that parts of CK+ might have been used for training; see Section 4.4.4 for details.

behaviour is also visible in the analysis of videos from the first EmotionBike experiment (see Section 6.4.2).

4.4.4. Detailed Performance Analysis

The evaluation of all four AFER systems was performed in batch mode, so the videos were completely analysed before the output was saved to a file. A runtime analysis for frame-by-frame processing was not in the focus of this benchmark[10], although all systems are in principal capable of soft real-time (<1s) processing. For the EmotionBike, the processing time (with Emotient) was measured to approx. 130 ms, as described in Section 3.5.2.

Due to the six classes of basic emotions, the minimal detection rate required is the probability by chance: $\sim$16.67%. Table 4.7 presents the recognition rates for all databases and algorithms that apply the mean metric as the most effective key performance indicator.

The detection rates differed substantially between the databases, especially if a database was not applied for training in the AFER system. Recognition rates of over 90% or even 100% are an indication that these databases were (at least partially) used for training. For Emotient and CERT, this seems to be the case, for CERT it is known that they applied a subset of CK+ namely the original CK dataset as stated by Littlewort et al. (2011).

Potential training data should be used with care for benchmarking, as this introduces bias to the comparison. Nevertheless, there are compelling reasons to do so in this case:

1. Excellent results may indicate where particular data was applied as training-sets and provide insight into this particular case.

2. Benchmark provides blackbox-testing to detect software bugs and gauge over-training effects (e.g. CERT and CK+, Littlewort et al. (2011)).

3. CK+ is currently an established benchmark (e.g Happy and Routray (2015) and Lopes et al. (2017)).

[10] And would be impossible for CERT due to the missing runtime data export.

Database	**Emotion label**	**Number of valid videos**	**Number of videos (%) correctly detected by**			
			InSight	**Emotient**	**Affectiva**	**CERT**
CK+	Anger	45	26 (58%)	42 (93%)	16 (36%)	**43 (96%)**
CK+	Joy	69	53 (77%)	**69 (100%)**	51 (74%)	68 (99%)
CK+	Surprise	83	**81 (98%)**	80 (96%)	66 (80%)	74 (89%)
CK+	Sadness	28	17 (61%)	**25 (89%)**	15 (54%)	**25 (89%)**
CK+	Disgust	59	44 (75%)	56 (95%)	**59 (100%)**	57 (97%)
CK+	Fear	25	21 (84%)	20 (80%)	2 (8%)	**23 (92%)**
CK+	**overall**	**309**	**242 (78%)**	**292 (94%)**	**209 (68%)**	**290 (93%)**
BU4-DF	Anger	101	**64 (63%)**	61 (61%)	39 (39%)	54 (54%)
BU4-DF	Joy	100	49 (49%)	**97 (97%)**	88 (88%)	76 (76%)
BU4-DF	Surprise	101	72 (72%)	**86 (85%)**	57 (57%)	33 (33%)
BU4-DF	Sadness	101	21 (21%)	67 (67%)	25 (25%)	**77 (76%)**
BU4-DF	Disgust	101	9 (9%)	**73 (72%)**	72 (72%)	48 (48%)
BU4-DF	Fear	101	8 (8%)	21 (21%)	4 (4%)	**29 (29%)**
BU4-DF	**overall**	**605**	**223 (37%)**	**405 (70%)**	**285 (47%)**	**317 (52%)**
AFEW	Anger	157	**34 (22%)**	14 (9%)	8 (5%)	26 (17%)
AFEW	Joy	178	9 (9%)	49 (28%)	**97 (54%)**	**97 (54%)**
AFEW	Surprise	102	33 (33%)	**51 (50%)**	17 (17%)	10 (10%)
AFEW	Sadness	145	29 (29%)	34 (34%)	7 (7%)	**123 (85%)**
AFEW	Disgust	100	1 (1%)	12 (12%)	**46 (46%)**	15 (15%)
AFEW	Fear	101	6 (6%)	4 (4%)	0 (0%)	**8 (8%)**
AFEW	**overall**	**912**	**112 (12%)**	**164 (18%)**	**172 (19%)**	**279 (31%)**
ALL	**overall**	**1826**	**577 (32%)**	**861 (47%)**	**666 (36%)**	**886(49%)**

Table 4.7: Corresponding matches between labels for the six basic emotions and identified primary expressions for all three databases with the mean metric. The best correspondences are highlighted.

4. CK+ provides the benchmark with expected maximum results which can be used for comparison to all other datasets.

As CK+ provides an easy frontal setup, the positive results are not surprising. As the conditions (e.g. lighting, camera angle) of recorded videos from AFEW increase the challenge of detection, the results decrease. Interestingly, for AFEW, 'sadness' was correctly identified by CERT with twice the frequency of the other algorithms.

4.5. Identified Challenges for AFER Algorithms

In addition to the general performance, some important results from the benchmark should be highlighted and are therefore discussed in more detail in this section.

4.5.1. Uncertainty with Anger and Fear

While the recognition rate for ’fear’ and ’anger’ on videos from CK+ was usually high (overall ∼84%) for all algorithms except Affectiva as seen in Table 4.7[11], for most videos from AFEW, it dropped constantly to nearly zero (maximum 8%). Instead of recognising ’fear’, AFEW videos often seemed to misinterpreted as ’surprise’ and ’sadness’.

The same tendency can be observed with ’anger’, but the results are not as clear as with ’fear’. While the detection rate of ’fear’ decreases to half the rate of chance (8%), the rate of anger was still above this (22%). The lower recognition rate of ’fear’ has been reported in the literature, Valstar et al. (2011a) for instance mentioned the same outcomes in the preparation the baseline for their facial expression recognition and analysis challenge. The poor results of Affectiva for ’fear’ and ’anger’ were also reported by Stöckli et al. (2017) for their benchmark using different databases. It was also noted by Lewinski, Uyl, and Butler (2014) for their benchmark of FaceReader.

As Stöckli et al. (2017) mentioned, the categorisation of ’fear’ and ’anger’ is also difficult for a human observer. According to Shan, Gong, and McOwan (2005), confusing ’anger’, ’fear’, ’sadness’ and ’disgust’ is a common case described in literature.

[11] Affectiva only detected 8% of the ’fear’ files.

A possible solution to overcome this limitation is to cluster emotions as explained in detail in Chapter 5.

4.5.2. More Natural Recording Setup: Difficulties of Analysis

Out of the 912 processed videos from AFEW labelled with the six basic emotions, for 156, no usable result was generated by any of the four AFER systems when the mean metric was applied (226 with *meanp5*, 229 with *binaryp5*). The average rate of no detectable expressions varied between 10% ('disgust') and 17% ('fear').

The analysis revealed that InSight could not recognise a face in 129 videos. Affectiva was unable to identify a face in 104 of the files, while 16 videos resulted in face detection for only a small sequence of images. Compared to Emotient as successor of CERT which failed to recognize in 129 videos, CERT could detect faces in all but 17 files. The intersection of these videos led to the 156 files mentioned above.

Two reasons why several videos could not be analysed by the AFER algorithms are plausible: no face was detected or the AFER algorithms delivered very low values, so that the mean value was rounded to zero (if the face was only recognised for a few frames, the latter is often the case).

Investigation of the conditions that made detection difficult led to the following results: videos often contained faces that were partially covered by long hair, beards, glasses, strong shadows, or featured head positions at extreme angles in relation to the camera. Sometimes, the faces appeared to be very small due to their distance to the camera. Reflections in the background also seemed to interfere with the recognition of faces; some videos included bright – but not directly dazzling – background lights. As, for example, the Viola-Jones (Viola and

Jones 2001) method is based on edge detection by strong contrasts (haar-like features), this is a possible explanation for the observed behaviour.

The problematic conditions are also mentioned in the related literature: ‘The main challenges in automatic affect recognition are head-pose variations, illumination variations, registration errors, occlusions and identity bias. [...] Illumination variations can be problematic even under constant illumination due to head movements.’ (Sariyanidi, Gunes, and Cavallaro 2015)

Because of these conditions, the AFEW videos are particularly challenging as they closely simulate ’real life’ environments.

4.5.3. Handling of Neutral Expressions

As defined by Lewinski (2015), ‘a neutral face should indicate a lack of emotion’. The handling of neutral expressions is a case for special consideration, since each AFER algorithm handles it very differently. One possibility is to train separate machine-learning instances only for the detection of neutral faces. This is the case for the Emotient and CERT algorithms, in which all different expressions are built as separate SVMs. Another possibility is to calculate neutrality as ’1.0 minus all other emotions’. In the case of InSight, 1.0 is the maximum value of the probabilities for the sum of all expressions (illustrated in Figure 4.7). Affectiva does not explicitly provide the neutral probability value, but it may be computed with a similar method as InSight: 1.0 minus all other probabilities.

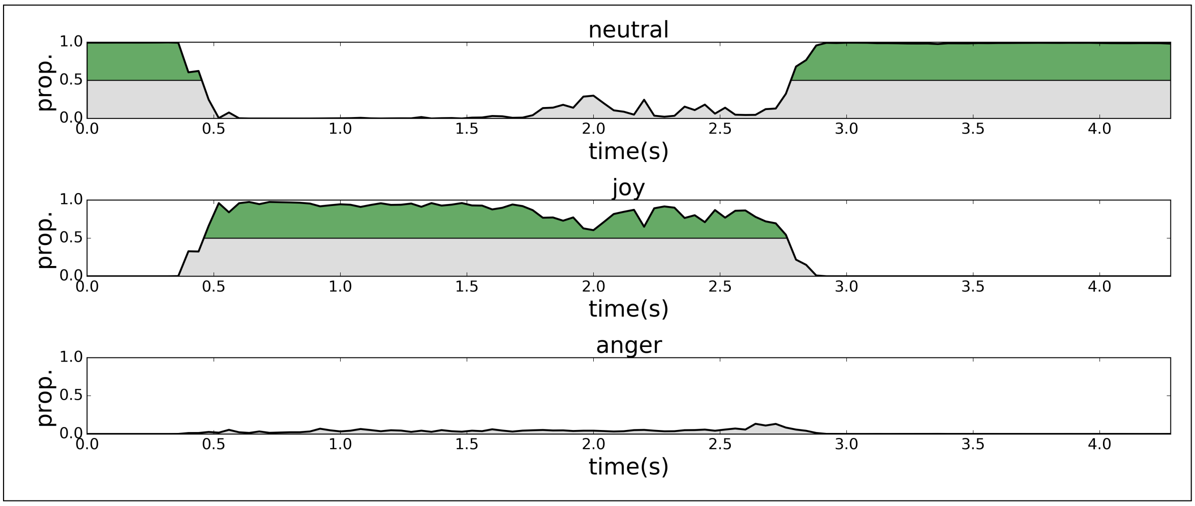


Figure 4.7: This graph displays the handling of the 'neutral' expression by the InSight AFER algorithm: the value of 'neutral' is 1.0 minus the sum of all other emotions. The probabilities of 'sadness', 'disgust', 'surprise' and 'fear' are omitted because the values are close to zero.

4.6. Summary and Discussion of Benchmarking Results

4.6.1. Summary of Chapter

In this chapter, the evaluation approach for determining a suitable AFER solution for the EmotionBike was presented. Starting with the discussion of literature related to testing and benchmarking of AFER systems, databases and the challenging task to find the primary expression in a video sequence. With the background of this review, an metric-based evaluation method for comparing AFER algorithms based on three labeled databases which share common characteristics with the cockpit-scenario of the EmotionBike has been developed. This approach was chosen, since no labeled database for the cockpit-scenario was available. Four state-of-the art AFER algorithms were benchmarked with the three metrics and the results were presented, revealing the accuracy of the approaches and challenges identified in the results, especially uncertainty in the interpretation of negative emotions and the recognition of faces in environments with challenging view angles and lighting conditions.

4.6.2. Discussion of Benchmarking Results

Robust AFER is critical for the realisation of affective applications. Although the problem of identifying expressions for the six basic emotions may be considered to be solved (Gunes and Hung 2016; Sariyanidi, Gunes, and Cavallaro 2015), this only seems to be true for the CK+ results. However, even this may be biased when CK+ was applied to the training of the systems.

As the setup conditions become more difficult (BU-4DF to AFEW), differences of performance become more visible. For AFEW in particular, accurately labelling videos that provide multiple facial expressions seems as difficult for human observers as automatic classification does for machines. Yet, besides the limitations of data provided by AFEW, it seems closer to practical application with non-desktop scenarios and is therefore beneficial.

The ’joy’ expression performed best in terms of recognition results over all data-sets with 100% for CK+, 97% for BU-4DF and 54% for AFEW. For CK+ and BU-4DF these results are similar to the results of ∼90%, when frontal facial expressions were categorised by human observers in cross-cultural settings (Russell 1994). This is not surprising, since ’joy’ brings about significant changes in the facial expression (e.g. open mouth or corners of the mouth pointing upwards). Although a correct categorisation of ∼50% over all datasets seems low, it is still three times the rate of pure chance with six output channels.

Information about the internal operations of AFER systems must be known by developers and researchers in order to understand the results and respond appropriately; for example, how 'neutrality' is handled and whether the algorithm applies several independent state machines for each expression (e.g. Emotient) or a *one vs all* processing scheme (e.g. Insight). Benchmarking enables this knowledge to be obtained for trained systems based on machine-learning, which are all black-boxes by design.

On the base of this benchmark, the AFER algorithm for the EmotionBike was selected. The best two candidates from the result Table 4.7 are CERT, with an overall accuracy of 49% and Emotient with 47%. Since CERT does not enable interactivity due its limitation of batch processing video data, Emotient was chosen as primary source for AFER in the EmotionBike, as it is capable of online processing with a networked API. For the integration of Emotient into the EmotionBike Framework, an AFER component was developed as described in Section 3.5.2.

Since Emotient only offers an accuracy of 47% across all databases, the question arises as to how and whether this result may be improved. The next chapter offers an approach based on the dimensionality of the algorithms output[12] with grouping of expressions.

[12] refer to Section 4.4.1 for an example.

5. Enhancing Facial Expression Recognition Robustness with Grouping

Benchmark on state-of-the-art AFER algorithms presented in Chapter 4 revealed, that there is problem with the accuracy of the algorithms when classifying facial expressions. This is on the one hand caused by the demanding environments (e.g. camera view angle and lighting conditions) and on the other hand by the wrong classification of facial expressions. The confusion in the interpretation of facial expressions which also occurs among humans poses a sever problem for AFER[1]. Two main causes for this are, firstly the improper interpretation by the AFER algorithms trained into the system by falsely labeled training data, especially for closely related emotions (e.g. disgust and contempt). Secondly, this may caused by the missing context in which an expression occurred, such as false interpretation of a smile as 'joy' occuring in an frustratrion provoking scenario[2].

This chapter presents a novel approach by **enhancing existing AFER systems overall**

[1] refer to Section 2.6.7 and Section 4.5.1.

[2] Smiles often occur during natural frustration (D'Mello and Kory 2015).

accuracy with post-classification application-specific grouping (Bernin et al. 2017) on the example of the benchmark results presented in Chapter 4 and experimental data of the EmotionBike experiments.

5.1. Introduction

For affective application development, it is not necessarily the most important which emotion theory or model is generally more advanced, but which provides the best performance in relation to the application's meaningful categories.

Many algorithms that have been developed to recognise emotions in facial expressions use discrete emotion models such as the six basic emotions of anger, surprise, joy, fear, disgust and sadness (described in Section 2.3.3). There are also different numbers of discrete emotions (from two to 20) proposed in the literature, but the underlying problem remains.

According to basic emotion theories, discrete emotions are mutually exclusive, and there are no mixed emotions such as surprise combined with fear[3]. However, in practical implementations, AFER algorithms are often designed as set of multiple independently trained machine learning instances (e.g. CERT and Emotient). In this case, the AFER system does not prevent the presence of mixed emotions, and conflicting output is a possible result (as described in Section 4.2.3).

This problem of **emotions being *non-exclusive*** has two main sources: firstly, fine **granularity of the underlying emotional model may not be robust enough** for practical solutions because the differences between the emotions (or at least their corresponding facial expressions) are too minor to be detected. As an example, if contempt and disgust are both part

[3] Although this may be questionable, it is still a basic concept of the underlying model of basic emotions.

of a model[4], they may be distinguished in theory, but the resulting facial expressions are quite similar and difficult to separate (Aleman and Swart 2008). Secondly, machine learning instances are trained with human-categorised data as ’ground truth’. If the human observers are not able to distinguish consistently between certain expressions, this **error is trained into the system** and also occurs in the test data.

The database-based performance analysis of four AFER systems described in Chapter 4 revealed, that **false categorisation is a common problem** in AFER systems (e.g. the confusion of ’fear’ and ’anger’ in Section 4.5.1). Often, this uncertainty occurs with closely related emotions and their expressions. In addition, if the number of possible categories rises, the uncertainty in the results may also increase (as described in Section 5.3).

One possible solution to overcome this limitation of AFER systems is based on the idea of grouping facial expressions. There are two ways to achieve this: firstly, by building groups and therefore reduce the dimensions of the emotion model before building and training an AFER system. Secondly, by relying on existing (but imprecise) AFER solutions and grouping the output of these systems to reduce the effect of uncertainty from the trained ML instances induced with incorrect recognition of similar expressions.

[4] as provided by Emotient

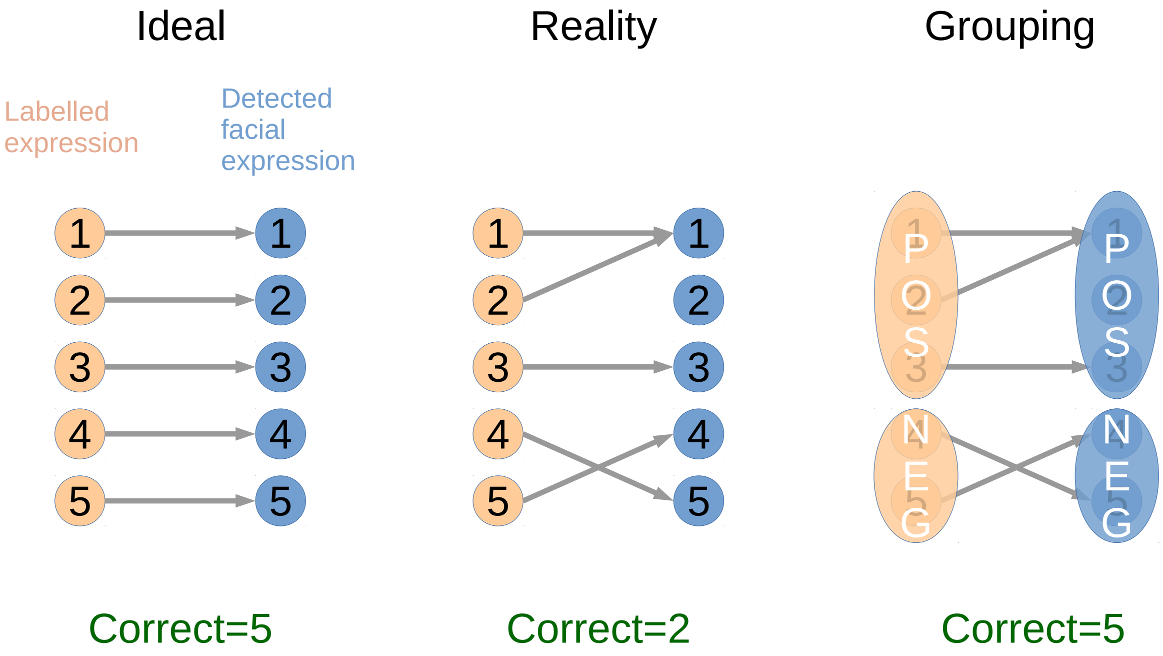


Figure 5.1: The basic concept of grouping expressions: with an *ideal* emotion theory and detection, labels of image data and categorisation of AFER systems would match one-to-one. In *reality*, matches are often lower due to incorrect categorisation. *Grouping* expressions on both sides can maximise the matches. In this example, 'positive' and 'negative' are grouping within the five expressions, which increases the rate of correct classifications.

This second approach is depicted in Figure 5.1: in an ideal world, categorizations of human observers and AFER algorithms are always accurate, allowing a correct one-to-one match from labelled to detected expression. In reality, categorisation includes the aforementioned uncertainty. Tailored grouping for the relevant categories of applications can reduce the effect of incorrect classification. Since **this work focuses on the integration of existing AFER solutions into an affective application and not on the development of a new AFER system, the second approach of grouping the results was applied.**

The approach presented here improves robustness by tailoring the emotion categories (and thus reduce the dimensions and complexity) to the specific application context using post-processing grouping. This step is executed after the classification for the primary expression (see Chapter 4) is performed. Figure 5.2 depicts the corresponding pipeline.

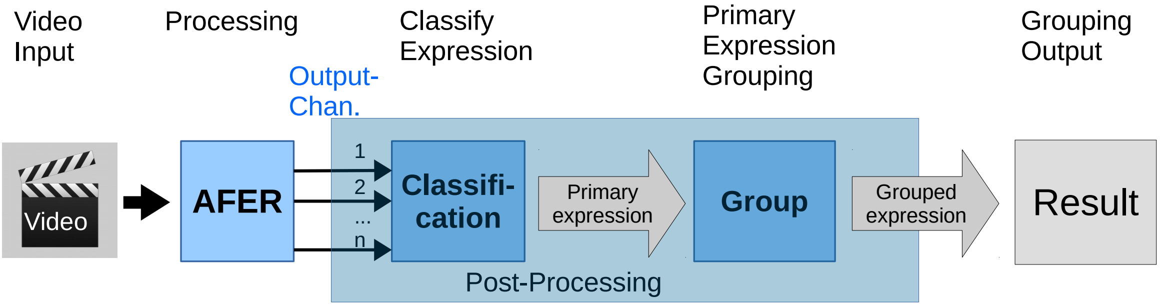


Figure 5.2: Pipeline for primary expression grouping: the output of AFER channels is classified (as described in Chapter 4) in search of the dominant emotion. The expression identified is grouped using a grouping table.

5.2. Background

The following section describes literature related to the grouping of emotions, facial expressions and context-aware adaption of emotion models.

5.2.1. Grouping Emotion and Expression

Grouping of emotions and their expression has a long tradition in theories of emotion. It was mainly applied in the creation of new emotional models, their application to facial expressions or to increase reliability of the emotion categories that form a model.

Creating Emotion Theory and Models

According to Becker-Asano (2008), grouping of natural language terms that describe emotions has been a common technique for the generation of broader clusters in emotion theory. For example, Ortony, Clore, and Collins (1990) created ‘representative groups or clusters’ forming six clusters for the OCC-Model (see Section 2.3.4).

A similar approach to generate an emotional model by building broader clusters was described by Russell (1994) when explaining the method of Woodworth (Woodworth and Schlosberg 1954):

> Woodworth first grouped together synonyms and words for closely related emotions (e.g., wonder, astonishment, amazement, and surprise). He then joined the groups into even broader clusters (e.g., joining the love, happiness, and mirth groups into one cluster, and joining the fear and suffering groups into another). Scoring any response within the resulting broad cluster as correct eliminated much of the disagreement. (Russell (1994))

Forced List Grouping of Expressions

Forcing facial expressions into a constrained list of possible emotions has been a common technique for labelling data and creating AFER algorithms since the (six) basic emotions were declared 'universal' (for a detailed discussion, refer to Section 2.3.3). Under the premise that there are more than these six possibilities, this can also be interpreted as a form of (forced) grouping of potential broader expressions.

Grouping for Reliability

Li et al. (2013) grouped basic emotions in three output classes: *positive* (happy), *negative* (disgust, fear, anger, sadness) and *surprise*, (surprise) while recording their database for short facial expressions (also called microexpressions). The participants were provoked with various videos and had to answer a self-report that revealed problems in distinguishing between negative emotions for the probands. The differentiation of the four negative emotions (sad, disgust, anger and fear) was difficult, although the videos were initially selected with separate emotions in mind. Grouping of the negative expressions improved reliability (Li et al. 2013).

Zeng et al. (2007) created a positive/negative cluster with corresponding AUs to classify their multimodal (audio/video) spontaneous emotional expressions recorded in a human conversation setting. Focussing on these two groups of emotions 'can be used as a strategy to improve the quality of interface in HCI' (Zeng et al. 2007).

5.2.2. Context-aware Expression Recognition

Context-aware modelling of emotions is often applied in the area of implementing computational agents' emotions (Marsella, Gratch, and Petta 2010). For the recognition of emotional expressions, the context in which these affective computer techniques are applied can have a significant influence on the underlying computational model, since not all emotions occur equally frequently in all environments. Context in this sense is meant as a prior known application domain rather than a spontaneously changing dynamic environment, or as 'elements of the user's environment which the computer knows about' (Baldauf, Dustdar, and Rosenberg 2007).

According to Broekens, Bosse, and Marsella (2013) there are two ways to incorporate the specific application context: modelling emotions with a domain-specific or a domain-independent approach. As an example, they describe 'modelling emotion in teaching situations versus modelling appraisal and applying the model to teaching situations' (Broekens, Bosse, and Marsella 2013). Unfortunately, no standard protocols or procedures were defined for evaluation scenarios or application-specific emotion models.

