TEENA CHAKKALAYIL HASSAN

TOWARDS ROBUST AND INTERPRETABLE PRACTICAL APPLICATIONS OF AUTOMATIC MENTAL STATE ANALYSIS USING A DYNAMIC AND HYBRID FACIAL ACTION ESTIMATION APPROACH

TOWARDS ROBUST AND INTERPRETABLE PRACTICAL APPLICATIONS OF AUTOMATIC MENTAL STATE ANALYSIS USING A DYNAMIC AND HYBRID FACIAL ACTION ESTIMATION APPROACH

TEENA CHAKKALAYIL HASSAN

Submitted in partial fulfilment of the requirements for the degree of Doctor of Natural Sciences (Dr. rer. nat.) of the University of Bamberg

Advisor and Reviewer:
Prof. Dr. Ute Schmid

External Reviewer:
Prof. Dr. Jürgen Beyerer
Karlsruhe Institute of Technology

Further Committee Members:
Prof. Dr. Diedrich Wolter
Prof. Dr. Daniela Nicklas

Submitted: February 24, 2020

Defended: May 27, 2020

TOWARDS ROBUST AND INTERPRETABLE PRACTICAL APPLICATIONS OF AUTOMATIC MENTAL STATE ANALYSIS USING A DYNAMIC AND HYBRID FACIAL ACTION ESTIMATION APPROACH

TEENA CHAKKALAYIL HASSAN

Fraunhofer IIS

Intelligent Systems Group
Electronic Imaging Department
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Germany

URN: urn:nbn:de:bvb:473-irb-486414
DOI: https://doi.org/10.20378/irb-48641

Manuscript prepared using `classicthesis` LaTeX template version 4.6 available at https://ctan.org/pkg/classicthesis under the GNU General Public License version 2 or newer.

Final version for online publication edited and compiled on September 17, 2020.

'Like, the joyous, sprightly spring,
Forever follows an icy winter streak,
When things are going not quite well,
Wait! Life flips, sooner than we can tell.'

From the poem *Braving the Weather*

— Myself

Dedicated to the cosmos that destined and acted through its elements and agents to make this happen.

ABSTRACT

Facial expressions constitute one of the main channels through which humans convey a rich variety of non-verbal cues that facilitate communication and interaction with other humans. Affective computing systems usually analyse human facial expressions in order to recognise the affective, mental, or psychophysiological states of humans. This has the potential to enhance human-machine interaction as well as facilitate the development of assistance systems for improving the quality of life of humans. The Facial Action Coding System (FACS) is used by psychologists as the standard for describing facial expressions objectively in terms of their constituent facial muscle movements, known as Action Units (AUs). Analysis of the semantics and nuances of facial expressions is then performed in terms of AUs.

Computer vision researchers have developed several approaches for automatically recognising AUs and their intensities from facial images or videos. One category of approaches focuses on using data-driven machine learning methods to detect AUs based on patterns in visual input data. Another category of approaches focuses on using deformable face models that describe AUs semantically in terms of the facial shape deformations that they cause. While the former category of approaches can attain good predictive performance by learning robust patterns covering large variance in input training data, the latter category of approaches facilitates interpretability by virtue of using AU-based deformable face models. Therefore, a combination of both categories of approaches could help in building interpretable systems for automatic facial expression analysis with good predictive performance.

This dissertation presents a probabilistic state estimation framework for integrating data-driven machine learning models and a deformable facial shape model in order to estimate continuous-valued intensities of 22 different AUs. A practical approach is proposed and validated for integrating class-wise probability scores from machine learning models within a Gaussian state estimation framework. Furthermore, driven mass-spring-damper models are applied for modelling the dynamics of facial muscle movements. Both facial shape and appearance information are used for estimating AU intensities, making it a hybrid approach.

Several features are designed and explored to help the probabilistic framework to deal with multiple challenges involved in automatic AU detection. On the human front, these features calibrate the person-specific facial shape and appearance, and enable adaptation to the viscoelastic properties of different facial muscles. On the technical

front, these features (i) deal with the similarities in facial shape deformations caused by various AUs, (ii) handle the disproportional shape deformations caused by subtle and pronounced AUs, and (iii) confine the estimated AU intensities to a valid range. On the practical front, these features enhance robustness by handling missing or anomalous information.

The proposed AU intensity estimation method and its features are evaluated quantitatively and qualitatively using three different datasets containing either spontaneous or acted facial expressions with AU annotations. The proposed method produced temporally smoother estimates that facilitate a fine-grained analysis of facial expressions. It also performed reasonably well, even though it simultaneously estimates intensities of 22 AUs, some of which are subtle in expression or resemble each other closely. The estimated AU intensities tended to the lower range of values, and were often accompanied by a small delay in onset. This shows that the proposed method is conservative. In order to further improve performance, state-of-the-art machine learning approaches for AU detection could be integrated within the proposed probabilistic AU intensity estimation framework.

In addition to AU intensity estimation, this dissertation explores the applicability of the estimated AU intensities for automatic analysis of mental states such as pain and distraction. A survey of automatic pain detection approaches, conducted as part of this dissertation, highlights the progress and deficits in this field. Several AU-based rules are designed for pain intensity estimation based on psychological evidence, and their performance is evaluated empirically. The potential of these AU-based rules to automatically generate explanations for pain detections is also illustrated. Furthermore, a preliminary analysis of the estimated AU intensities shows differences in facial actions between various distraction scenarios during simulated driving.

Facial expressions are not the only channel through which mental states are expressed. Physiological changes also accompany changes in mental states. However, these physiological changes vary between persons, and are influenced by a multitude of other factors. This dissertation presents some initial efforts made towards dealing with these challenges. The results of this dissertation show that more interdisciplinary research is needed to address the open challenges in the field of automatic mental state analysis, particularly to build reference datasets, to model interpersonal differences, and to generate human-comprehensible explanations of predictions.

ZUSAMMENFASSUNG

Mimik ist einer der wichtigsten Kanäle, über die Menschen eine Vielzahl von nonverbalen Signalen vermitteln, die die Kommunikation und Interaktion mit anderen Menschen erleichtern. Affektive Computersysteme analysieren in der Regel menschlichen Gesichtsausdrücke, um die affektiven, mentalen oder psychophysiologischen Zustände von Menschen zu erkennen. Dies hat das Potenzial sowohl die Mensch-Maschine-Interaktion zu verbessern als auch die Entwicklung von Assistenzsystemen zur Verbesserung der Lebensqualität von Menschen zu ermöglichen. Das Facial Action Coding System (FACS) wird von Psychologen als Standard zur objektiven Beschreibung von Gesichtsausdrücken auf Basis von konstituierenden Gesichtsmuskelbewegungen verwendet, die als Action Units (AUs) bezeichnet werden. Die Analyse der Semantik und der Nuancen von Gesichtsausdrücken wird dann mithilfe von AUs durchgeführt.

Forscher im Bereich des maschinellen Sehens haben mehrere Ansätze entwickelt, um AUs und ihre Intensitäten automatisch aus Gesichtsbildern oder Videos zu erkennen. Eine Gruppe von Ansätzen konzentriert sich auf den Einsatz datengetriebener Methoden des maschinellen Lernens, um AUs auf der Grundlage von Mustern im visuellen Eingabedaten zu erkennen. Eine andere Gruppe von Ansätzen konzentriert sich auf die Verwendung von veränderbaren Gesichtsmodellen, die AUs semantisch, in Bezug auf Gesichtsformveränderungen, beschreiben. Während die erste Gruppe von Ansätzen eine gute Vorhersageleistung durch das Lernen robuster Muster erreichen kann, die eine große Varianz in den Trainingsdaten abdecken, erleichtert die zweite Gruppe von Ansätzen die Interpretierbarkeit durch die Verwendung von AU-basierten veränderbaren Gesichtsmodellen. Daher würde eine Kombination beider Kategorien von Ansätzen dabei helfen, interpretierbare Systeme für die automatische Gesichtsausdrucksanalyse mit guter Vorhersageleistung zu bauen.

Diese Dissertation präsentiert ein probabilistisches Framework zur Zustandsschätzung, das datengetriebene Modelle des maschinellen Lernens und eines veränderbaren Gesichtsformmodells integrieren, um kontinuierlicher Intensitäten von 22 verschiedenen AUs zu schätzen. Es wird ein praktischer Ansatz zur Integration von Wahrscheinlichkeiten aus Klassifikatoren innerhalb eines Gaußschen Zustandsschätzungs-Frameworks vorgeschlagen und validiert. Darüber hinaus werden Masse-Feder-Dämpfer-Modelle, die durch äußere Kraft angetrieben werden, zur Modellierung der Dynamik von Gesichtsmuskelbewegungen eingesetzt. Sowohl Gesichtsform als auch Texturmerkmale werden zur Schätzung der AU-Intensitäten verwendet, so dass es sich um einen hybriden Ansatz handelt.

Es wurden mehrere Funktionen entwickelt und untersucht, um die verschiedenen Herausforderungen bei der automatischen Erkennung der AUs zu bewältigen. Erstens wurden Lösungen entwickelt, die die personenspezifische Gesichtsform und das Aussehen kalibrieren und die Anpassung an die viskoelastischen Eigenschaften verschiedener Gesichtsmuskeln ermöglichen. Zweitens wurden Lösungen konzipiert, die (i) die Ähnlichkeiten der durch verschiedene AUs verursachten Gesichtsformveränderungen behandeln, (ii) die durch subtile und markante AUs verursachten disproportionalen Formveränderungen berücksichtigen und (iii) die geschätzten AU-Intensitäten auf einen gültigen Bereich beschränken. Drittens wurden Lösungen entworfen, die fehlende oder anomale Informationen robust behandeln.

Die vorgeschlagene Methode zur Schätzung der AU-Intensitäten wird quantitativ und qualitativ unter Verwendung von drei verschiedenen Datensätzen bewertet, die entweder spontane oder gespielte Gesichtsausdrücke mit AU-Annotationen enthalten. Die vorgeschlagene Methode führte zu zeitlich glatteren Schätzungen, die eine feinkörnige Analyse der Gesichtsausdrücke ermöglichen. Sie hat auch eine recht gute Leistung erbracht, obwohl sie gleichzeitig Intensitäten von 22 AUs schätzt, von denen einige im Ausdruck subtil oder einander sehr ähnlich sind. Die geschätzten AU-Intensitäten tendierten dazu, in einem relativ niedrigen Wertebereich zu bleiben, und wiesen oft einen etwas verzögerten Beginn auf. Dies zeigt, dass die vorgeschlagene Methode konservativ ist. Um die Leistung weiter zu verbessern, könnten die modernsten Ansätze des maschinellen Lernens zur AU-Erkennung in den vorgeschlagenen probabilistischen Framework zur Schätzung der AU-Intensitäten integriert werden.

Zusätzlich zur Schätzung der AU-Intensitäten, wird in dieser Dissertation auch die Anwendbarkeit der geschätzten AU-Intensitäten für die automatische Analyse von mentalen Zuständen wie Schmerz und Ablenkung untersucht. Ein im Rahmen dieser Dissertation durchgeführter Survey über die Ansätze zur automatischen Schmerzerkennung zeigt die Fortschritte und Defizite in diesem Bereich auf. Mehrere AU-basierte Regeln werden für die Abschätzung der Schmerzintensität auf der Grundlage psychologischer Evidenz konzipiert und empirisch bewertet. Das Potenzial dieser AU-basierten Regeln zur automatischen Generierung von Erklärungen für die Schmerzerkennung wird ebenfalls dargestellt. Darüber hinaus zeigt eine vorläufige Analyse der geschätzten AU-Intensitäten, dass es Unterschiede in den Gesichtsausdrücken zwischen verschiedenen Ablenkungsszenarien beim simulierten Autofahren gibt.

Die Mimik ist nicht der einzige Kanal, durch den mentale Zustände ausgedrückt werden. Physiologische Veränderungen gehen auch mit Veränderungen der mentalen Zustände einher. Diese physiologischen Veränderungen sind jedoch von Person zu Person unterschiedlich und werden von einer Vielzahl anderer Faktoren beeinflusst. In die-

ser Dissertation werden erste Arbeitsansätze zur Bewältigung dieser Herausforderungen vorgestellt. Die Ergebnisse dieser Dissertation zeigen, dass mehr interdisziplinäre Forschung erforderlich ist, um die Herausforderungen auf dem Gebiet der automatischen Analyse der mentalen Zustände anzugehen, insbesondere um Referenzdatensätze zu erstellen, zwischenmenschliche Unterschiede zu modellieren und für den Menschen verständliche Erklärungen zu generieren.

PUBLICATIONS

I. The following publications contain some of the scientific contributions, visualisations and results produced as part of this doctoral research:

- **Teena Hassan**, Dominik Seuß, Johannes Wollenberg, Katharina Weitz, Miriam Kunz, Stefan Lautenbacher, Jens-Uwe Garbas, and Ute Schmid. "Automatic Detection of Pain from Facial Expressions: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–17. DOI: 10.1109/TPAMI.2019.2958341.
- Miriam Kunz, Dominik Seuss, **Teena Hassan**, Jens U. Garbas, Michael Siebers, Ute Schmid, Michael Schöberl, and Stefan Lautenbacher. "Problems of video-based pain detection in patients with dementia: a road map to an interdisciplinary solution." In: *BMC Geriatrics* 17.33 (2017). DOI: 10.1186/s12877-017-0427-2.
- **Teena Hassan**, Dominik Seuss, Johannes Wollenberg, Jens Garbas, and Ute Schmid. "A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework." In: *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August - 2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. Frontiers in Artificial Intelligence and Applications. The Hague, The Netherlands: IOS Press, 2016, pp. 1812–1817. DOI: 10.3233/978-1-61499-672-9-1812.
- **Teena Hassan**, Dominik Seuß, Andreas Ernst, and Jens Garbas. "A Kalman Filter with State Constraints for Model-based Dynamic Facial Action Unit Estimation." In: *Forum Bildverarbeitung 2018*. Ed. by Thomas Längle, Fernando Puente León, and Michael Heizmann. KIT Scientific Publishing, 2018. DOI: 10.5445/KSP/1000085290.
- Dominik Seuss, Anja Dieckmann, **Teena Hassan**, Jens-Uwe Garbas, Johann Heinrich Ellgring, Marcello Mortillaro, and Klaus Scherer. "Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array." In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 35–41. DOI: 10.1109/ACII.2019.8925458.
- Bhargavi Mahesh, **Teena Hassan**, Erwin Prassler, and Jens-Uwe Garbas. "Requirements for a Reference Dataset for

Multimodal Human Stress Detection." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 492–498. DOI: 10.1109/PERCOMW.2019.8730884.

- Pelin Genc and **Teena Hassan**. "Analysis of Personality Dependent Differences in Pupillary Response and its Relation to Stress Recovery Ability." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 505–510. DOI: 10.1109/PERCOMW.2019.8730779.
- Katharina Weitz, **Teena Hassan**, Ute Schmid, and Jens-Uwe Garbas. "Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods." In: *tm-Technisches Messen* 86.7-8 (2019), pp. 404–412. DOI: 10.1515/teme-2019-0024.

II. The following preprints are planned for submission and contain some of the scientific contributions, visualisations and results produced during this doctoral work:

- Martin Gjoreski, Matjaž Gams, Mitja Luštrek, Pelin Genc, Jens-Uwe Garbas, and **Teena Hassan**. "Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals." Preprint. (*Update 02.08.2020:* Published in IEEE Access in April 2020.)
- **Teena Hassan**, Dominik Seuss, Jens Garbas, and Ute Schmid. "Automatic Facial Action Unit Detection for Psychological Research: Comparison between a Data-Driven and a Probabilistic Approach." Preprint.

III. The following patent also contains some of the ideas generated and tested as part of this doctoral research:

- Determining Facial Parameters, by Dominik Seuss, **Teena Chakkalayil Hassan**, Johannes Wollenberg, Andreas Ernst, and Jens-Uwe Garbas. (2019, Apr. 30). *Patent* US 10,275,640 B2. Accessed on: Jan. 26, 2020. [Online]. Available: USPTO PatFT Databases.

The scientific and written contents contributed by me in each of these publications, preprints and patent are listed in detail in the Appendices B, C and D.

Update 02.08.2020: The preprint Gjoreski et al. has been published in IEEE Access in April 2020. The full reference is given below:

Martin Gjoreski, Matjaž Gams, Mitja Luštrek, Pelin Genc, Jens-Uwe Garbas, and **Teena Hassan**. "Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals." In: *IEEE Access* 8 (2020), pp. 70590–70603. DOI: 10.1109/ACCESS.2020.2986810.

"When I started counting my blessings,
my whole life turned around."

— Willie Nelson

ACKNOWLEDGEMENTS

This thesis is a result of the continuous encouragement and support that I received from a wide network of people. First of all, I thank Fraunhofer IIS, for the opportunity to do this thesis within the premises of an industry-funded project. This was an enriching and at the same time challenging experience, whereby I learned to balance the demands of research and practical applicability. I thank Prof. Dr. Ute Schmid, my doctoral advisor at University of Bamberg, for accepting me as an external doctoral candidate and for the crucial role that she played during the entire doctoral phase. She guided me, always with my best interests in mind. I cannot thank her enough, for supporting me, when my self-confidence took a blow. I also thank Prof. Dr. Jürgen Beyerer, Director of Fraunhofer IOSB and Professor at Karlsruhe Institute of Technology, for his kind consent to write the second review for my thesis, and for sharing important suggestions and research relevant to this topic. I also thank Prof. Dr. Daniela Nicklas and Prof. Dr. Diedrich Wolter from University of Bamberg, for their willingness to be part of the doctoral committee.

Fraunhofer IIS played an instrumental role in the initial phase of my research career and throughout this doctoral work. I thank Dr. Jens-Uwe Garbas, then Head of Intelligent Systems Group, for believing in my potential and for giving me opportunities that were sometimes above the normal profile of a PhD student. His strategic vision opened up surprising, new research paths for me, which helped me uncover my hidden scientific skills and encouraged me to take important research initiatives. I also thank my colleagues Andreas Ernst, Dominik Seuß, Johannes Wollenberg and Sebastian Hettenkofer for all the intellectually stimulating discussions, for their collegiality, and for the support they gave me during personally difficult times. Many thanks to Dr. Nadine Lang, my mentor at Fraunhofer IIS, under whose guidance and grooming, I learned the basics about writing major national research project proposals. A very special and heartfelt thanks to Dominik Seuß, Dr. Jens-Uwe Garbas and Stephan Gick, for their trust in me and for their unfailing support during the decisive last phase of my PhD. Without their strong leadership and timely intervention, I would not have been able to complete this PhD work as an external researcher.

I extend my gratitude to the Fraunhofer Society for awarding me the two-year Fraunhofer TALENTA *start* scholarship for female researchers in 2015. This funding helped me find dedicated time for preparing my initial publications, and helped to fund visits to summer schools and scientific and networking conferences. I also thank Fraunhofer IIS and Fraunhofer Society for the opportunity to participate in the Young Research Class programme during the two year period 2016-2017. This helped me to start my research on multimodal stress recognition, some results of which find a place in this doctoral thesis.

Several aspects of this doctoral work resulted from projects and collaborations with the Nuremberg Institute for Market Decisions (formerly GfK Verein), the Swiss Center for Affective Sciences, University of Bamberg, and the Jožef Stefan Institute in Slovenia. I thank the Nuremberg Institute for Market Decisions, especially Dr. Anja Dieckmann and Dr. Matthias Unfried, for the insightful discussions and valuable feedback during the entire course of the project on inferring appraisal dimensions. I also express my gratitude to them for the kind consent to use their proprietary market-research database for the purpose of evaluating the approach developed in this doctoral work. I thank Dr. Marcello Mortillaro and Prof. Dr. Klaus Scherer for the face models that were used in this doctoral work, and for the support with the psychological aspects pertaining to facial actions. My deepest gratitude to the Cognitive Systems Group and the Faculty of Human Sciences and Education of University of Bamberg, for the collaborative work on the topics of automatic and explainable pain detection from facial actions or facial images. I thank Michael Siebers and Katharina Weitz from the Cognitive Systems Group headed by Prof. Dr. Ute Schmid, for the fruitful discussions and joint research. I place on record my deepest gratitude to Prof. Dr. Miriam Kunz from University of Augsburg and Prof. Dr. Stefan Lautenbacher from the Faculty of Human Sciences and Education of University of Bamberg, for their expert advice and guidance on the psychophysiological aspects of pain in cognitively healthy and cognitively impaired individuals. Discussions with them revealed immense knowledge about pain research, in general, and the facial expressions of pain, in particular. The stress detection research that was started during the Young Research Class programme, was continued in a joint collaboration with the Jožef Stefan Institute in Slovenia. I thank Martin Gjoreski, Prof. Dr. Matjaž Gams and Mitja Luštrek for the scientific collaboration on multimodal stress detection using facial and physiololgical data. I would also like to thank Dr. Arnaud Dapogny and Dr. Kévin Bailly of Sorbonne University, Paris, for providing access as well as consenting to the use of their facial action recognition system as a benchmark for evaluating the approach developed in this thesis.

I have had the very good fortune of being guided by exceptional teachers throughout my life. The knowledge and wisdom they gifted me, continue to guide and illuminate my life. I would like to specifically mention my Maths teacher, Ms. Vijayalakshmi, who told me at a crucial juncture in my school life: "Teena, you are good the way you are." Those words continue to motivate me, everytime self-doubt kicks in. Germany was no exception, in giving me wonderful teachers. I express my gratitude to my professors at Bonn-Rhein-Sieg University of Applied Sciences, especially Prof. Dr. Erwin Prassler, Prof. Dr. Paul Plöger, and late Prof. Dr. Gerhard Kraetzschmar for fostering my research skills, and for supporting me through ups and downs in my research and personal lives. Bielefeld University came as a pleasant surprise into my life, and gave me the much-needed opportunity to rediscover myself. I thank Prof. Dr. Stefan Kopp and my colleagues at Bielefeld University, especially Dr. Hendrik Buschmeier, Sonja Stange, Christopher Ritter and Jan Pöppel for their encouragement, suggestions, and continued support during the crucial, last phase of my doctoral research. I also thank each member of my peer support network at Bielefeld University, especially Jasmin Bernotat and Eva Nunnemann, for the constant support and motivation during the writing phase. My special thanks to Niels Diekmann from Bielefeld University of Applied Sciences, for his light-hearted, yet persistent motivation, during the final months of writing. I also express my heartfelt thanks to all those who proofread my manuscript and gave invaluable suggestions that improved the quality of this dissertation. This includes Dr. Jens-Uwe Garbas, Jan Pöppel, Dr. Matias Valdenegro, Sonja Stange, Dr. Hendrik Buschmeier and Lina Varonina.