Several domain-specific emotional modelling systems have been proposed in the area of learning/tutoring. For example, Rodrigo and Baker (2011) applied a model containing confusion, delight, engaged concentration, frustration, surprise and neutral to evaluate their automatic tutor and an educational game. Grafsgaard et al. (2013) applied their online learning-specific model based on engagement, frustration and confusion. For the classification, they relied on AU detection with CERT.

Jack, Garrod, and Schyns (2014) presented their research on the time-based segmentation of the human perception of facial expressions with a focus on perceiving ’danger’. They grouped domain-specific expressions referring to basic emotions with fear/surprise into ’fast-approaching danger’ and disgust/anger into ’stationary danger’. Although stated, that these four danger-approaching expressions ‘may be a from a simpler system of communication in early man developed to subserve developing social interaction needs’ (Jack, Garrod, and Schyns 2014) and though a predecessor of basic emotion theory, their grouping might also be applied to AFER output post-processing.

Studies that combine context and the human interpretation of facial expressions have found that there is a mutual connection between facial expression and known context:

> when particular stimulus contexts were created, the *contempt* expression was judged as *disgust*, the *surprise* expression was judged as a *surprise-fear* blend, and the *anger* expression was judged as *sad*. Also expressions claimed to be *neutral* have been judged to be *happy* when seen embedded in one stimulus set and as *sad* when seen embedded in another. (Russell (1994))

While this may indicate problems with classification by human observers, it also reveals that certain expressions may have different meanings in different contexts.

In the following paragraphs, the method of applying a standard emotion model to a domain specific context (Broekens, Bosse, and Marsella 2013) is enhanced by utilising the novel method of context-specific post-processing.

5.3. Challenges of AFER in Applications

The following challenges emerge from the analysis of the performance evaluation described in Chapter 4, particularly concerning given conditionality of the results by the number of expressions recognised by the AFER algorithms.

5.3.1. Different Number of Expressions

The selection of facial expression channels from the AFER output has a significant effect on the results for the detection of primary expressions. Figure 5.3 depicts an example output with six and nine channels from analysis with Emotient. The output displays conflicting interpretations depending on the utilised expression channels: when only the output of the six basic emotions is analysed, the outcome can be correctly classified as ’anger’. Applying more expressions results in a misinterpretation as ’confusion’ or ’frustration’, as these start to increase earlier.

5.3.2. More than Six Expressions: Confusing the Matrix

In addition to the single example depicted in Figure 5.3, this behaviour is also displayed in larger samples. The confusion matrix in Figure 5.4 depicts this for the expressions of the

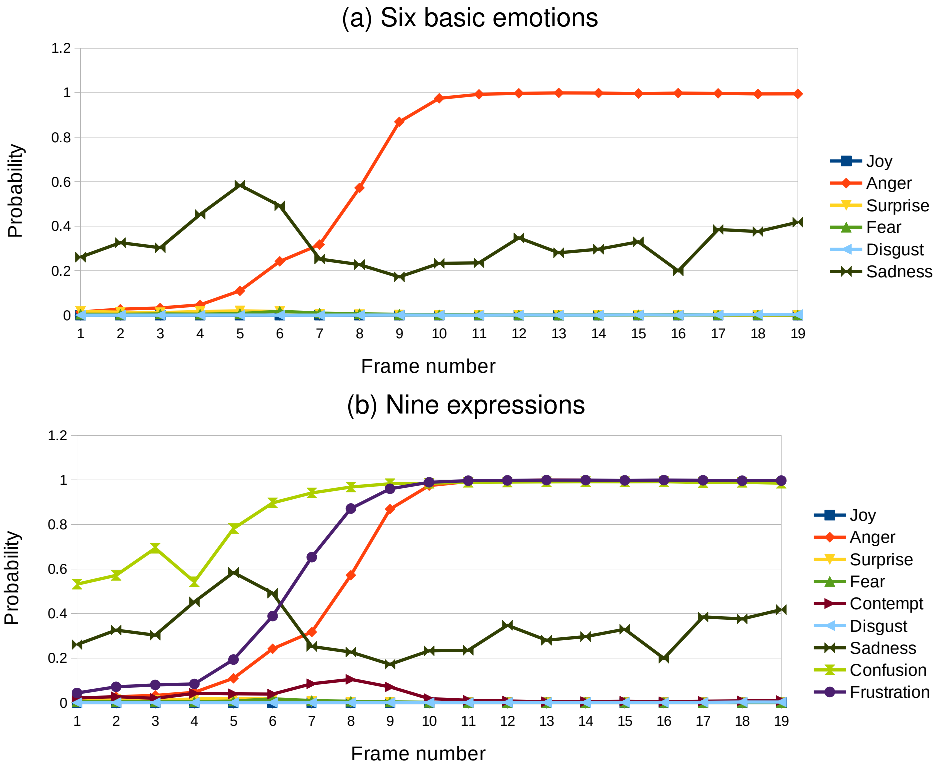


Figure 5.3: Sample output of subject S010 from CK+ database labelled with 'anger'. The video has been analysed with the Emotient AFER algorithm. (a) displays the results for six basic emotions, while (b) also depicts the additional expressions 'frustration', 'contempt' and 'confusion'. In (b), 'anger' is present, but 'confusion' and 'frustration' begin to increase earlier and they therefore have higher average values.

six basic emotions and displays high rates of agreement with the CK+ database. However, if the output of other facial expressions – such as 'contempt', 'confusion', 'frustration', which are available when using algorithms such as Emotient – is included, the results are different; 'confusion' is recognised as the most common expression in nearly 69% of the videos. The second most common expression is 'frustration', with 20% and only after that does the expression of 'anger' follow in about 4% of cases. When using the six basic emotions as the only possible target emotions, 'anger' was classified as the primary expression with a rate of 93%.

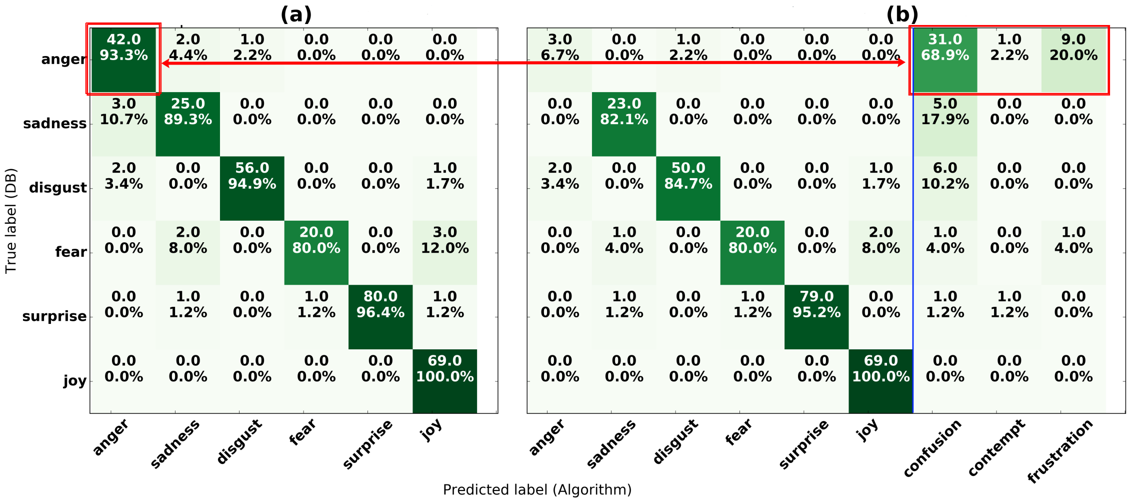
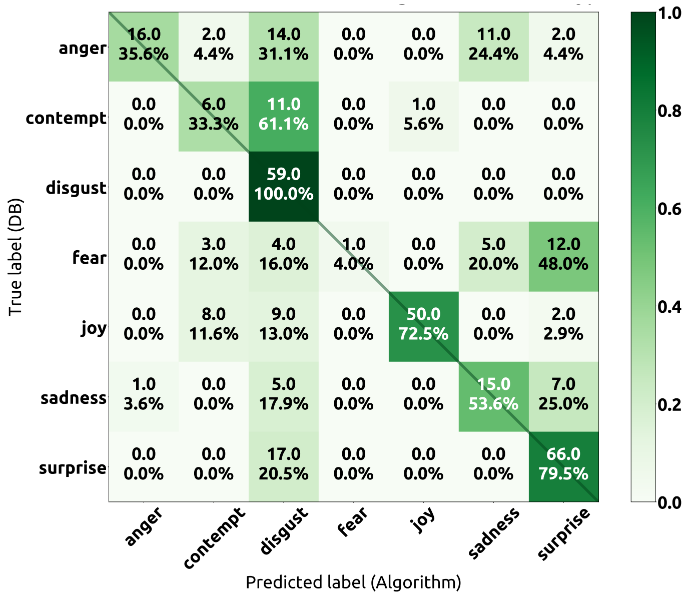


Figure 5.4: Examples of confusion matrices for Emotient with data from the CK+ database analysed. (a) displays the natrix with only the six basic emotions. In (b), three additional expressions provided by Emotient are included. The first row displays, that adding these additional expressions leads to a massive shift from (a) 'anger' to (b) 'confusion' and 'frustration' leading to a wrong categorisation because the videos are labelled as 'anger' in the database.

Two possible explanations seem plausible: the videos could have been initially mislabelled when classified by human observers[5], or the output signal of the AFER algorithms independent channels could have led to a misinterpretation. An example of the latter case is displayed in Figure 5.3 where the channels for 'confusion' and 'frustration' demonstrate the corresponding SVMs' faster reactions of to the input.

Figure 5.5 displays the confusion matrix for Affectiva. It contains the videos labelled as six basic emotions together with 'contempt' from the CK+ database. When analysing the matrix, two results are obvious. Firstly, the weakness in the recognition of 'fear', as described in Section 4.5.1, becomes obvious again: most 'fear' videos are recognised as 'surprise'. Secondly, videos labelled as 'contempt' are often interpreted as 'disgust'.

[5] The 'accuracy problem' for labelled datasets as described in Section 6.2.2



	anger	contempt	disgust	fear	joy	sadness	surprise
anger	16.0 35.6%	2.0 4.4%	14.0 31.1%	0.0 0.0%	0.0 0.0%	11.0 24.4%	2.0 4.4%
contempt	0.0 0.0%	6.0 33.3%	11.0 61.1%	0.0 0.0%	1.0 5.6%	0.0 0.0%	0.0 0.0%
disgust	0.0 0.0%	0.0 0.0%	59.0 100.0%	0.0 0.0%	0.0 0.0%	0.0 0.0%	0.0 0.0%
fear	0.0 0.0%	3.0 12.0%	4.0 16.0%	1.0 4.0%	0.0 0.0%	5.0 20.0%	12.0 48.0%
joy	0.0 0.0%	8.0 11.6%	9.0 13.0%	0.0 0.0%	50.0 72.5%	0.0 0.0%	2.0 2.9%
sadness	1.0 3.6%	0.0 0.0%	5.0 17.9%	0.0 0.0%	0.0 0.0%	15.0 53.6%	7.0 25.0%
surprise	0.0 0.0%	0.0 0.0%	17.0 20.5%	0.0 0.0%	0.0 0.0%	0.0 0.0%	66.0 79.5%

Figure 5.5: Confusion matrix for Affectiva and CK+, labelled with the six basic emotions and 'contempt'. The expression 'contempt' is often misinterpreted as 'disgust' and 'fear' as 'surprise'.

While in the first observation, the 'fear-weakness' might be caused by a systematic problem with Affectiva[6], the second observation is an excellent example of potential grouping of expressions, as they are closely related in emotion theory: 'Contempt and disgust are closely related emotions, that have been considered to be 'moral emotions'. [...] Both emotions involve rejection, disapproval and a degree of hostility' (Aleman and Swart 2008).

The results depicted in Figure 5.5 are an example of how a confusion matrix can form the base of a structured analysis for finding candidates used in application-tailored AFER grouping.

[6] also described by Stöckli et al. (2017).

5.4. Tailoring AFER Results to a Specific Application

The following post-processing methods may be applied to affective applications to increase the robustness of AFER-based interaction. They were evaluated using the benchmarking results from Chapter 4 and the EmotionBike experiments (Chapter 3). The methods consist of excluding or changing the weight of individual expressions and examples for the grouping of expressions.

5.4.1. Excluding Irrelevant Expressions

As noticed in Section 5.3.2, the number of expressions searched for in the AFER output, can lead to misinterpretation. A question arises from this is, what emotions are irrelevant to the scenario? The irrelevant expressions should be excluded from the grouped output to improve recognition of the relevant ones to avoid effects of misinterpretation.

5.4.2. Grouping of Similar Expressions

Grouping of Anger, Frustration and Confusion

In the example depicted in Figure 5.3, grouping ’anger’, ’frustration’ and ’confusion’ as negative emotions would lead to a more robust classification. As mentioned in the related work (Section 5.2.1), this is a feasible solution.

Grouping of Smile and Joy

Figure 5.6 visualises the output of all algorithms for a video from AFEW and the expression of joy from the CK+ database. Affectiva seems to apply a stricter definition of ’joy’ than for the 'smile' facial expression, which results in reduced detection sensitivity. The grouping of the two emotions 'joy' and 'smile' increases the detection rate.

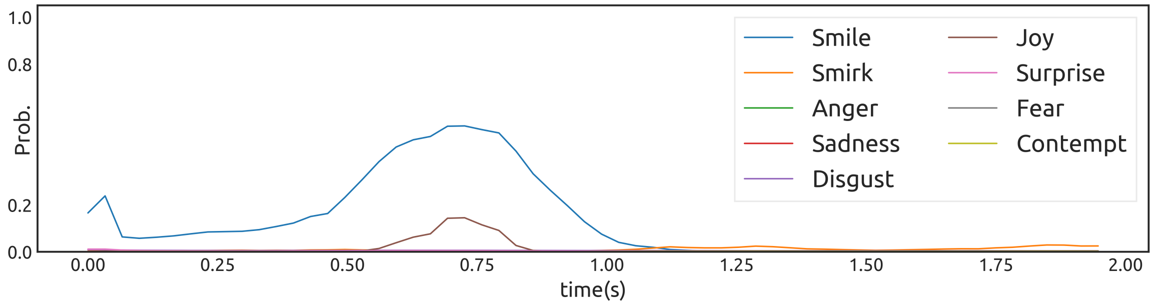


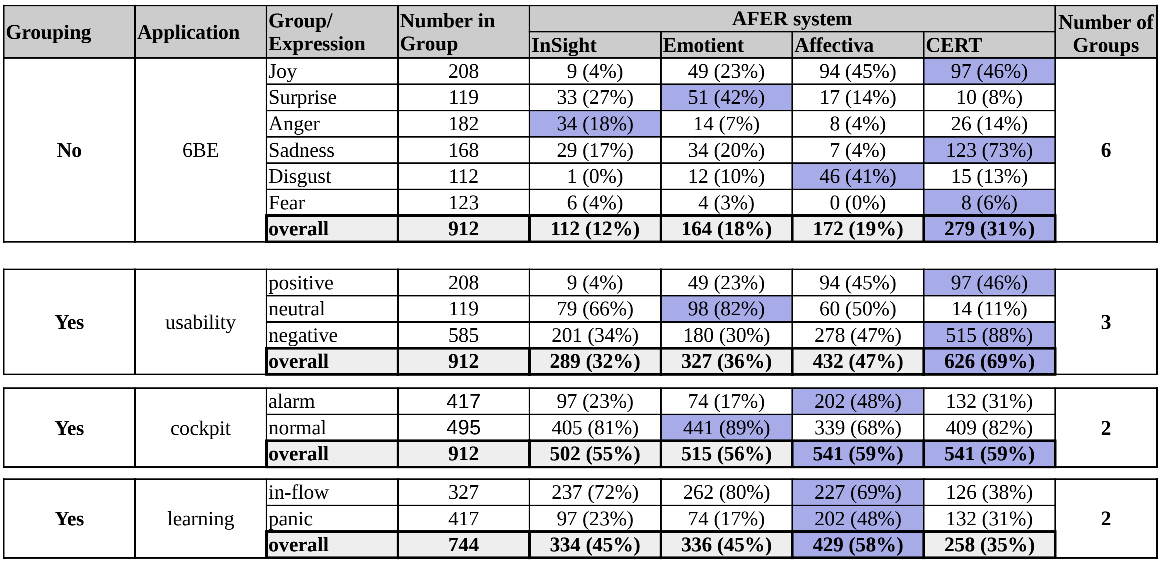
Figure 5.6: An example of an Affectiva output of all expressions for videos labelled 'joy' from the AFEW dataset. If grouping of 'smile' and 'joy' is applied, the output is correctly classified.

5.4.3. Boost Robustness by Application-specific Output Grouping

To introduce the new method of application-specific output grouping of expressions, three examples of possible applications are explored. This method enables the integration of application-specific expression groups and the exclusion of emotional states that are irrelevant to the particular context of usage in one step.

Application dependent emotional states (such as engagement or frustration) have been proposed in emotion theory, as described in Section 5.2.2. However, in contrast to the proposed change of emotional models before the algorithms are trained, the application-specific groups the outputs of AFER algorithms and includes handling the case of *no expression detected*.

To illustrate the application-specific grouping, three scenarios have been defined and are evaluated using the classification of AFEW videos as depicted in Table 5.1.



Grouping	Application	Group/ Expression	Number in Group	AFER system				Number of Groups
				InSight	Emotient	Affectiva	CERT	
No	6BE	Joy	208	9 (4%)	49 (23%)	94 (45%)	97 (46%)	**6**
		Surprise	119	33 (27%)	51 (42%)	17 (14%)	10 (8%)	
		Anger	182	34 (18%)	14 (7%)	8 (4%)	26 (14%)	
		Sadness	168	29 (17%)	34 (20%)	7 (4%)	123 (73%)	
		Disgust	112	1 (0%)	12 (10%)	46 (41%)	15 (13%)	
		Fear	123	6 (4%)	4 (3%)	0 (0%)	8 (6%)	
		overall	**912**	**112 (12%)**	**164 (18%)**	**172 (19%)**	**279 (31%)**	
Yes	usability	positive	208	9 (4%)	49 (23%)	94 (45%)	97 (46%)	**3**
		neutral	119	79 (66%)	98 (82%)	60 (50%)	14 (11%)	
		negative	585	201 (34%)	180 (30%)	278 (47%)	515 (88%)	
		overall	**912**	**289 (32%)**	**327 (36%)**	**432 (47%)**	**626 (69%)**	
Yes	cockpit	alarm	417	97 (23%)	74 (17%)	202 (48%)	132 (31%)	**2**
		normal	495	405 (81%)	441 (89%)	339 (68%)	409 (82%)	
		overall	**912**	**502 (55%)**	**515 (56%)**	**541 (59%)**	**541 (59%)**	
Yes	learning	in-flow	327	237 (72%)	262 (80%)	227 (69%)	126 (38%)	**2**
		panic	417	97 (23%)	74 (17%)	202 (48%)	132 (31%)	
		overall	**744**	**334 (45%)**	**336 (45%)**	**429 (58%)**	**258 (35%)**	

Table 5.1: Exemplary application-specific grouping for the six basic emotions (6BE). The detection rate for correct classification (accuracy) is shown in relation to the videos in AFEW labelled to one of the six basic emotions (912). The best result for each row is marked. The overall detection rate increases from 31% (6BE, without grouping) to 69% (usability, with grouping) for CERT based analysis when grouping of expressions is applied. This major difference demonstrates, that there is a problem with the accuracy of the AFER algorithms and/or the labeling of the databases, as otherwise, no major difference in overall recognition rates should occur.

Usability study: For a smart environment usability test of a new application, developers are curious about whether and when people display positive, neutral or negative expressions while testing the software. They group with the following classes: *neutral* means that ’no expression’ or ’surprise’ was detected, *negative* consists of ’sadness’, ’anger’, ’fear’ and ’disgust’ and the *positive* class contains the ’joy’.

Driver and Operator Assistance: In this scenario, the driver or operator should receive more automatic support in the event of negative stress (*alarm* class). Under normal conditions (*normal* class), control is left to the driver/operator. The *normal* class consists of

’joy’, ’surprise’, ’sadness’ and ’no expression’. ’Disgust’, ’fear’ and ’anger’ form the class of *alarm*.

Learning system: In an automatic and intelligent learning system, the state of the learners is to be monitored to determine whether they can understand the material or whether they are overwhelmed. For this purpose, the classes *in-flow* and *panic* are formed. *Panic* consists of ’fear’, ’anger’ and ’disgust’, *in-flow* of ’surprise’, ’joy’ and ’no expression’. The expression ’sadness’ is ignored because this emotion is not triggered expectedly and unrelated to learning.

To illustrate these scenarios with data, the analysis of videos from the AFEW database was selected. While these videos do not reflect the above scenarios in detail, they are the most challenging and have certainly not been applied to train any of the tested algorithms. The significantly increased detection rates for the grouping are depicted in Table 5.1 with a comparison to the six basic emotions. The recognition rates are at least doubled with the application-based approach; for example, from 19% (6BE) to 47% (usability), 59% (cockpit) and 58% (learning) for the Affectiva algorithm. As a result, grouping enables a much more stable detection and application-oriented reaction.

To define the default behaviour (fallback) in the sense of no emotion detected (’none’), it is possible to assign it to a class.

5.5. Grouping Results in the EmotionBike Context

As a practical example for grouping, data from the second series of experiments (refer to Section 3.4) were utilised to evaluate application-specific grouping. As the different levels were designed to provoke different emotions, grouping was applied on a per-level basis.

The results of two game levels, jump-scare/surprise and challenge/falling are presented here. In the jump-scare level, the probands experienced a sudden exposure to a group of monsters while riding through a dark forest (for details see Figure 6.10). This provoking event was randomly triggered for two or three times during the level.

In the challenge level, the task was to use a ski-jump between two islands to complete the level (as depicted in Figure 6.11). The probands often failed and fell off the island. Both levels are described in more detail in Section 6.4.2. For both levels, the recorded time at which events occurred was used to define a time window for evaluation.

Table 5.2 describes the different recognition rates of fear and surprise for the jump-scare event applying the AFER algorithms Emotient. The first occurrence of the monsters was perceived as most scary. In addition, the rates dropped with the number of times when the event occurred, since it was expected to happen again by the probands after first exposure. 92.5% of the probands stated that they felt surprised and frightened when the zombies occurred for the first time, when they were asked in the context of a self-assessment after the level.

For the jump-scare event, the expressions of ’fear’ and ’surprise’ were grouped, which **increased the detection rate by 12.5% for the first event and compared to the best single expression of fear which was originally detected in only 62.5% of the cases**.

Event Occurrence	FER Surprise	FER Fear	FER Combined	Improvement
1	**58.3%**	**62.5%**	**75.0%**	**12.5%**
2	12.5%	45.8%	50.0%	4.2%
3	13.6%	50.0%	50.0%	0.0%
Mean	**28.1%**	**52.8%**	**58.3%**	**5.6%**

Table 5.2: Grouping effects on jump-scare level results for one to three occurrences of the jump-scare event. The improvement is at a maximum for the first occurrence of the event, since it surprised the and therefore had the biggest effect.

As reacting with 'anger' instead of 'fear' to a potentially threat ('flight or fight' reaction), in a second step, 'anger' was added to the grouped expressions, as well as 'disgust' as there is also a known similarity of 'anger' and 'disgust' in the face (Susskind et al. 2007). Which is also consistent with the findings about perceiving danger described by Jack, Garrod, and Schyns (2014)[7].

With this extended grouping, the overall **detection rate increased to 90% for the first and second occurrence of the jump-scare event and 75% for the third occurrence** which additionally boosted the detection of the expected and reported emotional reaction to the provoking event to nearly the self-reported felt emotions (92.5%).

In the challenge level, 'frustration' and 'joy' were grouped, as smiling is often encountered in natural frustration (D'Mello and Kory 2015). The grouping resulted in an **increase of 4.7%: 95.3% for grouped detection instead of 90.6% for only 'joy'**. This is consistent with the self-reports of the probands as nearly all of them stated, that they were challenged by the level and felt frustration when failing to achieve the goal.

[7] refer to Section 5.2.1.

5.6. Summary and Discussion of Results

In this chapter, methods to improve AFER results based on the tailored grouping of algorithms output have been presented. In particular, the approach of application-specific grouping was examined based on extensive empirical experiments described in Chapter 4. Additionally, results for grouping expressions during the second EmotionBike experiment were presented, significantly improving detection results.

To tailor post-processing grouping, the follow three-step approach is recommended: firstly, decide potential grouping candidates in the context of the application. Secondly, select dataset-based benchmarking to check the potential for this grouping (both steps are described in Section 5.4.3). Thirdly, verify the grouping with application-specific data (see Section 5.5 for an example).

Although AFER systems are generalised utilities for the analysis of facial expressions, **tailored grouping** offers the possibility to significantly increase their accuracy for specific applications. This **is a simple but practical solution** to reduce the limitations of current AFER solutions caused by misinterpretation of the algorithms' results.

By using the methodology of application-specific grouping, the **shift of the recognised primary expression** (see Section 5.3.2) **becomes an advantage instead of a constraint**; due to the more broadly grouped recognition, the application-related detection (and thus response possibility) becomes more robust.