My circle of friends is amongst the choicest blessings in my life. In times of need, they found me, stood with me, and boosted my morale. Their unconditional support helped me find joy and face every challenge with courage. I would like to especially thank my dearest friends Dr. Matias Valdenegro, Iman Awaad, Edith Holzwarth, Ranjana Rajendran, Johannes Wollenberg, Pelin Genc, Bhargavi Mahesh and Dr. Liz Babu Neelicattu, for weaving a web of emotional support for me during the crucial period spanning the last few years, and for guiding me on personal and academic fronts. I express my heartfelt thanks also to all my other innumerable friends in Germany and around the world, for their encouragement, love and selfless support. With a joyful heart, I thank my doting neighbours, especially Mrs. Grete Schmidt in Bielefeld and Mr. Otto Pickl in Erlangen, who rejoiced with me at every milestone in this journey, and whose friendly enquiries made sure that I did not lose sight of my goal, amidst the hassles of everyday life.

I owe what I am in my life to my beloved parents–my mother, Saaby Ismail, and my father, Hassan Lebba. My mother is my first teacher,

and she cultivated in me the capacity for self-reliance and reflection, especially in academic matters. Her creativity and resourcefulness quietly blended into my thought-fabric over time, without me being conscious about it. She taught me that treasures and beauty are sometimes hidden in the details and in the smallest of things that we might tend to overlook. Thank you, Mama, for ingraining me with these skills that are so crucial for research! While my mother taught me to pay attention to details, my father showed me through his own life, the transformative power of vision, positive thinking and principles. He taught me to dream big and beyond, and at the same time, instilled in me the spirit to get up and run, everytime I fell down. His wisdom helps me dispel my self-limiting doubts, fears and tears. Thank you, Papa, for being my light in darkness, for steadying my shaking feet, and for teaching me to embrace life and its unpredictabilities with confidence, optimism and courage! My parents always gave me and my brother, Thanzil, equal opportunities in life, and never considered my gender to be a barrier for my intellectual pursuits, or for making personal decisions. I am thankful for growing up in a home where gender equality, individual freedom and mutual respect are upheld.

The journey towards this doctoral thesis was one of wins and losses. But I believe that every struggle that I faced along the way, made me a bolder person and gave me new insights about life. Even for only this, this journey was indeed worth taking, and I am thankful for being fortunate enough to have overcome the obstacles on the way. Therefore, I dedicate this thesis to the cosmos that helped me at every juncture, when I could not see the way forward!

CONTENTS

LIST OF TABLES

ACRONYMS

AAM	Active Appearance Model
AI	Artificial Intelligence
ASM	Active Shape Model
AU	Action Unit
AUC	Area Under ROC Curve
BPS	Behavioral Pain Scale
CLM	Constrained Local Models
CNN	Convolutional Neural Network
DBN	Dynamic Bayesian Network
FACS	Facial Action Coding System
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients

HRV	Heart Rate Variability
LBP	Local Binary Patterns
LGBP	Local Gabor Binary Patterns
LIME	Local Interpretable Model-Agnostic Explanations
LPQ	Local Phase Quantization
LRP	Layer-wise Relevance Propagation
LSTM	Long Short-Term Memory
NIPS	Neonatal Infant Pain Scale
PAINAD	Pain Assessment in Advanced Dementia
PDM	Point Distribution Model
PSPI	Prkachin-Solomon Pain Intensity
RVR	Relevance Vector Regression
SC	Skin Conductance
SU	Shape Unit
SVM	Support Vector Machine
SVR	Support Vector Regression
TOP	Three Orthogonal Planes

Part I

SYNOPSIS OF THESIS

1 INTRODUCTION

Humans are social beings and engage in social interactions, which require them to infer the mental states of others. Psychologists and cognitive scientists have been (i) studying the socio-cognitive processes in the brain that provide humans with this ability to infer the mental states of others [7, 178], (ii) developing theories on how external or internal events influence mental states [131, 132, 162], and (iii) investigating how the mental state affects the behaviour and physiology of the individual [73, 91, 97, 106, 157]. Computer scientists have applied these pieces of knowledge to develop methods to automatically detect mental states based on observable behavioural or measurable physiological signals [88, 170, 188].

Such automatic mental state analysis finds application in many fields. In human-robot interaction, automatic analysis of the human user's mental state could help to adapt the robot's behaviour. For example, in [112], a robot basketball coach adapts the game's difficulty level based on the anxiety levels of the player. Further, automatic mental state analysis could contribute to preventive healthcare, by monitoring and detecting unhealthy stress levels at workplaces or in everyday life [15]. In the medical field, mental state analysis could assist professionals in the diagnosis of psychological conditions such as depression [151], or assist caregivers in efficient detection and subsequent treatment of pain [68, 198]. In the field of market research, automatic mental state analysis could be used to study a subject's emotional response to advertisements (e.g. [60]). As can be seen from these examples, the mental states (e.g. anxiety, stress, depression, pain, emotions) that are analysed depend on the application or use case.

Mental states can be analysed using different behavioural and physiological signals. Behavioural signals include, for example, facial expressions, vocalisations and body movements. Physiological signals include electrocardiogram, electromyogram, electroencyphalogram and body temperature, to name a few. The signals that are analysed depend on the mental state that is to be detected. Facial expressions have been found to be important for non-verbal communication [66, 126, 142], which includes communication of information about mental states such as emotions and pain [31, 45, 47, 97]. Body movements and vocalisations are also useful for assessing pain (cf. [33, 91, 102, 192]). Speech or auditory signals have been found to be useful for communicating emotions [51] or attitude [126]. Stress influences physiological signals [49], and can be observed through physiological parameters such as Heart Rate Variability (HRV) [73] and Skin Conductance (SC)

[106]. The analysis of eye movements could be useful in detecting distractions [89, 160]. Although certain signals might contribute more information than others, combining information from multiple signals could improve the performance of automatic mental state analysis by making use of redundant as well as complementary information (cf. [57, 182, 196]).

Given the prominence of facial expressions in the communication of mental states, the automatic analysis of facial expressions using images and videos has received a great deal of attention from the computer vision research community. Over the last two decades, good progress has been made in this field [124, 159]. Several datasets with annotated facial videos and images have been published (see [124, 159]), several data-driven machine learning (e.g. [35, 109, 118]) and deformable face model-based (e.g. [37, 38]) approaches have been investigated, and international competitions (e.g. FERA Challenge [185], AVEC Challenge [151], EmotiW [34]) have been organised to promote the development of efficient and robust methods for facial expression analysis and facial expression based mental state analysis.

Psychologists have analysed facial expressions using two approaches: *message judgment* and *sign judgment* [21, 22]. This terminology can be adopted to categorise the computer vision methods for automatic facial expression analysis (cf. [188]). The methods for message judgment detect the mental state or 'message' communicated through facial expressions. The most common goal is either the recognition of the basic emotions identified by Paul Ekman and colleagues[1][43–45] (categorical model) (e.g. [35]) or the estimation of the valence and arousal dimensions of emotions proposed by Russell [155] (dimensional model) (e.g. [53, 150]). The methods for sign judgment (e.g. [109]) detect basic facial movements known as Action Units (AUs) that are defined in the Facial Action Coding System (FACS) [46], and thus produce an objective description of the facial expression.

Message judgment methods learn the target categories either directly from visual data (*one-step*) (e.g. [92]) or from the output produced by sign judgment methods (*two-step)* (e.g. [10]). In the latter case, the complexities involved in the automatic detection of mental states are divided into two parts. The first part–AU detection–deals with the complexities and challenges at the level of facial image/video processing, in order to produce a high-level, semantic description of facial expressions in terms of AUs. The second part–mental state detection–deals with the complexities at the level of inferring mental states by applying domain knowledge and by using noisy estimations of facial muscle movements (i.e. AUs). Since AUs can be used to describe any facial expression at a fine-grained level, AU detection systems can support the analysis of a broader range of facial expressions, beyond

1 Basic emotions identified by Paul Ekman and colleagues [44] are fear, anger, sadness, disgust, happiness, contempt and surprise.

the prototypical facial expressions associated with basic emotions (cf. [22]). Consequently, AU detection systems can be used in a variety of application domains, for example, human-robot interaction, market research and healthcare, where different context-dependent 'messages' are conveyed through diverse, subtle and spontaneous facial expressions. The use of automatically learned or empirically determined rules for inferring these messages (see [133, 172, 202]) could improve the interpretability of the system's decisions.

This work focuses on the automatic estimation of intensities of AUs (sign judgment), and its application to the detection of experimentally induced acute pain in laboratory settings and experimentally induced distractions in a simulated driving setting (message judgment). For automatic AU intensity estimation, computer vision researchers have used data-driven machine learning approaches (e.g. [82, 87, 161] as well as deformable face model-based approaches (e.g. [37, 38]). While deformable face models are interpretable, data-driven models are capable of covering large variance in input data. Therefore, a combination of these two types of methods could take the field of automatic mental state analysis towards strong and ultra-strong models–as envisioned by Michie [129]–that have good predictive performance and are, at the same time, comprehensible to humans.

In this thesis, a novel combination of data-driven machine learning and deformable face model-based approaches is developed to estimate AU intensities. A Gaussian state estimation based framework[2] integrates (i) an AU-based deformable face model, (ii) a viscoelastic facial muscle motion model, (iii) several data-driven AU classification models based on appearance information, and (iv) a data-driven facial landmark detection model that provides shape information. Due to the use of both facial shape and appearance information, the proposed method is a hybrid approach for AU intensity estimation. Several enhancements to this framework are designed and explored for improving the quality of estimates and for enhancing the robustness in real-world applications. These enhancements take person-specific, facial muscle-specific, and deformable face model-specific properties into consideration, and handle cases of missing or anomalous information. The estimated AU intensities are then used to detect pain on the basis of AU-based rules, and to analyse facial expressions under different types of distractions during simulated driving.

In addition to the above, this dissertation also examines several challenges that need to be addressed in order to build practically useful automatic mental state analysis systems. As mentioned earlier, mental states also cause changes in the physiology of a person. Multimodal systems for automatic mental state analysis that combine evidence

2 An earlier version of this framework, developed during my master's thesis [69], integrates the deformable, AU-based face model, a constant velocity motion model, and facial landmarks detected by a deformable face model fitting algorithm.

from facial and physiological signals could be helpful in improving predictive performance, as shown in [196]. However, the quality and characteristics of the data used for developing such systems play a crucial role in determining whether these systems generalise well to new subjects, when applied in the real world. In addition, a thorough investigation and modelling of interpersonal differences in the facial and physiological expression of mental states is necessary. Furthermore, in order to ensure ethical and responsible use of automatic mental state analysis systems, it is crucial that humans can comprehend the underlying Artificial Intelligence (AI) models. Some initial work was done within the scope of this doctoral research in order to deal with these challenges. This includes the gathering of requirements for reference datasets for mental state analysis, the modelling or inspection of interpersonal differences in responses to painful or arousing stimuli, and the generation of different explanations for automatic detections of pain.

1.1 SCOPE OF THESIS

This doctoral thesis is restricted to the following scope:

- There are 44 AUs defined in FACS [46] and each has a numerical code. In consultation with psychologists at the Swiss Center for Affective Sciences, 22 of these AUs were selected for this work. The selection was based on three criteria[3], guided by evidence from literature (see [27, 127, 163]): (i) AUs that have been frequently reported in the literature as being important for expressing emotions; (ii) AUs that have been predicted to be related to appraisal inferences[4]; (iii) AUs that actors could display in the laboratory. The AU intensity estimation system developed in this thesis estimates the intensities of these 22 AUs. However, this system can be extended easily to include additional AUs or additional sources of information without changing the existing parameter configuration.

- Only visual information in the form of time-sequences of 2D facial colour/greyscale images is considered in this work. Extensions to other forms of visual input (e.g. depth, thermal, infrared) as well as integration of auditory or physiological signals (e.g. facial electromyogram) could be considered in future work.

- This work is mainly concerned with nearly-frontal head poses involving minimal out-of-plane head rotations. Some robustness

3 Credits to Dr. Marcello Mortillaro, Swiss Center for Affective Sciences, University of Geneva, for clarifying these three AU selection criteria and for sharing the references.

4 Appraisal inference refers to the process by which an observer detects an expressor's emotional appraisals (evaluations) of a stimulus on the basis of the expressed facial actions [163].

to fast head movements and facial occlusions is built-in in the proposed AU intensity estimation method and its anomaly detection feature. However, a systematic evaluation of its performance under non-frontal head poses and facial occlusions is out of scope of this work.

- To illustrate the applicability of the estimated AU intensities for inferring mental states, two use cases were selected: (i) experimental acute pain; and (ii) driver distraction during simulated driving sessions.
- Practical applicability was a key motivator for the project, which this work was part of. Therefore, several design decisions pertaining to the system were taken with the objective of enabling robustness in practical and uncontrolled settings, and for practical convenience.

1.2 RESEARCH QUESTIONS

This thesis primarily deals with the estimation of intensities of 22 AUs from 2D facial colour/greyscale image sequences. Following this, automatic mental state (pain, distraction) analysis based on AU intensities is explored. The main research questions investigated in this doctoral thesis are as follows:

1. How can data-driven machine learning approaches be combined with deformable face model-based approaches for AU intensity estimation inside a Gaussian state estimation framework? Would such a combined approach improve predictive performance?
2. How can the viscoelastic facial muscle motion be modelled as a first-order Markov process for integration within a Gaussian state estimation framework that estimates AU intensities? How can this model be adapted to the properties of the particular facial muscle or group of facial muscles that produces a specific AU?
3. How do the differences in the physical expressiveness of AUs influence the quality of AU intensity estimation? How can negative effects be mitigated?
4. How can the estimated AU intensities be ensured to be in conformance with FACS as well as with the design of the deformable face model?
5. How can the geometric/semantic similarities and dissimilarities between AUs inform the design of the components of the Gaussian state estimation framework?

6. How does the proposed method for AU intensity estimation fare qualitatively and quantitatively against a state-of-the-art method on AU recognition and AU intensity estimation tasks?
7. How can the proposed AU intensity estimation method be made robust against anomalous detections made by the data-driven face and facial landmark detection models?
8. Which approaches have been explored so far by computer vision researchers for automatic detection of pain from facial expressions?
9. How can pain be defined in the form of simple mathematical expressions involving AU intensities? How can these mathematical expressions or rules be used to create verbal explanations for pain detections in terms of AUs?
10. Can facial activity, described in terms of AUs, provide clues about different types of distraction while driving in simulation settings?
11. Which technical and annotation requirements should be fulfiled by a reference dataset for mental state analysis in order to enable the creation of generalisable, reliable and comparable models?
12. How can interpersonal differences in the expression of mental states be considered during the development of automatic mental state detection systems?

1.3 SCIENTIFIC CONTRIBUTIONS

The investigation of the above-mentioned research questions led to scientific contributions that can be grouped into three areas: (i) automatic facial action estimation, (ii) automatic mental state analysis, and (iii) addressing open challenges in automatic mental sate analysis. The scientific contributions are listed in the following subsections.

1.3.1 *Automatic Facial Action Estimation*

In this doctoral work, a Gaussian state estimation based approach is developed as the basic framework for estimating AU intensities. This framework models the dynamics of facial muscle motion and fuses shape and appearance information about AUs. Several enhancements are proposed to the basic framework to enhance robustness and to handle the properties of AUs. The main scientific contributions that are related to the basic framework and its enhancements are listed below:

- In Preprint C.2.1, I categorised the automatic AU intensity estimation methods into data-driven machine learning methods and deformable face model-based methods.

- In Publication B.2.1, I proposed a novel and practical method to integrate categorical probability outputs from a data-driven classifier within a continuous, Gaussian state estimator. I applied the proposed method to fuse facial shape and appearance information for AU intensity estimation. The integration of categorical probability outputs from AU classifiers that were trained on facial appearance features improved the AU recognition performance of a Gaussian state estimator that previously used only shape information in the form of facial landmark positions.

- To model the dynamics of facial muscle movements, I applied a set of driven mass-spring-damper models. Each AU was modelled using a separate driven mass-spring-damper model. These were used as process models in the Gaussian state estimator, in order to model how AU intensities change over time. The values of the model parameters and the integration of the models in the Gaussian state estimator are elaborated in Patent D.1.1.

- In order to adapt each AU-specific driven mass-spring-damper model to the viscoelastic properties of the corresponding facial muscle or muscle group, I proposed a method that is based on the facial muscle fibre composition. According to this method, the facial muscles (a.k.a. mimic muscles) containing more Type I fibres [50, 64] are modelled using stiffer and more strongly damped springs. The proposed method and qualitative results are presented in Section 3.3.4.

- In order to prevent the state estimator from using subtle AUs such as AU06-CheekRaiser to correct errors in predicted facial landmark positions, I proposed the use of a fully dependent facial landmark noise configuration, in addition to the independent landmark noise configuration. Under the fully dependent noise configuration, the noise in each landmark is assumed to be dependent on the noise in every other landmark. Consequently, the proposed AU intensity estimation framework uses two state estimators, one with the full landmark noise configuration, and the other with the independent landmark noise configuration. The decision about which state estimator is to be used for which AUs, was made on the basis of empirical analysis.

- In Publication B.2.2, I extended the state constraints on AU intensities to the driven mass-spring-damper process models. In order to ensure that the AU intensities belonged to the valid range of values, additional constraints were applied on the driving forces acting on the mass-spring-damper systems. The effect of the constraints was validated through a histogram analysis of the estimated AU intensities. A reinterpretation and visualisation of the operation of the state constraints was also done.

- To quantify the similarities in the facial shape deformations caused by different AUs, I computed the cosine of the angle between each pair of mean-normalised AU-based facial shape deformation vectors. I proposed to use these cosines as correlation coefficients for modelling the correlations between noise in AU intensities. These correlation coefficients were used to determine the AU noise covariances in the noise covariance matrices associated with the initial state, the process model and the AU constraint model. This is described in detail in Patent D.1.1. Based on empirical evaluation results, the use of correlation coefficients has been selectively turned off for specific AUs.

- In order to deal with anomalies in detected facial landmark positions, I proposed a solution involving three steps. In the first step, I applied the anomaly detection method based on normalised innovation squared [136], in order to identify anomalous facial landmark detections. The threshold needed to flag a set of detected facial landmark positions as an anomaly, was determined empirically. In the second step, the AU intensity estimates were prevented from being updated, in the event of anomalous facial landmark detections. Finally, if the number of consecutive anomalous detections crossed a pre-defined upper limit, state estimation was suspended and the state estimator was later reset, when the suspension exceeded a pre-defined duration of time and facial landmark detections were available. Anomaly handling is presented in detail in Section 3.3.3. To handle missing facial landmark detections, a strategy similar to the last two steps of anomaly handling was devised. This is also mentioned in Section 3.3.3.

- I compared the performance of the proposed AU intensity estimation system with a state of the art system from Dapogny et al. [30]. The AU recognition performance as well as the AU intensity estimation performance were compared using different performance metrics. Qualitative analysis of the outputs from the two systems was also performed. For the evaluation, the Actor Study Database [168], the UNBC-McMaster Shoulder Pain Expression Archive Database [121], and a proprietary market-research database [70] were used. The results, along with a discussion of the pros and cons of the two systems, are presented in detail in Preprint C.2.1. Benchmark results were computed on the Actor Study Database [168] using the system developed by Dapogny et al. [30], and are included in Publication B.2.3.

1.3.2 *Automatic Mental State Analysis*

In this doctoral work, the AU intensities estimated by the proposed Gaussian state estimation based approach are used to detect pain and analyse driver distraction. The main scientific contributions that are related to these mental state analyses are listed below:

- In Publication B.1.1, I surveyed the state of the art in automatic pain detection from facial expressions. The papers published during the period 2006–2018 were reviewed. I categorised the pain detection approaches into *one-step* and *two-step* approaches, depending on whether or not an intermediate step of AU detection was involved. I also categorised the extracted features, learning tasks and machine learning methods. Deficits in the state of the art were identified and future research directions were proposed. A very short summary of the state of the art is presented in Publication B.1.2, where I categorised the methods into single-level and two-level methods, and additionally, grouped them on the basis of learning goals.

- I defined and evaluated several rules for automatic pain detection using AU intensities (two-step approach). The pain detection rules are defined on the basis of psychological evidence from studies conducted by Kunz and Lautenbacher [97], as well as Prkachin and Solomon [146]. While the work of Prkachin and Solomon [146] has inspired a few two-step approaches [125, 202], the pain clusters identified by Kunz and Lautenbacher [97] have not yet been applied for automatic AU-based pain detection. The proposed pain detection rules and their evaluation are described in Preprint C.2.1.

- I performed a preliminary analysis of the facial activity under the influence of different sources of distraction during simulated driving sessions. Facial activity was determined using the AU intensities (and their time-derivatives) estimated by the proposed Gaussian state estimation based approach. Between different distraction conditions, differences in the facial activity were found. The details are presented in Section 2.2.1. Based on these insights, I conceptualised and supervised a research study to create machine learning models to predict driver distraction based on AU intensities. This study also explored the combination of facial and physiological signals for driver distraction detection. The results of this study are presented in Preprint C.1.1 (Update 02.08.2020: This is now published in IEEE Access. See [56]).

1.3.3 *Addressing Open Challenges in Automatic Mental State Analysis*

Three key challenges in the field of automatic mental state analysis include: (i) a lack of reference datasets for benchmarking algorithms, (ii) interpersonal differences in responses to stimuli, and (iii) a lack of interpretable models for automatic mental state detection. As part of this doctoral work, some basic steps towards addressing these challenges were explored, mainly in the form of research performed by master-level students, (co-)supervised by me. As a result, (i) a set of requirements for building a multimodal reference dataset for detecting human stress was developed, (ii) interpersonal differences in pupillary responses to an arousal stimulus were examined, and (iii) a qualitative comparison of different explainable AI methods was performed, to interpret and explain deep Convolutional Neural Network (CNN) models for distinguishing pain from happiness and disgust. The main scientific contributions made by me in these research tasks are listed below:

- I conceptualised several requirements to be fulfiled by reference datasets, in order to promote the building of sensor-independent and generalisable models for automatic mental state analysis. More specifically speaking, I conceptualised the requirements related to documenting sensor calibration and sensor noise characteristics, gathering annotation data using different methods, and including 'signature' modalities for mental states (for example, heart activity and electrodermal activity for stress). In Publication B.3.1, these requirements are formulated for reference datasets for human stress detection. These requirements can be generalised to multimodal reference datasets for any mental state, as described in Section 4.1.

- In Publication B.3.2, I conceptualised and supervised a personality-based analysis of pupillary responses to an arousal stimulus, in order to examine interpersonal differences.

- Using the AU-based rules, I created verbal explanations for pain detection in terms of AU activations as well as in terms of discretised AU intensities. The terminology defined in FACS [46] for discrete AU intensities was adopted for use in the explanations. The generation of these explanations is illustrated in Preprint C.2.1. In order to facilitate a future comparison of AU-based explanations with image-based explanations, I conceptualised and co-supervised a research work that applied different explainable AI methods to generate explanations for predictions made by deep CNN models for automatic pain detection. The results are published in Publication B.3.3.

1.4 STRUCTURE OF THESIS

The rest of the thesis is organised as follows: Chapters 2 to 4 provide a synopsis of the research and development performed as part of this doctoral work. Section 2.1 summarises the state of the art in automatic pain detection from facial expressions, and then presents the proposed two-step approach for automatically detecting pain. Section 2.2 describes the preliminary analysis of facial activity during different types of distractions while driving. Chapter 3 summarises the state of the art in automatic analysis of facial actions, and describes in detail, the AU intensity estimation method developed in this doctoral work. The basic framewok is explained in Section 3.2 and the enhancements are explained in Section 3.3. A brief summary of the quantitative and qualitative evaluations of the proposed AU intensity estimation method is also included in Chapter 3. Chapter 4 summarises the research contributions towards addressing three challenges in the field of automatic mental state analysis that pertain to multimodal reference datasets, interpersonal differences and interpretability of machine learning models. Chapter 5 concludes this thesis, providing an outlook for future research. Appendix A presents some additional results of the evaluation and analysis performed in this work. In Appendix B, all the publications that resulted from this doctoral work are attached, and in Appendix C, all the preprints or planned submissions are attached. In Appendix D.1, the patent that contains contributions from this doctoral work is attached. A list of the scientific and written contents contributed by me is provided in the Appendix, along with each attached publication, preprint, or patent.

2 AUTOMATIC MENTAL STATE ANALYSIS

2.1 AUTOMATIC PAIN DETECTION

Pain is defined as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage"' [128, p. 209]. Pain is assessed by clinical staff by observing behavioural cues such as facial expressions, vocalisations and body movements (cf. pain scales such as Behavioral Pain Scale (BPS) [33], Pain Assessment in Advanced Dementia (PAINAD) [192], Neonatal Infant Pain Scale (NIPS) [102]). Among these cues, facial expressions have been found to be a valid indicator [99, 171]. In this section, the work on automatic pain detection[1] from facial expressions is described.

2.1.1 *Summary of State of the Art*

Over the last decade, researchers have applied automatic facial expression analysis methods to automatically detect pain. In the process, several datasets consisting of facial expressions of pain have been created. A survey of these automatic pain detection methods and the pain datasets is provided in Hassan et al. [68] (Publication B.1.1). The key findings are summarised in this section.

Automatic pain detection from facial expressions has followed two types of approaches, namely *one-step* and *two-step*. The difference between the two types of approaches lies in whether or not an intermediate AU-based representation of facial expressions is learned before learning the pain-related targets. While one-step approaches involve only pain-related learning tasks, two-step approaches involve learning of AU-related targets followed by learning of pain-related targets. So far, one-step approaches predominate the field of automatic pain detection, and the binary classification task of detecting the presence or absence of pain has received the most attention. Besides pain versus no pain (e.g. [5, 12, 92]), other classification tasks such as distinguishing pain from other emotions or states (e.g. [13, 62, 135]), distinguishing facial expressions of genuine pain from those of faked pain (e.g. [10, 110, 111]), and detecting different discrete levels of pain intensity (e.g. [55, 77, 197, 202]) have been investigated. The regression task of estimating continuous-valued pain intensity has also been investigated in the published literature (e.g. [40, 87, 152]). Automatic pain detection methods almost always employed supervised learning strategy that re-

1 The term 'detection' is used here to refer to both classification and regression tasks.

quires ground truth annotations for every learning instance. The pain datasets provided ground truth about pain at frame-level, segment-level, or sequence-level (see Publication B.1.1 for details). Frame-level annotations included labels of pain or distress states, a set of AUs and their intensities, or discrete pain intensity levels computed from AU intensities using the Prkachin-Solomon Pain Intensity (PSPI) scale [146]. Sequence-level annotations were mostly pain scores collected from self or observer reports. Segment-level annotations consisted of AUs and their intensities as well as self-reports. The intensity level of pain stimulus applied in experimental settings, or the type of stimulus (genuine pain, posed pain, different induced emotions) was also sometimes provided as segment-level or sequence-level annotation. The most widely used pain dataset is the publicly available UNBC-McMaster Shoulder Pain Expression Archive Database [121].

Among supervised machine learning methods, Support Vector Machines (SVMs) were used most widely for pain-related classification tasks. Random forests, neural networks and variants of AdaBoost have been used for pain-related supervised classification and regression tasks. More recently, deep learning methods such as CNNs and Long Short-Term Memory (LSTM) recurrent neural networks have been used for supervised pain intensity estimation (see Publication B.1.1 for details). Weakly supervised and unsupervised methods were used very rarely. Semi-supervised methods were not explored. Pain detection tasks were performed at frame-level (e.g. [5, 12, 152]) or sequence-level (e.g. [10, 114, 173, 197]). Occassionally, frame-level predictions were aggregated to obtain sequence-level predictions (e.g. [5, 125]).

To train pain detection models, features were first extracted from the visual input (single image or a sequence of images). The extracted features can be categorised into *spatial* or *spatiotemporal* features, depending on whether temporal information was included in the features, in addition to spatial information. Spatial features are extracted from single images and include geometric or textural features that describe the facial shape or appearance, respectively. Spatiotemporal features refer to geometric or textural features extracted from a sequence of images. These features were used either alone or in different combinations to develop automatic pain detection approaches. For example, [87] used a combination of spatial geometric and textural features; [197] used a combination of spatiotemporal geometric and textural features; [40] used a combination of spatial and spatiotemporal features.

Facial landmark positions, or distances and angles between facial landmarks are examples of spatial geometric features. Statistical features extracted from a series of these geometric features constitute spatiotemporal geometric features. Gabor filter coefficients [32, 48, 52], Local Binary Patterns (LBP) [137, 138] and Histogram of Oriented Gradients (HOG) [29] are often used to describe facial texture. Spatiotemporal variants of LBP and HOG extract features from Three

Orthogonal Planes (TOP) defined by the time and 2D pixel axes. These dynamic features are therefore known as LBP-TOP [203] and HOG-TOP [19], respectively. Hankel matrix representation of time series of spatial textural features (cf. [115, 145]) is another example for spatiotemporal textural features. Deep learning methods automatically extract spatial or spatiotemporal textural features from the visual input. In one-step approaches, these features were used for pain detection, whereas in two-step approaches, these were used for AU detection. For pain detection, the two-step approaches used the AU detection outputs directly as features (e.g. [119]) or extracted temporal features from a time-series of AU detection outputs (e.g. [10]).

It was noted that the approaches for automatic pain detection were predominantly based on data-driven machine learning methods. Models based on expert and interdisciplinary knowledge about pain and facial expressions were limited to the use of the PSPI scale for pain intensity estimation based on AU intensities, in two-step approaches (e.g. [202]). Data-driven machine learning approaches are capable of covering large variance in input training data, without the need for modelling each variance by hand. However, computational models designed by experts would not only allow to incorporate interdisciplinary knowledge about facial muscle properties and dynamics, possible facial shape deformations, and the facial expressions of pain, but also facilitate human comprehension of their predictions. A combination of such model-based and data-driven machine learning based approaches is developed in this doctoral work, in order to bring together the strengths of both types of approaches. The next subsection provides an overview of these models and components that are brought together in a two-step approach for pain detection.

2.1.2 *A Two-Step Approach*

Kunz et al. [100] provides a roadmap for developing solutions for automatic pain detection from facial expressions through joint interdisciplinary research. In addition, the outline of a two-step approach for automatically detecting pain is presented in Fig. 1 in [100]. A more detailed view of this approach is given in Figure 2.1, showing the combination of model-based constructs and data-driven machine learning methods. The components of the proposed approach are described below:

- **Shape model:** In this work, the facial shape is described by a vector of 68 facial landmark positions. A 3D, linear, parameterised, deformable model constitutes the facial shape model.[2] This is similar to the Point Distribution Model (PDM) proposed by Cootes et al. [25]. The facial shape model consists of a vector

2 This facial shape model was also used in the master's thesis [69] that preceded this doctoral work.

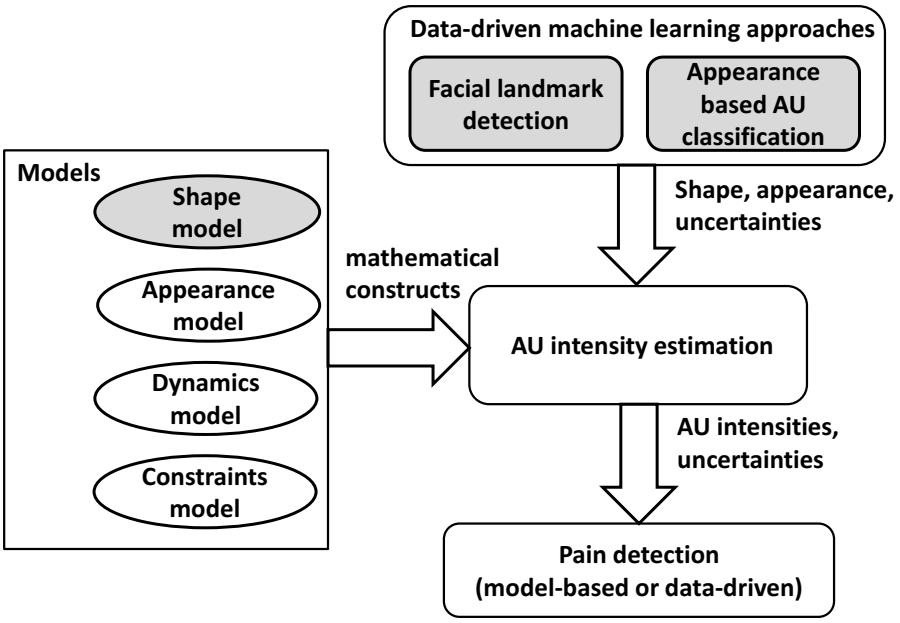


Figure 2.1: A two-step approach for automatic pain detection: The approach combines data-driven machine learning approaches with different mathematical models representing interdisciplinary knowledge. The components marked in grey were not developed as part of this doctoral work.

representing the shape of a mean, neutral face and 83 additional vectors representing various deformations that can be caused to the neutral facial shape. Twenty-two of these deformation vectors represent transient shape deformations caused by 22 different AUs, and the remaining 61 deformation vectors represent the relatively permanent person-dependent facial shape variations. Table 2.1 provides the list of 22 AUs used in this face model. The AU deformation vectors represent the direction and magnitude of maximum anatomically possible facial shape changes that can be caused by the AUs. These deformation vectors are derived from the high-poly 3D mesh models of AU expressions designed by psychologists [96, 153]. The Shape Unit (SU) deformation vectors represent person-dependent facial shape variations, and were derived statistically from the facial morphs in the software FaceGen Modeller [75]. Equation 2.1 represents the 3D deformable facial shape model. In this equation, $\mathbf{m}$ represents the mean, neutral shape as a column vector; $\mathbf{A}$ represents the matrix of the 22 AU deformation vectors and $\mathbf{x_a}$ represents the column vector of 22 *AU parameters*, equivalent to AU intensities; $\mathbf{S}$ represents the matrix of 61 SU deformation vectors and $\mathbf{x_s}$ represents the column vector of the *SU parameters*. The elements of each deformation vector in $\mathbf{S}$ and $\mathbf{A}$ are distances in meters, and the AU and SU parameters in $\mathbf{x_a}$ and $\mathbf{x_s}$ are unitless intensities/weights. Rotation, translation and scaling can be applied to the model

in Equation 2.1 in order to account for global head motion or pose changes, as shown in Equation 2.2. The rotation, translation and scaling parameters $\boldsymbol{\Theta}$, $\mathbf{t}$ and p constitute the rigid parameters, and the AU and SU parameters in $\mathbf{x_a}$ and $\mathbf{x_s}$ constitute the non-rigid parameters of the facial shape model. By applying the camera model to Equation 2.2, the 3D facial shape model can be perspective-projected onto 2D image space. Figure 2.2 provides an illustration of the perspective-projected deformable facial shape model. In Patent D.1.1, Figure 3 shows the facial shapes resulting from the application of the AU deformation vectors to the mean, neutral face.[3]

$$\boldsymbol{\Omega} = \mathbf{m} + \mathbf{A}\mathbf{x_a} + \mathbf{S}\mathbf{x_s} \tag{2.1}$$

$$\boldsymbol{\Omega}' = p(\boldsymbol{R}_{\boldsymbol{\Theta}}\boldsymbol{\Omega}) + \mathbf{t} \tag{2.2}$$

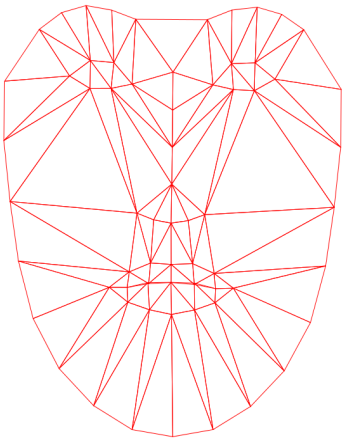
- **Appearance model:** AUs cause not only transient changes in the shape of facial features, but also transient changes in the appearance of the face. Changes in the form of transient wrinkles and folds on different regions on the face are characteristic of specific AUs. For example, AU04-BrowLowerer causes vertical folds in between the eyebrows, and AU06-Cheek Raiser creates wrinkles–commonly known as "crow's feet"–at the outer corners of the eyes [42, 46]. Appearance of the face is affected by several factors, for example, age, skin texture, skin colour and illumination conditions. Therefore, a data-driven machine learning approach is necessary to cover these large variances and learn a robust model to discriminate between different AUs based on appearance or appearance changes. The outputs of the data-driven machine learning approach are then mapped to AU parameters by the appearance model. In this work, the identity function is used as the appearance model in order to map the output probabilities produced by data-driven AU classifiers onto AU parameters or AU intensities.

- **Dynamics model:** Facial expressions are caused by facial muscle movements. Different mechanical models can be used to model the dynamics of facial muscle motion. Very simple models like Gaussian random walk (see [37, 76]) or autoregressive models (see [38]) have been used in the context of facial action intensity estimation. Complex models involving several interconnected springs (see [180, 193]) have been used to model facial deformations, especially in the fields of computer graphics and computer animation. Hill [72] proposed a mathematical model for muscles, which consists of a spring connected in parallel with a viscous

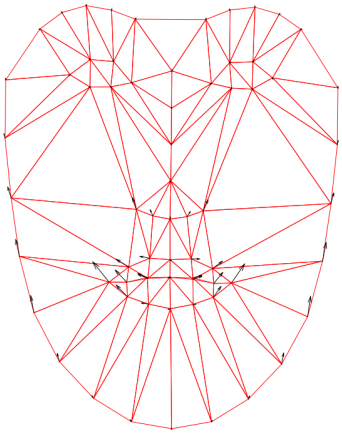
3 Both Figure 2.2 in this work and Figure 3 in Patent D.1.1 were created as part of the master's thesis [69] that preceded this doctoral work.

Table 2.1: List of 22 AUs included in the deformable facial shape model used in this thesis. The numerical codes and names of these AUs as well as the facial muscles that generate these AUs–as defined in FACS [42, 46]–are listed here (see also [22]). Images showing the shape and appearance patterns associated with these AUs are available in [22].

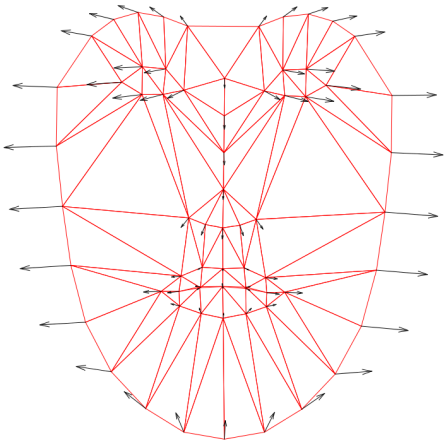
AU Code	AU Name	Facial Muscle(s)
01	Inner Brow Raiser	Frontalis (pars medialis)
02	Outer Brow Raiser	Frontalis (pars lateralis)
04	Brow Lowerer	Depressor supercilii, corrugator supercilii
05	Upper Lid Raiser	Levator palpebrae superioris
06	Cheek Raiser	Orbicularis oculi (pars orbitalis)
07	Lid Tightener	Orbicularis oculi (pars palpebralis)
09	Nose Wrinkler	Levator labii superioris alaeque nasi
10	Upper Lip Raiser	Levator labii superioris
11	Nasolabial Deepener	Zygomaticus minor
12	Lip Corner Puller	Zygomaticus major
13	Sharp Lip Puller	Levator anguli oris (a.k.a. Caninus)
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris (a.k.a. Triangularis)
16	Lower Lip Depressor	Depressor labii inferioris
17	Chin Raiser	Mentalis
20	Lip Stretcher	Risorius with platysma
23	Lip Tightener	Orbicularis oris
24	Lip Pressor	Orbicularis oris
25	Lips Part	Depressor labii inferioris or relaxation of mentalis, or orbicularis oris
26	Jaw Drop	Masseter, relaxed temporalis and internal pterygoid
27	Mouth Stretch	Pterygoids, digastric
43	Eyes Closed	Relaxation of levator palpebrae superioris; orbicularis oculi (pars palpebralis)



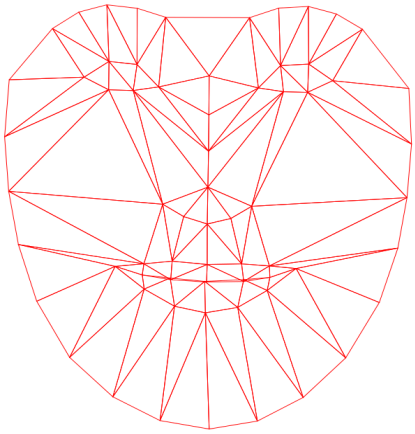
(a) Mean neutral face



(b) AU12 deformation vector



(c) An SU deformation vector



(d) Combined shape

Figure 2.2: 2D perspective projections of three components of the 3D deformable facial shape model and the combination of these components. As can be seen, a round, smiling face was created when the mean neutral face, AU12-LipCornerPuller, and the SU that produces a round face were combined. The AU12 deformation vector is derived from [96].

damper and in series with another spring (see also [169]). This model has been applied to study the force-length responses of gastrocnemius muscles in frogs [169]. In this doctoral work, a single mass-spring-damper system, consisting of a spring attached in parallel to a viscous damper, is used to model the viscoelastic properties of the facial muscle or muscles group associated with each AU. This model was chosen by keeping practical convenience in mind, to reduce the complexity of the developed system and the number of unknown parameters to be modelled. A similar model was used by Liu and Ebbini [113] to model the viscoelastic properties of stiff tissues. More details about the modelling of facial dynamics is provided in Section 3.2.1.

- **Constraints model:** As mentioned above, the shape model consists of 22 anatomically based AU deformation vectors. That is, each of these vectors define the anatomically possible maximum facial shape deformations that can be caused by the corresponding AU. Consequently, the AU parameters in the model have a valid range of $[0, 1]$. Negative values for AU parameters are not conformant with the definitions of AUs and their intensities in FACS [42, 46]. In this work, a linear model is applied to constrain the AU intensities to this range. More details are provided in Section 3.3.2.

- **Facial landmark detection or face alignment:** A data-driven machine learning approach is used to locate the 2D positions of 68 facial landmarks in an image. This process of determining the geometry/shape of the face and facial features in terms of positions of specific fiducial points/landmarks is referred to as *face alignment* [16, 63, 90]. The 2D facial landmark positions used in this work are detected using the face alignment method proposed by Kazemi and Sullivan [90]. The empirical modelling of uncertainties in these detected facial landmark positions is explained in Section 3.2.2.

- **Appearance-based AU classification:** Data-driven machine learning approaches are used to learn appearance patterns associated with different AUs. Textural features such as those specified in Section 2.1.1 can be used to describe facial appearance in an image or facial appearance changes in a sequence of images. Classifiers trained to identify the presence or absence of individual AUs, or regression models trained to estimate intensities of different AUs can be used. In this work, SVMs trained on LBP or HOG features extracted from each frame in an image sequence are used as appearance-based AU classifiers. The output from these classifiers are probabilities for the presence or absence of specific AUs or AU combinations. More information about

these classifiers and the modelling of the uncertainties in their predictions is provided in Section 3.2.3.

- **AU intensity estimation:** This is the first step in the two-step approach for automatic pain detection. A Gaussian state estimation framework integrates the above-mentioned mathematical models of facial shape, appearance, dynamics and constraints, with the outputs from the data-driven facial landmark detection and appearance-based AU classification methods, in order to produce continuous-valued estimates of AU intensities. The uncertainties associated with these components are modelled as Gaussian noise, and are integrated in the state estimation process. A detailed description of this AU intensity estimation framework is provided in Chapter 3.

- **Pain detection:** This is the second step in the two-step approach for automatic pain detection. It can either be based on simple models that are PSPI scale-like (e.g. [125, 202]), or use empirically determined thresholds (cf. [77]) or learned grammar rules (e.g. [164, 172]). More complex machine learning models based on SVMs or neural networks could be considered, along with appropriate methods for enhancing the interpretability and explainability of models. In this doctoral work, a set of PSPI scale-like rules is used for pain detection. These rules are described and evaluated in the to-be-submitted Preprint C.2.1, which also illustrates the verbal explanations generated for pain detected using these rules.

Among the components described above, facial shape model, facial landmark detection, and appearance-based AU classification were not developed by me. Only the final models and the outputs from these components were applied in this doctoral work.

2.2 AUTOMATIC DISTRACTION DETECTION

Distraction is defined in the Cambridge Dictionary as "something that prevents someone from giving their attention to something else" [36]. Distractions during driving increase safety risks [140]. Regan et al. [148] used the term "Driver Diverted Attention" to refer to distraction, and defined it as "diversion of attention away from activities critical for safe driving toward a competing activity, which may result in insufficient or no attention to activities critical for safe driving" [148, p. 1776]. The causes of distraction may or may not be related to driving tasks. Olson et al. [140] quantified the safety risks contributed by different sources of distraction. Mobile phone usage, reading, writing, and checking route maps were noted as activities that increase safety risks. Automation levels 2 and 3 defined in SAE International's standard J3016 [156] require human attention and readiness to take

over control in difficult situations or when the vehicle requests. As the level of automation increases, the probability of distraction due to engagement in non-driving tasks increases. Therefore, driver distraction monitoring is necessary even in vehicles with partial automation. In this section, the use of facial expression analysis for driver distraction detection is discussed.

2.2.1 *Driver Distraction and Facial Activity*

Driver distraction is typically detected on the basis of head and eye behaviours, or using physiological signals [88]. Eye gaze patterns (cf. [65, 105]), eye blink frequency (cf. [104]), head movements (cf. [130]), and skin temperature at nose tip (cf. [78]) and around the eyes (cf. [200]) have been reported to reveal information about cognitive distraction of the driver. Apart from eye and head movements, facial expressions have been generally used less in driver distraction detection. However, facial expressions might have the potential to reveal cues about driver distraction, especially when the distraction is caused by emotional stressors.