Grouping of expressions applies a domain-specific modification of generalised emotion model to an application, thus combining both proposed methods for including the applications context (Broekens, Bosse, and Marsella 2013) which was described in Section 5.2.2.

Grouping of expressions also enabled a significant increased accuracy of detecting the level-dependent emotion in the context of the EmotionBike, providing an increased robustness in detection. This increased robustness in detections is important for providing an adequate system response to the emotional state of the user.

6. Emotional Shift Analysis for Automatic Categorisation of Facial Reactions

While the previous chapter focussed on the dimensionality of AFER algorithms output to increase accuracy for applications such as the EmotionBike, this chapter focusses on another approach, evaluating the timing of facial expressions. The evaluation and modelling of timing and its dynamics are still considered as open challenges for AFER algorithms, as described in Section 2.6.8.

6.1. Introduction

One strategy to increase reliability and robustness of event-based emotional provocation is to enable plausibility checks and subject independent response patterns. The approach presented in this chapter develops and evaluates post-processing methods to facilitate the detection of *emotional shift*. Emotional shift is defined as a *fast transition of emotional facial* *expressions*[1] *within a time window for analysis* (Bernin et al. 2018). This shift usually consists of the fast offset (falling edge) of one expression and a corresponding fast onset (rising edge) of another expression. This enables to utilise the timing information inside the analysis-window, rather than to analyse a single channel of expression. This is particularly relevant for interactive systems with provoke-response patterns, such as the EmotionBike.

The challenges of integrating the output of AFER software into functional interactive applications are not only related to the identification of the current dominant or primary expression utilised in the post-processing of the AFER algorithm's output, as described in Chapter 4. **Due to the complex and subjective nature of facial expressions, the detected dominant expression caused by an event varies between subjects and even AFER algorithms which makes one-to-one mapping from specific expressions to reactions difficult**. Figure 6.1 depicts typical data of three different channels.

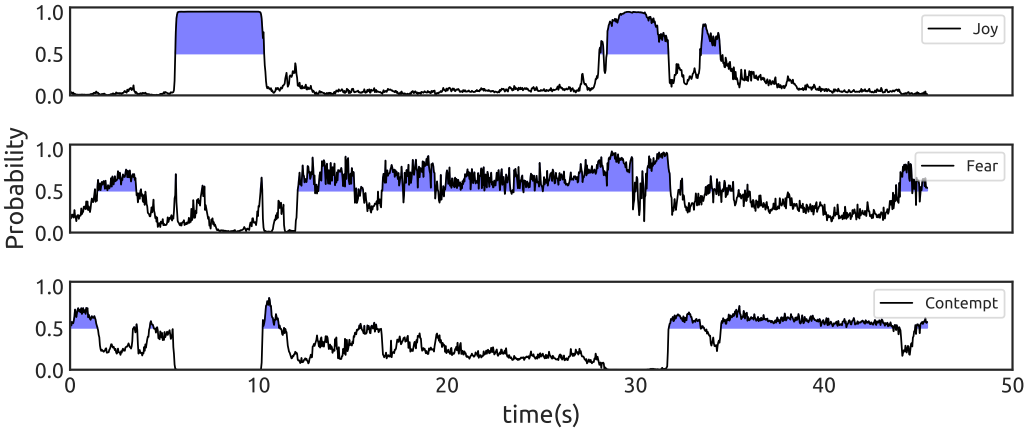


Figure 6.1: This figure depicts three example output channels (complete 12 channels are described in Section 6.4.2) of the AFER algorithm Emotient and displays data from the EmotionBike experiments. Processing and interpretation of this output can be considered a multi-channel signal processing problem since the different channels for each expression are usually independent of each other.

[1] Facial expressions linked to emotions as in basic emotion theory.

The goal is to determine whether the user perceives an event – based on their reaction and emotional expressions – by applying the recognition of emotional shift, which is independent of the actual expressions involved.

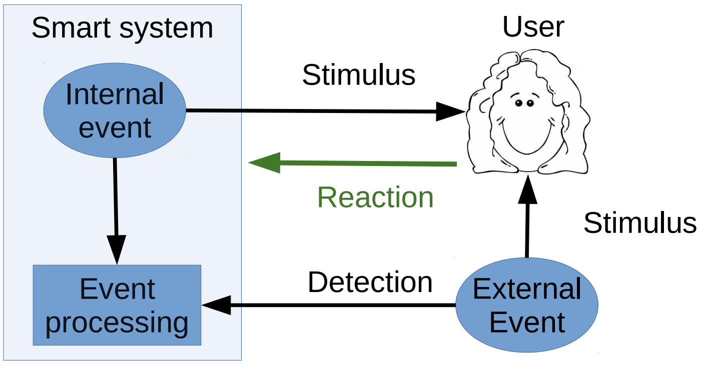


Figure 6.2: Overview of a smart system that is capable of processing internal and external events and the reaction of the user.

To implement the emotional shift analysis, three different algorithmic methods developed and evaluated are introduced: **fixed window mean bisection (FWMB), pattern matching peak (PMP), and Bayesian change point detection (CP)**. With these three approaches, it is possible to transform the data in the analysis window into simple, two-digit binary patterns (e.g. '10'[2]), which enables easy shift detection in a following step. The evaluation of these methods is performed on the experimental data from the interactive cycling simulation EmotionBike described in Chapter 3, in which participants were provoked with game elements and their facial expressions were recorded. In the evaluation, the outputs of the three approaches are compared to an expert human observer's categorisation of the signals in the analysis window.

In this chapter, a comparison of approaches for the automatic recognition of responses from the output of AFER algorithms is explained and a benchmark for this post-processing step is

[2] refer to Table 6.1.

provided. The post-processing methods are demonstrated with the state-of-the-art Emotient AFER system.

Although this work is based on an emotion provoking exergame, the findings may be applied to any affective scenario in which an application setting provides internal or external events to fix search windows that occur in smart and assisting environments.

6.2. Related Work for this Chapter

Complementary to the related works in Chapter 2, specific literature is discussed here, including: examples of application areas of emotional shift, and metrics for measuring the agreement or reliability of categorisations.

6.2.1. Application: Driver and Operator Assistance

One application domain for emotional shift is systems, such as the EmotionBike (Chapter 3), which apply an event-based emotion processing scheme. Figure 6.3 illustrates a generalised affective smart system that is able to process internal and external events. Internal events enable the opportunity to provoke emotional responses, while external events should be detected by the system to pinpoint user reactions to certain time windows.

A possible scenario in the area of 'cyber physical systems' (Lee 2008) that is event-aware and presupposes a user's reaction is a car-driver assistance system: after a potentially harmful external event (e.g. an obstacle in front of the vehicle) occurs, the choice between waiting for an appropriate reaction from the driver and initiating an automatic response is crucial. A shift in the driver's facial expressions is one indication to wait, whereas if no reaction is detected, the system could react at once.

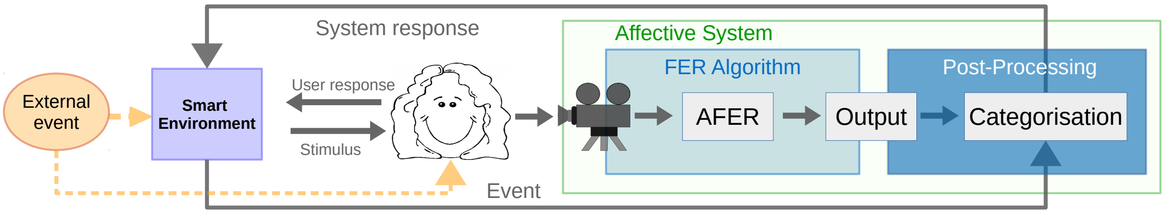


Figure 6.3: Example of a pipeline for an affective event-aware smart system for driver and operator assistance. A camera records the user's face and AFER algorithms analyse the facial expressions. Internal and external events define the position of time windows in the AFER algorithm's output for post-processing analysis, which is the focus for of chapter (figure from Bernin et al. (2018)).

Cockpit-based scenarios such as the NAVIEYES system (Mihai, Florin, and Gheorghe 2015) provide a lightweight architecture for a driver assistance system, that could benefit from facial expression shift detection as an additional input source to improve the detection of driver's intentions. Intelligent brake assistance may adapt there behaviour based on situational severity and driver attentiveness (McCall and Trivedi 2007) as 'driver distraction is a contributing factor to many crashes' (Ahlstrom, Kircher, and Kircher 2013).

6.2.2. Reliability and Inter-rater Agreement

In order to quantify and test the reliability of facial expression detection in the domain of affective computing it is important to note that there **is seldom a real ground truth for data**, as labelling of the videos is normally based on human observers and used as training input for automated approaches, such as when training with the CK+ database (Lucey et al. 2010). For the measurement of reliability between observers (algorithmic or human) the methods of inter-rater agreement are applicable.

According to Gendron and Barrett (2009), **inter-rater agreement is always an implicit part of labelling emotional facial expression databases**: 'in fact, the extent to which perceivers agree with one another in their judgements of emotion when looking at other people's faces (especially when perceiver and target are not from the same cultural context) is taken as an index of *accuracy* during *emotion recognition* rather than an index of *inter-rater agreement* during *emotion perception* '. Calvo and D'Mello (2010) added that calculating performance and reliability is a challenging task for which there is 'no objective gold standard', as 'emotions are notoriously fuzzy, ill defined, and possibly indeterminate'.

Applying the proposed analysis method of emotional shift to a series of facial expressions changes the scale of data from a probability of expression to data with a nominal two-digit scale (described in 6.3.1). Different approaches for the calculation of an inter-rater agreement coefficient for nominal scales are presented below:

Lewinski, Uyl, and Butler (2014) state that their inter-coder reliability agreement index is based on the work of Ekman (FACS Manual) and Wexler. This index is defined in Equation 6.1. Although applied in the context of AUs, it could easily be adopted for the categorisation of expressions.

$$ICRAI = \frac{NumAUsAgreed * 2}{TotalNumScored} \tag{6.1}$$

Equation 6.1: Inter-coder reliability agreement index (ICRAI) with NumAUsAgreed as the number of AUs that both coders agree upon, and TotalNumScored as the total number of AUs scored by the two coders.

An often-proposed enhancement over this simple coefficient – which only calculates the **agreement** ratio – is a correction of values by chance. As an example, Calvo and D'Mello (2010) suggested to apply Cohen's kappa (Cohen 1960).

Cohen's kappa (κ) is intended to calculate the **reliability of two raters**: the relative rate of

agreements is calculated between raters and the probability of chance agreement is applied to correct the values, as depicted in Equation 6.2.

$$\kappa = \frac{Po - Pe}{1 - Pe} \tag{6.2}$$

Equation 6.2: Calculation of Cohen's kappa with Po as agreement (often named 'accuracy') between raters, and Pe as the probability of agreement by chance.

For $\kappa = 1$, complete agreement between the raters can be assumed, while $\kappa = 0$ means total disagreement. Calvo and D'Mello (2010) suggested, that in the field of affective computing, the following scores for kappa apply: 'scores ranging from 0.4-0.6 are typically considered to be fair, 0.6-0.75 are good, and scores greater than 0.75 are excellent'.

Another established metric for measuring the inter-rater reliability is Krippendorff's alpha (Hayes and Krippendorff 2007; Krippendorff 2011). While usually applied to measure the reliability of a complete group of observers, it can also compare subgroups where the total **number of coders is greater than or equal to two**. Krippendorff's alpha can be **applied to nominal, ordinal and interval scaled data** (Hayes and Krippendorff 2007).

As depicted in Equation 6.3, Krippendorff's alpha does not rely on the *agreement* – as Cohen's kappa does – but on the **disagreement** between observers.

$$\alpha = 1 - \frac{Do}{De} \tag{6.3}$$

Equation 6.3: Calculation of Krippendorff's alpha with Do as the disagreement rate between raters, and De as the probability of disagreement by chance.

Krippendorff (2013) suggested, that values for true reliability should be above $\alpha >=$0.800, and values with $\alpha <$0.800 and $\alpha > 0.667$ *indicates* reliability. Values of $\alpha <$ 0.667 should be handled with care and should not normally be relied on.

One advantage of utilising Krippendorff's alpha is the ability to compare more than two reviewers, and the ability to calculate with missing data. The flexibility in the number of raters enables the possibility to apply the same coefficient for comparison to pairs of coders (e.g. algorithm 1 vs. human observer) as the total measurement (algorithm 1-3 and human observer). Because of this flexibility, Krippendorff's alpha was chosen for this work.

6.3. Developed Categorisation Algorithms

Three different algorithms for the categorisation of the output of AFER algorithms are presented here. All three were developed for the event-based interaction scheme of the EmotionBike setup. The main goal was to automate the categorisation of data, since this task was previously conducted partially manually.

The first algorithm, FWMB, is based on a binary search and mean-based comparison with a fixed window size. The second method, PMP, utilises a matched filter with a fixed size, which is a common approach in digital signal processing. The third algorithm, CP, is based on Bayesian changepoint detection and has[3] never been previously utilised in the AFER context.

6.3.1. Categorisation of AFER Output Data

Two symbols form the base of the categorisation as presented in Table 6.1. The symbols consist of binary digits and '?' for inconclusive data; '01', as an example, categorises a dataset in which a rising edge follows low values (see Fig 6.4).

[3] To the knowledge of the author

Category	Example Data
00	‾‾‾‾
01	___/‾‾
10	‾‾\___
11	‾‾‾‾‾
??	/‾\/---/

Table 6.1: Basic binary two-digit categorisation of the AFER output for the first and second half of the analysis window: ’00’ denotes that no expression could be detected, ’01’ a rising edge, ’10’ a falling edge, ’11’ a stable signal near to 1.0 and ’??’ that the data could not be categorised due to a noisy signal.

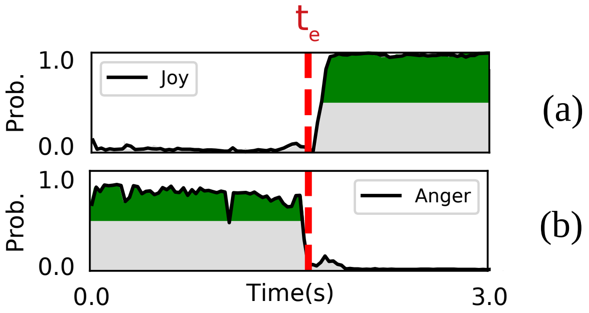


Figure 6.4: Example of an AFER algorithm output for emotional shift in two channels (expressions): (a) fast onset (’01’) of the ’joy’ expression and (b) corresponding offset (’10’) of ’anger’. The event position is marked by the red vertical line at t_e.

Two characters were chosen to reflect the analysis windows before and after an event, and the data was subdivided into these sub-windows. For categorisation, four approaches were applied: a human expert as an observer for comparison, and the three developed algorithms. To ensure comparability, the human observer had the same categorisation choices as the algorithms. Using the two symbols, detection of the *emotional shift* was an easy-to-implement second step: if ’01’ and ’10’ are both present for a window, a shift has been found.

6.3.2. Peak Detection

Peak detection is an essential part of all three algorithms to improve the categorisation results. For this work, the peak detection method by Billauer (2005) has been integrated with a minimum delta value of 0.25. The look-ahead was set to 1. As the rather small threshold of 0.25 results in in the detection of a number of false-positive minima and maxima, the outcome had to pass an additional threshold-filter with a value of 0.5. This method was superior to a peak detection delta of 0.5 itself, as prior tests demonstrated.

Preliminary testing also revealed, that the new peak detection method produced slightly better results with this dataset im comparison to the peak detection by Bergman (Negri and Vestri 2017) which had been utilised in previous work(e.g. Müller et al. (2015)).

6.3.3. Edge-Detection-Based Algorithms (CP, PMP)

Two of the developed algorithms (CP, PMP) utilise edge detection and share a common design by applying a multi-step approach as presented in Figure 6.5. Smoothing of the data, edge detection and categorisation of the edges are shared (marked blue) while the method for pre-processing of data and the thresholds for peak and edge detection (marked green) differ. Details about the steps are explained below.

Smoothing of Input Data

The input data is pre-processed with a modified, block-based (n=4) single-pass moving average filter: instead of applying the mean value to the block, a threshold check with t=0.5 is executed. If the average of the block is above the threshold, the maximum value of the block

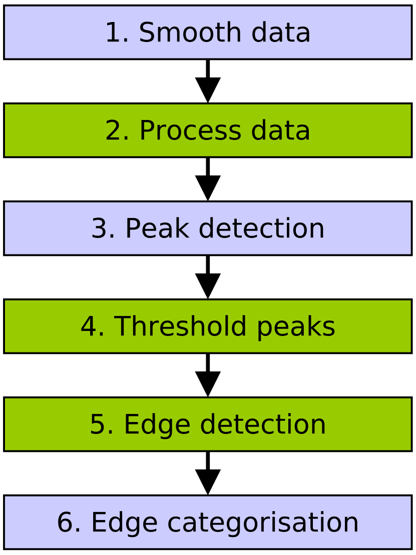


Figure 6.5: Basic steps for the edge-based CP and PMP algorithms. Identical steps that are implemented in both algorithms are coloured blue, while steps that differ are coloured green. The smoothed input data (1) is processed by either CP or PMP (2), providing the likelihood of a changepoint throughout the data (Figure 6.7 depicts an example). This likelihood is analysed in a two-step peak detection (3,4). When the threshold is exceeded, either a falling or rising edge was detected (5) and this edge is categorised (6). The final binary edge categorisation is determined under the conditions explained in Table 6.2.

is applied; otherwise, the minimum value is applied. This approach maximises the spread within the block data which improves the following categorisation and change detection.

Processing Data with Bayesian Change Point Detection (CP)

The CP method utilises Bayesian change point detection to identify rising and falling edges. This method is based on Bayesian inference for multiple change point problems as proposed by Fearnhead (2006). The improved version by Xuan and Murphy (2007) was applied to the developed algorithm with a *constant prior* of $1/len(data)$ and a *truncate value* of -20. These values had achieved the best results in preliminary tests on the data set. The original algorithm was improved by adding a smoothing preliminary step and an additional peak-detection of the algorithms' output as additional threshold, thus avoiding false-positive detection of change points. Section B.3.1 provides the source code of the algorithm.

Processing Data with Pattern Matching Peak (PMP)

The PMP algorithm utilises 1D cross-correlation (see Equation 6.4) as the base of a matched filter with a mask, which was proposed by Jaynes (2003). This simple mask is depicted in Figure 6.6 for a filter length of eight frames (l=8) before and after the peak. The filter length was deliberately selected as a multiple of 2, which enables the possibility to easily halve and double the length. The initial filter length chosen was l=16, which results in an overall length of $cl = 2 * l = 32$. This filter length of 32 was selected due to its proximity to the video frame rate of 30 fps, which is similar to a normal facial expression (>1s, according to Yan et al. (2013)). Different sizes of the filter with (l=8) and (l=24) were added, which designated the variants of the algorithm: PMP8, PMP16 and PMP24. Section B.2.1 provides the corresponding source code.

$$(f \; * \; g)(t) \overset{def}{=} \int_{-\infty}^{+\infty} f^*(\tau) \; g(t + \tau) \; d\tau \tag{6.4}$$

Equation 6.4: General cross correlation between two signals according to Turin (1960).

A simple approach in signal processing – applying a binary filter with threshold – is not adequate to process the data, because it still generates a signal that requires additional pattern filtering to detect edges. Therefore, the reverse approach with pattern filtering and a threshold to detect rising and falling edges, was applied.

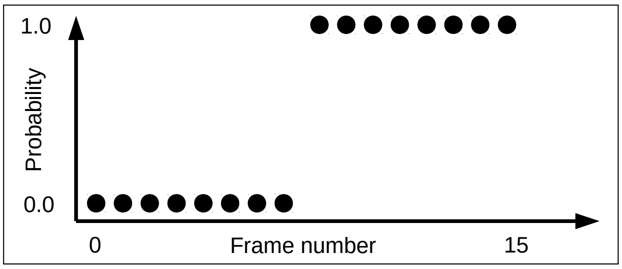


Figure 6.6: Figure depicting the pattern for the matched filter utilised for edge detection in PMP8. Signal and filter mask are spread from 0.0-1.0 to -1.0-1.0 to avoid negative effects near 0.0. On a rising edge, the correlated signal is near 1.0, and is near -1.0 on a falling edge.

Peak Detection

As described in Section 6.3.2, the peak detection process starts with a halved threshold of 0.25 for the first phase and rechecks the results with a correct threshold of 0.5.

Edge Detection

The cross-correlation applied to PMP for the detection of edges, the correlate function (see Appendix B.2.2) already returns separate lists for rising and falling edges. For CP-based detection, the edges are differentiated by comparing the two data values preceding and following the actual edge, which results in a decision: falling or rising edge.

Edge Categorisation

Table 6.2 describes the conditions utilised to categorise the data according to the number of rising and falling edges. The table also covers cases of doubled falling or rising edges, such as if a smaller rising edge is followed by another one before the next falling edge occurs. This is an extension of the simple categorisation described in Table 6.1.

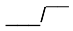
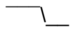
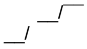
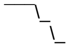
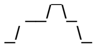
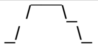
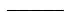
# Rising edges	# Falling edges	Condition	Result	Example data
1	0		01	
0	1		10	
2	0	rise[0] < rise[1]	01	
0	2	fall[0] < fall[1]	10	
2	2	rise[0] < rise[1] < fall[0] < fall[1]	10	
1	2	rise[0] < fall[0] < fall[1]	10	
0	0	mean (left) > 0.5 and mean (right) > 0.5	11	
0	0	mean (left) < 0.5 and mean (right) < 0.5	00	

Table 6.2: Extended binary two-digit categorisation result for the edge-based CP and PMP algorithms based on the number of identified edges in the data segment. The corresponding constrained conditions guarantee the correct sequence of falling and rising edges (source code in Section B.3.4).

6.3.4. Fixed Window Mean Bisection (FWMB)

The fundamental idea behind this algorithm is founded on binary search: the search window is divided by half and the mean value is calculated for both sides. The distance of both mean values is compared with a threshold of 0.5, if 0.5 is exceeded, a direct categorisation (e.g. ’01’) is returned. Otherwise, two further subdivision steps are performed to narrow the search window for each side of the original partition. If comparison of the mean values does not produce success, an additional peak detection is applied to determine the minimum and maximum on each side and compare their distances with the threshold of t=0.5. If the threshold between minimum and maximum on different sides of the event exceeds 0.5, a categorisation of ’01’ or ’10’ is returned. Refer to Section B.3.5 for the source code.

6.3.5. Example Emotional Shift Post-Processing Output

An example of the outputs of all three algorithms is illustrated in Figure 6.7. All three methods agree on a ’01’ categorisation, but the graph demonstrates the main distinction between them, which is the influence of the window size on categorisation. While FWMB always centres on the half data length (1 second in case of a 2 seconds window as depicted in Figure 6.7), PMP and CP rely on edge detection which makes them more flexible to the position of the actual edges.

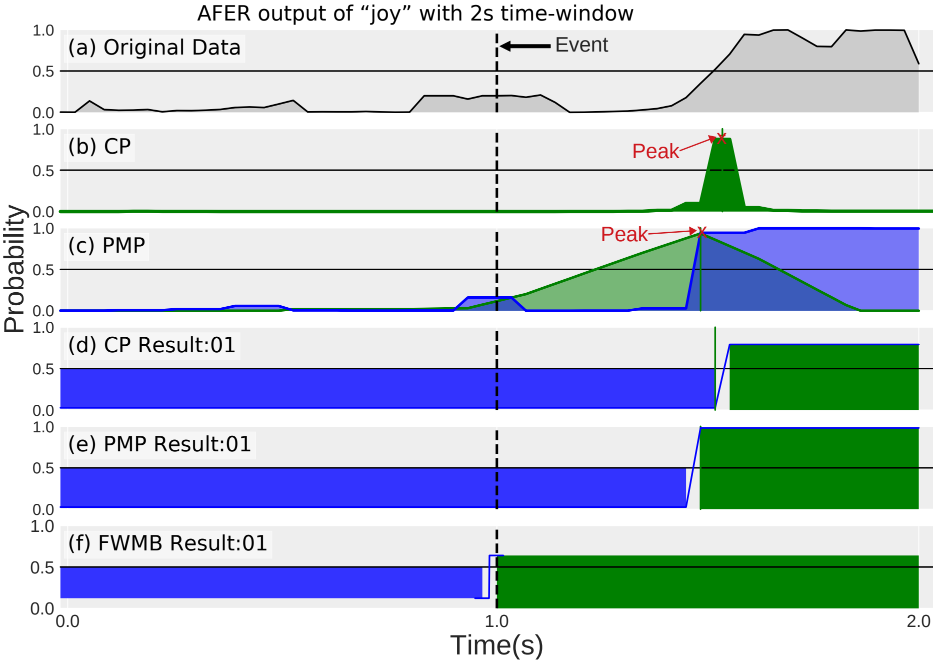


Figure 6.7: Example of the categorisation of the ’joy’ output channel with a time window of 2 seconds (1 second before and after the marked event). (a) displays the original data set, (b) displays output from the change point detection with a rising edge (green) and peak (red), (c) depicts the pattern matching with the smoothed data (blue), edge (green line) and peak detection (red). The categorisation results (’01’) for all three algorithms (d, e, f) also display mean values coloured for values > 0.5 (green) and values <= 0.5 (blue). While CP and PMP are flexible in the length of the analysis window, FWMB uses a fixed window size.

6.3.6. Pipeline for Automatic Detection of Emotional Shift

Figure 6.8 depicts the overall pipeline for the automatic detection of emotional shift; with video and event-times as input, window-based analysis of the AFER output, binary categorisation of the window data and a possible shift detected as output. The multi-channel signal from the AFER system Emotient consists of the 12 frame-based probabilities for the presence of a certain facial expression (refer to Section 6.4.2 for a list)..

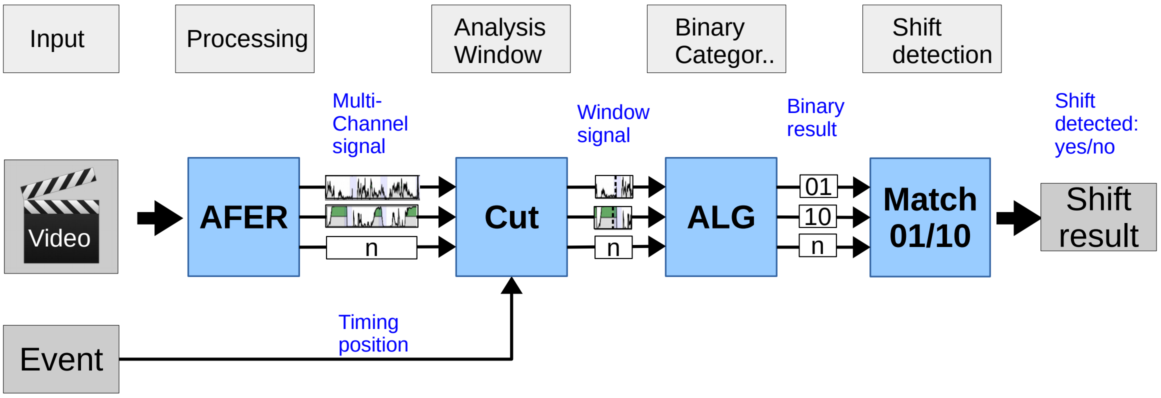


Figure 6.8: Pipeline for the automatic detection of emotional shift: the multi-channel results (Emotient provides 12 channels as described in Section 6.4.2) of an AFER algorithm are divided into analysis windows based on the timing information of an event. This window is processed by one of the three algorithms, which produces a binary, two-digit categorisation. If a matching ’01’ is found to a ’10’, a shift of expressions has been detected.

6.4. Evaluation of Emotional Shift Detection Algorithms

6.4.1. Evaluation Process

The data for the evaluation of the three algorithms was subdivided from the original AFER output with an analysis window around the event of 2 seconds, 4 seconds and 8 seconds. With video data recorded at 30 fps, this results in 60, 120 and 240 frames in a data set.

Each analysis window was then categorised by applying one of the three algorithms and the results were then compared with those of the human observer (see Figure 6.9).

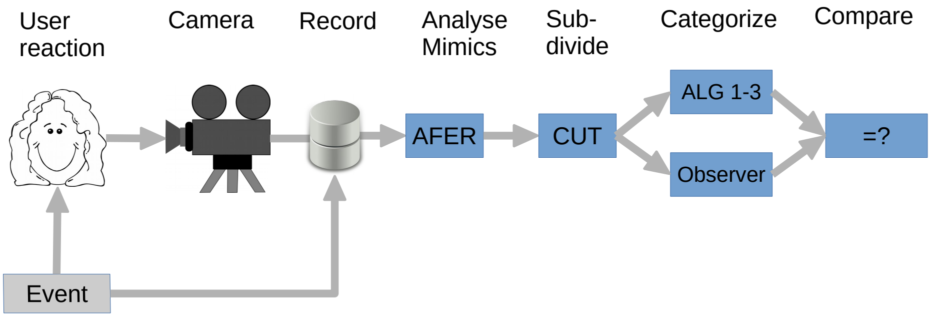


Figure 6.9: Post-processing evaluation pipeline: the recorded face of the proband is analysed by an AFER algorithm and cut to the event window size. All post-processing algorithms categorise the output of AFER just as a human observer does as reference. The categorisation results of the observer and algorithms are compared, and additional inter-rater agreement checks (see Section 6.2.2) are applied.

6.4.2. Experimental Data for Evaluation

In order to test the metrics for detecting emotional shift in a real application scenario, data from the first series of EmotionBike experiments [4]) was analysed. The EmotionBike setup provides three different types of events, with the first two providing event notifications as they can be pin-pointed to the time of the event occurring:

1. **Surprising events**: users were not warned before the occurrence of the event, which resulted in a smaller facial reactions detection window. An example event of this type is the jump-scare, as depicted in Figure 6.10.

2. **Fuzzy events**: this event is (to an extent) predictable, and so users might estimate its occurrence, especially when fulfilling the same task again. This increases the possible

[4] ES1, refer to Chapter 3.

window size for detecting responses. Falling/task-failure is an example of this event type and is depicted in Figure 6.11.

3. **Continuous events**: in this case, a constant condition is present for as longer time, such as during the complete game level. As a result, no event time can be determined and it is not a suitable case for the evaluation of shift detection; therefore, these types were ignored. An example of this type of event is an ascending mountain slope, which leads to an increase in pedal resistance for the participants over a longer time period.

The evaluation data was collected during the first series of experiments in the EmotionBike project. The video and event data from two of the game levels was analysed: jump-scare (js) from the night level as a surprise event, and falling (fa) from the challenge level to indicate task success or failure.

Figure 6.10 and Figure 6.11 also illustrate the boosting of probabilities, as explained in Section 4.4.3, even for less intense expressions; although the subject in Figure 6.10 only displays a slight smile at position 1, the probability value is almost the same as for a smile with more intensity at position 3 in the Figure 6.10.

Eleven participants provided 92 events which where subdivided into 3,312 sequences (92 * 12 channels * 3 window-sizes = 3312). Refer to Table 6.3 for further details.

The recorded video data was processed by the Emotient AFER algorithm. This algorithm produced good results in the previous benchmarking (see Chapter 4 and Bernin et al. (2017)). **Emotient provides 12 channels of emotional expressions: the six basic emotions of joy, anger, surprise, fear, disgust and sadness; and additionally, contempt, confusion,**

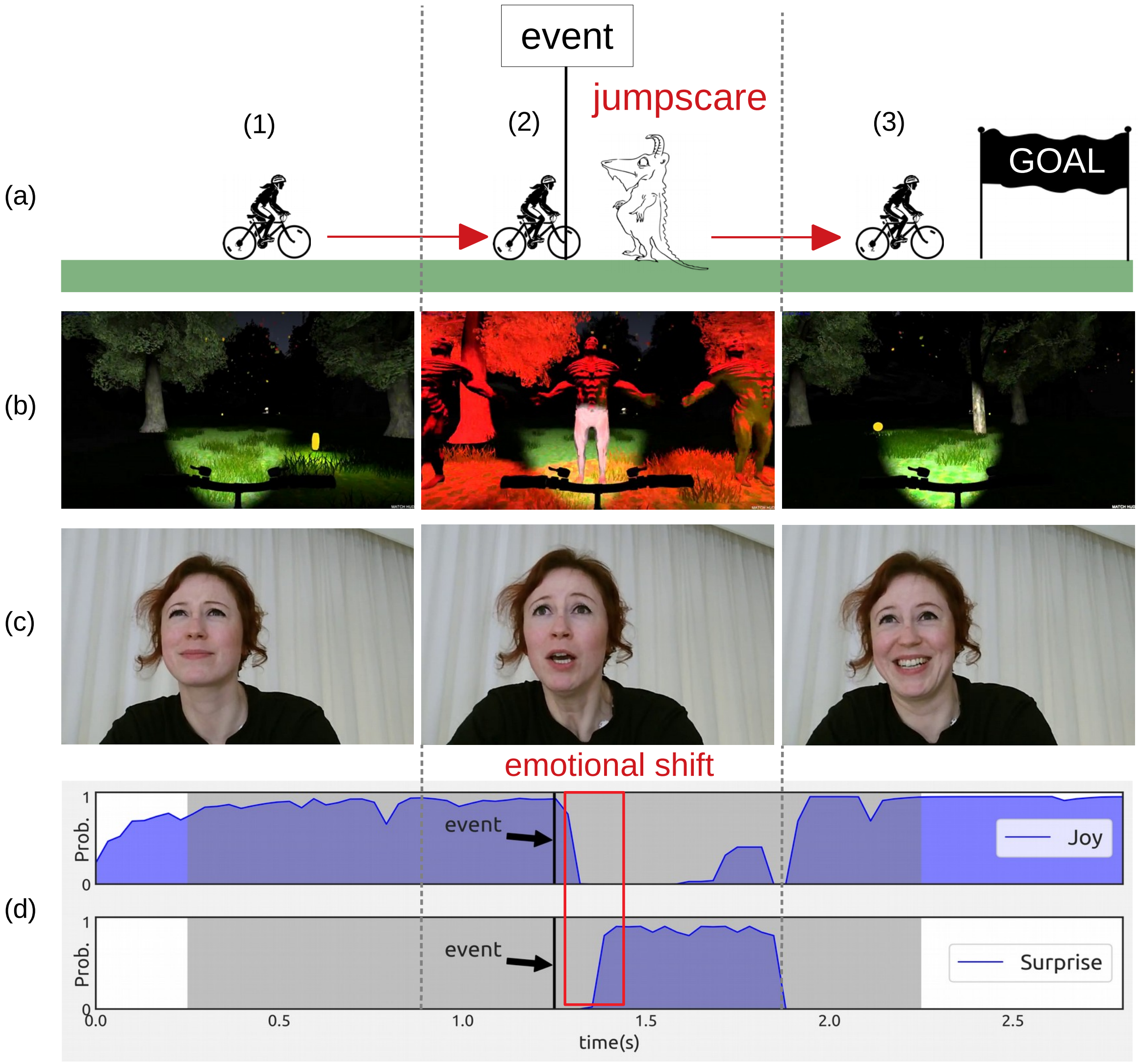


Figure 6.10: Example of details for a jump-scare event. (a) depicts the general task: to complete the level, the participants ride through a dark forest, as displayed in (b). Suddenly, a group of monsters appears in front of the participant. Besides the user's general suspicion that something surprising might happen in such an environment, the user was given no prior warning of this event. Three different bike positions (1-3) are marked and (c) displays the corresponding facial expressions. The data of two expressions, 'joy' and 'surprise' are depicted in (d), with a clear shift of expressions in a small time period close to the actual event.

frustration and neutral, positive and negative emotions. While in previous benchmark-

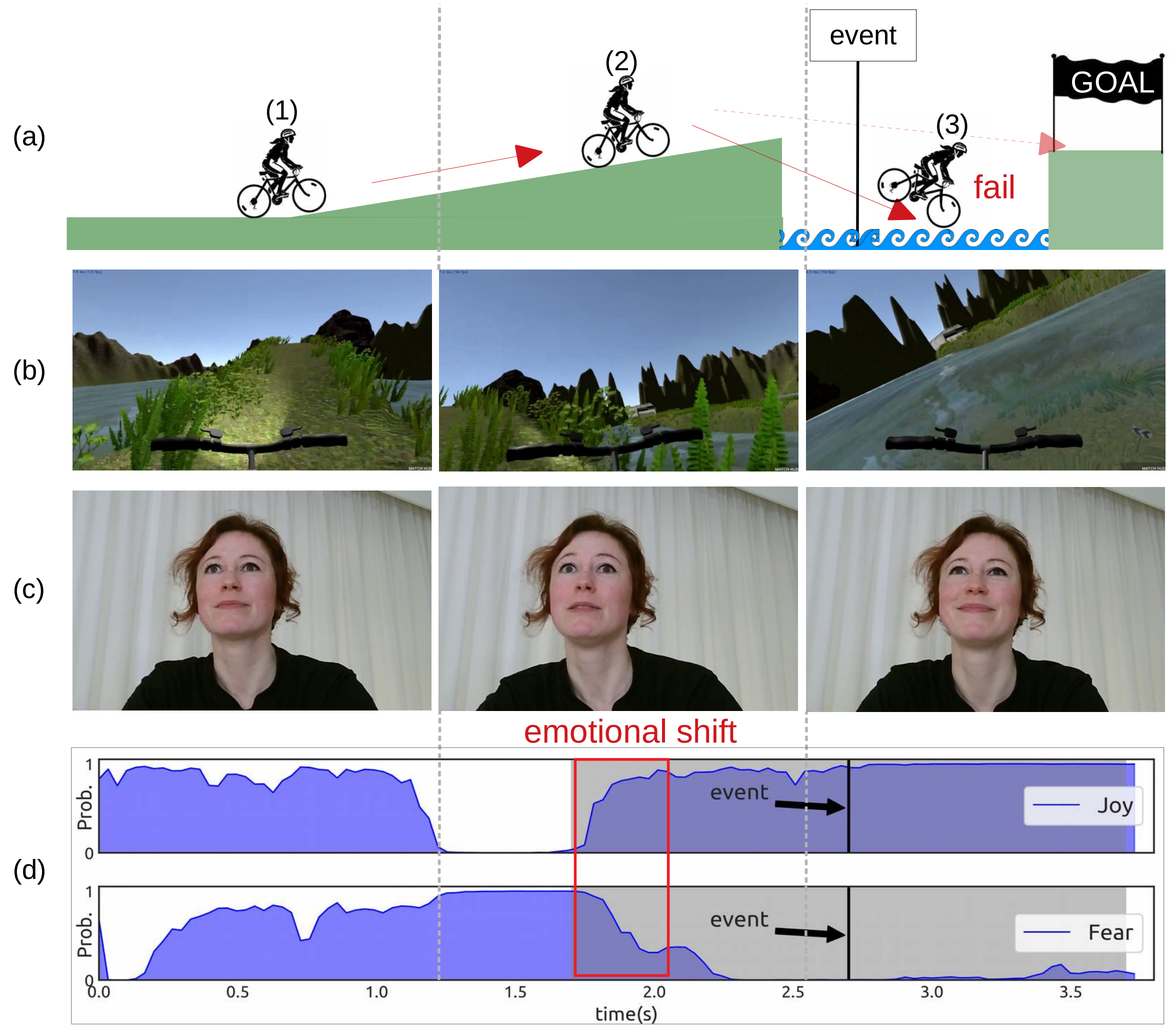


Figure 6.11: Example of details for the task failure/falling event. (a) depicts the general task: to complete the level, a large gap must be crossed using a ski jump (1). If the jump attempt is too short or drifts to one side, the participant will fail the task and must start again (3). Some participants fell even before reaching the ski jump, as in this case. Usually participants realised their failure before it actually occurred(2). Three different bike positions are marked (1-3) and the corresponding view as seen by the participant is presented in (b), with the face of the participant in (c). The data of two expressions, 'joy' and 'fear', representing emotional shift, are depicted in (d). The EmotionBike system automatically generates an event notification in the case that the user falls off the ramp.

ing, the focus was on the six basic emotions as they were present in all tested algorithms, now all 12 channels were utilised for categorisation and shift detection.

Event	Num of events	Num of expressions	Category	Analysis window size		
				2s	4s	8s
falling/task-fail	81	972	00	535	444	375
			01	167	251	263
			10	196	218	276
			11	65	17	10
			??	9	42	48
jump-scare/ surprise	11	132	00	73	60	52
			10	22	27	36
			10	16	36	30
			11	21	7	4
			??	0	2	10

Table 6.3: Categorisation results for both events by the human observer with a total number of 1104 expressions to analyse. The different analysis window sizes of two, four and eight seconds result in a total number of 3,312 rated sequences of AFER output data. The number of expressions that are categorised as '00' decreases with the increase of the window size length, as the possibility of edges also increases.

6.5. Results of Categorisation and Shift Detection

This sections present the two-digit categorisation results for the analysis window and the following shift detection. Figure 6.12 depicts an example of when not all algorithms agree in their results. The data for this analysis was the same as in Figure 6.10 for the 'joy' expression.

6.5.1. Categorisation Evaluation Results Table

The categorisation results for every approach[5], window size and event are displayed in Table 6.4 as the rate of agreement compared to the human observer. The calculation of an

[5] Only PMP16 for PMP, because it performed best.

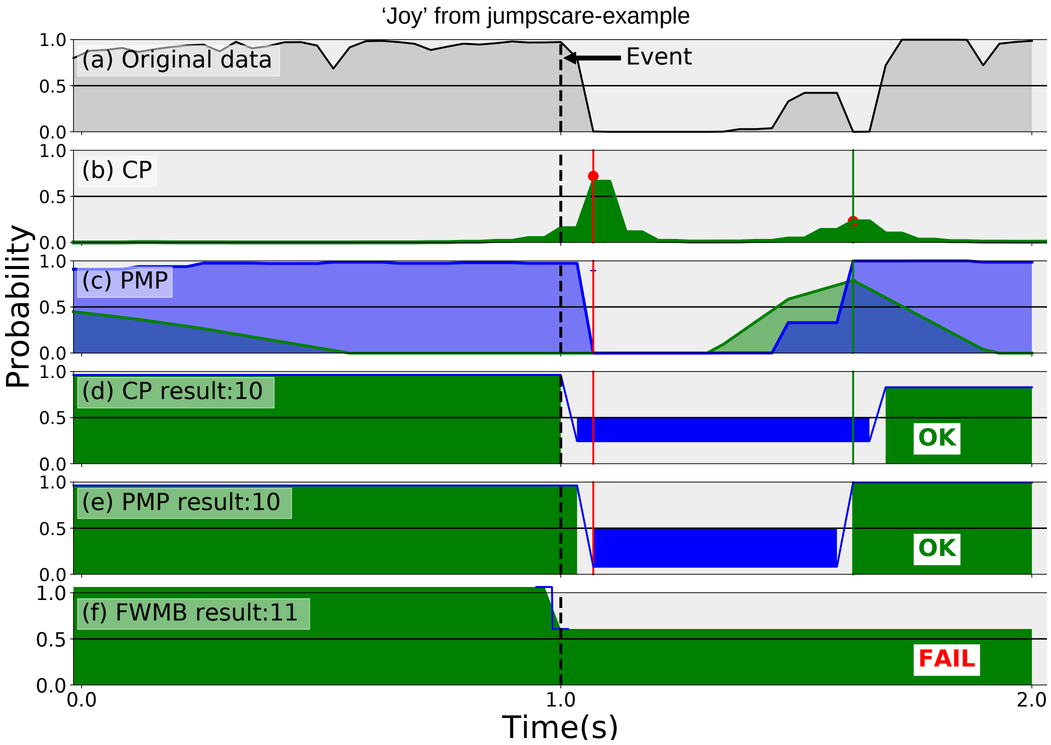


Figure 6.12: Example analysis and categorisation for 'joy' data from Figure 6.10 that demonstrates the case of when FWMB fails because the difference between mean values on both sides of the event is too small. CP and PMP correctly categorise this as '01', while FWMB assumes '00'.

additional overall mean enables the identification of the best overall results for all three approaches. A mean value over all four types of successful categorisation (excluding the results categorised as unsolvable) is calculated and the best match is highlighted.

For window lengths of four seconds or less, the CP-based algorithms produced the best results. Data categorised by the observer as '11' often failed in the longer eight seconds analysis window due to small peaks and spikes that were often ignored by the human observer but not the algorithms.

Krippendorff's alpha was calculated to compare categorisation agreement of the human observer with all three algorithms. For this purpose, results of the individual algorithms were

Event	Categorisation	Algorithm	Window			Mean
			2s	4s	8s	
falling	00	CP	0.98	0.98	0.90	0.95
falling	01	CP	0.71	0.76	0.59	0.69
falling	10	CP	0.79	0.85	0.69	0.78
falling	11	CP	0.91	0.82	0.30	0.68
falling	**mean**	**CP**	**0.85**	**0.85**	0.62	**0.77**
falling	00	PMP16	1.00	1.00	0.99	1.00
falling	01	PMP16	0.51	0.48	0.47	0.49
falling	10	PMP16	0.57	0.61	0.51	0.56
falling	11	PMP16	0.98	0.94	0.90	0.94
falling	mean	PMP16	0.77	0.76	0.72	0.75
falling	00	FWMB	0.97	0.97	0.92	0.95
falling	01	FWMB	0.55	0.59	0.59	0.58
falling	10	FWMB	0.54	0.58	0.57	0.56
falling	11	FWMB	0.86	0.94	1.00	0.93
falling	mean	FWMB	0.73	0.77	**0.77**	0.76
jump-scare	00	CP	0.99	1.00	0.88	0.96
jump-scare	01	CP	0.64	0.81	0.64	0.70
jump-scare	10	CP	0.88	0.97	0.83	0.89
jump-scare	11	CP	0.86	1.00	0.75	0.87
jump-scare	**mean**	**CP**	**0.84**	**0.95**	0.78	**0.85**
jump-scare	00	PMP16	1.00	1.00	1.00	1.00
jump-scare	01	PMP16	0.45	0.67	0.61	0.58
jump-scare	10	PMP16	0.25	0.36	0.60	0.40
jump-scare	11	PMP16	0.90	1.00	1.00	0.97
jump-scare	mean	PMP16	0.65	0.76	**0.80**	0.74
jump-scare	00	FWMB	0.90	0.88	0.81	0.86
jump-scare	01	FWMB	0.73	0.70	0.61	0.68
jump-scare	10	FWMB	0.44	0.47	0.53	0.48
jump-scare	11	FWMB	0.90	1.00	1.00	0.97
jump-scare	mean	FWMB	0.74	0.76	0.74	0.75

Table 6.4: Abbreviated categorisation results for all three algorithms (CP, PMP, FWMB) compared to the human observer. The results include different window sizes of two seconds, four seconds and eight seconds for the falling and jump-scare event. For the PMP algorithm, only PMP16 showed better results than PMP8 and PMP24. In addition, a mean value for all three window sizes and all classes has been calculated and the best result is **marked** with a bold font. The CP-based algorithm shows the best overall accuracy.

compared to the result of the human observer. As Emotient provides 12 channels, for jump-scare, 132 (12 channels * 11 events) analysis windows were categorised by the human observer and the three algorithms, which resulted in a mean over all 11, events as depicted in Figure 6.13. For falling events, 972 plots were analysed.

As mentioned in Section 6.2.2, α should reach 0.8 to ensure reliability, but this is only true for CP and a time window of two seconds for falling and jump-scare. Falling and the time window of four seconds nearly reached 0.8; all others were below.

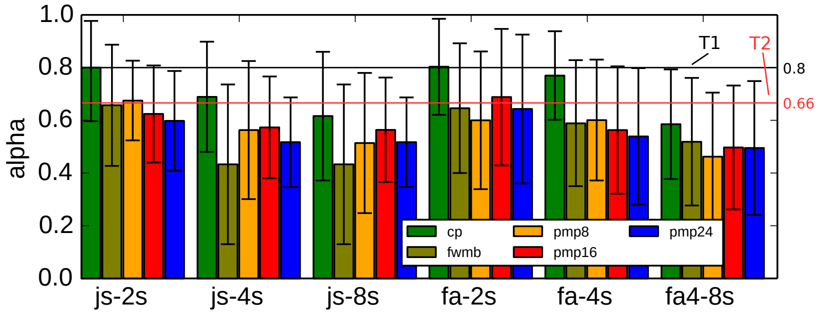


Figure 6.13: Krippendorff's alpha for inter-rater agreement as described in Section 6.2.2, calculated for all algorithms and variants compared separately with the human observer ratings for each jump-scare (js) and falling (fa) event within the time windows (two seconds, four seconds, eight seconds). The y-axis displays the mean value over alpha values for each algorithm with the standard deviation (SD). T1 (0.8) and T2 (0.66) mark the proposed thresholds for minimum reliability of α as described in Section 6.2.2. The CP-based method performed best when applied to smaller window sizes of two or four seconds. Using large window sizes of eight seconds degraded the performance of CP.

6.5.2. Emotional Shift Detection Results

Table 6.5 displays results of the shift detection. The shift was first categorised with the human observer output for both event types and the three detection window sizes. The findings were compared to the outcome of the algorithms to identify corresponding matches. In nearly all cases, the CP-based approach achieved the best results (from 82% to 100%). In only one scenario, with a window of two seconds, PMP8 performed slightly better than CP (93% compared to 90%).

Event type	Window	Observer	CP	PMP8	PMP16	PMP24	FWMB	Max. Success rate
falling	2s	58	52	54	37	36	35	0.93
falling	4s	68	62	47	50	51	50	0.91
falling	8s	72	59	52	48	51	62	0.82
jump-scare	2s	5	5	3	2	2	3	1.00
jump-scare	4s	8	8	6	5	4	6	1.00
jump-scare	8s	9	8	6	8	8	7	0.89

Table 6.5: Shift detection results for the falling and jump-scare events. The results display the number of categorised shifts by the human observer and the corresponding matches in finding the shifts by all algorithm variants (e.g. 54 out of 58 were found for PMP8 and falling with two seconds). The best match is highlighted in light green. The success rate of the maximum result is displayed in the last column. The overall success rate of the CP-based approach is almost 0.92.