In order to study the potential of facial expressions in detecting different types of distractions during driving, an initial, coarse analysis was performed on the facial videos provided in the dataset from Taamneh et al. [177]. This dataset contains data recorded from 68 human subjects while they were driving on predefined routes in a driving simulator under different conditions. Cognitive, emotional and sensorimotor stressors were used to induce distraction during driving. In addition, a case of unexpected system failure leading to loss of vehicle control was also introduced. Practice, normal and relaxed drives, as well as a baseline where the subjects did not drive are also included in the dataset. Facial videos, various physiological signals, and drive parameters were recorded using multiple wearable and contactless sensors. The facial videos were analysed using the AU intensity estimation approach introduced in Section 2.1.2 and elaborated in Chapter 3. This produced a multivariate time series consisting of intensities of 22 AUs for each facial video. Time-derivatives of these AU intensities were then computed, resulting in a time series of AU velocities for each facial video. The AU intensities and AU velocities were then averaged for each facial video, and grouped based on the driving condition. The AU intensity and AU velocity pairs were used to represent facial activity. A 2D histogram of these pairs was plotted for each driving condition, in order to visually examine the distribution of facial activity. Figure 2.3 shows the 2D histograms for two conditions, namely, driving under a cognitive stressor and driving under an emotional stressor. The plots show differences in the distribution of AU intensities and AU velocities under the two conditions. Facial activity during the other six driving conditions is presented in Appendix A.1.

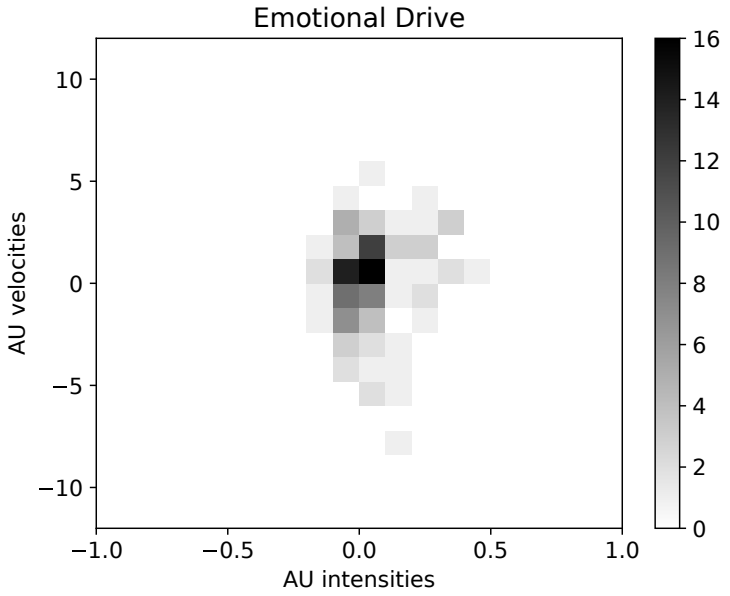
On the whole, the average facial activity is not very high during the different driving sessions. However, facial videos recorded during drives with emotional stressors showed more spread in AU intensities and AU velocities than the facial videos recorded during drives with cognitive stressors (see Figure 2.3). This hints at more facial activity during distraction caused by emotional stressors. This preliminary analysis shows that an AU-based two-step approach could have the potential for detecting driver distraction under different conditions.

These initial observations are based on the average facial activity over all 22 AUs and over all frames in a sequence. A closer look at the level of individual AUs is necessary, in order to gain more insights into the facial activity that might accompany different types of stressors. For example, startle and fear expressions might occur in the case of unexpected system failure, yawns might occur during a relaxed drive, and lip movements caused by speech might accompany cognitive and emotional stress. A research study was conceptualised to explore various data-driven classical machine learning methods as well as different deep neural network architectures to detect driver distraction.[4] This study is described in detail in Preprint C.1.1 (Update 02.08.2020: This is now published in IEEE Access. See [56]). It was found that the intensities of AU25-LipsPart produced the most informative feature for recognising cognitive distraction. In addition, it was observed that AU intensities estimated by the proposed state estimation framework performed well in recognising distracted driving sessions (F1-scores above 90%). Combinations of facial and physiological signals were also explored in this study.

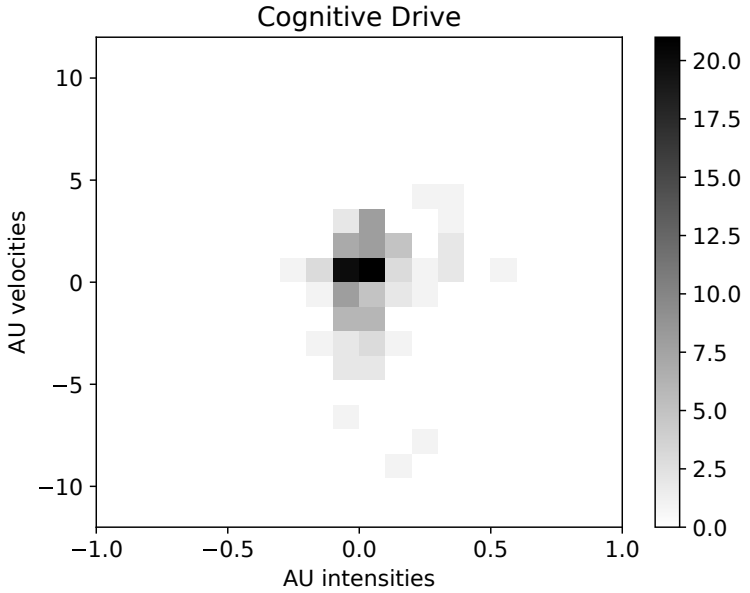
2.3 CHAPTER SUMMARY

This chapter summarised the state of the art in automatic pain detection from facial expressions, and presented an overview of the two-step approach for AU-based pain detection that was developed in this doctoral work. This two-step approach first estimates AU intensities by fusing facial shape and appearance information within a Gaussian state estimation framework. Afterwards, it feeds these AU intensities into pre-defined rules, in order to produce pain intensity estimates. This chapter also presented the analysis of facial activity during different types of distractions induced during simulated driving. The facial activity, described on the basis of AU intensities from the Gaussian state estimation framework, was found to differ between different types of distractions. The next chapter will elaborate the proposed Gaussian AU intensity estimation framework and its features.

4 The machine learning models were designed, trained and validated by Martin Gjoreski, as part of a research collaboration between Fraunhofer IIS, Germany, and Jožef Stefan Institute, Slovenia.



(a) Facial activity under an emotional stressor



(b) Facial activity under a cognitive stressor

Figure 2.3: 2D histograms showing the distribution of sequence-level facial activity during drives under emotional and cognitive stressors, computed on the dataset from Taamneh et al. [177]. Facial activity is represented by pairs of AU intensities (horizontal axis) and AU velocities (vertical axis). AU intensities are unitless and AU velocities have the unit seconds^{-1}.

3 AUTOMATIC FACIAL ACTION ESTIMATION

FACS identifies and codes 44 AUs, which are the basic facial muscle movements that can be visually distinguished by an observer [46]. Automatic detection of AUs from facial images and videos is the first step of two-step approaches for automatic mental state analysis. Apart from this, automatic AU detection is also useful as a stand-alone tool for faster annotation of facial expressions, in order to support behavioural and psychological research. Research on automatic facial action (AU) analysis has generally focused on the multi-class, multi-label classification problem of recognising AUs in the visual input, or on the task of estimating the intensities of those AUs [124]. Apart from these, the task of identifying the temporal phases, i.e. neutral, onset, apex and offset phases of the portrayed AUs, has also been examined in the published literature (e.g. [183]). In this doctoral work, the focus is on the regression task of estimating intensities of 22 AUs. In this chapter, the proposed AU intensity estimation framework, its components and features, and performance evaluation are described.

3.1 SUMMARY OF STATE OF THE ART

Research on automatic facial expression analysis gained momentum following the success of the real-time object detection algorithm proposed by Viola and Jones [189] in 2001, which enabled fast and robust face detection in images. Initial research on automatic facial action analysis focused predominantly on designing geometric and textural features that would better describe facial expression patterns, and consequently improve performance [124]. Machine learning models were trained with one or more of these features. More recently, CNNs replaced hand-crafted features with self-learned features (e.g. [58]). In all of these approaches, the main focus was the improvement of predictive performance, often measured in terms of accuracy for classification tasks and in terms of error and correlation metrics for regression tasks. These approaches are referred to here as *data-driven machine learning* approaches [70].

Data-driven machine learning approaches have used hand-crafted textural features such as Gabor filter coefficients (e.g. [108]), LBP (e.g. [18]), Local Phase Quantization (LPQ) [139] (e.g. [83]) and HOG (e.g. [20]). Combinations of textural features such as Local Gabor Binary Patterns (LGBP) [195] have also been examined for automatic facial action analysis (see [167]). Some works used spatiotemporal textural features such as LBP-TOP (e.g. [84]), LPQ-TOP [84] and LGBP-TOP

[4]. Among machine learning methods, SVMs have been used widely for automatic facial action analysis [124]. In addition, boosting algorithms and artificial neural networks have also been used [124]. For continuous-valued AU intensity estimation, Support Vector Regression (SVR) (see [82, 161]) and Relevance Vector Regression (RVR) (see [87]) have been used [124]. Tong et al. [181] and Li et al. [103] used AU annotations to learn the structure and parameters of a Dynamic Bayesian Network (DBN) to represent the semantic and temporal relationships between AUs. Self-learned spatial features using CNN (see [58]) as well as self-learned spatiotemporal features using a combination of CNN and LSTM recurrent neural networks (see [80]) have been used for the task of recognising AUs.

In parallel, another line of research explored the use of statistically or empirically determined parameterised and deformable models of facial shape and appearance for facial action analysis (e.g. [37, 38]). These are referred to here as *deformable face model-based* approaches. Different model-fitting methods to optimise the fitting of 3D shape and appearance models to 2D images of objects (e.g. faces) have been developed. These include Active Shape Model (ASM) [24], Active Appearance Model (AAM) [23, 26], Constrained Local Models (CLM) [28], and several variants of CLM, such as the regularised landmark mean-shift method [158] and the discriminative response map fitting method [6]. Such model-fitting methods have been developed with the objective of minimising the face alignment or face reconstruction errors, and are static methods that do not model the dynamic properties of rigid or non-rigid facial motion. Therefore, such model-fitting methods have been combined with state estimation methods such as Kalman filter and particle filter, which allow the modelling of the dynamics of model parameters separately (see [37, 38, 144]).

The statistically learned facial shape models, like the 79-point PDM of face used in [144], contain orthogonal vectors representing the most common non-rigid shape deformations of the face. These vectors need not necessarily provide direct semantic information about shape deformations caused by individual facial actions or AUs. Therefore, models such as CANDIDE-3[1] [2], which contain semantic deformation vectors representing AUs, were used for facial action analysis (see [37, 38, 76]). As already mentioned, the algorithms for model-fitting aimed to minimise fitting and reconstruction errors, and cannot–without additional modifications in some cases–deal with domain-specific semantic relationships or constraints between model parameters. For example, FACS does not allow negative intensity codes for AUs. The lowering of eyebrows can therefore not be coded as the raising of eyebrows in the opposite direction. It would be difficult to model such constraints on selected model parameters using the existing model-fitting algorithms. Therefore, these model-fitting methods were more often used for pre-

1 CANDIDE model is available for download at [3].

processing facial images in data-driven machine learning approaches, rather than as standalone solutions for automatic AU detection. For example, Chew et al. [20] used CLM as well as AAM for face alignment and tracking. Subsequently, they used the results of these model-fitting methods to define pixel-level appearance features for AU detection. Prabhu et al. [144] used the face model fitting results of ASM as noisy observations in their Kalman filter based facial landmark tracking approach. With respect to facial action analysis, the results from deformable face model-based approaches have been limited to either qualitative results or to a very limited number of AUs. Furthermore, very simple dynamic models such as autoregressive model (see [38]) or Gaussian random walk (see [37, 76]) have been used for modelling AU dynamics. None of these existing approaches have explored the use of viscoelastic models for modelling facial muscle dynamics. In addition, it was noted that those methods that attempted to estimate AU parameters (e.g. [38, 76]), did not enforce any range constraints on these parameters to resolve semantic ambiguties and to ensure FACS conformance. This makes the AU intensity estimates produced by these approaches difficult to compare or interpret consistently across different sequences.

Data-driven machine learning methods have the advantage that they can learn general models covering large variance in the training data, resulting in very good predictive performance. In contrast, the performance of deformable face model-based approaches could be limited by the accuracy and correctness of the models and the assumptions involved. However, the strength of deformable face model-based approaches lie in the interpretability of the models and its parameters, and the possibility to integrate interdisciplinary and human expert knowledge. Combining the two approaches could help in moving towards strong or ultra-strong AI systems [129] for facial action and mental state analysis that have good predictive performance and facilitate comprehensibility of decisions.

In this doctoral work, a Gaussian state estimation based method for estimating continuous-valued AU intensities is developed. This method combines an AU-based, deformable facial shape model, a viscoelastic model of facial muscle motion, and data-driven machine learning models for facial landmark detection and AU classification. Some researchers have already integrated probability outputs from data-driven machine learning methods in discrete state estimation frameworks. For example, Krüger et al. [95] combined SVM outputs within a Hidden Markov Model (HMM) for speech recognition; Valstar and Pantic [184] combined SVM and Hidden Markov Model (HMM) for recognising the temporal phases of AUs; Tong et al. [181] integrated outputs from a set of AdaBoost classifiers as noisy measurements (or, observations) within a DBN for recognising the presence of 14 AUs; Li et al. [103] integrated outputs from a set of SVMs as observations in a

DBN for estimating six discrete AU intensity levels for 12 AUs. Unlike these methods, in this dissertation, the fusion of the class-wise probability outputs from machine learning methods such as SVM, within a continuous, Gaussian state estimation framework is explored. The proposed AU intensity estimation framework is implemented mainly in the programming language Lua[2] [74]. The following subsections provide more details about the design of the proposed AU intensity estimation method.

3.2 BASIC FRAMEWORK: PROBABILISTIC, DYNAMIC, HYBRID

A dynamic, probabilistic state estimation framework is used to estimate continuous-valued intensities of the 22 AUs listed in Table 2.1. State estimation methods track the state of a system with the help of a model describing how the state changes over time and by measuring the observable properties of the system. Mathematically, the *state* of a system is represented by a vector of variables, of which, some, all, or none may be directly observable. A dynamic model of the process that causes the state changes is called the *transition model*. This model is often not known accurately, and therefore, is a noisy approximation of the actual process. An *observation model* defines a mapping of the state to the system's observable/measurable properties. Actual observations are also usually ridden with noise. State estimation methods allow modelling of noise in *parametric* or *non-parametric* ways. Parameterised noise models describe noise in the form of a mathematical function. An example of a parameterised noise model is a single or multivariate Gaussian model with mean(s) and covariances as parameters. Non-parametric noise models represent or encode noise as a collection of points in a corresponding sample space. For example, noise in a state estimate could be represented as a set of plausible state estimates. In the non-parametric case, the underlying noise distribution is encoded non-parametrically in the distribution of the sample points. In this work, a state estimation framework with the parameterised Gaussian noise model is chosen, for practical convenience. As will be seen later, the state space is high-dimensional. Therefore, a large set of sample points would be required to model noise non-parametrically. This could prove to be computationally expensive and therefore, unsuitable for practical, real-time applications.

There are two main steps in the dynamic state estimation process [175], as shown in Figure 3.1: (i) *prediction* and (ii) *correction*. Starting with a pre-defined initial state at time step $k = 0$, the prediction step predicts the state at the current time step k based on the state at the previous time step $k-1$, by applying the transition model that describes how the state changes over time. In other words, the evolution of the state during the time interval $(t_{k-1}, t_k]$ is computed

2 https://www.lua.org/

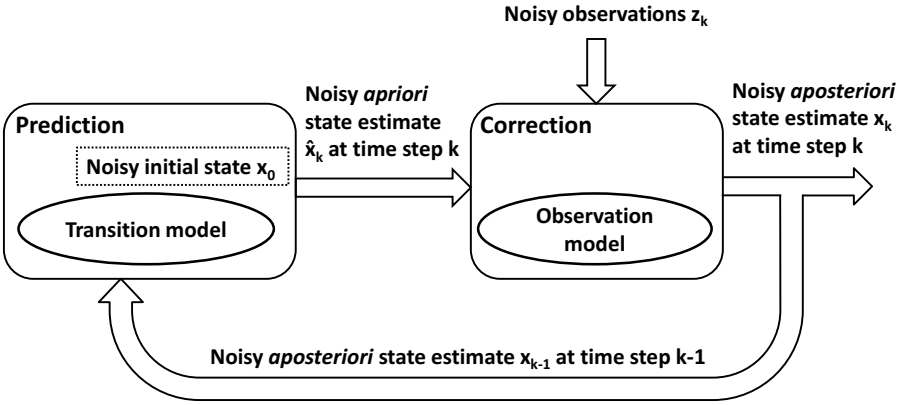


Figure 3.1: State estimation involves two steps: prediction and correction. In the prediction step, the *aposteriori* state estimate from the previous time step $k-1$ or the initial state at time t_0, if $k=1$, is transformed through the transition model to obtain the *apriori* estimate for the state at the current time step k. This is followed by a correction step, in which the *apriori* state estimate is corrected based on the evidence provided by the noisy observations. The observation model predicts the observable properties based on the *apriori* state estimate. After correction, the less noisy *aposteriori* state estimate for the current time step k is obtained.

using the transition model[3]. The predicted state is referred to as the *apriori* estimate at time step k, and is denoted as $\hat{\mathbf{x}}_\mathbf{k}$. The noise in $\hat{\mathbf{x}}_\mathbf{k}$ is a combination of the noise in the state estimate from the previous time step $k-1$ and the noise in the transition model that was applied. In the correction step, measurements or observations about the state are used to correct $\hat{\mathbf{x}}_\mathbf{k}$. The amount of correction applied depends on two factors: (i) the difference between the actual observations $\mathbf{z}_\mathbf{k}$ and the observations predicted based on $\hat{\mathbf{x}}_\mathbf{k}$; and (ii) the amount of noise in the actual and predicted observations. To predict observations, observation models that map the state to observations are used. The new state estimate resulting from the correction step is referred to as the *aposteriori* or *filtered* estimate at time step k, and is denoted as $\mathbf{x}_\mathbf{k}$. It is used in the prediction step of the next time step $k+1$.

The Kalman filter, as described in [85, 86], is a linear state estimation framework that models the noise in the state transition model and the noise in the observations in terms of a unimodal, zero-mean, Gaussian distribution. The Kalman filter or one of its variants that is used with non-linear systems–the extended Kalman filter [81]–has the following features:

- **Predictions based on dynamic models:** $\hat{\mathbf{x}}_\mathbf{k}$ is predicted using a noisy state transition model that approximates the dynamics of internal process(es) that cause(s) the observed system behaviour.

3 It is assumed that there are no external control inputs to the system to influence the state estimate.

This model is also known as *process model*. The true process models of real-world systems are usually unknown or difficult to model accurately. Therefore, approximate models, defined as *state transition functions* $\mathbf{f_k}$ are used, and the mismatch with the real system behaviour is modelled as *epistemic noise* or *process noise*. This noise is represented using an error covariance matrix referred to as *process noise covariance matrix*, $\mathbf{Q_k}$. Thus, $\mathbf{f_k}$ and $\mathbf{Q_k}$ together constitute the process model (see Equation 3.1). $\hat{\mathbf{x}}_\mathbf{k}$ is computed as shown in Equation 3.5.

- **Reasoning based on uncertainty:** $\hat{\mathbf{x}}_\mathbf{k}$ is computed using a noisy process model and a noisy $\mathbf{x_{k-1}}$. The noise in $\hat{\mathbf{x}}_\mathbf{k}$ is therefore a sum of $\mathbf{Q_k}$ and the noise transferred from $\mathbf{x_{k-1}}$, and is represented using the *apriori state estimation error covariance matrix* $\hat{\mathbf{P}}_\mathbf{k}$ (see Equations 3.3 and 3.6). The methods used for observing or measuring system properties or its changes are also prone to noise. This noise is modelled using the *measurement noise covariance matrix* $\mathbf{R_k}$. While computing $\mathbf{x_k}$, the noisy $\hat{\mathbf{x}}_\mathbf{k}$ is corrected on the basis of the noisy $\mathbf{z_k}$ in a proportion that is determined by $\hat{\mathbf{P}}_\mathbf{k}$ and $\mathbf{R_k}$ (see Equations 3.7 and 3.8). The greater the noise $\mathbf{R_k}$ in the measured observations $\mathbf{z_k}$ compared to the noise in the observations predicted using $\hat{\mathbf{x}}_\mathbf{k}$, the smaller is the applied correction. The noise in $\mathbf{x_k}$ is represented by the *aposteriori state estimation error covariance matrix* $\mathbf{P_k}$ (see Equation 3.4), and is computed as shown in Equations 3.7 and 3.9. As a result of the correction step, the overall noise or uncertainty in the state estimate is reduced, and therefore, $\mathbf{x_k}$ is less noisy or more certain than $\hat{\mathbf{x}}_\mathbf{k}$. The computation of $\hat{\mathbf{x}}_\mathbf{k}$, $\hat{\mathbf{P}}_\mathbf{k}$, $\mathbf{x_k}$ and $\mathbf{P_k}$ is elaborated in [14, 81] as well as in Patent D.1.1.