For inter-ratio reliability testing, values for Krippendorff's alpha were also calculated for shift detection agreement (see Figure 6.14). Only two values reached a reliability level of $\alpha = 0.8$ (jump-scare with CP and a window size of two or eight seconds). However, the small number of data points should be considered for this case.

The CP-based algorithm always (with the exception for js-4s) provided the best results. The results support the conclusion that the CP-based algorithms achieved the best overall results.

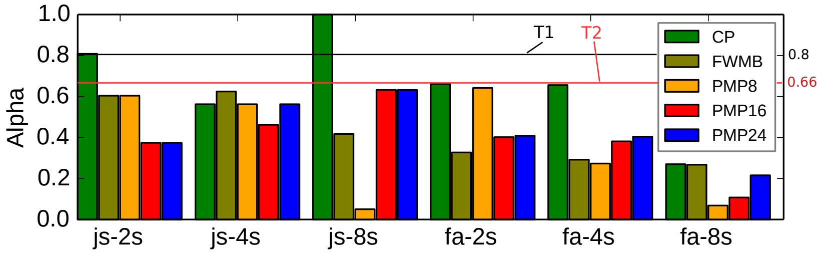


Figure 6.14: Comparison of Krippendorff's alpha for the shift detection results of algorithms, including PMP variants, to the human observer. As shift is searched in all channels simultaneously, only one detection is possible for each event and window size. Due to the small number of jump-scare events (only 11 in total), no SD was calculated.

6.5.3. Example Confusion Matrices for Categorisation

(a)

CP jump-scare	00	01	10	11	???
00	1.00	0.00	0.00	0.00	0.00
01	0.07	0.81	0.11	0.00	0.00
10	0.00	0.03	0.97	0.00	0.00
11	0.00	0.00	0.00	1.00	0.00
???	0.00	0.50	0.50	0.00	0.00

(b)

CP falling	00	01	10	11	???
00	0.90	0.03	0.06	0.00	0.00
01	0.06	0.59	0.21	0.00	0.15
10	0.01	0.11	0.69	0.00	0.19
11	0.00	0.10	0.50	0.30	0.10
???	0.06	0.21	0.46	0.00	0.27

Table 6.6: Example of confusion matrices from the binary categorisation for the CP-based approach, which depicts (a) values for a four second jump-scare/surprise event as the best overall result in all categorisations. The true-positive rate (ignoring the '??' categorisation) has a mean value of 0.95. The matrix (b) displays an eight seconds task failure/falling event as the lowest result found for the CP-based method (with a mean of 0.55, including the '??' categorisation). For eight second windows, the '??' categorisation often increases, while data categorised as '00' decreases (see Table 6.3).

Examples of confusion matrices for good (blue), best (green) and worst (red) case categorisation results of the CP-based approach are provided in Table 6.6. Two cases are demonstrated: firstly, the best case found for the CP-based algorithm for a jump-scare event with a four second window and success rates of nearly 100%; except for in the event that data was labelled inconclusive ('??') by the human observer: CP always categorised this as '01' or '10'. Secondly, it presents the case of an overly wide-ranging analysis window (eight seconds) where the detection rates of '01', '10' and especially '11' dramatically decrease, while the rate of data categorised as inconclusive by CP increases from 0.0 to 0.27.

6.5.4. Performance and Limitations of Categorisation Algorithms

The processing time of all three approaches was measured and demonstrated soft real time (processing in < 1s) capabilities. Although this behaviour was expected for PMP and FWMB, CP is also capable of this, although it was developed for offline processing and had the longest computing time. CP has a computational complexity of $O(n^2)$, which makes the runtime of CP sensitive to significant sample enlargements. This must be considered when increasing the window size or analysing datasets without windowing.

The three categorisation approaches presented in this chapter are based on the assumption that change of facial expressions occur rapidly. **They are therefore not suitable for detecting smooth transitions**. Since users usually reacted quickly to events provoked in the EmotionBike setup, this did not pose a problem to the data processed.

All developed algorithms assume, that the complete window is present. While for FWMB this is inherent, reactions to flexible windows is a possible future extension for CP and PMP to shorten the reaction time of the affective system as both are more flexible than FWMB.

Figure 6.15 demonstrates the case of processing borderline data. It illustrates the challenge: the signal is often not as clear as the optimum depicted in Figure 6.4. CP and FWMB categorised this data as '10', as did the human observer did. PMP was unable to detect the smaller edges and failed to find anything other than '00'.

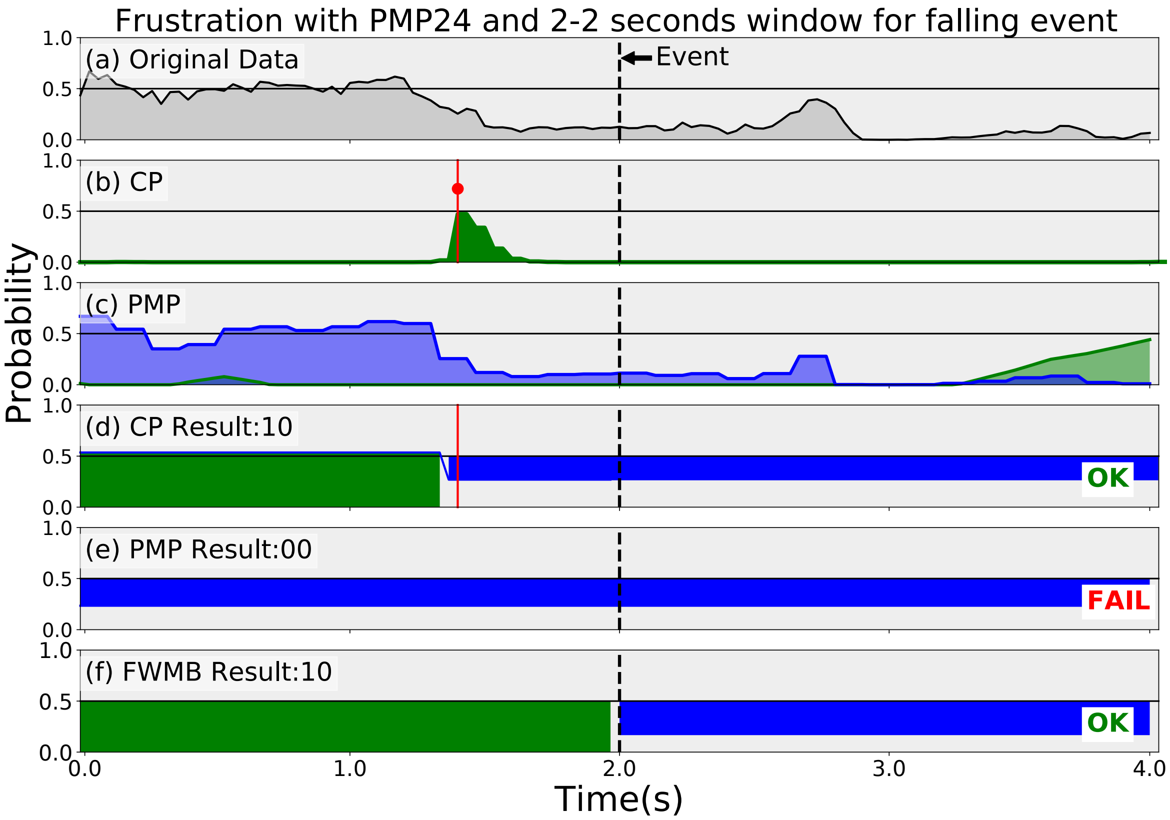


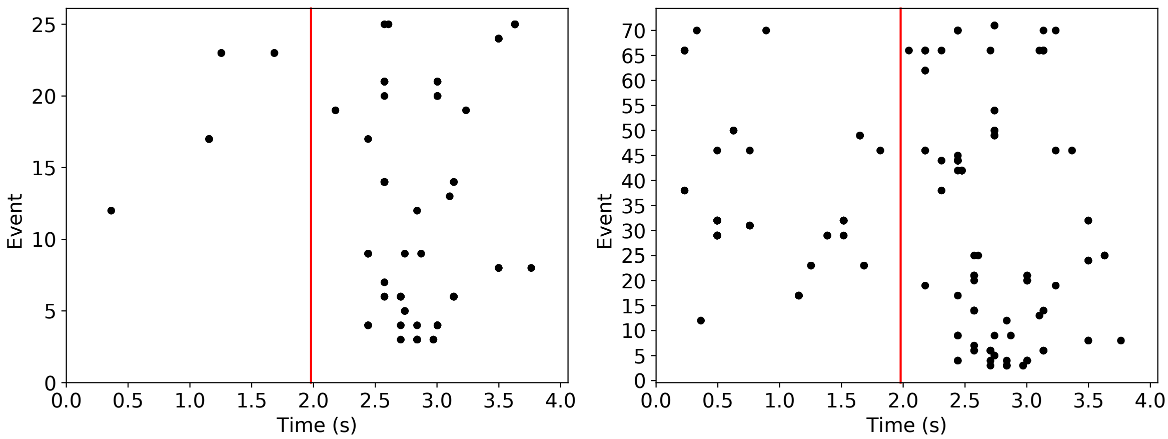
Figure 6.15: Challenging example of data to categorise, in which the x-axis display frame number and y-axis displays probability: the observer categorised this with '10', as did the CP and FWMB methods. In contrast, PMP categorised this with '00' for all matched filter sizes.

6.6. CP-based Edge Detection as Metric for Emotion Provoking Events

The outcome of evaluating the different algorithms for detecting shift of expressions revealed, that the CP-based algorithm performed best. It is also capable of detecting shifts without a fixed time window enabling to search for multiple shifts near an event.

This enables to analyse the timing of provoke-response patterns as a new metric to evaluate whether the event provokes as expected. As described in Section 6.4.2, events in the EmotionBike setup, are unexpected, expected or continuous. The type of the provocation should therefore be reflected in the timing of corresponding changes in facial expressions, at least for non-continuous events.

As an enhancement to the method proposed in Section 6.3, for this metric, not only complete shifts, but also single falling or rising edges – without the need for a corresponding counterpart in another channel – was utilised. To evaluate its feasibility, this method was applied to analyse the jump-scare and challenge level from ES2, with the results depicted in Figure 6.16 and 6.17.



(a) (b)

Figure 6.16: Rising and falling edges analysed for jump-scare events (event time marked in red, edge as dot) for the different measures of ES2. (a) depicts only the detected edges for the first occurrence (n=25) of the jump-scare event, while (b) depicts the detected edges for all measures (n=71). Especially (a) depicts that edges of expressions especially occurred after the unexpected event.

For the first jump-scare event, 89% of the edges were detected after the event(Figure 6.16a), which is also consistent to observation and self-reporting during the experiments. Probands were **surprised by the unexpected appearing group of monsters, resulting in a change**

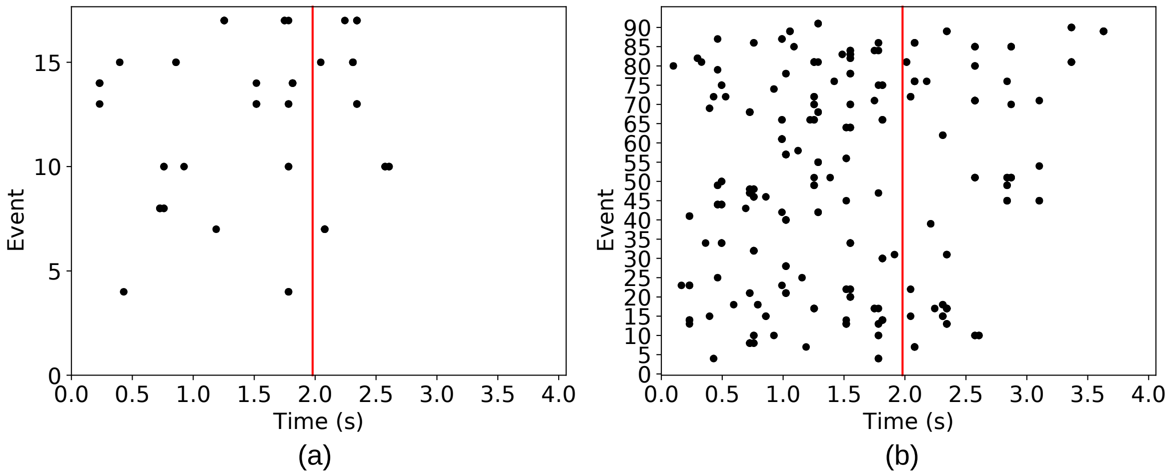


Figure 6.17: Rising and falling edges analysed for falling events (event time marked in red, shift as dot) for the different measures of ES2. (a) depicts the detected edges for the first occurrence of the falling event (n=17), while (b) depicts the detected edges for all measures (n=91). Both a and b show that edges of expressions especially occurred before the expected event.

of expression after the event. As most of them were expecting more jump-scare events, the rate of after event edges drops to 75% for all events (Figure 6.16b).

For the first falling event, 71% of the edges occurred before the actual falling event was triggered (Figure 6.17a), which increases to 74% for all events (Figure 6.17b). **This is consistent with the observation, that probands knew prior to falling from the island that they will fail to complete the task**, which was also reported due the self-reporting questions after the game level.

The different count of measures for the first occurrence of events between jump-scare and falling events are caused by the fact, that in a few cases probands were able to complete the challenge level in the first attempt, without falling down.

These results underline, that **applying CP-based edge detection as a metric to the ES2 data enables to evaluate the reaction to the provoking event**. For jump-scare, the effects are located mostly on the time after the event, while for falling, most reactions occurred before the event.

Utilising the CP-based edge detection as a metric provides a method of evaluating the *quality* of provoking events – whether they provoked a response as expected – for the EmotionBike and independent of the actual expression channels by exploiting the timing of the facial expression.

6.7. Discussion of Shift Analysis Results

This chapter focused on evaluation of the timing of facial expressions in environments, which are capable of providing a provoke-response pattern for interaction. As explained in Chapter 3, the EmotionBike system utilises such a pattern and provides time-sensitive automatic annotation of data based on the game events, providing a prerequisite for the analysis of emotional shift around provoking events.

The emotional shift analysis is based on two methods: firstly, an event-based interaction scheme that pinpoints an affective context with a time analysis window around an event. Secondly, by exploiting an shift in facial emotion expression, that finds the transition behaviour within this window, which leads to a more robust and subject-independent outcome. This detection pattern is useful for applications in affective environments where the environment is able to process internal or external events to temporally pinpoint user responses (Bernin et al. 2018).

For detecting the shift in the AFER algorithms output, the three developed approaches CP, FWMB and PMP were applied and evaluated in the practical application context of the EmotionBike. The CP method demonstrated the best performance and was closest to the human perception of the curves and observer results. CP also had the highest values for the inter-rater reliability metric of Krippendorff's alpha, although only reached the recommended range of α above 0.8 a few times.

The CP-based algorithm provided the best results for the detection of emotional shifts within a fixed time-window with a accuracy of 92% compared to human observer-based categorisation, which suggests that the automatic processing of shift events as an additional tool for coping with subject variances is promising.

In addition, the CP-based algorithm provided a starting point to investigate event-based facial responses without fixed window sizes. To further evaluate the variability in timing and dynamics, data from the ES2-set of EmotionBike experiments were analysed with CP-based edge detection.

This evaluation revealed, that the timing-related position of changed facial expressions is consistent with the expected behaviour of probands in response to the provocation. **The CP-based edge/shift detection therefore provides a novel metric for analysing the nature of the provoke-response pattern for the game levels of the EmotionBike.**

The presented evaluation may be applied more generally (see Figure 6.2) when events are triggered by internal provocations of the system, such as audio-visual-haptic stimuli in games, or by events external to the system, such as real world incidents for driver or operator assistance. In the latter case, the system must detect the external event in order to determine the time-frame for the users' reaction.

7. Summary and Conclusion

Affective technologies have matured in recent years, creating new possibilities for user interface applications in many areas, such as entertainment, medical diagnosis and education. As one key technology, AFER represents a modality with useful properties in terms of dynamics and a wide range of reactions[1].

Most contemporary AFER approaches utilise **generalised emotion models**, which present a **challenge** for the requirements of **application-specific scenarios**. As a complementary approach, this thesis examined the adaptation and tailoring of **existing AFER algorithms** and emotion models from an **application-centric perspective**. On the basis of general approaches, special application requirements and the integration of existing AFER algorithms in application contexts were evaluated by the example of the affective exergame and cockpit-scenario EmotionBike.

[1] As explained in Section 2.5.

7.1. Overall Research Questions and Findings

This section describes this thesis' findings in a research question-and-answer format. The numbering of the research questions corresponds to the related chapters as described in Section 1.5.

RQ 2.1: What is an emotion and how is it modelled for human-computer interaction?

In the last 150 years, more than 150 emotion models from different fields of science (e.g. psychology, neurobiology) have been proposed (Reisenzein et al. 2013). There is a wealth of models from which to choose, but a final consensus has yet to be reached with regard to which of these models most accurately reflects reality, the exact nature of an emotion (Gross 2010) and how it can be modelled. As a consequence, different emotion models have been applied in the field of computer science. Predominantly, discrete models have been utilised in the recognition of emotions, while for emotion models that represent the internal state of, for example, virtual avatars, hierarchical or multimodal approaches have often been applied. For some scenarios, including learning and pain recognition, application-specific models based on the coding of the expressions with FACS have been demonstrated (Corneanu et al. 2016). However, in most scenarios generic approaches without application-specific adaptations have been applied.

RQ 2.2: What is the relationship between emotion and facial expression?

In general, there is a close connection between emotions as an abstract state and physical facial expressions in most emotion models. Depending on the model type, there is a discrepancy in whether an emotion triggers a facial expression (e.g. basic emotions), a facial expression triggers an emotion (e.g. body response), or both are independent components of an underlying affective system (e.g. core affect). These differences in the model are important for the model itself (e.g. timing and sequence), but play only a secondary role in the case of AFER, as if the expression is visible, AFER is applicable. From a generalised AFER perspective, the underlying emotional model is interchangeable and partly application-specific (e.g. for learning applications).

RQ 2.3: What are the current application domains for facial expression based human-computer interaction?

Various research projects have demonstrated AFER's effect on improving the learning effectiveness of automated learning systems, recognition of medical conditions, assessment of therapy (Valstar 2014), social training companions to improve social abilities (Anderson et al. 2013) and adaptive entertainment solutions (Moniaga et al. 2018). There are still challenges for applying AFER, such as including and modelling the general context (e.g. task, environment and topic) or the current context (e.g. user actions, external events) in an application scenario.

RQ 3.1: What are the requirements for building a cross-platform framework for the EmotionBike?

For the EmotionBike project, a cockpit scenario was chosen, as interaction and emotional state are important factors for passengers' safety and drivers' performance. System requirements were defined that are not only applicable to this specific scenario, but also in the more generalised context of multimodal affect recognition frameworks. In the process, seven main requirements (see Section 3.2.1) were identified: interactivity, concurrency, recording, logging, metrics, integrability and communication. These requirements were implemented in the framework design.

RQ 3.2: In order for a system to satisfy the requirements from RQ 3.1, how would it have to be designed?

An important requirement for a multimodal affective system is interactivity; to ensure responsive reactions that mimic responsiveness between humans. Therefore, time-dependent adequate behaviour must be modelled when developing a time-sensitive system. To enable runtime and post-experiment processing of data, time-sensitive recording in combination with the event-based design of the EmotionBike enables automatic annotation of the recorded data set. The focus of this thesis is on the image-processing and interpretation (utilising AFER) components of the EmotionBike system. Nevertheless, as the framework is flexible enough to be applied to other applications (e.g. the investigation of emotional reactions of car drivers), it represents a generalised approach that enables inter-application comparability. To ensure this flexibility, a distributed system approach with a service-oriented architecture was chosen.

RQ 3.3: What are appropriate metrics to measure whether the defined framework design fulfils the requirements and was the evaluation successful?

Several metrics were defined for the framework and evaluated during the EmotionBike experiments. Although these metrics were primarily designed for the EmotionBike, they may be generalised and also applied accordingly for other multimodal experiments in the field of affective applications. The results of the experiments demonstrated that the metrics were suitable to analyse the frameworks' performance, and the developed framework fulfilled the necessary requirements in terms of frame rates, latency, responsiveness, and interactivity, as described in Section 3.5.

RQ 4.1: How effectively do existing systems for AFER perform? What is an appropriate method to generate an AFER algorithm benchmark that reflects characteristics of the EmotionBike?

Although many AFER approaches have been benchmarked against common datasets such as CK+, no general standard protocol for AFER comparison has been defined in the existing literature. In particular, the fact that the method used to determine the primary expression (e.g. apex or mean in an analysis window) is not defined in most published work makes comparisons challenging. Therefore, it is necessary to quantify and interpret the algorithms' output using metrics to ensure comparability between AFER approaches, and a dynamic and effective reaction to the displayed facial expressions (refer to Section 4.4). Selecting datasets with characteristics that are close to the expected data is suitable for a comparison of AFER considering application-specific settings. To identify this characteristics,

To summarise the benchmark results, recognising facial expressions with discrete emotion-based AFER remains a major challenge in real-life or in-the-wild application scenarios and the EmotionBike setup.

RQ 4.2: What are the challenges and limitations in state-of-the-art algorithms?

The underlying question in relation to the focus of this work is 'are existing AFER systems sufficient to be used in applications?' Current AFER solutions show accurate results only for discrete emotions in posed datasets, with a frontal view similar to a desktop scenario.

However, the application of AFER to real-life or in-the-wild scenarios – with spontaneous expressions and changing lighting conditions or view-angles – remains an open field. Most AFER approaches are based on a generalised emotion model, while application-specific adaptations provide room for improvements.

Analysis has revealed that few studies have investigated how the output of existing algorithms can be interpreted, or what approaches exist for improvement through post-processing. In addition, dimensionality of models and timing of facial expressions are open issues, especially in the context of an application.

Based on this knowledge, two application-specific approaches were developed that are able to improve the performance of state-of-the-art AFER and enable its customisation to specific applications without retraining the AFER algorithm.

RQ 5.2: What is a suitable method to improve the performance of state-of-the-art AFER in a specific application context by exploiting the dimensionality of the algorithms output?

One method to improve the performance of AFER is to apply the newly developed method of application-related grouping. Clustering of emotions has been proposed in the literature as a method to develop emotion theories (e.g. clustering of emotion terms into categories), but not to group the algorithms' post-processed output. For the three demonstrated prototypical applications of usability testing, learning and cockpit-based environments, the accuracy may significantly increase with the developed method of application-related grouping when compared to the 6BE performance. This method was adapted for the EmotionBike by exploiting the event-based annotation of the data which provides the current context of the user. By evaluating experimental data from the EmotionBike experiments, an significant increase in accuracy[2] was demonstrated.

RQ 6.2: What is a suitable method of exploiting the event-based provocation in the EmotionBike to advance the subject-independent, time-sensitive analysis of AFER?

The new method of emotional shift analysis is another approach to increase the performance of AFER systems and is also based on the post-processing of the algorithms' output. To enable a time-sensitive analysis, an event-based analysis that utilises an analysis window is applied. As subjects react individually to events, they display individual facial expression patterns. Emotional shift provides a subject-independent approach, as only the rapid change of expressions is considered, not the expressions themselves. Three algorithms were developed and evaluated to fulfil this task. A Bayesian changepoint detection-based approach exhibited the best results, with a correspondence of 82% to 100% when compared with a human observer.

Analysing the EmotionBike video data with a CP-based edge detection demonstrated, that this is a feasible metric to evaluate the quality of the provoke-response pattern, as it allows to determine subject-independent facial reactions based on the timing of the response in relation to the provoking event.

Besides answering the research questions presented above, the contributions of the thesis are summarised in the next section.

[2] From 58.3% to 90% for first jump-scare event.

7.2. Contributions and Impact

Figure 7.1 summarises this thesis' main contributions. Each contribution presented in this thesis is part of the effort to achieve the goal of enhancing existing AFER solutions more effectively for applications such as the EmotionBike.