- **Fusion of multiple information sources:** The measurements or observations of the system could come from one or more sources. The sources may be either sensors or data processing methods. Fusion of multiple observation sources is enabled by (i) augmenting the observations vector $\mathbf{z_k}$ and the noise covariance matrix $\mathbf{R_k}$, and (ii) by defining observation models, for each observation source. The observation models are denoted as $\mathbf{h_k}$ (see Equation 3.2).

$$\hat{\mathbf{x}}_\mathbf{k} = \mathbf{f_k}(\mathbf{x_{k-1}}) + \mathcal{N}(\mathbf{0}, \mathbf{Q_k}) \tag{3.1}$$

$$\mathbf{z_k} = \mathbf{h_k}(\hat{\mathbf{x}}_\mathbf{k}) + \mathcal{N}(\mathbf{0}, \mathbf{R_k}) \tag{3.2}$$

$$\hat{\mathbf{x}}_\mathbf{k} \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{P}}_\mathbf{k}) \tag{3.3}$$

$$\mathbf{x_k} \sim \mathcal{N}(\mathbf{0}, \mathbf{P_k}) \tag{3.4}$$

$$\hat{\mathbf{x}}_\mathbf{k} = \mathbf{f_k}(\mathbf{x_{k-1}}) \tag{3.5}$$

$$\hat{\mathbf{P}}_\mathbf{k} = \mathbf{F_k}\mathbf{P_{k-1}}\mathbf{F_k}^T + \mathbf{Q_k}, \textit{ where } \mathbf{F_k} \textit{ is Jacobian of } \mathbf{f_k} \textit{ at } \mathbf{x_{k-1}} \tag{3.6}$$

$$\mathbf{K_k} = \frac{\hat{\mathbf{P}}_\mathbf{k}\mathbf{H_k}^T}{\mathbf{H_k}\hat{\mathbf{P}}_\mathbf{k}\mathbf{H_k}^T + \mathbf{R_k}}, \textit{ where } \mathbf{H_k} \textit{ is Jacobian of } \mathbf{h_k} \textit{ at } \hat{\mathbf{x}}_\mathbf{k} \tag{3.7}$$

$$\mathbf{x_k} = \hat{\mathbf{x}}_\mathbf{k} + \mathbf{K_k}(\mathbf{z_k} - \mathbf{h_k}(\hat{\mathbf{x}}_\mathbf{k})) \tag{3.8}$$

$$\mathbf{P_k} = \hat{\mathbf{P}}_\mathbf{k} - \mathbf{K_k}\mathbf{H_k}\hat{\mathbf{P}}_\mathbf{k} \tag{3.9}$$

The Gaussian state estimation framework used in this doctoral work to estimate AU intensities has the following assumptions and features (see Patent D.1.1 for more details):

- **State estimation method:** In this work, it is assumed that the process models are first-order Markov processes, i.e. the *apriori* state estimate at time step k depends only on the *aposteriori* state estimate at the previous time step $k-1$. The noise is assumed to follow parametric, unimodal, zero-mean Gaussian distributions. As will be seen later, the process models are non-linear, are continuous in time, and are represented using differential equations. But, the observations are available at discrete time steps. Therefore, a continuous-discrete extended Kalman filter (see pages 290–293 in [14]) is used as the state estimation method.[4] The parameters of the 3D, deformable facial shape model, namely head pose parameters, AU parameters and SU parameters, constitute the main state variables. Depending on the process models used, additional parameters or variables are also included in the state vector.

- **Process models:** In this work, first-order Markov processes based on constant position model, constant velocity model and driven mass-spring-damper model are used as noisy process models. Just like in the preceding work [69], constant position model is used for the relatively constant SU parameters, and constant velocity model is used for the rigid, head motion parameters, i.e. 3D translation and 3D rotation parameters.[5] In contrast to [69], in this doctoral work, a driven mass-spring-damper model is used to model the AU parameters that are controlled by facial muscle motion. Taken together, these three process models describe how different elements of the 3D, deformable facial shape model evolve over time. For simplicity, it is assumed that the noise in these three categories of process models are independent of each other. The state vector contains a total of 185 elements. In Section 3.2.1, the driven mass-spring damper model and the associated Gaussian noise model are explained in detail. The noise in the process models correspond to *epistemic* uncertainty.

4 The preceding work [69] had used the discrete form of extended Kalman filter, because the process model and process noise used in that work modelled the cumulative effect over the time interval $(t_{k-1}, t_k]$.

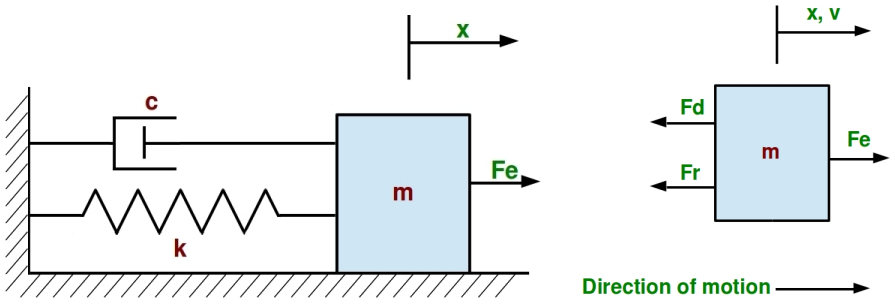
5 The scaling factor is set to 1.0 and kept constant. In effect, the scaling of the face model in the 2D perspective projection is effected via the translation along the z-axis.

- **Noisy observations:** The two sources of observation used in this work comprise facial landmark detection and appearance-based AU classification. As mentioned in Section 2.1.2, facial landmark detection provides the 2D positions of 68 facial landmarks in a facial image, and appearance-based AU classification uses textural features to detect the presence of specific AUs. The noise in the detected facial landmark positions is determined empirically, and the noise in the AU detections is modelled mathematically. The noise in the measurements represent *aleatoric* uncertainty. More details about the noise modelling are given in Sections 3.2.2 and 3.2.3.

- **Observation models:** The 3D deformable facial shape model in Equations 2.1 and 2.2 serves as the basis for the observation model to map state variables to facial landmark positions. Perspective projection of this facial shape model converts the state variables into 2D facial landmark positions by applying camera parameters. The rotation parameters as well as the perspective projection introduce non-linearity in the observation model associated with facial landmark positions. For the AU detections from the classifiers, an identify function maps the probability output for each AU to the corresponding AU parameter. More details are provided in Section 3.2.4, including a description of how the two sources of noisy observations are fused.

3.2.1 *Driven Mass-Spring-Damper Model*

AUs are caused by facial muscle movements that are characterised by the viscoelastic properties of the associated facial muscles or muscle groups. To model these viscoelastic properties, a mass-spring-damper system is used. It consists of a mass m attached to a spring having spring constant k and a damper having viscous damping coefficient c, and assumed to be moving on a frictionless surface. Compressing or extending the spring displaces it from its resting position. On the one hand, this creates a *restoration force* $\vec{F}_r$ in the spring that acts along the axis of the spring and in a direction opposite to that of the displacement of the spring. On the other hand, the damper opposes the motion of the spring by creating a viscous friction (*damping force* $\vec{F}_d$) that acts in the direction opposite to the direction of motion of the spring. The restoration force tries to bring the spring back to its resting position and is proportional to the displacement $\vec{x}$ of the spring (Hooke's law), provided the displacement falls within the limit of proportionality of the spring, i.e. $\vec{F}_r = -k\vec{x}$. The damping force tries to slow down the motion of the spring, and is proportional to the velocity $\vec{v}$ of motion, i.e. $\vec{F}_d = -c\vec{v}$. When an external force is applied to the mass, it elongates or compresses the spring, and thereby creates opposing restoration and damping forces. In the absence of a



(a) A mass-spring-damper system with an external force $\vec{F}_e$ acting on it.

(b) Free-body diagram showing the forces acting on mass m.

Figure 3.2: A mass-spring damper system driven by external force $\vec{F}_e$ is used to model the facial muscles that cause AUs. In this model, a mass m is attached to a spring with spring constant k and to a damper with viscous damping coefficient c. The external force $\vec{F}_e$ acting on the mass displaces the system from equilibrium, and thereby causes motion. Here, $\vec{x}$ is the displacement vector, $\vec{v}$ is the velocity vector, $\vec{F}_r$ is the restoration force, and $\vec{F}_d$ is the damping force.

damper, under ideal conditions, a displaced spring will produce non-stop simple harmonic oscillation, after the external force is removed. The frequency of this oscillation is referred to as the natural frequency ω_0 of the mass-spring system and is given by $\omega_0 = \sqrt{\frac{k}{m}}$. It is the viscous friction provided by the damper that enables a mass-spring-damper system to eventually settle to equilibrium. Three types of damping effects are possible: overdamped, underdamped and critically damped. This is determined by the damping ratio ζ, which is defined as $\zeta = \frac{c}{2\sqrt{km}}$. If $\zeta < 1$, the oscillation is underdamped; if $\zeta > 1$, the oscillation is overdamped; if $\zeta = 1$, the oscillation is critically damped.

Figure 3.2a shows a mass-spring-damper system and Figure 3.2b shows the free-body diagram of the forces acting on the mass m, when an external force $\vec{F}_e$ is applied to it. Based on the forces acting on m, the net acceleration $\vec{a}$ of m is the vector sum of the accelerations caused by $\vec{F}_e$, $\vec{F}_r$ and $\vec{F}_d$. The net acceleration $\vec{a}$ can be derived as given in Equations 3.10 to 3.13. Figures 3.3 and 3.4 show the response of different configurations of mass-spring-damper systems to a driving force having the form of a square pulse or a trapezoidal pulse. These figures reveal the non-linearity in the motion of mass-spring-damper systems.

To model AU dynamics using a driven mass-spring-damper system, the direction of elongation of the spring is assumed to be identical to the direction defined by the corresponding AU deformation vector. The external force $\vec{F}_e$ is modelled to act in this direction in order to create a positive displacement of the mass-spring-damper system.

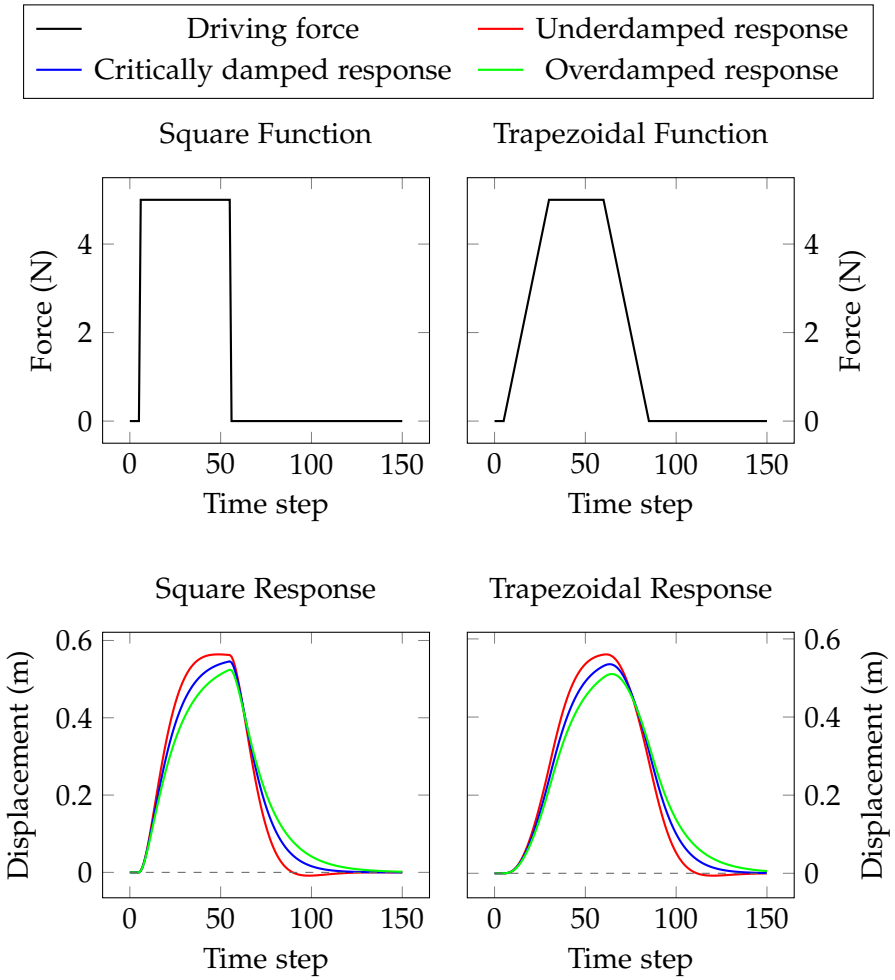


Figure 3.3: The plots in the bottom row show the responses (displacement in meters (m)) of three mass-spring damper systems to an external driving force (unit: newton (N)) having the form of a square pulse or a trapezoidal pulse. The square pulse includes a positive and a negative step (top left). The trapezoidal pulse includes a positive and a negative ramp (top right). The delay between successive time steps was set to 40 ms. All three mass-spring-damper systems had a natural frequency of 3 Hz, but differed in the damping ratio: 0.8 for underdamped, 1.2 for overdamped, and 1 for critically damped. It can be seen that during the positive step or positive ramp, the underdamped system produced the highest peak displacement, and the overdamped system produced the lowest peak displacement. Moreover, the underdamped system showed faster motion than the overdamped system. During the negative step or negative ramp, the underdamped system was the fastest to touch the equilibrium position (displacement = 0), but also overshot it. In contrast, the overdamped system needed more time to approach the equilibrium position, but did not overshoot. The responses of the critically damped system lie in between that of the other two systems.

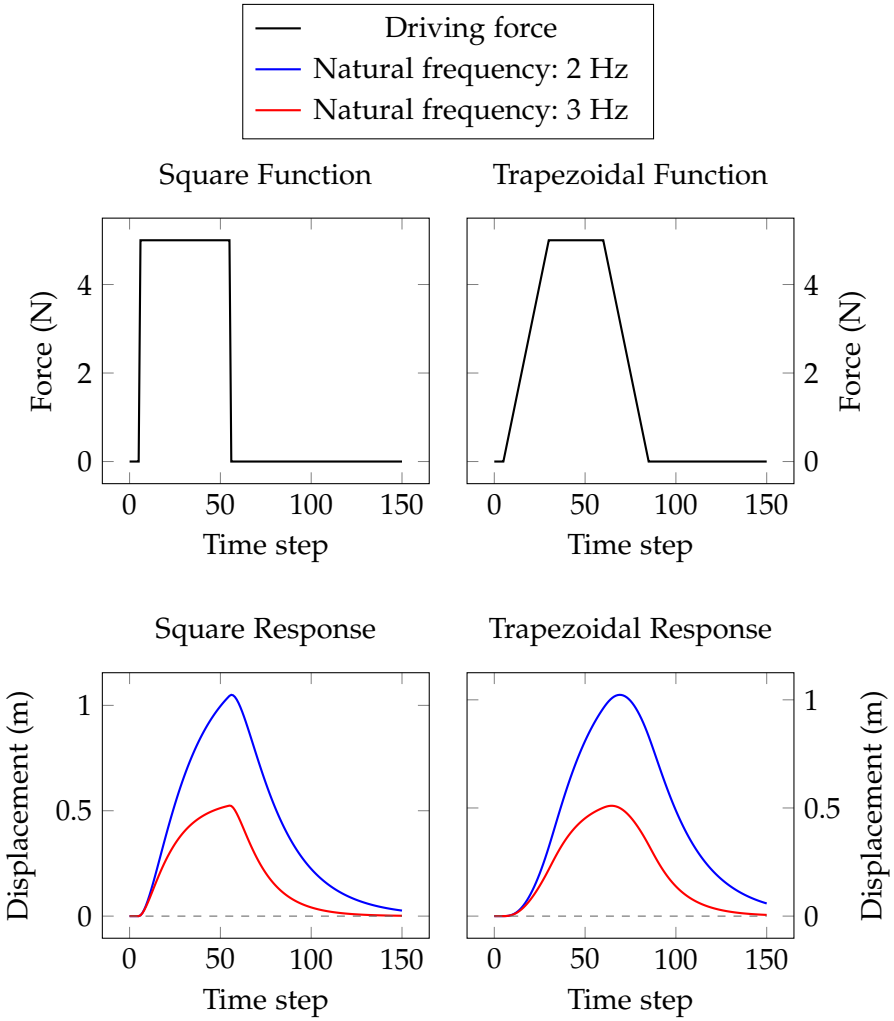


Figure 3.4: The plots in the bottom row show the responses (displacement in meters (m)) of two overdamped mass-spring damper systems to an external driving force (unit: newton (N)) having the form of a square pulse or a trapezoidal pulse. The square pulse includes a positive and a negative step (top left). The trapezoidal pulse includes a positive and a negative ramp (top right). The delay between successive time steps was set to 40 ms. Both mass-spring-damper systems had a damping ratio of 1.2, but different natural frequencies of 2 Hz and 3 Hz. The stiffer the spring, the higher is its natural frequency. It can be seen that the stiffer system (natural frequency 3 Hz) was displaced less and slower than the less stiff system (natural frequency 2 Hz), under the influence of the same driving force. Similarly, after the removal of the driving force, the stiffer system showed a slower rate of decrease in displacement. However, the stiffer system returned to the equilibrium position earlier due to it being closer to the equilibrium position.

Consequently, the magnitude of the displacement $\vec{x}$ represents the intensity of the corresponding AU. In order to prevent the mass-spring-damper system from overshooting the equilibrium position, when the external force is removed, a critically damped or overdamped configuration can be used. This, together with a restriction of $\vec{F}_e$ to non-negative values, ensures that the AU intensity estimates given by this dynamic model are non-negative. More details on constraining the range of AU intensity estimates are provided in Section 3.3.2.

$$\vec{F} = \vec{F}_e + \vec{F}_r + \vec{F}_d \tag{3.10}$$

$$\implies m\vec{a} = m\vec{a_e} - k\vec{x} - c\vec{v} \tag{3.11}$$

Diving both sides by m,

$$\implies \vec{a} = \vec{a_e} - \frac{k\vec{x}}{m} - \frac{c\vec{v}}{m} \tag{3.12}$$

Introducing ω_0 and ζ,

$$\implies \vec{a} = \vec{a_e} - \omega_0^2\vec{x} - 2\zeta\omega_0\vec{v} \tag{3.13}$$

A separate driven mass-spring-damper model is used for each AU. To integrate these models in the Gaussian state estimation method, five entities need to be defined for each driven mass-spring-damper model: (i) state vector $\mathbf{x}$; (ii) state transition function $\mathbf{f_k}$; (iii) Gaussian noise in process model $\mathbf{Q_k}$; (iv) initial state $\mathbf{x_0}$; and (v) Gaussian noise in initial state $\mathbf{P_0}$.[6] The state vector representation includes all factors on the right-hand side of Equation 3.13, namely x, v, a_e, ω_0 and ζ.[7] Since the model parameters ω_0 and ζ are not known exactly, and the external force F_e, which represents the facial muscle activation, changes dynamically according to the facial expression, ω_0, ζ, and a_e are included in the state vector and estimated along with the variables x and v. The state transition function $\mathbf{f_k}$ takes the form of a set of continuous-time ordinary differential equations (see Equation 3.14). Each equation defines the time-derivative of one element of the state vector. The elements a_e, ω_0 and ζ are assumed to be constant; hence, their time-derivatives are set to zero. However, these are updated during the correction step, based on the evidence provided by the observations $\mathbf{z_k}$. To compute $\hat{\mathbf{x}}_\mathbf{k}$, these differential equations are integrated numerically, with initial conditions set as $\mathbf{x}(0) = \mathbf{x_{k-1}}$. Fourth-order Runge-Kutta method with a single iteration (as per Table 6.5 in [79]) is used as the numerical method to solve these differential equations. The step size is set to $t_k - t_{k-1}$. The definitions of the state vector as well as $\mathbf{f_k}$ are based on the work of Liu and Ebbini [113]. However, this doctoral work differs from [113] in the

6 The suffix k in $\mathbf{f_k}$ and $\mathbf{Q_k}$ is used to indicate time-dependence, in general. However, these entities may also be unchanging or independent of time.

7 For simplicity, the vector notation $\vec{}$ is ignored here, and the vectors are represented as scalars, since the positive direction is defined implicity by the corresponding AU deformation vector.

treatment of the external force F_e. In [113], F_e is treated as a control input. In this doctoral thesis, F_e is treated as an unknown/hidden model parameter to be estimated by the state estimation method, and is included in the state vector (in the form of acceleration a_e). Moreover, only one iteration of fourth-order Runge-Kutta method is used, in contrast to the 32 iterations in [113]. This is because no significant differences were found in the empirical results when 32 iterations were used instead of one iteration. In addition, fewer iterations reduce computational cost. The application domain also differs between the two works. In [113], a driven mass-spring-damper system (referred to as "forced harmonic oscillator") is used for modelling the dynamics of viscoelastic tissue displacements. In this doctoral work, mass-spring-damper models are applied to model the dynamics of different facial actions or viscoelastic facial muscle movements.

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = a; \quad \frac{da_e}{dt} = 0; \quad \frac{d\omega_0}{dt} = 0; \quad \frac{d\zeta}{dt} = 0 \tag{3.14}$$

The process noise $\mathbf{Q_k}$ associated with the mass-spring-damper model is modelled using non-zero variances for each of the differential terms given in Equation 3.14, excluding v. Since v is the time integral of a, the noise in v is contributed by the noise in a. Therefore, no separate noise is modelled for v in $\mathbf{Q_k}$ (i.e. the variance of Gaussian noise in v is set to zero). In this doctoral work, the values for these noise variances were determined empirically as $0; 0.03; 0.0001; 0.0001; 100$, respectively, for the differential terms in Equation 3.14. Furthermore, the noise in each differential term is considered to be independent of the noise in the other terms (i.e. noise covariances between these terms are set to zero).

As mentioned above, $\mathbf{f_k}$ computes the time-derivatives of the elements of the state vector. Accordingly, the continuous-discrete Kalman filter defines a differential form of the state estimation error covariance matrix, for use in the prediction step (see pages 290–293 in [14]). This is given in Equation 3.15. This equation is also numerically integrated using a single iteration of fourth-order Runge-Kutta method with initial conditions set as $\mathbf{P}(0) = \mathbf{P_{k-1}}$, in order to compute $\hat{\mathbf{P}}_\mathbf{k}$. $\mathbf{F_k}$ represents the Jacobian matrix of $\mathbf{f_k}$ computed at the state $\mathbf{x_{k-1}}$.

$$\frac{\mathbf{d}\hat{\mathbf{P}}}{\mathbf{dt}} = \mathbf{F_k P} + \mathbf{P}{\mathbf{F_k}}^T + \mathbf{Q_k} \tag{3.15}$$

The initial state $\mathbf{x_0}$ was set as $x = 0$; $v = 0$; $a_e = 0$; $\omega_0 = 3$ Hz; $\zeta = 1.2$ (overdamped). The noise covariance matrix $\mathbf{P_0}$ was configured as given in Equation 3.16, where Δt is the interval between two consecutive time steps ($\Delta t = t_k - t_{k-1}$). In the context of this thesis, Δt represents the interval between two consecutive image frames in a video.

So far, the process of incorporating a single driven mass-spring-damper model in the Gaussian state estimation method was described.

In order to extend it to include several driven mass-spring-damper models–one for each of the 22 AUs–the vectors and matrices mentioned above ($\mathbf{x_k}$, $\mathbf{f_k}$, $\mathbf{Q_k}$, $\mathbf{x_0}$, $\mathbf{P_0}$) are augmented appropriately with the elements for each mass-spring-damper model. In this way, the process models for all 22 AUs can be integrated into the same Gaussian state estimation framework. See also Patent D.1.1 for details.

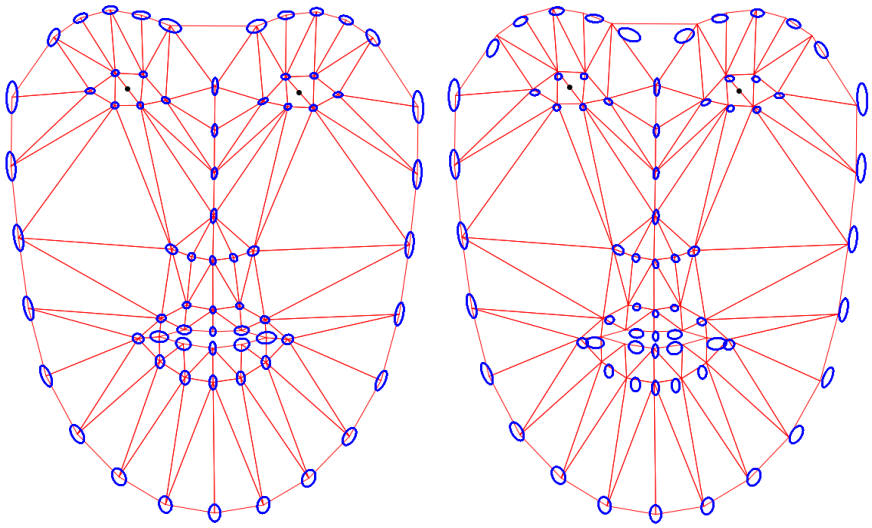
$$\mathbf{P_0} = \begin{bmatrix} 0.3 & 0.3(\Delta t)^{-1} & 0 & 0 & 0 \\ 0.3(\Delta t)^{-1} & 0.3(\Delta t)^{-2} & 0 & 0 & 0 \\ 0 & 0 & 0.0001 & 0 & 0 \\ 0 & 0 & 0 & 0.0001 & 0 \\ 0 & 0 & 0 & 0 & 100 \end{bmatrix} \tag{3.16}$$

3.2.2 *Noise in Facial Landmark Detection*

Similar to the preceding work [69], the noise in the detected facial landmark positions was computed empirically. The CK+ dataset [120] was used for this purpose. It contains annotations of 68 landmark positions in 2D pixel coordinates.[8] In this dissertation, a state-of-the-art face alignment method based on [90] is used to detect the facial landmark positions. The facial landmark positions detected for the images in the CK+ dataset were compared with the annotated positions, and subsequently, the sample mean and sample variance of the errors normalised by the distance between the eyes were computed. The noise in the landmark positions were computed under two different assumptions: (i) The noise in the pixel coordinates of each landmark are correlated with each other, but are uncorrelated with the noise in the pixel coordinates of other landmarks (independent noise configuration); (ii) The noise in the pixel coordinates of each landmark is correlated with the noise in the pixel coordinates of every other landmark (full noise configuration). These two noise configurations lead to two different state estimators. The independent noise configuration is used for AU05, AU10, AU13, AU16, AU17, AU23, AU27 and AU43. The full noise configuration is used for AU01, AU02, AU04, AU06, AU07, AU09, AU11, AU12, AU14, AU15, AU20, AU24, AU25 and AU26. The full noise configuration lends greater robustness or noise resistance to the state estimator, since individual landmarks cannot as easily influence the state estimate as in the independent noise configuration. Therefore, it is good for subtle AUs such as AU06-CheekRaiser and for AUs such as AU02-OuterBrowRaiser that might be incorrectly used to cover interpersonal shape variations in facial features such as eyebrows. The independent noise configuration can be visualised graphically in the form of 1-σ error ellipses, as shown in Figure 3.5a.

8 In the pixel coordinate system, the x-axis goes from left to right, and y-axis from top to bottom, with the origin at the top, left corner of the image.

The error ellipses represented in this figure correspond to the face alignment method based on [90][9], and are therefore different from the noise ellipses shown/used in the master's thesis [69] preceding this work. It can be seen from Figure 3.5a that the noise is normally higher along the edges than across them. The landmarks located on the facial boundary are more noisy than those located on the eye lids, nose and upper lip boundary. Figure 3.5b shows the same error ellipses shifted by adding the mean noise. The mean noise represents the systematic error in the facial landmark detections, and is subtracted from the facial landmark detections, before they are used to correct the state estimate.



(a) 1-σ error ellipses

(b) 1-σ error ellipses shifted by adding the mean errors

Figure 3.5: 1-σ error ellipses representing the empirically computed noise in the facial landmark detections from the face alignment method [90] applied in this work. The empirical noise was computed on the CK+ dataset [120]. The ellipses and the mean errors have been rescaled based on the 2D distance between the centers of the eyes in this perspective-projected wireframe model of the neutral face deformed with a mild parting of lips (AU25). The mean errors represent the empirically determined systematic error in the facial landmark detections.

9 The face alignment method based on [90] was not implemented by me. Only the outputs from this method were used in this doctoral work.

3.2.3 *Noise in Action Unit Classification*

In this dissertation, SVMs trained on HOG or LBP features are used as appearance-based classifiers.[10] The integration of these SVM classifier outputs in the state estimation method is an important contribution of this thesis, and is published in [70] (Publication B.2.1).

The appearance-based AU classifiers used in this work are of three different types: two-class, multiclass Type-I and multiclass Type-II. Two-class classifiers predict the probability of presence or absence of a single AU. In this work, two-class classifiers are used for detecting AU02, AU06, AU09, AU17 and AU25. Mutliclass Type-I classifiers predict the probabilities of all Boolean combinations of a selected number of AUs. In this work, a multiclass Type-I classifier is used for two AUs, namely AU01 and AU04. This classifier predicts probabilities for four classes: (i) both AU01 and AU04 present; (ii) AU01 present, but not AU04; (iii) AU04 present, but not AU01; (iv) neither AU01 nor AU04 present. From a multiclass Type-I classifier, the probabilities for individual AUs can be computed using the marginalisation technique. In the above example, the probability of presence of AU01 can be obtained by adding the probabilites for classes (i) and (ii). Similarly, the probability of presence of AU04 can be obtained by adding the probabilities for classes (i) and (iii). Multiclass Type-II classifiers predict probabilities of a set of individual AUs. In this work, a four-class SVM that detects the presence of AU12, presence of AU15, presence of AU26, and absence of AU12, AU15 and AU26 is used. This is an example of a multiclass Type-II classifier. With multiclass Type-II classifiers, the probability of presence of an AU is the probability output for that class.[11]

In order to integrate these AU probability outputs in the Gaussian state estimation method, the probabilities should first be converted to Gaussian noise variances. The probability p of presence of an AU A defines a Bernoulli distribution, where the Boolean random variable A takes value 1 (presence of AU) with probability p and value 0 (absence of AU) with probability $1 - p$. The second moment of this distribution is utilised as the Gaussian noise variance, for integration in the state estimation method. Figure 3.6 shows the second moment or variance of Bernoulli distribution as a function of the probability p that the Boolean random variable takes the value 1. As can be seen, the variance is highest when the probability is 0.5. This is the point of

10 The design, training and testing of the SVM classifiers were not done by me. Only the outputs from the SVMs were used in this doctoral work.

11 All the SVM classifiers used in this work have been created using the LIBSVM [17] software library. To convert the real-valued scores given by SVMs into pairwise class probabilities, LIBSVM uses the method proposed by Lin et al. [107], which is an improvement of Platt's method [143]. Individual class probabilities are computed from the pairwise class probabilities via pairwise coupling based on the second approach proposed by Wu et al. [201].

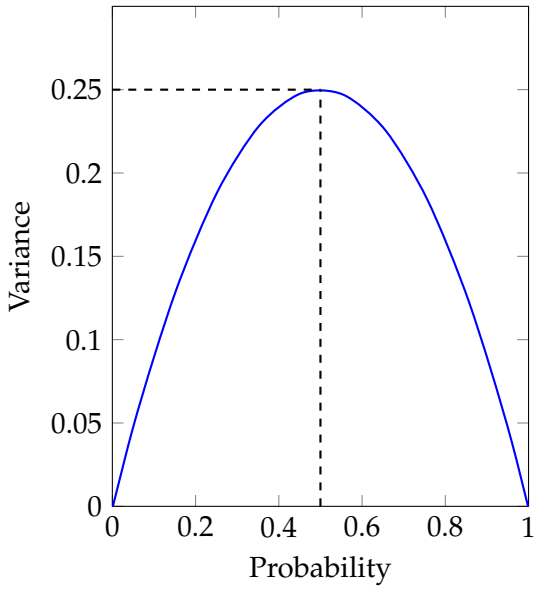


Figure 3.6: Variance of Bernoulli distribution, as a function of the probability of the Boolean random variable taking the value 1. This figure is also published in [70].

highest uncertainty about the value taken by the random variable, i.e. whether the AU A is present or absent. More details can be found in the Publication B.2.1.

3.2.4 *Fusion of Multiple Noisy Observations*

As mentioned earlier, the observation model for facial landmark positions is given by the perspective projection of the 3D, deformable facial shape model. The observation model for appearance-based AU classification is the identity function that interprets the probabilities as AU intensities [70]. This is based on the rationale that AUs expressed at higher intensities will create stronger appearance changes that lead to higher classification probabilities. A similar rationale was adopted in other works [9, 61]. Since probabilities as well as valid AU intensities belong to the range $[0, 1]$, a rescaling of probabilities is not necessary. In order to fuse the noisy observations from facial landmark detection and appearance-based AU classification, the observation vector $\mathbf{z_k}$, the observation model $\mathbf{h_k}$, and the measurement noise covariance matrix $\mathbf{R_k}$ are augmented. The performance improvement through the fusion of facial shape and appearance information was successfully demonstrated on three upper face AUs, namely AU01-InnerBrowRaiser, AU04-BrowLowerer and AU06-CheekRaiser, using image sequences from a proprietary market-research database (see Table 2 and Figure 3 in [70] or Publication B.2.1).

3.3 ENHANCEMENTS

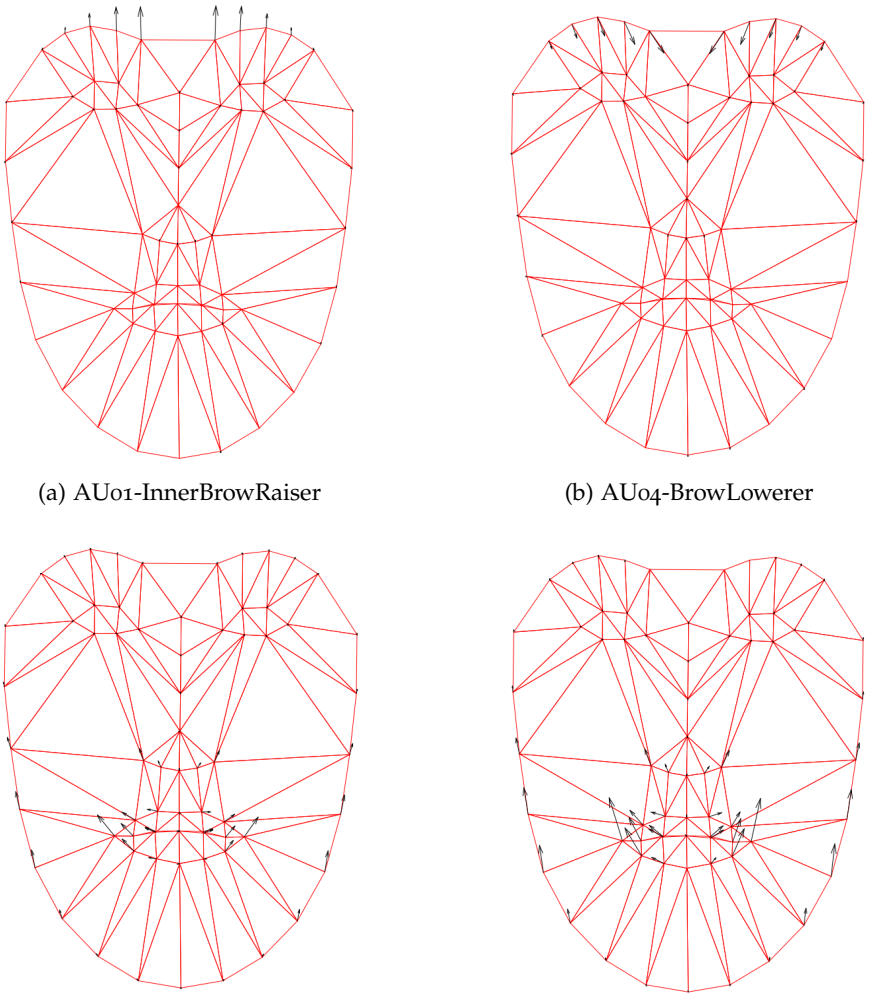
In this section, several enhancements made to the basic state estimation framework in order to deal with practical as well as domain-specific challenges are discussed.

3.3.1 *Action Unit Correlations in Noise Models*

The AU deformation vectors in Equation 2.1 represent the displacements in the positions of facial landmarks that are caused when AUs are displayed at maximum anatomically possible intensities. A visual inspection of these deformation vectors would reveal that several AUs, especially those in the same facial region, cause facial shape deformations that appear to be very similar to each other and involve the movement of similar facial landmarks, either in the same or in the opposite direction. For example, Figures 3.7a and 3.7b show that the deformation vector for AU04-BrowLowerer pulls the inner corners of the eyebrows downwards, whereas the deformation vector for AU01-InnerBrowRaiser pushes them upwards. From Figures 3.7c and 3.7d, it can be seen that AU12-LipCornerPuller and AU13-Sharp LipPuller vectors raise the lip corners upwards, but in slightly different angular directions. Computationally, the degree of similarity between any two AU deformation vectors can be determined by calculating the Pearson's correlation coefficient ρ. This is computed by first mean-normalising each of the 3D vectors, and then by computing the cosine of the angle between each pair of mean-normalised 3D vectors (cf. [41]). This correlation coefficient ρ is invariant to both scale and translation. It takes a value in the interval $[-1, 1]$, with negative values indicating negative correlation and positive values indicating positive correlation between the vectors. The higher the magnitude of ρ, the stronger is the correlation (positive/negative similarity). A value of zero represents independence (no similarity) between the vectors, in which case, the vectors are orthogonal to each other.[12] The closer the magnitude of ρ is to zero, the higher is the dissimilarity between the two vectors. Between the deformation vectors for AU01 and AU04, ρ had the value -0.78, indicating a strong negative correlation. Between the deformation vectors for AU12 and AU13, ρ had the value 0.94, indicating a strong positive correlation. When two AUs involve the movement of different facial landmarks, correlation coefficients are close to zero. For example, between AU01 and AU12, ρ had the value -0.06.

The values for ρ that were computed for each pair of AUs are visually depicted in Figure 3.8 as a matrix of ellipses. This visualisation of correlations as ellipses is inspired by [134]. In Figure 3.8, the blue ellipses represent positive correlations, and the red ellipses represent negative correlations. This information is also represented by the slopes of

12 Correlation coefficients are computed between two non-zero vectors.



(a) AU01-InnerBrowRaiser (b) AU04-BrowLowerer

(c) AU12-LipCornerPuller (d) AU13-SharpLipPuller

Figure 3.7: 2D perspective projections of the deformation vectors for AU01, AU04, AU12 and AU13 that are part of the face model in Equation 2.1. The vectors are derived from [96].

the (semi-)major axes that are marked using black arrows. Negative slopes of (semi-)major axes denote negative correlations, and positive slopes denote positive correlation. In addition, the length of a minor axis is inversely related to the strength of the correlation. That is, the stronger the correlation (positive or negative) between the corresponding pair of AUs, the shorter is the minor axis, and the narrower is the ellipse. In Figure 3.8, the strong negative correlation between AU04-BrowLowerer and AU01-InnerBrowRaiser is indicated by the narrow, red ellipse, whereas the strong positive correlation between AU12-LipCornerPuller and AU13-SharpLipPuller is indicated by the narrow, blue ellipse. An AU has the maximum positive correlation of 1 with itself, in which case, the minor axis of the ellipse has length zero. This reduces the ellipse to a straight line segment with a slope of 1. In contrast, the more uncorrelated the corresponding AU vec-

tors (ρ values closer to zero), the more circular are the ellipses, with major and minor axes of more or less equal lengths. For example, AU01-InnerBrowRaiser and AU12-LipCornerPuller are uncorrelated, as represented by the red circle.

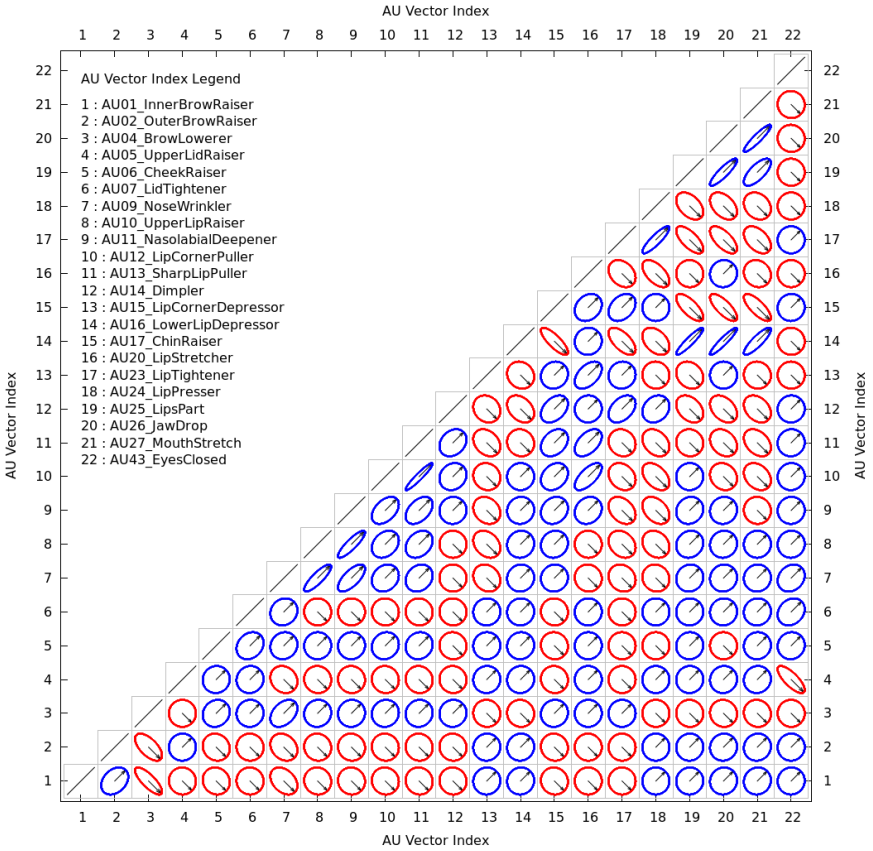


Figure 3.8: Correlation coefficients computed between pairs of AU deformation vectors. Positive correlations are shown using blue ellipses, and negative correlations are shown using red ellipses. The black arrows represent the semi-major axes of the ellipses. Positive or negative slope of a semi-major axis indicates positive or negative correlation between the corresponding AUs. Narrower ellipses represent stronger correlations. The more circular an ellipse, the more uncorrelated are the corresponding AU vectors. A black-and-white version of this figure is also published in the Patent D.1.1.

As can be seen in Figure 3.8, many of the AU deformation vectors have strong positive or negative correlations with each other. Therefore, the AU intensities that represent the displacements along these vectors cannot be considered as being uncorrelated to each other. Consequently, these AU correlation coefficients were introduced in the process noise covariance matrix $\mathbf{Q_k}$, in the initial state error covariance matrix $\mathbf{P_0}$, as well as in the AU constraint models, so that the noise in each AU intensity contributes a proportional amount of noise in

another AU intensity. The proportionality constant was set to the correlation coefficient between the corresponding AU deformation vectors. It is to be noted that the covariances between AU intensities computed in this way represent the face model-based noise and do not reflect the probability of co-occurrence of AUs. These covariances operate in such a way that higher uncertainties in one AU intensity estimate increase the uncertainties in the intensity estimates of other positively correlated AUs, and decrease the uncertainties in the intensity estimates of other negatively correlated AUs.

Another way to deal with the non-orthogonality of the facial shape deformations caused by AUs would have been the use of principal components of the original set of AU deformation vectors. However, this removes the semantic information about AUs and therefore, muscle-specific modelling of AU dynamics would no longer be possible. To learn the process model for the parameters associated with the principal components, a good amount of data with reliable AU intensity annotations would be required. However, this still remains an unresolved challenge.

3.3.2 *Constraints on Action Unit Intensity Range*

The AU deformation vectors defined in Equation 2.1 represent the maximum shape deformations that are anatomically possible when expressing the AUs. By virtue of this, the maximum value that can be taken by the AU parameters or AU intensities is 1. Furthermore, AU intensities are non-negative by definition in FACS [46]. Hence, the valid AU intensities are limited to the range $[0, 1]$. In order to ensure that the estimated AU intensities belonged to this valid range, state constraints were introduced in the extended Kalman filter framework. There are several ways of doing this, as mentioned in [174]. In the preceding work [69] that used a constant velocity model for facial motion, soft constraints were applied to the AU parameters during the correction step. Soft constraints do not have to be strictly met, in contrast to hard constraints. The underlying idea was to introduce an effect of pushing (or, equivalently, pulling) the AU intensities towards zero. The negative estimates were pushed more strongly than the positive estimates. In order to realise this, (i) additional measurements of zero–one for each AU parameter–were introduced in the observation vector $\mathbf{z_k}$, (ii) a zero-mean Gaussian noise was defined to control the softness of the constraint, and (iii) a linear observation model was defined to convert the noisy *apriori* AU intensity estimates to predicted observations. The same principle was adopted and extended in this doctoral work, in which the constant velocity models are replaced by driven mass-spring-damper models. In order to enforce the AU parameter range constraints, it is necessary to also constrain the direction in which the driving force is applied to the mass-spring-damper system, so that it

causes displacements only along the direction of the AU deformation vector. The modelling, working and impact of the constraints are described in [67] (Publication B.2.2).

3.3.3 *Handling of Anomalies in Face Alignment*

In this work, a face alignment method [90] is used to detect the positions of 68 facial landmarks in an image frame. These positions are used directly as observations in the AU intensity estimation method. These are also used to define the face region(s) from which textural features are extracted for AU classification. Occasionally, unseen[13] or sudden variations in facial shape and texture introduced by changes in illumination, face and head motion, or facial expressions can result in anomalous facial landmark detections that differ significantly from the predicted facial landmark positions. Furthermore, illumination changes or body movements can affect the texture of the clothing or the background. This could cause a sudden displacement of the facial landmark detections to a completely different region in the image frame. Since the accuracy and precision of face alignment are central to the performance of the AU intensity estimation method, anomalies in face alignment should be detected and prevented from being used to correct the *apriori* AU intensity estimates.