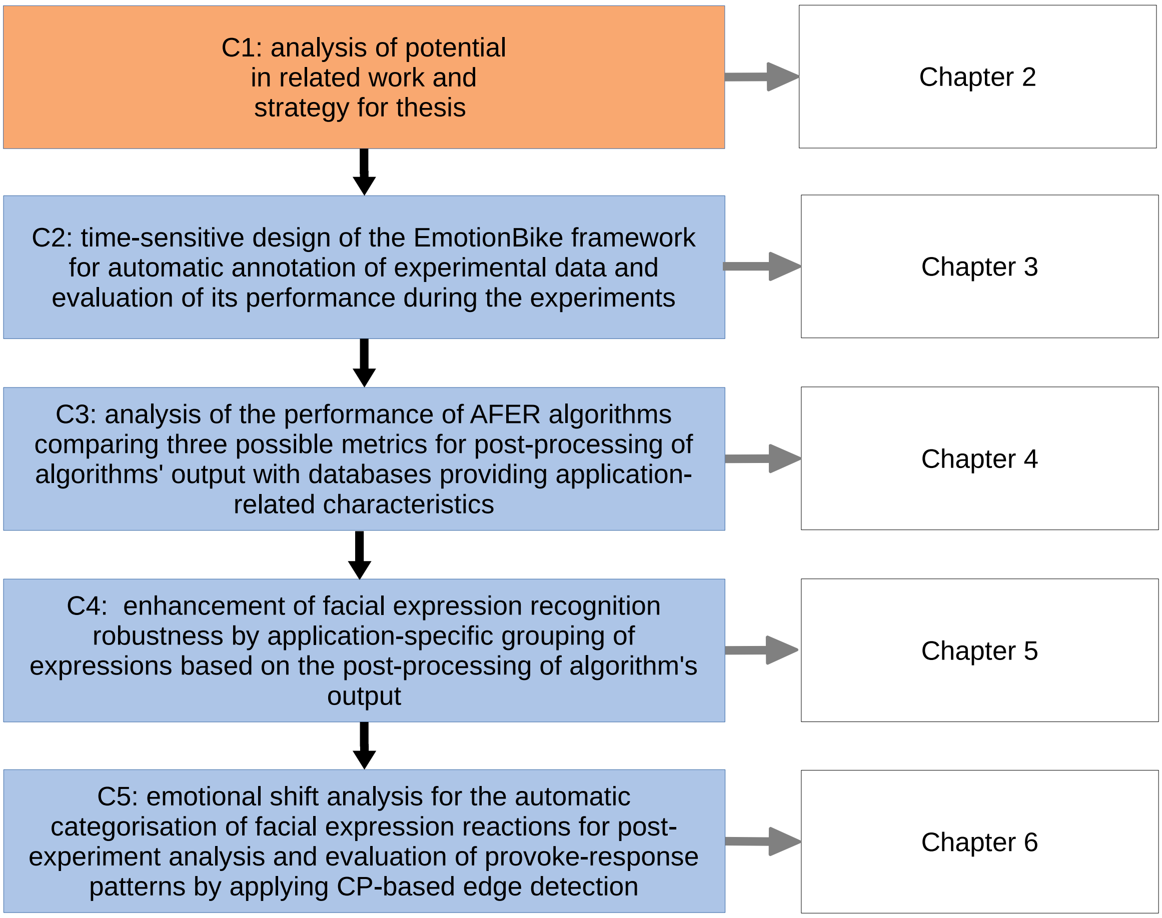

C1: analysis of potential in related work and strategy for thesis
Chapter 2
C2: time-sensitive design of the EmotionBike framework for automatic annotation of experimental data and evaluation of its performance during the experiments
Chapter 3
C3: analysis of the performance of AFER algorithms comparing three possible metrics for post-processing of algorithms' output with databases providing application-related characteristics
Chapter 4
C4: enhancement of facial expression recognition robustness by application-specific grouping of expressions based on the post-processing of algorithm's output
Chapter 5
C5: emotional shift analysis for the automatic categorisation of facial expression reactions for post-experiment analysis and evaluation of provoke-response patterns by applying CP-based edge detection
Chapter 6

Figure 7.1: The five main contributions of this thesis, their logical progression and relationship to the main chapters of this thesis. The four contributions coloured blue are related to the corresponding objectives described in Section 1.4.

The general objective of this thesis was to bridge the gap between general and application-specific AFER solutions by benchmarking, adapting and tailoring state-of-the-art algorithms to match application-specific settings by the application setting of the EmotionBike. The main contributions (C2-5) that are developed and presented in this thesis provide different methods to fulfil this goal and the four objectives (OBJ1-4)presented in Section 1.4.

The main contributions presented in this thesis are developed for the camera-based recognition of facial expressions, although some approaches are transferable to other sensors and modalities of automated emotion recognition. Application-related grouping, benchmarking and detection of emotional shift may also be applied to other modalities, such as thermographic image analysis or physiological signals.

7.3. Conclusion and Future Directions

This work successfully demonstrates an application-centric approach to affective computing by introducing a software framework design and a performance analysis of AFER systems. Furthermore, two novel approaches to improve facial expression recognition by applying post-processing were developed. In contrast to the often applied, generalised solutions for AFER, the advantages of the application-oriented approach has been demonstrated to provide a superior solution.

Figure 7.2 presents the methodology that can be applied to improve AFER systems; instead of improving the general AFER solution, application-specific solutions are developed that foster a better understanding and focus on the application requirements. This knowledge may then be applied retrospectively to improve general approaches. Other than for AFER, this method can be applied as a **general approach** to improve overall computational affect recognition.

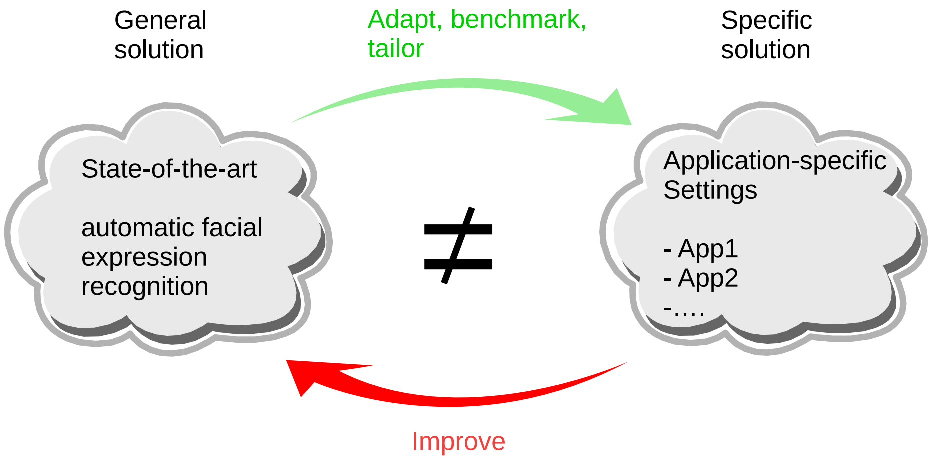


Figure 7.2: General approach for future AFER development and applications. Based on the existing general solutions, these are tailored to applications and applied. The resulting knowledge can in turn be used to improve the general solution.

In order to improve state-of-the-art AFER, computational models of emotion must evolve: they must include subject- and application-dependent knowledge, and the context of the scenario and situation. Application-specific solutions offer the opportunity to acquire this knowledge and later incorporate it into advanced emotion models.

The required facial expression recognition technology is already widespread; for example, commercial products currently include AFER (e.g. iPhone X). This offers the potential for the results of this work to be deployed in applications including e-learning systems, driver assistance, exercise and entertainment. For an evolving future field of applications, personal smart robots[3], AFER offers the possibility to enable socially intelligent human interaction, especially since robots already include and often require camera systems that can be utilised for AFER. It is a possibility that in the future, robots or virtual agents may have their own

[3] Gartner suggests that smart robots require another five to ten years to become established (https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/).

emotions in addition to being able to perceive them, allowing bi-directional emotional interaction.

New opportunities also arise from the shifting general computer interface paradigm to mobile devices (Gunes and Hung 2016). Devices such as fitness trackers provide additional information that combines well with AFER. Combining AFER with other modalities is especially useful in challenging environments and with incomplete data, as multimodal affect recognition may compensate for weaknesses in individual modalities and improve interactive applications to increase their overall coverage for detecting emotions.

As the final statement of this thesis, referring to the opening cite of Umberto Eco who postulated more than 40 years ago that the ‘language of the face’ was no longer unspeakable. Since then, significant success has been achieved in automating the recognition of facial expressions in order to teach this ‘language’ to machines. The goal to have computers become fluent in the language of facial expressions has not yet been achieved, and further work lies ahead, but the foundations exist that may enable machines to match the challenge in the foreseeable future.

Bibliography

Adam, C., A. Herzig, and D. Longin (2009). 'A Logical Formalization of the OCC Theory of Emotions'. In: *Synthese* 168.2, pp. 201–248. ISSN: 00397857, 15730964.

Ahlstrom, Christer, Katja Kircher, and Albert Kircher (2013). 'A gaze-based driver distraction warning system and its effect on visual behavior'. In: *IEEE Transactions on Intelligent Transportation Systems* 14.2, pp. 965–973.

Aleman, André and Marte Swart (2008). 'Sex differences in neural activation to facial expressions denoting contempt and disgust'. In: *PloS one* 3.11, e3622.

Anderson, Keith et al. (2013). 'The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews'. In: *Advances in Computer Entertainment*. Ed. by Dennis Reidsma, Haruhiro Katayose, and Anton Nijholt. Cham: Springer International Publishing, pp. 476–491. ISBN: 978-3-319-03161-3.

Anzalone, Salvatore M. et al. (2014). 'IMI2S: A Lightweight Framework for Distributed Computing'. In: *Simulation, Modeling, and Programming for Autonomous Robots*. Ed. by Davide Brugali et al. Cham: Springer International Publishing, pp. 267–278. ISBN: 978-3-319-11900-7.

Arguel, Amaël et al. (2017). ‘Inside out: detecting learners’ confusion to improve interactive digital learning environments’. In: *Journal of Educational Computing Research* 55.4, pp. 526–551.

Arnold, M.B. (1960). *Emotion and Personality*. Emotion and Personality Bd. 1. Columbia University Press.

Ashkanasy, Neal M, ed. (2008). *Research Companion to Emotion In Organizations (New Horizons in Management)*. Edward Elgar Pub. ISBN: 1-84542-637-1.

Bahreini, Kiavash, Rob Nadolski, and Wim Westera (2016). ‘Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning’. In: *International Journal of Human-Computer Interaction* 32.

Baldauf, Matthias, Schahram Dustdar, and Florian Rosenberg (2007). ‘A survey on context-aware systems’. In: *International Journal of ad Hoc and ubiquitous Computing* 2.4, pp. 263–277.

Bannach, David et al. (2006). ‘Distributed Modular Toolbox for Multi-modal Context Recognition’. In: vol. 3894, pp. 99–113.

Bannach, David et al. (2010). ‘Integrated Tool Chain for Recording and Handling Large, Multimodal Context Recognition Data Sets’. In: *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct*. UbiComp ’10 Adjunct. Copenhagen, Denmark: ACM, pp. 357–358. ISBN: 978-1-4503-0283-8.

Barrett, Lisa and Kristen Lindquist (2008). ‘The embodiment of emotion’. In: *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*.

Bartlett, M. et al. (2008). ‘Computer Expression Recognition Toolbox’. In: *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–2.

Becker, Christian, Stefan Kopp, and Ipke Wachsmuth (2004). 'Simulating the Emotion Dynamics of a Multimodal Conversational Agent'. In: *Affective Dialogue Systems*. Ed. by Elisabeth André et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–165. ISBN: 978-3-540-24842-2.

Becker-Asano, Christian (2008). *WASABI: Affect simulation for agents with believable interactivity*. Vol. 319. IOS Press.

Becker-Asano, Christian et al. (2008). 'Virtual humans growing up: From primary toward secondary emotions'. In: *KI-Kuenstliche Intelligenz* 2008.1.

Bell, Charles (1844). *The anatomy and philosophy of expression.* London :John Murray, p. 280.

Bentham, Jeremy (1791). *Panopticon or the inspection house*. Vol. 2.

Berland, K. J. H. (1993). 'Reading character in the face: Lavater, Socrates, and physiognomy'. In: *Word & Image* 9.3, pp. 252–269.

Bernin, Arne (2011). 'Application of 3D Cameras of spatual gestures in the smart home environment'. MA thesis. Hamburg, Germany: University of Applied Sciences.

– (2012). 'A Framework Concept for Emotion Enriched Interfaces'. In: *Entertainment Computing - ICEC 2012 - 11th International Conference, ICEC 2012, Bremen, Germany, September 26-29, 2012. Proceedings*, pp. 482–485.

Bernin, Arne, Ralf Jettke, and Florian Vogt (2016). 'Zur Problematik der Gleichgewichts-Leistung im Handlungsbezug : Theorie - Messtechnik - Datenverarbeitung - Anwendungen'. In: ed. by Volker Lippens and Volker Nagel. Sportwissenschaft und Sportpraxis. FELDHAUS VERLAG GmbH und Co. KG, pp. 111–121. ISBN: 978-3-88020-639-7.

Bernin, Arne et al. (2017). 'Towards More Robust Automatic Facial Expression Recognition in Smart Environments'. In: *Proceedings of the 10th International Conference on PErvasive*

Technologies Related to Assistive Environments. PETRA ’17. Island of Rhodes, Greece: ACM, pp. 37–44. ISBN: 978-1-4503-5227-7.

Bernin, Arne et al. (2018). ‘Automatic Segmentation and Shift Detection of Facial Expressions in Emotion Provoking Environments’. In: *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’18. Island of Corfu, Greece: ACM, pp. 37–44.

Billauer, Eli (2005). *peakdet: Peak detection using MATLAB*. `http://www.billauer.co.il/peakdet.html`.

Birman, Ken and Thomas Joseph (1987). *Exploiting virtual synchrony in distributed systems*. Vol. 21. 5. ACM.

Blom, Paris Mavromoustakos, Sander Bakkes, and Pieter Spronck (2019). ‘Modeling and adjusting in-game difficulty based on facial expression analysis’. In: *Entertainment Computing* 31, p. 100307.

Blom, Paris Mavromoustakos et al. (2014). ‘Towards personalised gaming via facial expression recognition’. In: *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Bohus, Dan, Sean Andrist, and Mihai Jalobeanu (2017). ‘Rapid Development of Multimodal Interactive Systems: A Demonstration of Platform for Situated Intelligence’. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ICMI 2017. Glasgow, UK: ACM, pp. 493–494. ISBN: 978-1-4503-5543-8.

Borod, Joan C., Cornelia Santschi Haywood, and Elissa Koff (1997). ‘Neuropsychological aspects of facial asymmetry during emotional expression: A review of the normal adult literature’. In: *Neuropsychology Review* 7.1, pp. 41–60. ISSN: 1573-6660.

Bosch, Nigel et al. (2015). 'Automatic Detection of Learning-Centered Affective States in the Wild'. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI '15. Atlanta, Georgia, USA: ACM, pp. 379–388. ISBN: 978-1-4503-3306-1.

Boucher, Patrice et al. (2016). 'PHYSIOSTRESS: A multimodal corpus of data on acute stress and physiological activation'. In:

Boulogne, G. B. Duchenne de (1990). *The Mechanism of Human Facial Expression*. Ed. by R. AndrewEditor Cuthbertson. Studies in Emotion and Social Interaction. Cambridge University Press.

Boychuk, Vasiliy et al. (2016). 'An exploratory sentiment and facial expressions analysis of data from photo-sharing on social media: the case of football violence'. In: *Procedia computer science* 80, pp. 398–406.

Brauer, Henrik, Christos Grecos, and Kai von Luck (2014). 'Robust False Positive Detection for Real-Time Multi-target Tracking'. In: *Image and Signal Processing: 6th International Conference, ICISP 2014, Cherbourg, France, June 30 – July 2, 2014. Proceedings*. Ed. by Abderrahim Elmoataz et al. Cham: Springer International Publishing, pp. 450–459. ISBN: 978-3-319-07998-1.

Broekens, Joost, Tibor Bosse, and Stacy C. Marsella (2013). 'Challenges in Computational Modeling of Affective Processes'. In: *IEEE Trans. Affect. Comput.* 4.3, pp. 242–245. ISSN: 1949-3045.

Burrell, Jenna (2016). 'How the machine 'thinks': Understanding opacity in machine learning algorithms'. In: *Big Data & Society* 3.1, p. 2053951715622512.

Buschmann, Frank et al. (1996). *Pattern-Oriented Software Architecture - Volume 1: A System of Patterns*. Wiley Publishing. ISBN: 0471958697, 9780471958697.

Buzby, D. E. (1924). 'The Interpretation of Facial Expression.' In: *The American Journal of Psychology* 35, pp. 602–604.

Cadwalladr, Carole (2018). *'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower*. URL: `https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump` (visited on 01/28/2019).

Calvo, R. A. and S. D'Mello (2010). 'Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications'. In: *IEEE Transactions on Affective Computing* 1.1, pp. 18–37. ISSN: 1949-3045.

Cambria, Erik, Andrew Livingstone, and Amir Hussain (2012). 'The Hourglass of Emotions'. In: *Cognitive Behavioural Systems*. Ed. by Anna Esposito et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 144–157. ISBN: 978-3-642-34584-5.

Cambria, Erik et al. (2012). 'Sentic Computing for Social Media Marketing'. In: *Multimedia Tools Appl.* 59.2, pp. 557–577. ISSN: 1380-7501.

Castellano, Ginevra et al. (2010). 'A Blueprint for Affective Computing: A sourcebook and manual'. In: ed. by J. Fagerberg, D. C. Mowery, and R. R. Nelson. Oxford: Oxford University Press. Chap. Body gesture and facial expression analysis for automatic affect recognition.

Chander, Anupam (2016). 'The racist algorithm'. In: *Mich. L. Rev.* 115, p. 1023.

Cheng, Shiyang et al. (2017). '4DFAB: a large scale 4D facial expression database for biometric applications'. In: *arXiv preprint arXiv:1712.01443*.

Claypool, Mark and Kajal Claypool (2006). 'Latency and player actions in online games'. In: *Communications of the ACM* 49.11, pp. 40–45.

Clore, Gerald L. and Andrew Ortony (2013). 'Psychological Construction in the OCC Model of Emotion'. In: *Emot Rev* 5.4. 25431620[pmid], pp. 335–343. ISSN: 1754-0739.

Cohen, Jacob (1960). ‘A Coefficient of Agreement for Nominal Scales’. In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Conati, Cristina (2002). ‘Probabilistic assessment of user’s emotions in educational games’. In: *Applied artificial intelligence* 16.7-8, pp. 555–575.

Cook, Diane and Sajal Kumar Das (2004). *Smart environments: Technology, protocols and applications*. Vol. 43. John Wiley & Sons.

Cook, Diane J. and Sajal K. Das (2007). ‘How smart are our environments? An updated look at the state of the art’. In: *Pervasive and Mobile Computing* 3.2. Design and Use of Smart Environments, pp. 53 –73. ISSN: 1574-1192.

Corneanu, C. A. et al. (2016). ‘Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8, pp. 1548–1568. ISSN: 0162-8828.

Craglia, Max et al. (2018). *Artificial Intelligence: A European Perspective*. ISBN: 978-92-79-97217-1.

Crockford, D. (2006). *The application/json Media Type for JavaScript Object Notation (JSON)*. RFC 4627. `http://www.rfc-editor.org/rfc/rfc4627.txt`. RFC Editor.

Dabrowski, James R and Ethan V Munson (2001). ‘Is 100 milliseconds too fast?’ In: *CHI’01 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 317–318.

Damian, Ionut, Michael Dietz, and Elisabeth Andre (2018). ‘The SSJ Framework: Augmenting Social Interactions Using Mobile Signal Processing and Live Feedback’. In: *Frontiers in ICT* 5, p. 13. ISSN: 2297-198X.

Darwin, Charles (1872). *The expression of the emotions in man and animals /*. https://www.biodiversitylibrary.org/bibliography/4820 — Includes index. New York ;D. Appleton and Co., p. 406.

Dhall, A. et al. (2012a). ‘Collecting Large, Richly Annotated Facial-Expression Databases from Movies’. In: *IEEE MultiMedia* 19.3, pp. 34–41. ISSN: 1070-986X.

Dhall, Abhinav et al. (2012b). ‘Finding happiest moments in a social context’. In: *Asian Conference on Computer Vision*. Springer, pp. 613–626.

Di Mauro, Dario et al. (2017). ‘A Framework for Distributed Interaction in Intelligent Environments’. In: *Ambient Intelligence*. Ed. by Andreas Braun, Reiner Wichert, and Antonio Maña. Cham: Springer International Publishing, pp. 136–151. ISBN: 978-3-319-56997-0.

Dick, Matthias, Oliver Wellnitz, and Lars Wolf (2005). ‘Analysis of Factors Affecting Players’ Performance and Perception in Multiplayer Games’. In: *Proceedings of 4th ACM SIGCOMM Workshop on Network and System Support for Games*. NetGames ’05. Hawthorne, NY: ACM, pp. 1–7. ISBN: 1-59593-156-2.

Ding, Changxing and Dacheng Tao (2016). ‘A Comprehensive Survey on Pose-Invariant Face Recognition’. In: *ACM Trans. Intell. Syst. Technol.* 7.3, 37:1–37:42. ISSN: 2157-6904.

D’Mello, Sidney, Arthur Graesser, and Rosalind W. Picard (2007). ‘Toward an Affect-Sensitive AutoTutor’. In: *IEEE Intelligent Systems* 22.undefined, pp. 53–61. ISSN: 1541-1672.

D’Mello, Sidney, Rosalind W Picard, and Arthur Graesser (2007). ‘Toward an affect-sensitive AutoTutor’. In: *IEEE Intelligent Systems* 22.4, pp. 53–61.

D’Mello, Sidney K. and Jacqueline Kory (2015). ‘A Review and Meta-Analysis of Multimodal Affect Detection Systems’. In: *ACM Comput. Surv.* 47.3, 43:1–43:36. ISSN: 0360-0300.

Dormann, Claire (2003). ‘Affective experiences in the Home: measuring emotion’. In: *HOIT*. Vol. 3.

Doshi, A. and M. M. Trivedi (2011). ‘Tactical driver behavior prediction and intent inference: A review’. In: *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*.

Du, Shichuan, Yong Tao, and Aleix M Martinez (2014). ‘Compound facial expressions of emotion’. In: *Proceedings of the National Academy of Sciences* 111.15, E1454–E1462.

Dufner, Michael et al. (2018). ‘Does Smile Intensity in Photographs Really Predict Longevity? A Replication and Extension of Abel and Kruger (2010)’. In: *Psychological science* 29.1, pp. 147–153.

Dumas, Bruno, Denis Lalanne, and Sharon Oviatt (2009). ‘Multimodal Interfaces: A Survey of Principles, Models and Frameworks’. In: *Human Machine Interaction: Research Results of the MMI Program*. Ed. by Denis Lalanne and Jürg Kohlas. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3–26. ISBN: 978-3-642-00437-7.

Eco, Umberto (1976). ‘Das Gesicht ist der Spiegel der Seele’. In: *Zeit* 33. translated from german by A. Bernin.

Ekman, P. and W. Friesen (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press.

Ekman, P., W. V. Friesen, and P. Ellsworth (1982). ‘What emotion categories or dimensions can observers judge from facial behavior?’ In: *Emotions in the human face*. Ed. by Paul Ekman, pp. 39–55.

Ekman, Paul (2009). ‘Darwin’s contributions to our understanding of emotional expressions’. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364.1535, pp. 3449–3451. ISSN: 0962-8436. eprint: `http://rstb.royalsocietypublishing.org/content/364/1535/3449.full.pdf`.

Ekman, Paul and Wallace V. Friesen (1986). 'A new pan-cultural facial expression of emotion'. In: *Motivation and Emotion* 10.2, pp. 159–168. ISSN: 1573-6644.

Eliav-Feldon, M., B. Isaac, and J. Ziegler (2009). *The Origins of Racism in the West*. Cambridge University Press. ISBN: 9780521888554.

Ellenberg, Jens et al. (2011). 'An environment for context-aware applications in smart homes'. In: *International Conference on Indoor Positioning and Indoor Navigation (IPIN), Guimaraes, Portugal*.

Fabian Benitez-Quiroz, C, Ramprakash Srinivasan, and Aleix M Martinez (2016). 'Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5562–5570.

Fankhauser, Péter et al. (2015). 'Kinect v2 for mobile robot navigation: Evaluation and modeling'. In: *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, pp. 388–394.

Fearnhead, Paul (2006). 'Exact and efficient Bayesian inference for multiple changepoint problems'. In: *Statistics and Computing* 16.2, pp. 203–213. ISSN: 1573-1375.

Fischer, Agneta H, Antony SR Manstead, and Ruud Zaalberg (2003). 'Social influences on the emotion process'. In: *European review of social psychology* 14.1, pp. 171–201.

Fontaine, Johnny et al. (2008). 'The World of Emotions is not Two-Dimensional'. In: 18, pp. 1050–7.

Foucault, Michel (2012). *Discipline & punish: The birth of the prison*. Vintage.

Fridlund, Alan J (1991). 'Sociality of solitary smiling: Potentiation by an implicit audience.' In: *Journal of personality and social psychology* 60.2, p. 229.

Frijda, Nico H (1986). *The emotions*. Cambridge University Press.

Gebhard, Patrick (2005). 'ALMA: a layered model of affect'. In: *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, pp. 29–36.

Gendron, Maria and Lisa Feldman Barrett (2009). 'Reconstructing the Past: A Century of Ideas About Emotion in Psychology'. In: *Emotion Review* 1.4, pp. 316–339.

Gonzalez-Sanchez, J. et al. (2011). 'ABE: An Agent-Based Software Architecture for a Multimodal Emotion Recognition Framework'. In: *2011 Ninth Working IEEE/IFIP Conference on Software Architecture*, pp. 187–193.

Gordon, Goren et al. (2016). 'Affective personalization of a social robot tutor for children's second language skills'. In: *Thirtieth AAAI Conference on Artificial Intelligence*.

Graesser, Arthur C. et al. (2016). 'Chapter 1 - Emotions in Adaptive Computer Technologies for Adults Improving Reading'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 3 –25. ISBN: 978-0-12-801856-9.

Grafsgaard, Joseph et al. (2013). 'Automatically recognizing facial expression: Predicting engagement and frustration'. In: *Educational Data Mining 2013*.