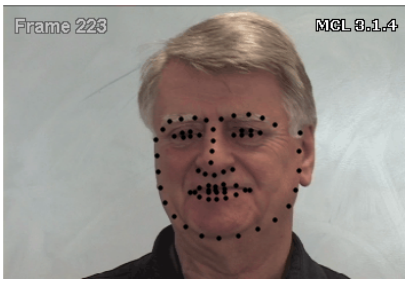
Detection of anomalies in measurements is a well-known problem in Kalman filter-based tracking applications. In [136], a solution based on normalised innovation squared is presented. Normalised innovation squared is a measure of divergence between the actual measurements and those predicted based on the *apriori* state estimates. Normalised innovation squared values follow chi-square distribution. The 68 2D facial landmarks contribute 136 degrees of freedom. Therefore, the normalised innovation squared values should be less than 192.707 to ensure a p-value of 0.001. However, in this doctoral work, the threshold for the normalised innovation squared values was set empirically to 325, by examining examples where face alignment produced anomalous results. The threshold was set very high, so that the sudden changes in facial landmark positions caused by facial muscle movements are not flagged wrongly as anomalies. This is more likely to happen in the case of AU43-EyesClosed, which involves a quick movement of eyelids, and in the case of AUs related to the opening of the mouth, namely AU25-LipsPart, AU26-JawDrop and AU27-MouthStretch. If the changes caused by these AUs are flagged as anomalies, then the state estimation method will miss to detect the activation of these AUs, thereby affecting its AU intensity estimation performance. With a high threshold, only extremely unlikely facial landmark detections would be flagged as anomalies. Figure 3.9 illustrates an anomalous and a non-anomalous set of facial landmark

13 Here, 'unseen' stands for 'not seen during the training of the face alignment method.'

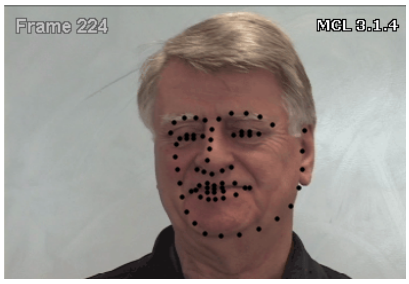
detections, in two consecutive image frames, and also provides the value of normalised innovation squared computed in each case.

(a) Frame no. 223

(b) Frame no. 224



(c) Frame no. 223 with detected landmarks



(d) Frame no. 224 with detected landmarks

Figure 3.9: An instance of anomaly in facial landmark detection in a sequence from the UNBC-McMaster Shoulder Pain Expression Archive Database [121]. On Frame 223, the value of normalised innovation squared was 135.441 (p-value > 0.001). This was well below the empirically determined threshold of 325. Therefore, the facial landmark detections for Frame 223 are marked as non-anomalous measurement. However, on Frame 224, the value of normalised innovation squared was 483.993, which exceeded the empirically set threshold of 325. Therefore, the facial landmark detections for Frame 224 are flagged as a highly unlikely or anomalous measurement. Facial images © Jeffrey Cohn.

If detected, anomalous facial landmark positions are not used to correct the *apriori* state estimate. The plot at the top in Figure 3.10 shows how an anomalous face alignment output in one frame corrupted the AU intensity estimates for several successive frames. The plot at the bottom in Figure 3.10 shows the AU intensity estimates for the same image sequence, when the anomalous facial landmark detections were ignored and not used to update the AU intensities. The noisy effect of the anomalies on AU intensity estimates were eliminated. If several successive frames are flagged as anomalies and dropped from the correction step, then the *apriori* state estimates will gradually drift over time and the normalised innovation squared based method would flag such situations as anomalous, even when the face alignment is correct. Furthermore, an update after several time steps would introduce noise in the AU intensity estimates. Therefore, if more than a specific number

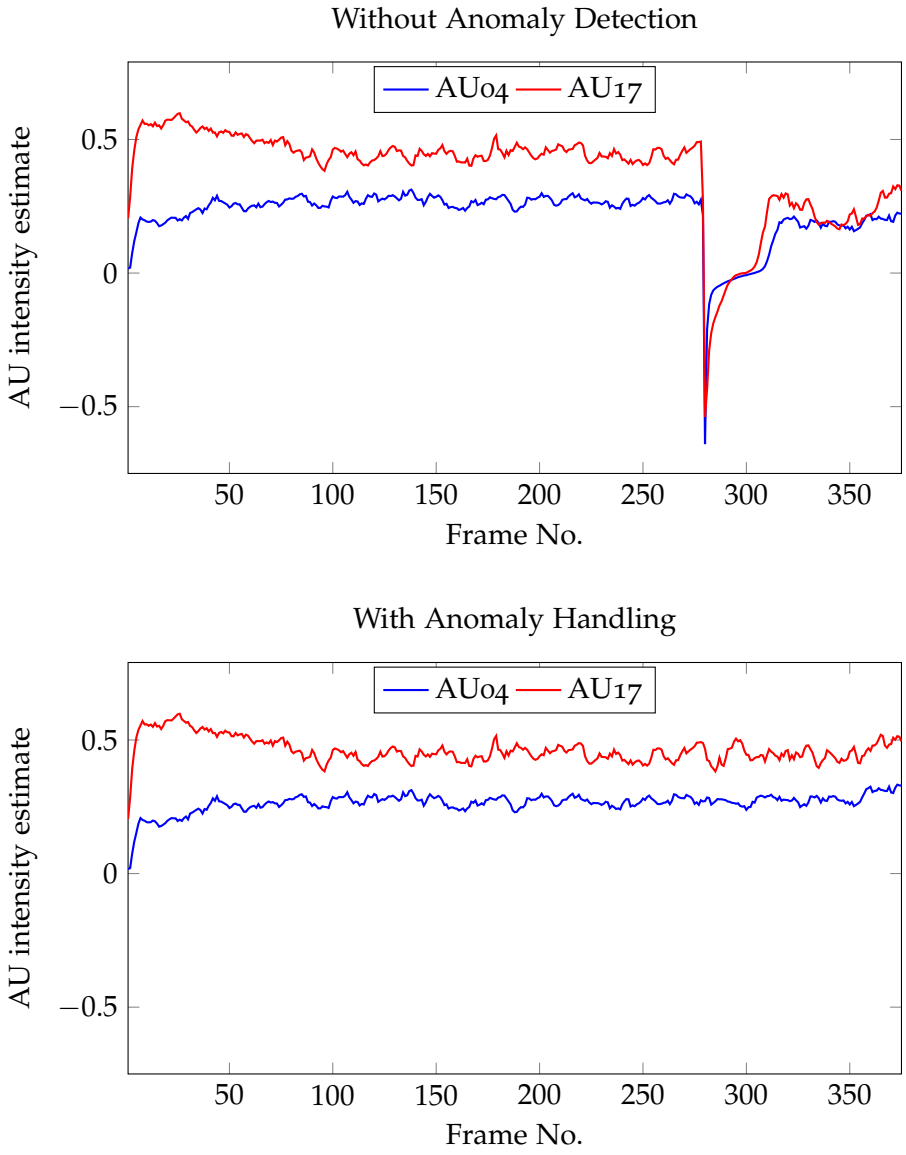


Figure 3.10: An illustration of the effect of the proposed strategy for handling anomalies in facial landmark detections. The effect is illustrated on a sequence from the proprietary market research database [70], recorded at 25 frames per second. For clarity, only two AUs, namely AU04 and AU17, are shown in the plots. The plot at the top shows the AU intensity estimates obtained when anomaly detection was not activated. The plot at the bottom shows the AU intensity estimates obtained when the anomaly was detected and removed from state estimation. This illustrates that, through anomaly handling, the noisy effects of anomalous facial landmark detections can be eliminated, which in turn would enhance the robustness of the system.

of consecutive frames produce anomalous facial landmark detections, then the state estimation should be suspended. The state estimator can later be reset, when a pre-defined timeout on suspension is exceeded

and facial landmark detections are available. The frames count for consecutive anomalies and the duration of timeout for resetting the state estimator could be determined empirically, as part of a future work. During the initial feasibility tests, the state estimation was suspended after three consecutive anomaly detections (frames count set to three), and reset when the next frame with facial landmark detections became available (timeout set to zero).

Sometimes, facial landmark detections are not available for a frame (missing measurement). In these cases, the correction step in state estimation is skipped. However, if no facial landmark detections are available for a pre-defined number of consecutive frames, then the state estimation is suspended until facial landmark detections become available again. In this work, the frames count for suspending state estimation due to missing facial landmark detections is set to three. If any of the AU classifiers fail to produce a prediction, then the corresponding elements are excluded from the observation vector, and the correction step is performed with the remaining measurements.

3.3.4 *Muscle-specific Models for Action Units*

As shown in Equation 3.13, the mass-spring-damper model for each AU has two internal parameters, namely the natural frequency of oscillation ω_0 and the damping ratio ζ. These parameters are part of the state vector, and are allowed to be dynamically updated, in order to account for epistemic modelling errors. These parameters were initialised identically for all AUs using values determined through trial and error. However, these parameters could be initialised differently for each AU, depending on the characteristics of the facial muscle(s) that produce(s) it. For this, a simple method based on facial muscle fibre composition is proposed here. Several histochemical studies [50, 64] have examined the composition of facial muscles. Happak et al. [64] grouped facial muscles based on the percentage of Type I muscle fibres that they contained. The subsequent work by Freilinger et al. [50] labelled these groups as phasic, intermediate and tonic muscles. Phasic facial muscles contain less proportion (14% to 15%) of Type I fibres. The orbicularis oculi muscle that is involved in AU06-CheekRaiser, AU07-LidTightener and AU43-EyesClosed belongs to this category. Intermediate facial muscles are composed of 28% to 37% Type I fibres. The muscles involved in several lip movements (AU10-UpperLipRaiser, AU12-LipCornerPuller, AU13-SharpLipPuller, AU15-LipCornerDepressor and AU20-LipStretcher) fall in this category (muscles: zygomaticus major, levator labii superioris, levator anguli oris, depressor anguli oris, and platysma [64]). Tonic facial muscles contain high proportions (41% to 67%) of Type I fibres. The occipitofrontal muscle involved in raising of eyebrows (AU01 and AU02) and the buccinator muscle involved in AU14-Dimpler belong

to this category. Phasic muscles respond fast, while tonic muscles are slow in responding to stimuli. As shown in Figures 3.3 and 3.4, mass-spring-damper systems with stiffer springs or stronger dampers are slower and milder in their responses to the driving force, whereas less stiff springs and weaker dampers respond faster and stronger to the driving force. Based on these pieces of information, different initial values could be configured for ω_0 and ζ for different AUs. In this work, a less stiff spring with ω_0 set to 0.8 and ζ set to 1.0 (critically damped) was used for AU06, AU07 and AU43, which are AUs that involve the action of the phasic orbicularis oculi muscle. Figure 3.11 gives a qualitative illustration of the effect of this adapted mass-spring

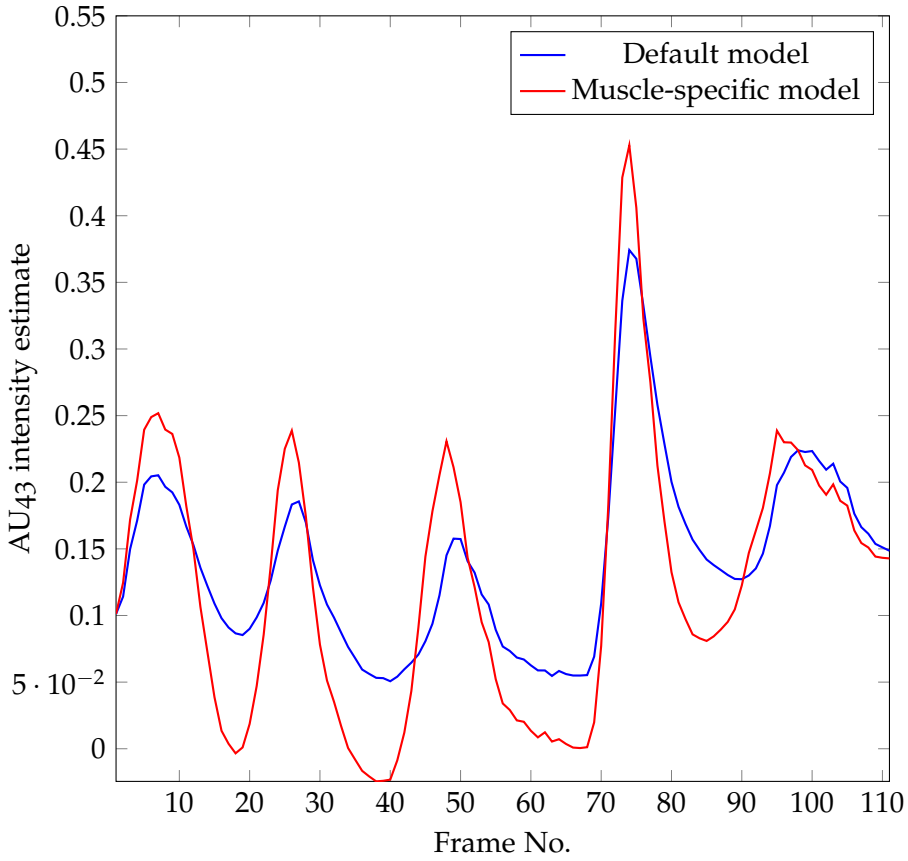


Figure 3.11: The plot illustrates the effect of using a muscle-specific model for AU43-EyesClosed on a sequence selected from the UNBC-McMaster Shoulder Pain Expression Archive Database [121]. The default model used a spring with ω_0 set to 3 Hz and a damper with ζ set to 1.2. The muscle-specific model used a spring with ω_0 set to 0.8 Hz and a damper with ζ set to 1.0. With this muscle-specific model: (i) the peaks of AU43 intensity estimates are sharper and higher; (ii) the intensities are able to come closer to zero, when the eyes are reopened; (iii) and the estimates increase or decrease at a faster rate. In sum, the responsiveness of the AU43 intensity estimates improved through the use of a muscle-specific model.

damper model on AU43 intensity estimates. It can be seen that the estimates obtained from the muscle-specific model show quicker and higher responsiveness than the estimates obtained from the default model. Empirical tuning of the parameters of the muscle-specific mass-spring-damper models could be investigated in a future work.

3.3.5 *Adapting to Person-Dependent Variations*

The facial properties such as shape and appearance vary between persons. These variations can affect the final shape and appearance of the face, when a facial expression is displayed by different persons. Figures 3.12 and 3.13 show some examples of facial shape or morphology variations between persons. Facial morphology includes the shape of face boundary as well as the shapes of facial features such as the eyebrows, eyes, nose and mouth. These variations are modelled using SU deformation vectors and SU parameters in the deformable facial shape model (see Equation 2.1). Facial appearance varies between persons due to the differences in the texture and tone of a person's skin. The textural variations include, for example, ageing-related wrinkles, presence of facial hair (beard and moustache), and facial occlusions caused by hairstyle. In this doctoral work, different methods have been applied to account for these interpersonal variations or to minimise the effect of these variations on AU intensity estimation. These methods can be categorised into (i) *one-time* calibration approaches, and (ii) *continuous* calibration approaches. In one-time calibration approaches, the person-specific characteristics are determined before the AU intensity estimation begins. In contrast, in continuous calibration approaches, the determination of person-specific characteristics and estimation of AU intensities are performed simultaneously.

One-time calibration of person-dependent facial shape and appearance can be performed on the first few frames of the sequence, if the person shows a neutral expression at the beginning of the sequence. Alternatively, it can be performed on a separate sequence, where the same person shows a neutral expression. First, let us look at one-time facial shape calibration. To determine the SU parameters in these cases, the state estimation process is run on these neutral (sub)sequences with AU intensity estimation disabled by setting the inital AU parameters, their error covariances, and the AU-related process noise to zero. This causes the state estimation process to estimate only the head pose (rigid) parameters and the SU parameters. The *aposteriori* SU parameter estimates obtained for the last frame of a neutral (sub)sequence are then used as fixed/calibrated SU parameters for the corresponding person. The AU parameter estimation is then activated by setting the AU-related process noise as modelled in Section 3.2.1. In the case where the SU parameters are determined on the first few frames of the sequence, the final *aposteriori* state estimation error covariances

of the SU parameters are transferred to the AU parameters, before the AU parameter estimation is activated. This is done to transfer the facial landmark fitting errors to the corresponding AU parameters. It is also done so as not to disturb the equilibrium of the system through sudden changes in the error covariances. This transfer of noise uses the Jacobian ($\mathbf{H}$) of the 2D perspective projection of the 3D deformable facial shape model, and the *aposteriori* state estimation error covariance matrix $\mathbf{P}$. The transfer of noise is described in Equations 3.17 to 3.19. The subscripts $\mathbf{r}$, $\mathbf{s}$, $\mathbf{a}$, and $\mathbf{l}$ represent the row/column indices of the rigid parameters, SU parameters, AU parameters, and facial landmark coordinates, respectively. Further adaptation of SU parameters is disabled during AU parameter estimation by setting the SU-related process noise and error covariances to zero.

(a) (b)

Figure 3.12: Different facial morphologies: Examples from Actor Study Database [168].

(a) (b)

Figure 3.13: Different facial morphologies: Examples from the UNBC-McMaster Shoulder Pain Expression Archive Database [121]. Images © Jeffrey Cohn.

$$\mathbf{H}_{\mathbf{l},\mathbf{a}}\ \mathbf{P}_{\mathbf{a},\mathbf{a}}\ \mathbf{H}_{\mathbf{l},\mathbf{a}}{}^{T} = \mathbf{H}_{\mathbf{l},\mathbf{s}}\ \mathbf{P}_{\mathbf{s},\mathbf{s}}\ \mathbf{H}_{\mathbf{l},\mathbf{s}}{}^{T} \tag{3.17}$$

$$\mathbf{H}_{\mathbf{l},\mathbf{r}}\ \mathbf{P}_{\mathbf{r},\mathbf{a}}\ \mathbf{H}_{\mathbf{l},\mathbf{a}}{}^{T} = \mathbf{H}_{\mathbf{l},\mathbf{r}}\ \mathbf{P}_{\mathbf{r},\mathbf{s}}\ \mathbf{H}_{\mathbf{l},\mathbf{s}}{}^{T} \tag{3.18}$$

$$\mathbf{H}_{\mathbf{l},\mathbf{a}}\ \mathbf{P}_{\mathbf{a},\mathbf{r}}\ \mathbf{H}_{\mathbf{l},\mathbf{r}}{}^{T} = \mathbf{H}_{\mathbf{l},\mathbf{s}}\ \mathbf{P}_{\mathbf{s},\mathbf{r}}\ \mathbf{H}_{\mathbf{l},\mathbf{r}}{}^{T} \tag{3.19}$$

Next, let us look at one-time facial appearance calibration. Variations in facial appearance influence the performance of the SVM-based AU classifiers that use textural features. This influence could be mitigated by debiasing the probability outputs produced by the AU classifiers.[14] This is done by computing the average of the probability outputs for each AU on the neutral (sub)sequence and subtracting these average values from the probability outputs produced by the classifiers during AU parameter esitmation. A rescaling of the debiased probablity outputs can also be performed to ensure that they belong to the range $[0, 1]$.

In contrast to one-time calibration approaches, the continuous calibration approaches determine the person-dependent facial shape and appearance variations during AU parameter/intensity estimation, and do not necessarily require a separate neutral (sub)sequence. In this doctoral work, the continuous approach to estimate person-dependent facial shape variations was also explored. It estimates SU parameters simultaneously with AU parameters as part of the same state estimator. In order to enable the slow adaptation of SU parameters, a constant position process model with low process noise is used. Constraints are also imposed on SU parameters to control the divergence of SU parameters from zero. These constraints act equally on positive and negative values of SU parameters. This continuous approach to deal with person-dependent facial shape variations was also part of my master's thesis [69] that preceded this doctoral research. A continuous approach for mitigating the effects of person-dependent facial appearance variations was not explored in this doctoral thesis, and could be examined in a future work.

There are evidences for interpersonal variations in muscle fibre composition of facial muscles such as those involved in jaw movements [93, 154]. These differences influence the dynamics of facial muscle motion. Therefore, the parameters ω_0 and ζ of the mass-spring-damper models used to model facial muscle motion should be able to adapt to these interpersonal variations. In order to do this, a continuous approach is used in this doctoral work. The parameters ω_0 and ζ are included in the state vector and estimated simultaneously with the AU intensities. A one-time approach could also be used, in which case, ω_0 and ζ are set as fixed after the calibration phase.

Figures 3.14 and 3.15 compare the results of one-time facial shape calibration and continuous facial shape calibration. Figure 3.14 examines the positions of facial landmarks estimated using the two approaches. It can be seen that there is no visually discernible differences in the facial landmark positions estimated by the two approaches. However, Figure 3.15 shows that the quality of AU intensity estimates improved when one-time facial shape calibration was used. This is due to the reduced interference from SU parameters, as they are estimated separately. Although the qualitative results speak in favour of one-time facial shape calibration, this approach requires a neutral

14 This approach was suggested by Dr. Jens-Uwe Garbas from Fraunhofer IIS.

(a) One-time shape calibration (b) Continuous shape calibration

Figure 3.14: Qualitative results showing the positions of facial landmarks estimated using one-time versus continuous facial shape calibration for the first frame of a sequence from the Actor Study Database [168]. In the one-time approach, only the SU and rigid parameters are used to fit the deformable facial shape model to the observed facial landmark detections. The resulting estimated facial landmark positions are shown in Subfigure 3.14a. In the continuous approach, the SU and AU parameters are estimated simultaneously, and are used along with the rigid parameters to fit the observed facial landmark positions. The facial landmark positions so estimated are shown in Subfigure 3.14b. The estimated facial landmark positions are almost similar in both cases. A few very subtle differences (which are not easily discernible via visual inspection) exist along the facial boundary, lips and eyelids.

(sub)sequence, to achieve best results. This is, however, difficult to mandate or control in real-world applications.

3.4 PERFORMANCE EVALUATION

Previous sections in this chapter already mentioned the qualitative or quantitative evaluation of various components and features of the proposed AU intensity estimation approach. In addition to these, the performance of the AU intensity estimation approach was compared with the performance of a state-of-the-art system for AU recognition [30]. The evaluation was performed on three datasets, namely the Actor Study Database [168], the UNBC-McMaster Shoulder Pain Expression Archive Database [121], and a proprietary market-research database that is not yet published. All datasets contain AU annotations at frame-level. The evaluation is described in detail in the to-be-submitted Publication C.2.1. It was found that the proposed approach performed quite well in recognising AUs, despite simultaneously estimating intensities for several closely resembling AUs. It also produced temporally smooth AU intensity estimates, which facilitates the identification of the different temporal phases of AUs.