Gratch, Jonathan and Stacy Marsella (2004). 'A domain-independent framework for modeling emotion'. In: *Cognitive Systems Research* 5.4, pp. 269–306.

Gray, Jeffrey A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system.* New York, NY, US: Clarendon Press/Oxford University Press, pp. 548–548. ISBN: 0-19-852109-X (Hardcover).

Grifoni, Patrizia et al. (2017). 'MIS: Multimodal Interaction Services in a cloud perspective'. In: *CoRR* abs/1704.00972. arXiv: `1704.00972`.

Gross, James J. (2010). 'The Future's So Bright, I Gotta Wear Shades'. In: *Emotion Review* 2.3, pp. 212–216.

Gunes, Hatice and Hayley Hung (2016). 'Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block'. In: *Image and Vision Computing* 55. Recognizing future hot topics and hard problems in biometrics research, pp. 6 –8. ISSN: 0262-8856.

Happy, SL and Aurobinda Routray (2015). 'Automatic facial expression recognition using features of salient facial patches'. In: *IEEE transactions on Affective Computing* 6.1, pp. 1–12.

Harley, Jason Matthew (2016). 'Chapter 5 - Measuring Emotions: A Survey of Cutting Edge Methodologies Used in Computer-Based Learning Environment Research'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 89 –114. ISBN: 978-0-12-801856-9.

Hartley, Lucy (2005). *Physiognomy and the meaning of expression in nineteenth-century culture*. Vol. 29. Cambridge University Press.

Hayes, Andrew F and Klaus Krippendorff (2007). 'Answering the call for a standard reliability measure for coding data'. In: *Communication methods and measures* 1.1, pp. 77–89.

Heinzl, Steffen et al. (2009). 'A scalable service-oriented architecture for multimedia analysis, synthesis and consumption'. In: *IJWGS* 5, pp. 219–260.

Hess, Ursula, Reginald B Adams Jr, and Robert E Kleck (2004). 'Facial appearance, gender, and emotion expression.' In: *Emotion* 4.4, p. 378.

Hjortsjö, Carl-Herman (1970). 'Man's Face and Mimic Language'. In: *Language*.

Hoda, M., R. Alattas, and A. E. Saddik (2013). 'Evaluating Player Experience in Cycling Exergames'. In: *2013 IEEE International Symposium on Multimedia*, pp. 415–420.

Hoque, Mohammed Ehsan, Daniel J McDuff, and Rosalind W Picard (2012). 'Exploring temporal patterns in classifying frustrated and delighted smiles'. In: *IEEE Transactions on Affective Computing* 3.3, pp. 323–334.

Hornschuh, Jonas (2015). 'Further development of a bicycle ergometer as an intuitive controller for virtual worlds'. Bachelor's Thesis.

Hou, Zonghao et al. (2003). 'Design and implementation of heartbeat in multi-machine environment'. In: *17th International Conference on Advanced Information Networking and Applications, 2003. AINA 2003.* IEEE, pp. 583–586.

Inoue, Kaoru, Kazuyoshi Wada, and Yuko Ito (2008). 'Effective Application of Paro: Seal Type Robots for Disabled People in According to Ideas of Occupational Therapists'. In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1321–1324. ISBN: 978-3-540-70540-6.

Izard, Carroll E. (2007). 'Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm'. In: *Perspectives on Psychological Science* 2.3. PMID: 26151969, pp. 260–280.

Izard, C.E. (1971). *The face of emotion*. Century psychology series. Appleton-Century-Crofts.

Izquierdo-Reyes, Javier et al. (2018). 'Advanced driver monitoring for assistance system (ADMAS)'. In: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 12.1, pp. 187–197. ISSN: 1955-2505.

Jack, Rachael E., Oliver G.B. Garrod, and Philippe G. Schyns (2014). 'Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time'. In: *Current Biology* 24.2, pp. 187–192.

Jack, Rachael E et al. (2012). 'Facial expressions of emotion are not culturally universal'. In: *Proceedings of the National Academy of Sciences* 109.19, pp. 7241–7244.

Jakobs, Esther, Antony SR Manstead, and Agneta H Fischer (2001). 'Social context effects on facial activity in a negative emotional setting.' In: *Emotion* 1.1, p. 51.

James, William (1884). 'What is an emotion?' In: *Mind* 9.34, pp. 188–205.

Jaques, Natasha et al. (2016). 'Understanding and predicting bonding in conversations using thin slices of facial expressions and body language'. In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 64–74.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Campbride: Cambridge University Press.

Jonell, Patrik et al. (2018). 'FARMI: A FrAmework for Recording Multi-Modal Interactions'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Kaltwang, Sebastian, Ognjen Rudovic, and Maja Pantic (2012). 'Continuous Pain Intensity Estimation From Facial Expressions'. In: vol. 7432.

Kanade, Takeo, Yingli Tian, and Jeffrey F Cohn (2000). 'Comprehensive database for facial expression analysis'. In: *fg*. IEEE, p. 46.

Kanjo, Eiman, Luluah Al-Husain, and Alan Chamberlain (2015). 'Emotions in context: examining pervasive affective sensing systems, applications, and analyses'. In: *Personal and Ubiquitous Computing* 19.7, pp. 1197–1212. ISSN: 1617-4917.

Kim, Yanghee, Diantha Smith, and Jeffrey Thayne (2016). 'Chapter 6 - Designing Tools that Care: The Affective Qualities of Virtual Peers, Robots, and Videos'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 115 –129. ISBN: 978-0-12-801856-9.

Kitchin, Rob (2017). 'Thinking critically about and researching algorithms'. In: *Information, Communication & Society* 20.1, pp. 14–29.

Koelstra, S., M. Pantic, and I. Patras (2010). 'A Dynamic Texture-Based Approach to Recognition of Facial Actions and Their Temporal Models'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.11, pp. 1940–1954. ISSN: 0162-8828.

Koskela, Hille (2000). 'The gaze without eyes: video-surveillance and the changing nature of urban space'. In: *Progress in Human Geography* 24.2, pp. 243–265.

Krippendorff, K (2013). *Content Analysis: An Introduction to its Methodology, 3 rd Sage Publications*. Sage Publications Sage UK: London, England, pp. 239–242.

Krippendorff, Klaus (2011). 'Computing Krippendorff's alpha-reliability'. In:

Krumhuber, Eva G, Arvid Kappas, and Antony SR Manstead (2013). 'Effects of dynamic aspects of facial expressions: A review'. In: *Emotion Review* 5.1, pp. 41–46.

Krumhuber, Eva G. et al. (2017). 'A Review of Dynamic Datasets for Facial Expression Research'. In: *Emotion Review* 9.3, pp. 280–292.

Kugurakova, Vlada, Maxim Talanov, and Denis Ivanov (2016). 'Neurobiological plausibility as part of criteria for highly realistic cognitive architectures'. In: *Procedia computer science* 88, pp. 217–223.

LaFrance, Marianne (2000). 'Nonverbal Communication'. In: *Encyclopedia of psychology* 5. Ed. by Alan E. Kazdin, pp. 463–466.

Lalanda, Philippe (1997). ‘Two complementary patterns to build multi-expert systems’. In: *Pattern Languages of Programs*. Vol. 25.

Landis, Carney (1924). ‘Studies of Emotional Reactions. II. General Behavior and Facial Expression.’ In: *Journal of Comparative Psychology* 4.5, p. 447.

Lazarus, R. S. (1966). *Psychological stress and the coping process.* New York, NY, US: McGraw-Hill.

Lee, Edward (2008). ‘Cyber Physical Systems: Design Challenges’. In: *Electrical Engineering and Computer Sciences*, pp. 363–369. ISBN: 978-0-7695-3132-8.

Lewinski, Peter (2015). ‘Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets’. In: *Frontiers in Psychology* 6.1386.

Lewinski, Peter, Tim M den Uyl, and Crystal Butler (2014). ‘Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader.’ In: *Journal of Neuroscience, Psychology, and Economics* 7.4, p. 227.

Li, X. et al. (2013). ‘A Spontaneous Micro-expression Database: Inducement, collection and baseline’. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6.

Liang, Paul Pu et al. (2018). ‘Multimodal language analysis with recurrent multistage fusion’. In: *arXiv preprint arXiv:1808.03920*.

Littlewort, G. et al. (2004). ‘Dynamics of Facial Expression Extracted Automatically from Video’. In: *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 80–80.

Littlewort, Gwen et al. (2011). ‘The computer expression recognition toolbox (CERT)’. In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, pp. 298–305.

Liu, Runpeng et al. (2017). 'Feasibility of an autism-focused augmented reality smartglasses system for social communication and behavioral coaching'. In: *Frontiers in pediatrics* 5, p. 145.

Lopes, André Teixeira et al. (2017). 'Facial expression recognition with convolutional neural networks: coping with few data and the training sample order'. In: *Pattern Recognition* 61, pp. 610–628.

Lövheim, Hugo (2012). 'A new three-dimensional model for emotions and monoamine neurotransmitters'. In: *Medical Hypotheses* 78.2, pp. 341 –348. ISSN: 0306-9877.

Lucey, P. et al. (2010). 'The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression'. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101.

Luck, K. von et al. (2010). *Living Place Hamburg - A place for concepts of IT based modern living*. Tech. rep. HAW Hamburg, Germany.

Lundström, Jens et al. (2016). 'Halmstad Intelligent Home - Capabilities and Opportunities'. In: *Internet of Things Technologies for HealthCare*. Ed. by Mobyen Uddin Ahmed, Shahina Begum, and Wasim Raad. Cham: Springer International Publishing, pp. 9–15. ISBN: 978-3-319-51234-1.

Lyons, M. et al. (1998). 'Coding facial expressions with Gabor wavelets'. In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205.

Magdin, Martin and F Prikler (2018). 'Real time facial expression recognition using webcam and SDK affectiva'. In: *IJIMAI* 5.1, pp. 7–15.

Maglogiannis, I. (2014). 'Human Centered Computing for the Development of Assistive Environments: The STHENOS Project'. In: *Proceedings of the 7th International Conference on*

PErvasive Technologies Related to Assistive Environments. PETRA '14. Rhodes, Greece: ACM, 29:1–29:7. ISBN: 978-1-4503-2746-6.

Marsella, Stacy, Jonathan Gratch, Paolo Petta, et al. (2010). 'Computational models of emotion'. In: *A Blueprint for Affective Computing-A sourcebook and manual* 11.1, pp. 21–46.

Martinez, Brais and Michel F Valstar (2016). 'Advances, challenges, and opportunities in automatic facial expression recognition'. In: *Advances in Face Detection and Facial Image Analysis*. Springer, pp. 63–100.

Martinez, Brais et al. (2017). 'Automatic analysis of facial actions: A survey'. In: *IEEE Transactions on Affective Computing*.

Mathias, Markus et al. (2014). 'Face Detection without Bells and Whistles'. In: vol. 8692.

Matsumoto, David (1991). 'Cultural influences on facial expressions of emotion'. In: *Southern Journal of Communication* 56.2, pp. 128–137.

Matthias, Andreas (2004). 'The responsibility gap: Ascribing responsibility for the actions of learning automata'. In: *Ethics and information technology* 6.3, pp. 175–183.

Mauro, D. Di and F. Cutugno (2016). 'A Framework for Interaction Design in Intelligent Environments'. In: *2016 12th International Conference on Intelligent Environments (IE)*, pp. 246–249.

McCall, J. C. and M. M. Trivedi (2007). 'Driver Behavior and Situation Aware Brake Assistance for Intelligent Vehicles'. In: *Proceedings of the IEEE* 95.2, pp. 374–387. ISSN: 0018-9219.

McDougall, William (1908). *An introduction to social psychology*. Methuen & Co.- London.

McDuff, Daniel et al. (2016). 'AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit'. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 2016*, pp. 3723–3726.

Mehrabian, Albert and James A Russell (1974). *An approach to environmental psychology.* the MIT Press.

Michel, Philipp and Rana El Kaliouby (2003). 'Real Time Facial Expression Recognition in Video Using Support Vector Machines'. In: *Proceedings of the 5th International Conference on Multimodal Interfaces*. ICMI '03. Vancouver, British Columbia, Canada: ACM, pp. 258–264. ISBN: 1-58113-621-8.

Mihai, Duguleană, Gîrbacia Florin, and Mogan Gheorghe (2015). 'Using Dual Camera Smartphones As Advanced Driver Assistance Systems: NAVIEYES System Architecture'. In: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '15. Corfu, Greece: ACM, 23:1–23:8. ISBN: 978-1-4503-3452-5.

Miller, Robert B. (1968). 'Response Time in Man-computer Conversational Transactions'. In: *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*. AFIPS '68 (Fall, part I). San Francisco, California: ACM, pp. 267–277.

Mittelstadt, Brent et al. (2016). 'The Ethics of Algorithms: Mapping the Debate'. In: In press.

Mone, Gregory (2015). 'Sensing Emotions'. In: *Commun. ACM* 58.9, pp. 15–16. ISSN: 0001-0782.

Moniaga, Jurike V et al. (2018). 'Facial Expression Recognition as Dynamic Game Balancing System'. In: *Procedia Computer Science* 135, pp. 361–368.

Moors, Agnes et al. (2013). 'Appraisal theories of emotion: State of the art and future development'. In: *Emotion Review* 5.2, pp. 119–124.

Mowrer, Orval (1960). 'Learning theory and behavior.' In:

Müller, L. et al. (2017). 'Emotional journey for an emotion provoking cycling exergame'. In: *2017 IEEE 4th International Conference on Soft Computing Machine Intelligence (ISCMI)*, pp. 104–108.

Müller, Larissa (2018). 'Affective Computing in Controlled Exergame Environments'. PhD thesis. University of the West of Scotland.

Müller, Larissa et al. (2012). 'Emotional Interaction with Surfaces - Works of Design and Computing'. In: *Entertainment Computing - ICEC 2012: 11th International Conference, ICEC 2012, Bremen, Germany, September 26-29, 2012. Proceedings*. Ed. by Marc Herrlich, Rainer Malaka, and Maic Masuch. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 457–460. ISBN: 978-3-642-33542-6.

Müller, Larissa et al. (2015). 'EmotionBike: a study of provoking emotions in cycling exergames'. In: *Entertainment Computing-ICEC 2015*. Springer, pp. 155–168.

Müller, Larissa et al. (2016). 'Physiological Data Analysis for an Emotional Provoking Exergame'. In: *Proceedings of the IEEE Symposium for Computational Intelligence*. Athens, Greece: IEEE.

Narayan Soni, Laxmi, Ashutosh Datar, and Shilpa Datar (2017). 'Viola-Jones Algorithm Based Approach for Face Detection of African Origin People and Newborn Infants'. In: *International Journal of Computer Trends and Technology* 51, pp. 75–81.

Negri, Lucas Hermann and Christophe Vestri (2017). *lucashn/peakutils: v1.1.0*.

Nelson, Nicole L. and James A. Russell (2013). 'Universality Revisited'. In: *Emotion Review* 5.1, pp. 8–15.

Noroozi, Fatemeh et al. (2018). 'Survey on Emotional Body Gesture Recognition'. In: *CoRR* abs/1801.07481.

Oatley, Keith and Philip N Johnson-Laird (1987). 'Towards a cognitive theory of emotions'. In: *Cognition and emotion* 1.1, pp. 29–50.

Olszanowski, M et al. (2008). 'Warsaw set of emotional facial expression pictures-Validation study'. In: *EAESP General Meeting, Opatija, Croatia*.

Ortony, Andrew, Gerald L Clore, and Allan Collins (1990). *The cognitive structure of emotions*. Cambridge university press.

Ortony, Andrew and Terence J Turner (1990). 'What's basic about basic emotions?' In: *Psychological review* 97.3, p. 315.

Orwell, George (1949). *Nineteen Eighty-Four*. London, UK: Secker and Warburg. Chap. 5.

Panksepp, Jaak (1982). 'Toward a general psychobiological theory of emotions'. In: *Behavioral and Brain sciences* 5.3, pp. 407–422.

– (2007). 'Neurologizing the Psychology of Affects: How Appraisal-Based Constructivism and Basic Emotion Theory Can Coexist'. In: *Perspectives on Psychological Science* 2.3. PMID: 26151970, pp. 281–296.

Pantic, Maja (2009). 'Machine analysis of facial behaviour: Naturalistic and dynamic behaviour'. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364.1535, pp. 3505–3513.

Park, Sanghoon (2016). 'Chapter 10 - Virtual Avatar as an Emotional Scaffolding Strategy to Promote Interest in Online Learning Environment'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 201 –224. ISBN: 978-0-12-801856-9.

Patton, Ron (2005). *Software Testing (2Nd Edition)*. Indianapolis, IN, USA: Sams. ISBN: 0672327988.

Picard, Rosalind W (1995). 'Affective Computing-MIT Media Laboratory Perceptual Computing Section Technical Report No. 321'. In: *Cambridge, MA* 2139.

Plass, Jan L. and Ulas Kaplan (2016). 'Chapter 7 - Emotional Design in Digital Media for Learning'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 131 –161. ISBN: 978-0-12-801856-9.

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper and Row. ISBN: 9780060452353.

Poria, Soujanya et al. (2017). 'A review of affective computing: From unimodal analysis to multimodal fusion'. In: *Information Fusion* 37, pp. 98 –125. ISSN: 1566-2535.

Ratneshwar, Srinivasan, David Glen Mick, and Cynthia Huffman (2003). *The Why of Consumption: Contemporary perspectives on consumer motives, goals, and desires*. Vol. 1. Psychology Press.

Reisenzein, R. et al. (2013). 'Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange'. In: *IEEE Transactions on Affective Computing* 4.3, pp. 246–266. ISSN: 1949-3045.

Reisenzein, Rainer (1992). In: Ashland, OH, US: Hogrefe & Huber Publishers. Chap. A structuralist reconstruction of Wundt's three-dimensional theory of emotion. Pp. 141–189. ISBN: 0-88937-100-8 (Hardcover); 3-456-82325-8 (Hardcover).

Reisenzein, Rainer and Achim Stephan (2014). 'More on James and the Physical Basis of Emotion'. In: *Emotion Review* 6.1, pp. 35–46.

Ringeval, Fabien et al. (2013). 'Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions'. In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, pp. 1–8.

Robinson, Peter (2014). ‘Modelling emotions in an on-line educational game’. In: *Control, Decision and Information Technologies (CoDIT), 2014 International Conference on*. IEEE, pp. 628–633.

Rodrigo, MMT and RSJd Baker (2011). ‘Comparing learners’ affect while using an intelligent tutor and an educational game’. In: *Research and Practice in Technology Enhanced Learning* 6.1, pp. 43–66.

Roth, Walton T (2010). ‘Diversity of effective treatments of panic attacks: what do they have in common?’ In: *Depression and anxiety* 27.1, pp. 5–11.

Ruiz, Natalie, Fang Chen, and Sharon Oviatt (2010). ‘Chapter 12 - Multimodal Input’. In: *Multimodal Signal Processing*. Ed. by Jean-Philippe Thiran, Ferran MarquÃ©s, and HervÃ© Bourlard. Oxford: Academic Press, pp. 231 –255. ISBN: 978-0-12-374825-6.

Rumpa, L. D. et al. (2015). ‘Validating video stimulus for eliciting human emotion: A preliminary study for e-health monitoring system’. In: *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, pp. 208–213.

Russell, James A. (1980). ‘A circumplex model of affect.’ In: *Journal of Personality and Social Psychology* 39.6, pp. 1161–1178.

Russell, James A (1994). ‘Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies.’ In: *Psychological bulletin* 115.1, p. 102.

Russell, James A. (1997). ‘The psychology of facial expression’. In: *The Psychology of Facial Expression*. Ed. by James A. Russell and Jose MiguelEditors Fernandez-Dols. Studies in Emotion and Social Interaction. Cambridge University Press.

Russell, James A (2003). ‘Core affect and the psychological construction of emotion.’ In: *Psychological review* 110.1, p. 145.

Russell, James A and Lisa Feldman Barrett (1999). ‘Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.’ In: *Journal of personality and social psychology* 76.5, pp. 805–819.

Ryan, Andrew et al. (2009). ‘Automated facial expression recognition system’. In: *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on*. IEEE, pp. 172–177.

Rychlowska, Magdalena et al. (2014). ‘Blocking mimicry makes true and false smiles look the same’. In: *PLoS One* 9.3, e90876.

Santos, Celso AS, Estêvão B Saleme, and Juliana CS de Andrade (2015). ‘A systematic review of data exchange formats in advanced interaction environments’. In: *development* 10.5.

Sariyanidi, E., H. Gunes, and A. Cavallaro (2015). ‘Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6, pp. 1113–1133. ISSN: 0162-8828.

Sariyanidi, Evangelos et al. (2013). ‘Local Zernike Moment Representation for Facial Affect Recognition.’ In: *BMVC*. Vol. 2, p. 3.

Schaefer, Alexandre et al. (2010). ‘Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers’. In: *Cognition and Emotion* 24.7, pp. 1153–1172.

Scherer, Klaus R. (2005a). ‘What are emotions? And how can they be measured?’ In: *Social Science Information* 44.4, pp. 695–729.

Scherer, Klaus R (2005b). ‘What are emotions? And how can they be measured?’ In: *Social science information* 44.4, pp. 695–729.

Scherer, Klaus R., Tanja Banziger, and Etienne Roesch (2010). ‘Outlook: IntegrationÂ andÂ futureÂ perspectives forÂ affectiveÂ computing’. In: *A Blueprint for Affective Computing: A Sourcebook and Manual*. 1st. New York, NY, USA: Oxford University Press, Inc. ISBN: 0199566704, 9780199566709.

Scherer, Stefan et al. (2013). ‘Automatic behavior descriptors for psychological disorder analysis’. In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, pp. 1–8.

Schlosberg, H (1941). ‘A scale for the judgment of facial expressions.’ In: *Journal of experimental psychology* 29.6, p. 497.

Schlosberg, Harold (1954). ‘Three dimensions of emotion.’ In: *Psychological review* 61.2, p. 81.

Schmidt, Douglas et al. (2000). *Pattern-Oriented Software Architecture: Patterns for Concurrent and Networked Objects, Volume 2*.

Schmidt, Karen L and Jeffrey F Cohn (2001). ‘Human facial expressions as adaptations: Evolutionary questions in facial expression research’. In: *American journal of physical anthropology* 116.S33, pp. 3–24.

Schuller, Björn et al. (2013). ‘ASC-Inclusion: Interactive emotion games for social inclusion of children with Autism Spectrum Conditions’. In: *Proceedings 1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2013) held in conjunction with the 8th Foundations of Digital Games 2013 (FDG)(B. Schuller, L. Paletta, and N. Sabouret, eds.), Chania, Greece*.

Schützwohl, Achim and Rainer Reisenzein (2012). ‘Facial expressions in response to a highly surprising event exceeding the field of vision: a test of Darwin’s theory of surprise’. In: *Evolution and Human Behavior* 33.6, pp. 657–664.

Sell, J. and P. O'Connor (2014). 'The Xbox One System on a Chip and Kinect Sensor'. In: *IEEE Micro* 34.2, pp. 44–53.

Shah, Miraj et al. (2013). 'Action unit models of facial expression of emotion in the presence of speech'. In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, pp. 49–54.

Shan, Caifeng, Shaogang Gong, and Peter W McOwan (2005). 'Appearance manifold of facial expression'. In: *International Workshop on Human-Computer Interaction*. Springer, pp. 221–230.

Shaver, Phillip et al. (2001). 'Emotion knowledge: Further exploration of a prototype approach'. In: *Emotions in social psychology: Essential readings*, pp. 26–56.

Shin, K. G. and P. Ramanathan (1994). 'Real-time computing: a new discipline of computer science and engineering'. In: *Proceedings of the IEEE* 82.1, pp. 6–24. ISSN: 0018-9219.

Spencer, Herbert (1855). *Principals of psychology*. London: Longman, Brown, Green &Longmans.

Stal, M. (2006). 'Using Architectural Patterns and Blueprints for Service-Oriented Architecture'. In: *IEEE Software* 23, pp. 54–61. ISSN: 0740-7459.

Steunebrink, Bas R, Mehdi Dastani, and John-Jules Ch Meyer (2009). 'The OCC model revisited'. In: *Proc. of the 4th Workshop on Emotion and Computing*.

Stöckli, Sabrina et al. (2017). 'Facial expression analysis with AFFDEX and FACET: A validation study'. In: 50.

Stratou, Giota and Louis-Philippe Morency (2017). 'MultiSense-Context-aware nonverbal behavior analysis framework: A psychological distress use case'. In: *IEEE Transactions on Affective Computing* 8.2, pp. 190–203.

Strongman, Kenneth T. (2003). *The Psychology of Emotion: From Everyday Life to Theory*. Wiley. ISBN: 0471485683.

Süssenbach, L. et al. (2014). 'A robot as fitness companion: Towards an interactive action-based motivation model'. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 286–293.

Susskind, JM et al. (2007). 'Human and computer recognition of facial expressions of emotion'. In: *Neuropsychologia* 45.1, pp. 152–162.

Suwa, Motoi (1978). 'A preliminary note on pattern recognition of human emotional expression'. In: *Proc. of The 4th International Joint Conference on Pattern Recognition*, pp. 408–410.

Tcherkassof, Anna et al. (2013). 'DynEmo: A video database of natural facial expressions of emotions.' In: *The International Journal of Multimedia & Its Applications* 5.5, pp. 61–80.

Tettegah, Sharon Y and Martin Gartmeier (2015). *Emotions, technology, design, and learning*. Academic Press.

Timmers, Monique, Agneta H Fischer, and Antony SR Manstead (1998). 'Gender differences in motives for regulating emotions'. In: *Personality and Social Psychology Bulletin* 24.9, pp. 974–985.

Tomkins, Silvan (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Tomkins, Silvan S (1984). 'Affect theory'. In: *Approaches to emotion* 163.163–195.

Turin, G. (1960). 'An introduction to matched filters'. In: *IRE Transactions on Information Theory* 6.3, pp. 311–329. ISSN: 0096-1000.

Turk, Matthew (2014). 'Multimodal interaction: A review'. In: *Pattern Recognition Letters* 36, pp. 189 –195. ISSN: 0167-8655.

Tussyadiah, Iis P and Sangwon Park (2018). 'Consumer evaluation of hotel service robots'. In: *Information and communication technologies in tourism 2018*. Springer, pp. 308–320.

Vallverdú, Jordi et al. (2016). 'A cognitive architecture for the implementation of emotions in computing systems'. In: *Biologically Inspired Cognitive Architectures* 15, pp. 34–40.

Valstar, M. F. et al. (2011a). 'The first facial expression recognition and analysis challenge'. In: *Automatic Face Gesture Recognition and Workshops (FG 2011)*, pp. 921–926.

Valstar, Michel (2014). 'Automatic behaviour understanding in medicine'. In: *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*. ACM, pp. 57–60.

Valstar, Michel and Maja Pantic (2010). 'Induced disgust, happiness and surprise: an addition to the mmi facial expression database'. In: *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, p. 65.

Valstar, Michel et al. (2013). 'AVEC 2013: the continuous audio/visual emotion and depression recognition challenge'. In: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, pp. 3–10.

Valstar, Michel et al. (2016). 'Avec 2016: Depression, mood, and emotion recognition workshop and challenge'. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pp. 3–10.

Valstar, Michel François et al. (2011b). 'The first facial expression recognition and analysis challenge'. In: *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pp. 921–926.

Van Der Schalk, Job et al. (2011). 'Moving faces, looking places: validation of the Amsterdam Dynamic Facial Expression Set (ADFES).' In: *Emotion* 11.4, p. 907.

Vela, Patricia, Patricio A. Vela, and Robert J. Jensen (2016). 'Chapter 9 - Robots, Emotions, and Learning'. In: *Emotions, Technology, Design, and Learning*. Ed. by Sharon Y. Tettegah and Martin Gartmeier. Emotions and Technology. San Diego: Academic Press, pp. 183 – 197. ISBN: 978-0-12-801856-9.

Viola, Paul and Michael Jones (2001). 'Robust Real-time Object Detection'. In: *International Journal of Computer Vision*.

Viola, Paul and Michael J Jones (2004). 'Robust real-time face detection'. In: *International journal of computer vision* 57.2, pp. 137–154.

Wagner, Johannes et al. (2013). 'The Social Signal Interpretation (SSI) Framework: Multimodal Signal Processing and Recognition in Real-time'. In: *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13. Barcelona, Spain: ACM, pp. 831–834. ISBN: 978-1-4503-2404-5.

Warburton, Darren ER et al. (2007). 'The health benefits of interactive video game exercise'. In: *Applied Physiology, Nutrition, and Metabolism* 32.4, pp. 655–663.

Watson, John B. (1919). 'A schematic outline of the emotions.' In: *Psychological Review* 26.3, pp. 165–196.

Watson, John Broadus (1930). 'Behaviorism, rev'. In:

Weber, Raphael, Catherine Soladie, and Renaud Seguier (2018). 'A survey on databases for facial expression analysis'. In: *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. VISAPP*. Funchal, Portugal.

Weiner, Bernard and Sarah Graham (1984). 'An attributional approach to emotional development'. In: *Emotions, cognition, and behavior*, pp. 167–191.

Weiser, Mark (1995). 'The computer for the 21st century: specialized elements of hardware and software, connected by wires, radio waves and infrared, will be so ubiquitous that no one will notice their presence'. In: *Readings in Human–Computer Interaction*. Elsevier, pp. 933–940.

Wells, Laura Jean, Steven Mark Gillespie, and Pia Rotshtein (2016). 'Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze'. In: *PloS one* 11.12, e0168307.

Wilkinson, P. (2013). 'Affective educational games: Utilizing emotions in game-based learning'. In: *2013 5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, pp. 1–8.

Woodworth, Robert Sessions and Harold Schlosberg (1954). *Experimental psychology*. Oxford and IBH Publishing.

Wu, Chung-Hsien, Jen-Chun Lin, and Wen-Li Wei (2014). 'Survey on audiovisual emotion recognition: databases, features, and data fusion strategies'. In: *APSIPA Transactions on Signal and Information Processing* 3.

Wundt, Wilhelm (1896). *Outlines of Psychology*. Leipzig:Engelmann.

Xuan, Xiang and Kevin Murphy (2007). 'Modeling Changing Dependency Structure in Multivariate Time Series'. In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: ACM, pp. 1055–1062. ISBN: 978-1-59593-793-3.

Yan, Wen-Jing et al. (2013). 'How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions'. In: *Journal of Nonverbal Behavior* 37.4, pp. 217–230. ISSN: 1573-3653.

Yarkoni, Tal and Jacob Westfall (2017). 'Choosing prediction over explanation in psychology: Lessons from machine learning'. In: *Perspectives on Psychological Science* 12.6, pp. 1100–1122.

Yin, L. et al. (2008). 'A high-resolution 3D dynamic facial expression database'. In: *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6.

Yin, Lijun et al. (2006). 'A 3D facial expression database for facial behavior research'. In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 211–216.

Zaalberg, Ruud, Antony Manstead, and Agneta Fischer (2004). 'Relations between emotions, display rules, social motives, and facial behaviour'. In: *Cognition and Emotion* 18.2, pp. 183–207.

Zafeiriou, Stefanos, Cha Zhang, and Zhengyou Zhang (2015). 'A survey on face detection in the wild: past, present and future'. In: *Computer Vision and Image Understanding* 138, pp. 1–24.

Zagaria, Sebastian (2017). 'Emotional adq̈uat reagierende KI-Agenten zur Erhöhung von-Immersion in Computerspielen'. MA thesis. HAW Hamburg: Faculty of Engineering and Computer Science.

Zeng, Z. et al. (2009). 'A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.1, pp. 39–58. ISSN: 0162-8828.

Zeng, Zhihong et al. (2006). 'One-class classification for spontaneous facial expression analysis'. In: pp. 281–286. ISBN: 0-7695-2503-2.

Zeng, Zhihong et al. (2007). 'Audio-visual spontaneous emotion recognition'. In: *Artifical intelligence for human computing*. Springer, pp. 72–90.

Zhang, Ligang et al. (2018). 'Facial Expression Analysis under Partial Occlusion: A Survey'. In: *ACM Computing Surveys (CSUR)* 51.2, p. 25.

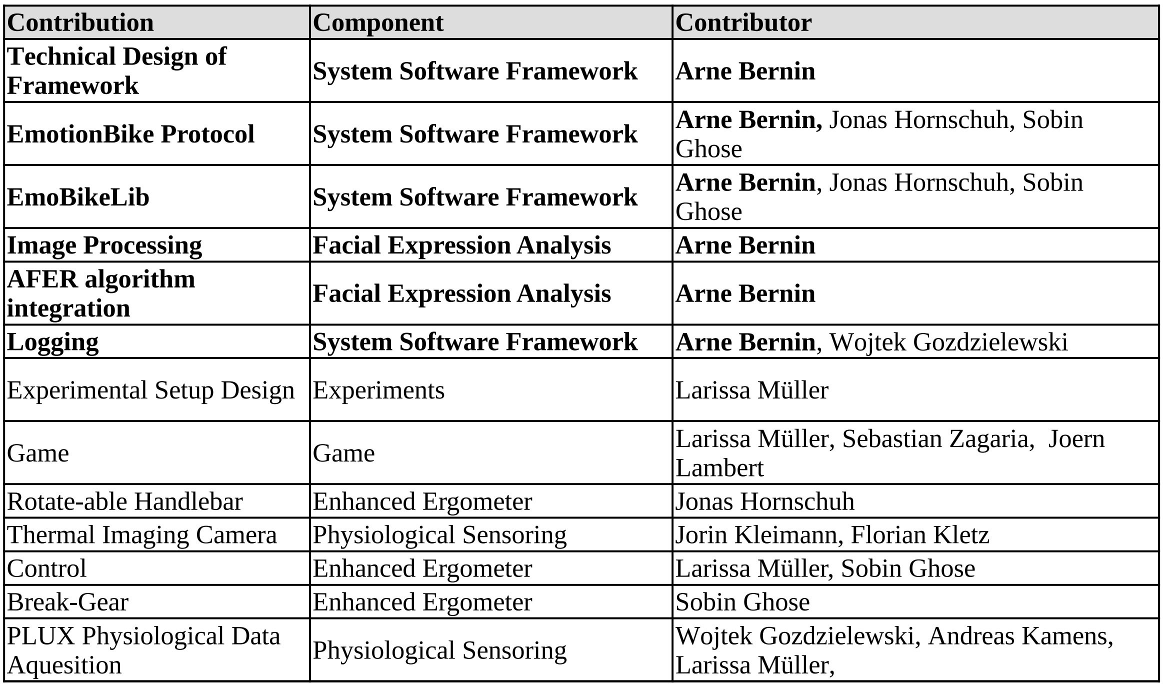
Zhang, Shen et al. (2010). *Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar*. Vol. 2010, pp. 109–132.

Zhu, Xiangxin and Deva Ramanan (2012). 'Face detection, pose estimation, and landmark localization in the wild'. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 2879–2886.

Appendices

A. Contributions to the EmotionBike Project

As a joint research project, many people contributed to the overall project. Table **??** lists the contributors and their contribution.



Contribution	**Component**	**Contributor**
Technical Design of Framework	**System Software Framework**	**Arne Bernin**
EmotionBike Protocol	**System Software Framework**	**Arne Bernin,** Jonas Hornschuh, Sobin Ghose
EmoBikeLib	**System Software Framework**	**Arne Bernin**, Jonas Hornschuh, Sobin Ghose
Image Processing	**Facial Expression Analysis**	**Arne Bernin**
AFER algorithm integration	**Facial Expression Analysis**	**Arne Bernin**
Logging	**System Software Framework**	**Arne Bernin**, Wojtek Gozdzielewski
Experimental Setup Design	Experiments	Larissa Müller
Game	Game	Larissa Müller, Sebastian Zagaria, Joern Lambert
Rotate-able Handlebar	Enhanced Ergometer	Jonas Hornschuh
Thermal Imaging Camera	Physiological Sensoring	Jorin Kleimann, Florian Kletz
Control	Enhanced Ergometer	Larissa Müller, Sobin Ghose
Break-Gear	Enhanced Ergometer	Sobin Ghose
PLUX Physiological Data Aquesition	Physiological Sensoring	Wojtek Gozdzielewski, Andreas Kamens, Larissa Müller,

Table A.1: Contributions to the EmotionBike project with the contributions by the author **marked**.

B. Emotional Shift Detection

B.1. Software Libraries Applied

The peakdetect Python package[1] by Sixten Bergman and Bayesian changepoint detection implemented by Johannes Kulick[2].

B.2. Shift Detection Algorithms

B.2.1. PMP Algorithm

```
def alg_pmp_x(data,mfsize=8,blocksize=4, crossthreshold=0.2):
    # generate pattern matching mask for rising edges,
    # will also be used for falling ([-1-,1-,1,1])
    mask=createFilter(mfsize)
    # use our correlate function
    bcorrrise, bcorrfall=matchCorrelate(data, mask)
    # get rising and falling temp indexes with half theshold
    risingIndexesTmp = peakutils.indexes(bcorrrise, thres=crossthreshold/2,
                                         min_dist=blocksize)
    fallingIndexesTmp = peakutils.indexes(bcorrfall, thres=crossthreshold/2,
                                          min_dist=blocksize)
```

[1] https://gist.github.com/sixtenbe/1178136

[2] https://github.com/hildensia/bayesian_changepoint_detection

```

    # recheck if value really > threshold
    risingIndexes=[]
    fallingIndexes=[]
    for x in risingIndexesTmp:
        if(bcorrrise[x] > crossthreshold):
            risingIndexes.append(x)
    for x in fallingIndexesTmp:
        if(bcorrfall[x] > crossthreshold):
            fallingIndexes.append(x)
    # return both
    return(risingIndexes,fallingIndexes)
```

B.2.2. Correlation for PMP

```
'''
matchCorrelate(a,v) data (a) with filter (v) and return positive and negative
correlation both cut to 1.0.
'''
def matchCorrelate(a,v):
    # be sure we have a numpy array as arguments for np.correlate later
    a=numpy.array(a)
    v = numpy.array(v)
    # fix shape for arguments
    a=numpy.reshape(a, len(a))
    v=numpy.reshape(v, len(v))
    # input is between 0.0 and 1.0 (propabilities and filter)
    # spread to -1. and 1 for correct correlation
    a = (a *2 -1)
    v = (v *2 -1)

    # use mode 'same' for return list should have same length as 'a'
    mode='same'
    # use numpy correlate
    out = numpy.correlate(a, v, mode)
    # use numpy correlate, make self correlation to mask to get maximum value
    # from https://docs.scipy.org/doc/numpy-1.9.0/reference/
    #                                       generated/numpy.correlate.html
    # c_{av}[k] = sum_n a[n+k] * conj(v[n])
    selfcorr=numpy.correlate(v, v, mode)
    val=max(numpy.abs(min(selfcorr)), numpy.abs(max(selfcorr)))
    # be sure to get a float
    val1=0.0 + (1/val)
    # normalize output to max val
    out=(out * val1)
    # generate pos and neg return arrays as empty
    pos=numpy.array(list(out))
    neg=numpy.array(list(out))
    # cut pos and neg to values between 0 and 1(pos) and 0 and -1 (neg)
    for i in range(0, len(out)):
        if out[i] < 0.0:
            pos[i] = 0.0
        if out[i] > 0.0:
            neg[i] = 0.0
    for i in range(0, len(out)):
        neg[i] = -1.0 * neg[i]
        # return both arrays
    return (pos,neg)
```

B.3. CP Algorithms

B.3.1. CP Enhanced with Peak Detection

```
def alg_cp_peak(data, trunc = -20, blocksize=4):
        risingIndexes = []
        fallingIndexes = []
        # smooth data for better results
        sdata=smooth(data,blocksize=blocksize)
        # call standard cp detection
        Q_full, P_full, Pcp_full = offcd.offline_changepoint_detection(sdata,
        partial(offcd.const_priorab, l=(len(data)+ 1)),
        offcd.gaussian_obs_log_likelihood, truncate=trunc)
        # get full data        showing likelyhood of change at this point
        cb = np.exp(Pcp_full).sum(0)
        # Enhancement: Peak detection
        #define threshold
        cpthreshold = 0.01
        # find peaks with above threshold (half of cpthreshold)
        rfindexes = peakutils.indexes(cb, thres=cpthreshold / 2, min_dist=1)
        # recheck all for value above threshold
        for x in rfindexes:
                x=int(x)
                if (cb[x] > cpthreshold):
                        # values left, values right of peak
                        left = (data[x - 1] + data[x - 2]) / 2
                        right = (data[x + 1] + data[x + 2]) / 2
                # we have a falling or a rising index?
                if left > right:
                        fallingIndexes.append(x)
                else:
                        risingIndexes.append(x)

        return (fallingIndexes, risingIndexes)
```

B.3.2. Original CP algorithm

```
def offline_changepoint_detection(data, prior_func,
                                  observation_log_likelihood_function,
                                  truncate=-np.inf):
    """Compute the likelihood of changepoints on data.

    Keyword arguments:
    data                                -- the time series data
    prior_func                          -- a function given the likelihood of a chang
    observation_log_likelihood_function -- a function giving the log likelihood
                                           of a data part
    truncate                            -- the cutoff probability 10^truncate to stop

    P                                   -- the likelihoods if pre-computed
    """

    n = len(data)
    Q = np.zeros((n,))
    g = np.zeros((n,))
    G = np.zeros((n,))
    P = np.ones((n, n)) * -np.inf

    # save everything in log representation
    for t in range(n):
        g[t] = np.log(prior_func(t))
        if t == 0:
            G[t] = g[t]
        else:
            G[t] = np.logaddexp(G[t-1], g[t])

#   print('G',G)

    P[n-1, n-1] = observation_log_likelihood_function(data, n-1, n)
    Q[n-1] = P[n-1, n-1]

    for t in reversed(range(n-1)):
        P_next_cp = -np.inf  # == log(0)
        for s in range(t, n-1):
            P[t, s] = observation_log_likelihood_function(data, t, s+1)

            # compute recursion
            summand = P[t, s] + Q[s + 1] + g[s + 1 - t]
            P_next_cp = np.logaddexp(P_next_cp, summand)

            # truncate sum to become approx. linear in time (see
            # Fearnhead, 2006, eq. (3))
            if summand - P_next_cp < truncate:
                break

        P[t, n-1] = observation_log_likelihood_function(data, t, n)

        # (1 - G) is numerical stable until G becomes numerically 1
        if G[n-1-t] < -1e-15:  # exp(-1e-15) = .99999...
            antiG = np.log(1 - np.exp(G[n-1-t]))
        else:
            # (1 - G) is approx. -log(G) for G close to 1

            antiG = np.log(-G[n-1-t])

        Q[t] = np.logaddexp(P_next_cp, P[t, n-1] + antiG)

    Pcp = np.ones((n-1, n-1)) * -np.inf
    for t in range(n-1):
        Pcp[0, t] = P[0, t] + Q[t + 1] + g[t] - Q[0]
        if np.isnan(Pcp[0, t]):
            Pcp[0, t] = -np.inf
    for j in range(1, n-1):
        for t in range(j, n-1):
            tmp_cond = Pcp[j-1, j-1:t] + P[j:t+1, t] + Q[t + 1] + g[0:t-j+1] \
                       - Q[j:t+1]
            Pcp[j, t] = logsumexp(tmp_cond.astype(np.float32))
            if np.isnan(Pcp[j, t]):
                Pcp[j, t] = -np.inf

    return Q, P, Pcp
```

B.3.3. Gaussian Log Likelihood for CP

```
def gaussian_obs_log_likelihood(data, t, s):
        s += 1
        n = s - t
        mean = data[t:s].sum(0) / n

        muT = (n * mean) / (1 + n)
        nuT = 1 + n
        alphaT = 1 + n / 2
        betaT = 1 + 0.5 * ((data[t:s] - mean) ** 2).sum(0) + ((n)/(1 + n)) \
                * (mean**2 / 2)
        scale = (betaT*(nuT + 1))/(alphaT * nuT)

        # splitting the PDF of the distribution up is /much/ faster.
        # (~ factor 20) using sum over for loop is even more worthwhile
        prob = np.sum(np.log(1 + (data[t:s] - muT)**2/(nuT * scale)))
        lgA = gammaln((nuT + 1) / 2) - np.log(np.sqrt(np.pi * nuT * scale))\
              - gammaln(nuT/2)
```

```python

return np.sum(n * lgA - (nuT + 1)/2 * prob)
```

B.3.4. Classifying edges to shift for PMP and CP

```python
def classifyEdgedata(riseindexes, fallingindexes, data):
    '''
    Function to categorise edges
    :param riseindexes: indices of rising edges
    :param fallingindexes: indices of falling edges
    :param data: data
    :return: classification
    '''
    category='??'
    if (len(riseindexes) == 1 and len(fallingindexes) == 0):
        category='01'
    elif (len(riseindexes) == 2 and len(fallingindexes) == 0):
            category = '01'
    elif (len(riseindexes) == 0 and len(fallingindexes) == 1):
        category='10'
    elif (len(riseindexes) == 0 and len(fallingindexes) == 2):
        category = '10'
    elif (len(riseindexes) == 1 and len(fallingindexes) == 1):
        if (riseindexes[0] < fallingindexes[0]):
            category='010'
        else:
            category = '101'
    elif (len(riseindexes) == 2 and len(fallingindexes) == 2):
            # case: 01010
            if (riseindexes[0] < fallingindexes[0] and fallingindexes[0]
                    < riseindexes[1] and riseindexes[1] < fallingindexes[1]):
                category = '101'
    elif (len(riseindexes) == 2 and len(fallingindexes) == 1):
                # case: 01010
                if (riseindexes[0] < fallingindexes[0] and fallingindexes[0]
                        < riseindexes[1]):
                    category = '101'
    elif (len(riseindexes) == 1 and len(fallingindexes) == 2):
        if (riseindexes[0] < fallingindexes[0] and fallingindexes[0]
                < fallingindexes[1]):
            category = '010'
    elif (len(riseindexes) == 0 and len(fallingindexes) == 0):
        mean=calcMean(data, 0, len(data))
        if (mean > 0.5):
            category = '11'
        else:
            category = '00'
    return category
```

B.3.5. FWMB algorithm for categorisation

```
def alg_fixed_win_mean_bisection(data, ax, threshold=0.5, diffthreshold=0.5,
                                 scansize=0):
    # algorithm with fixed window size using mean on left and right side of
    # event (in the middle)
    allmax = []
    allmin = []
    # default, nothing found
    category = '00'
    # first try: partition in left and right, get mean and see if mean diff is
    # above threshold
    # split data
    dleft = data[0:int(len(data) / 2)]
    dright = data[int(len(data) / 2):int(len(data))]
    # calc mean for every side
    dleftmean = calcMean(dleft, 0, len(dleft))
    drightmean = calcMean(dright, 0, len(dright))
    # get diff mean(pre) -mean(post)
    diffmean = dleftmean - drightmean
    # indices for plotting
    inds = []
    inds.append(int(len(data) / 2))
    # 1: check if we have simple case:
    if (diffmean > diffthreshold):
        # leftmean is greater than rightmean + threshold -> 10
        category = '10'
```

```python
    elif (diffmean < (-1 * diffthreshold)):
        # rightmean is greater than leftmean + threshold -> 01
        category = '01'
    elif (dleftmean > threshold and drightmean > threshold):
        # both sides above threshold -> 11
        category = '11'
    elif (dleftmean < threshold and drightmean > threshold):
        # dleftmean is smaller than threshold and right is above, but difference
        # is < threshold
        # get positions for second quarter of data (1/4 before event):
        half = int(len(data) / 2)
        hhalf = int(len(data) / 4)
        # get peakds over complete data but with half threshold
        # (easier for peak detection)
        (mins, maxs) = myPeakdetect(data, threshold / 2)
        # only look at minima
        for min in mins:
            if (min[0] >= hhalf and min[0] <= half):
                # minimum ist in second quarter of data, so before event,
                # meaning: there is a minimum before event, assume a valley
                category = '01'
                inds.append(hhalf)
    elif (dleftmean > threshold and drightmean < threshold):
        # dleftmean is greater than threshold and rightmean is below,
        # but difference is < threshold
        # get positions for third quarter of data (1/4 after event):
        half = len(data) / 2
        hhalf = len(data) * 3 / 4
        # get peakds over complete data but with half threshold
        # (easier for peak detection)
        (mins, maxs) = myPeakdetect(data, threshold / 2)
        # only look at mins
        for min in mins:
            if (min[0] >= half and min[0] <= hhalf):
                # min in front of our max side
                inds.append(hhalf)
                category = '10'
    else:
        # second try: use a smaller mean window and scan for maximum or minimum
        # means if not defined as argument, use half pre or post size
        if (scansize == 0):
            scansize = int(len(dleft) / 2)
        dleftall = []
        for i in range(len(dleft) - scansize):
            dleftall.append(calcMean(dleft, i, i + scansize))
        drightall = []
        for i in range(len(dright) - scansize):
            drightall.append(calcMean(dright, i, i + scansize))

        diffmaxmean = max(dleftall) - max(drightall)
        if (diffmaxmean > threshold):
            category = '10'
        elif (diffmaxmean < (-1 * threshold)):
            category = '01'
        else:
            # third try, search for peaks with different heights
            peakheights = [threshold / 2]

            # set last data point to zero for peak detection lookahead at end
            # of data,we need values lower the threshold for this
            # so make a copy and add 0.0
            peakdata = list(data)
            peakdata.append(0.0)
            peakdata.append(0.0)

            # and start with 0.0
            peakdata.insert(0, 0.0)
            peakdata.insert(0, 0.0)
            # scan for height, if no success, try smaller one
            for height in peakheights:
                # peakdetection for noisy data by
                # http://billauer.co.il/peakdet.html: peakdet
                _max, _min = peakdetect.peakdetect(peakdata, range(len(peakdata))
                                                   , lookahead=1, delta=height)
                allmax.extend(_max)
                allmin.extend(_min)
                lmax = []
                rmax = []
                for p in _max:
                    if (p[1] > threshold):
                        if p[0] < len(data) / 2:
                            lmax.append(p[0])
                        else:
                            rmax.append(p[0])
                if (len(lmax) > 0 and len(rmax) == 0):
                    category = '10'
                elif (len(lmax) == 0 and len(rmax) > 0):
                    category = '01'
    return category
```