In general, the estimates from the proposed approach are conservative, with delayed onsets and lower range of values for estimated

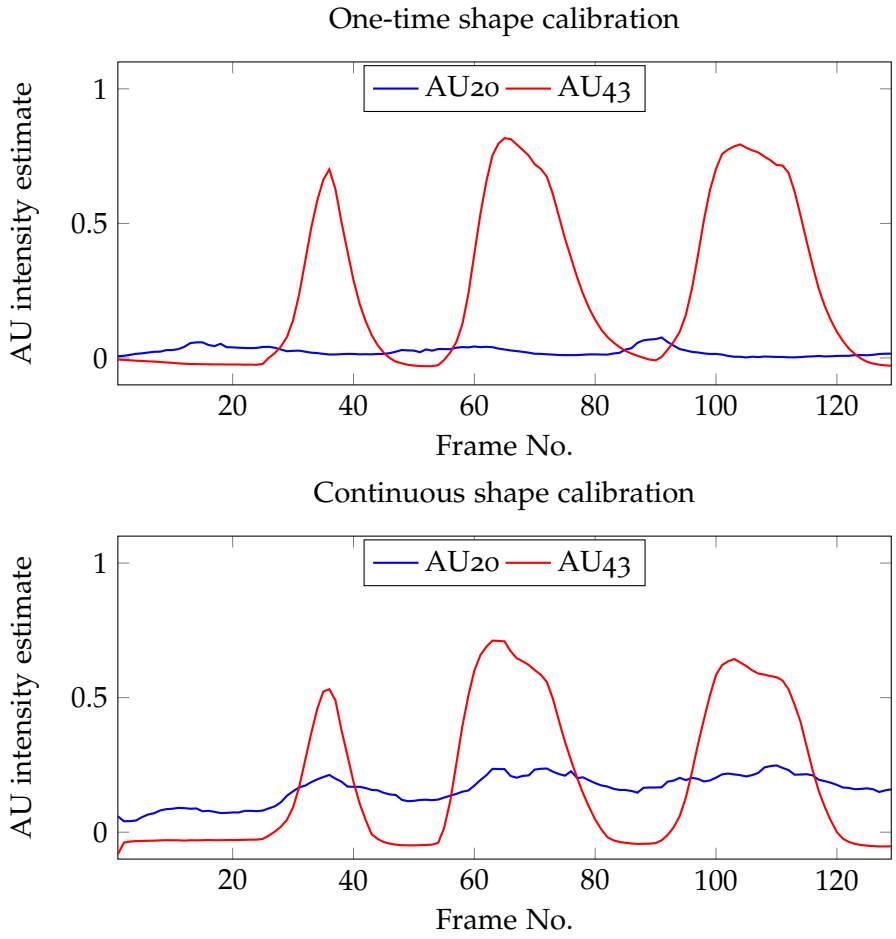


Figure 3.15: Qualitative results showing the effect of one-time versus continuous facial shape calibration on a sequence from the Actor Study Database [168], in which the actor displays AU43-EyesClosed. One-time shape calibration was performed iteratively on the first frame which was duplicated for 20 time steps. The *aposteriori* SU parameter estimates from the 20^{th} time step was used as the calibrated facial shape. After this calibration phase, the noise was transferred from SU parameters to AU parameters. In the plots above, it can be seen that continuous calibration created noisy intensity estimates in two ways. On the one hand, it used AUs for modelling person-dependent facial shape variations. For example, although the sequence did not inolve a stretching of lips, non-zero intensities are estimated for AU20-LipStretcher throughout the sequence. On the other hand, it used SU parameters to co-model AU-related facial shape changes. For example, it can be seen that the intensity estimates for the displayed AU43-EyesClosed became more prominent, when SU parameters were deactivated after the calibration phase of the one-time approach.

AU intensities. To make the intensity estimates less conservative as well as to improve the generalisation performance on unseen data, in the future, outputs from data-driven machine learning approaches such as [30] could be integrated within the proposed state estimation framework as additional evidence for AUs.

3.5 CHAPTER SUMMARY

This chapter described the probabilistic AU intensity estimation framework that was developed as part of this doctoral work. The framework is based on a Gaussian state estimation method, and it combines an AU-based deformable face model with data-driven AU classifiers. It models the facial muscle movements using driven mass-spring-damper systems, and fuses facial shape and appearance information to estimate continuous-valued intensities of 22 AUs. This chapter also presented the proposed practical approach to integrate probability scores from classifiers into a Gaussian state estimation framework. Several enhancements to deal with the person-specific and facial muscle-specific variations, and solutions for practical challenges such as anomalous and missing observations were also discussed. Additional features were presented to handle subtle AUs, to deal with the non-orthogonalities of facial shape deformations caused by AUs, and to enforce valid range constraints on AU intensity estimates. These components and features are summarised in Table 3.1. This chapter also provided a summary of the performance evaluation of the proposed AU intensity estimation approach on three facial expression datasets. The next chapter will discuss three open challenges in the field of automatic mental state analysis, and present some intial work done to address these challenges.

Table 3.1: Overview of the components and features of the proposed AU intensity estimation framework. These are grouped on the basis of the state estimation step where they are introduced. In addition, the features of the initial state are also listed.

Step	**Components and Features**
Prediction	Driven mass-spring-damper models for AUs
	AU correlations in process noise covariances
	Adaptation of facial shape and dynamics
Correction	Two different facial landmark noise configurations
	Fusion of facial shape and appearance measurements
	Constraints on range of AU intensities
	AU correlations in constraint noise covariances
	Handling of anomalous or missing measurements
	Adaptation of facial shape and dynamics
	Application of calibrated facial appearance
Initial state	Neutral facial expression
	Muscle-specific mass-spring-damper parameters
	AU correlations in initial state noise covariances

4

ADDRESSING OPEN CHALLENGES IN AUTOMATIC MENTAL STATE ANALYSIS

The multimodality and interpersonal variations associated with the expression of mental states call for building multimodal datasets and performing detailed analyses of the influence of various internal and external factors, such as personality, age, gender, diagnostic status, context and cultural background, on the activation and expression of mental states. Moreover, medical applications such as pain or depression detection [186] demand transparency in the decision models. That is, the learned models should be interpretable to humans, and it should be possible to generate comprehensible explanations of the predictions made by the models [59, 194].

In this chapter, some preliminary work done towards addressing three open challenges in the field of automatic mental state analysis is summarised. This includes the development of a set of requirements for reference datasets for mental state analysis, the analysis and modelling of interpersonal differences in the expression of mental states, and the generation of explanations for mental state detections.

4.1 REQUIREMENTS FOR MULTIMODAL REFERENCE DATASETS

Mental states such as stress and pain have been analysed via multiple modalities: physical (facial, vocal and muscular) and physiological (skin conductance, heart and breathing signals) [166, 170, 198]. Several datasets are available for mental state analysis that contain multimodal data (e.g. BioVid Heat Pain Database [190], WESAD [165][1], a multimodal distracted driving dataset [177]). However, the stimuli used to induce a mental state, the recorded modalities, the devices used to acquire multimodal data, and the self or observer-reporting methods used for annotating data vary between datasets [68, 166, 170, 198], making models developed on one dataset not comparable with those developed on another dataset. The use cases and settings that are considered for data collection also vary between datasets. Stress datasets have been collected for application contexts such as driving, computer-based work, or public speaking [123]. Pain datasets have been collected mostly under experimental conditions and occassionally under clinical settings [68]. Since there can be use case-specific opportunities and challenges in the acquisition of data, and use case-specific differences in the features and methods used for mental state detection, separate

1 WESAD is a publicly available database containing stress and affect data recorded using wearable sensors.

datasets for each use case is necessary. However, there is a need for defining a uniform framework for data collection and annotation, so as to improve the comparability of models. Earmarking a dataset or a group of datasets as a *reference dataset*, would enable more effective benchmarking.

A set of requirements that should be fulfilled by a multimodal reference dataset of human stress data is presented in [123] (Publication B.3.1).[2] These requirements can be generalised to mental state analysis as follows:

- **A representative sample size:** The number of subjects to be included in the reference dataset should be determined based on the statistics of the actual population cohort that is targeted. The distribution of the subcohorts within the chosen sample should be determined in a similar way.

- **An effective stimulus:** The stimuli chosen for inducing the target mental states in the subjects should be effective in eliciting the desired response. The characteristics of an effective stress stimulus have been described in [122]. The type and strength of stimuli should also be adapted to the individual sensitivity thresholds. For example, in the BioVid Heat Pain Database [190], the temperature levels for heat stimuli for pain induction were adapted to each individual's pain tolerance level.

- **Multiple modalities:** The reference dataset should include multiple modalities in order to cover possible interpersonal variations in the expression of the mental state. Moreover, the use of redundant and complementary information provided by multiple modalities could help in reducing false positives and false negatives during the detection of the mental state (cf. [57, 182, 196]). The reference dataset should include the modalities that have been shown to be reliable indicators of the mental state. For example, electrodermal activity and electrocardiogram should be included in stress datasets due to their correlation with cortisol levels [116, 187], and facial expressions should be included in pain datasets due to their validity [99, 171] and similarity across experimental and clinical pain conditions [98] as well as across different diagnostic status [11, 99, 101].

- **Information about personal and contextual factors**: Comprehensive information should be gathered about the main internal factors (e.g. age, gender, diagnostic status) and external factors (e.g. medicine intake, exercise, food intake, sleep pattern) that could influence the response to the administered stimuli. This is

2 These are results from the research and development work of Bhargavi Mahesh that was co-supervised by me, together with Prof. Dr. Erwin Prassler, Bonn-Rhein-Sieg University of Applied Sciences. My contributions are listed in Appendix B.3.1.

necessary to filter or model any known impact of these factors on the recorded multimodal response data.

- **Sensor specification (noise, calibration information):** In order to build more generalisable and robust models that are device-independent, it is essential to include information about the noise in the recorded signals as well as information about the calibration parameters of the sensors. These details should therefore be available in reference datasets.

4.2 ANALYSIS OF INTERPERSONAL DIFFERENCES

Interpersonal differences have been observed in the physical and physiological responses to various stimuli intended to cause mental state changes (cf. [54, 97, 176]). Taking these interpersonal differences into consideration is crucial for building generalisable automatic mental state analysis models. Some works attempted to implicitly learn different models for different cohorts by including personalised features as input to the learning algorithms. For example, Liu et al. [114] included personal information such as age, gender and complexion as input variables to learn models for estimating self-reported pain intensity levels, whereas Lopez-Martinez et al. [118] computed a personalised feature known as the individual facial expressiveness score, which was used as an input feature for the same task. Pain detection models have also been built to model each subject's response separately [117, 199].

In this doctoral work, interpersonal differences have been taken into consideration while designing the AU intensity estimation method and the AU-based pain detection rules. As mentioned in Section 3.3.5, several methods have been explored to account for the facial shape, facial appearance, and muscle-specific properties of a person during the AU intensity estimation process. The AU-based rules constructed for estimating pain intensities include the four clusters of facial expressions of pain (excluding the stoic response) identified by Kunz and Lautenbacher [97]. These rules are presented and validated in Preprint C.2.1. Figure 14 in Preprint C.2.1 illustrates the application of these rules to an image from the UNBC-McMaster Shoulder Pain Expression Archive Database [121]. Such composite rules based on individual clusters can help in detecting different types of facial expressions of pain, which would improve the generalisability of the pain detection system.

Besides the facial expression based mental state analysis, a preliminary investigation of the interpersonal differences in the pupil diameter changes caused by an arousal stimulus was conducted under my supervision. In this study [54] (Publication B.3.2)[3], pupil diameters

3 This study was conducted by Pelin Genc, under my supervision.

of subjects were recorded as they watched a video which contained a brief arousal stimulus. A qualitative analysis of the recorded data identified three types of pupillary response (i.e. changes in pupil diameter): increasing, decreasing and constant. Differences were also observed in the changes in the pupil diameters between different types of personalities. More details are provided in Publication B.3.2.

4.3 INTERPRETABLE MODELS AND DECISION EXPLANATIONS

In several real-world applications of automatic mental state analysis, especially in the medical field, it is crucial for humans to understand how an automated system makes its decisions. However, some of the most successful data-driven methods are based on SVMs and neural networks, which produce black-box models that are not comprehensible to humans [1, 59]. Therefore, these models–despite their advantages–cannot be deployed in critical applications. In order to improve the interpretability of black-box models and to make their decisions transparent to humans, several methods are being developed [59].

Following the success of deep CNNs in image recognition tasks [39, 94], more and more state-of-the-art approaches are applying CNNs for facial image analysis (e.g. [141]), including automatic recognition of pain from facial expressions (e.g. [179]). However, automatic pain detection approaches have seldom tried to distinguish pain from other emotions such as disgust and happiness [68]. Therefore, we trained a CNN model to discriminate between pain, happiness and disgust, and to apply explainable AI methods to make the predictions made by this CNN model transparent.[4] In [194] (Publication B.3.3), the use of two explainable AI methods, namely Local Interpretable Model-Agnostic Explanations (LIME) [149] and Layer-wise Relevance Propagation (LRP) [8], to explain the pain, happiness and disgust predictions made by the above-mentioned CNN model is illustrated with the help of image samples taken from the BioVid Heat Pain Database [190].[5] LIME and LRP can be used to make CNN models transparent, and subsequently help in identifying discrepancies in the models. In Publication B.3.3, qualitative explanations were generated for both correct and incorrect predictions using LIME and LRP. Facial regions around eyes, eyebrows, nose and lips were found to have contributed to the correct predictions of pain, disgust and happiness (see Figures 3, 5 and 6 in Publication B.3.3). For some input images, non-facial regions were found to improve prediction (see Figures 4 and 6 in Publication B.3.3), pointing to errors in the learned CNN model.

4 This research was conducted by Katharina Weitz, as part of her master's thesis that was supervised jointly by me and Prof. Dr. Ute Schmid, University of Bamberg.

5 Explainable AI methods other than LIME and LRP were also applied in Katharina Weitz's master's thesis to explain the predictions of the CNN model for recognising pain, happiness and disgust.

The two-step approach for AU-based pain detection (or, in general, mental state detection) presented in Figure 2.1 and described in Sections 3.2 and 3.3 has both interpretable and black-box components. These are described below:

- The 22 AU intensity estimates are interpretable, since they represent the amount of deformation along the facial shape deformation vectors defined in the deformable face model given in Equation 2.1.
- The components of the Gaussian state estimation framework such as the state transition functions and observation models are interpretable by virtue of design. The mass-spring-damper model of facial motion dynamics as well as the observation models for facial shape, facial appearance and AU intensity constraints can be presented graphically (see Figures 2.2 and 3.3 in this dissertation, and Figures 1.2 and 1.3 in Publication B.2.2). Such graphical representations aid in interpreting these models and their influence on AU intensity estimates.
- The uncertainties associated with the state estimates and observations can also be represented graphically (see Figures 3.5 and 3.6 in this dissertation, as well as Figures 6 and 7 in the Preprint C.2.1), which in turn facilitate the interpretation of their influence on AU intensity estimates.
- With regard to observations, the facial landmark detections can be visualised by superimposing them onto the input image. This helps in identifying any errors through visual inspection. As described in Section 3.2, the proposed AU intensity estimation approach uses predictions from SVM classifiers as additional observations. SVMs are, however, black-box models. To explain the decisions/predictions made by SVMs, the model-agnostic method LIME [149] or other explainable AI methods applicable to SVMs can be used (see [59]).
- In general, rule-based models that infer different mental states on the basis of AU intensities are interpretable. This is illustrated in Preprint C.2.1, where verbal explanations are generated for pain in terms of the detected AUs and their intensities, for a sample image from the UNBC-McMaster Shoulder Pain Expression Archive Database [121].

4.4 CHAPTER SUMMARY

This chapter identified three open challenges in the field of automatic mental state analysis, and summarised the initial work conducted towards addressing these challenges. First, a set of requirements that

should be fulfiled by multimodal reference datasets for automatic mental state analysis was identified. Second, the interpersonal differences in facial or physiological responses to pain or arousal stimuli were modelled or analysed. Third, the interpretable components of the proposed two-step approach for pain detection were examined, and sample explanations for pain detections were generated using AU-based rules. In addition, the use of explainable AI methods to explain the pain, happiness and disgust predictions from a CNN model was explored. More interdisciplinary research is required to address these challenges, in particular, (i) to gather and annotate multimodal reference datasets for mental state analysis, (ii) to systematically investigate the interpersonal differences in the expression of mental states across different modalities, and (iii) to generate and validate human-comprehensible explanations for automatic mental state predictions.

5 CONCLUSIONS AND OUTLOOK

5.1 AUTOMATIC FACIAL ACTION ESTIMATION

In this dissertation, a novel and hybrid approach for AU intensity estimation has been proposed and validated. This approach is based on a Gaussian state estimation framework, and it combines a deformable, AU-based facial shape model, a viscoelastic model of facial muscle motion, a set of appearance-based AU classifiers, and a facial landmark detection method. In order to integrate the probability scores from apperance-based AU classifiers in the Gaussian state estimation framework, a practical approach based on the variance of Bernoulli distribution and the marginalisation technique has been proposed and empirically evaluated. This integration of AU classifiers and AU-based deformable face model was found to improve AU recognition performance. The proposed AU intensity estimation approach provides both continuous-valued AU intensity estimates as well as the uncertainty associated with those estimates. This makes the proposed approach suitable for real-world applications that require a fine-grained analysis of facial expressions, accompanied by an uncertainty quantification, for probabilistic decision-making based on detected facial expressions.

Several enhancements have been proposed and integrated in the state estimation framework to deal with the person-specific properties, as well as technical and practical challenges. These enhancements contribute towards improving the quality of the estimated AU intensities. The Gaussian noise models and the anomaly detection feature enhance the robustness of the approach. Furthermore, several components of the proposed AU intensity estimation approach can be represented graphically, contributing to the enhancement of interpretability of the approach.

Qualitative and quantitative evaluations were performed on three facial expression datasets, in order to compare the performance of the proposed AU intensity estimation approach with that of a state-of-the-art, data-driven machine learning method for AU recognition [30]. It was found that the proposed approach produces temporally smoother estimates and performs quite well, even though it simultaneously estimates intensities for 22 AUs, many of which resemble each other closely. However, it was also found that the intensity estimates tend towards the lower range of values and are often characterised by delayed onsets. To overcome these limitations and to improve the generalisation performance, the outputs from the above-mentioned

state-of-the-art method could be integrated into the state estimation framework of the proposed approach.

Several other measures could be taken to further improve the performance of the proposed AU intensity estimation approach. The SVM classifiers could be replaced or augmented by better AU classifiers, probably based on deep CNNs, and trained on data containing more variance. The parameters of the facial muscle motion model could be tuned empirically with the help of sufficient amount of reliably annotated AU intensity data and additional evidence from the fields of biomechanics and histology. A deformable facial shape model of higher resolution, i.e. a model containing more than 68 facial landmarks, with denser coverage of cheeks, forehead and chin, combined with a face alignment method that can detect more facial landmarks could further improve AU intensity estimation performance. However, edge information is hard to obtain on cheeks, forehead and chin, making the annotation and detection of facial landmarks in these regions difficult. Therefore, new approaches for face alignment should be developed in the future, in order to overcome this challenge.

In addition, future work could explore whether human feedback can be integrated into the state estimation framework in the form of control commands in the process model. Such human feedback could help in correcting the current and future predictions on-the-fly. This could be a useful feature to speed up and improve the reliability of AU annotation of facial expression videos, and can be applied in psychological and behavioural research for facial expression annotation through human-computer collaboration.

A key characteristic of the proposed AU intensity estimation method is the assumption of unimodal Gaussian noise in the transition model and observations. However, Gaussian mixture models could be used for more realistic modelling of noise, and could be investigated in the future. It would also be interesting to explore the potential of Bayesian deep learning models [191] for probabilistic inference of AU intensities. A comparison of the performance of such models with the performance of the AU intensity estimation proposed in this dissertation might provide useful insights to extend the state-of-the-art.

In order to deal with practical challenges such as head rotations, illumination variations and occlusions, a combination of hardware and software solutions could be explored. Such solutions could use different active or passive strategies. Active strategies could involve repositioning of cameras or activation of artificial illumination. Passive strategies could involve the usage of alternate image analysis methods tailored for functioning under such challenging conditions. To deal with low resolution images, non-frontal faces, facial occlusions, or illumination variations, the use of face hallucination methods [147, 204] could be explored. Methods developed for face recognition from

low-quality images (e.g. Herrmann et al. [71]) might also be useful in processing low-resolution facial expression images and videos. Real-time performance is often important for real-world applications. The use of speed-optimised mathematical libraries and multithreading could bring gains in computational speed, and could be investigated in a future work.

5.2 AUTOMATIC MENTAL STATE ANALYSIS

In this dissertation, the AU intensity estimates produced by the proposed method have been applied for automatic detection of pain and for analysing facial activity during driver distractions. For automatic pain detection, a set of AU-based rules were defined based on evidence from experimental psychology [97, 146]. These rules were applied to the AU intensity estimates and the estimated pain intensities were evaluated empirically. In addition, the generation of verbal explanations for pain detections has been illustrated with the help of these AU-based rules. Future work could expand these explanations by integrating information about the uncertainty associated with the AU intensity estimates provided by the Gaussian state estimation method. In addition, these verbal explanations could be compared and contrasted with the explanations generated on the basis of image-based CNN models (e.g. [194]), and subsequently, a multi-level explanation strategy could be developed. The creation and evaluation of such explanation strategies would require interdisciplinary collaboration, especially between researchers from the fields of computer vision and psychology.

The preliminary analysis of facial activity during simulated driving sessions involving different types of distractions, showed differences in facial expressions between different sources of distraction. This inspired an effort to automatically detect driver distraction using the AU intensity estimates (see Preprint C.1.1. Update 02.08.2020: See publication [56]).

The results of automatic pain detection and driver distraction analyses show that the AU intensities estimated by the proposed approach are suitable for automatic mental state analysis. However, further research is necessary to develop AU-based pain rules that include the temporal characteristics of pain as well as contextual information such as intake of medication. Further research is also necesasry to identify the actual facial activity patterns associated with the different types of distraction. Future work could explore the use of AU intensity estimates and associated uncertainties for automatic mental state analysis in other application domains such as human-robot interaction.

To take the field of automatic mental state analysis forward, good quality data with reliable annotations are required. Interdisciplinary effort is needed to build reference datasets for benchmarking algorithms.

A few requirements for such reference datasets have been proposed as part of this doctoral thesis. Given the difficulty in manually annotating large volumes of data, the use of of semi-supervised and unsupervised methods for automatic mental state analysis should be investigated more extensively in the future. Interpersonal differences influence automatic mental state analysis. More interdisciplinary research is required in order to model interpersonal differences effectively, and to build systems that can dynamically adapt generic models to the expression characteristics of the individual. In cases where the identities of the persons are known, customised models can be built and later activated with the help of face recognition.

This thesis demonstrated that by combining data-driven machine learning approaches with deformable face model and state estimation methods, the predictive performance for AU detection can be improved, while facilitating robustness and interpretability in automatic mental state analysis based on facial expressions. Future research should focus on integrating different AI methods, in order to build strong and ultra-strong AU systems for automatic mental state analysis.

Part II

APPENDIX

A

ADDITIONAL RESULTS

A.1 MENTAL STATE ANALYSIS: DISTRACTION DETECTION

Figures A.1, A.2, A.3, and A.4 show additional results of the analysis of facial activity during different driving conditions. The results are obtained by analysing facial videos from the distracted driving dataset collected by Taamneh et al. [177]. Facial activity is represented by pairs of AU intensities and AU velocities. The AU intensities are obtained using the AU intensity estimation method developed in this doctoral work (see Sections 3.2 and 3.3). In the above-mentioned figures, practice drives, normal drives, relaxed drives, drives with unexpected system failure, drives under sensorimotor distraction, and the baseline (non-driving) conditions are included. The figures show that the distribution of facial activity is different under different driving and distraction conditions.

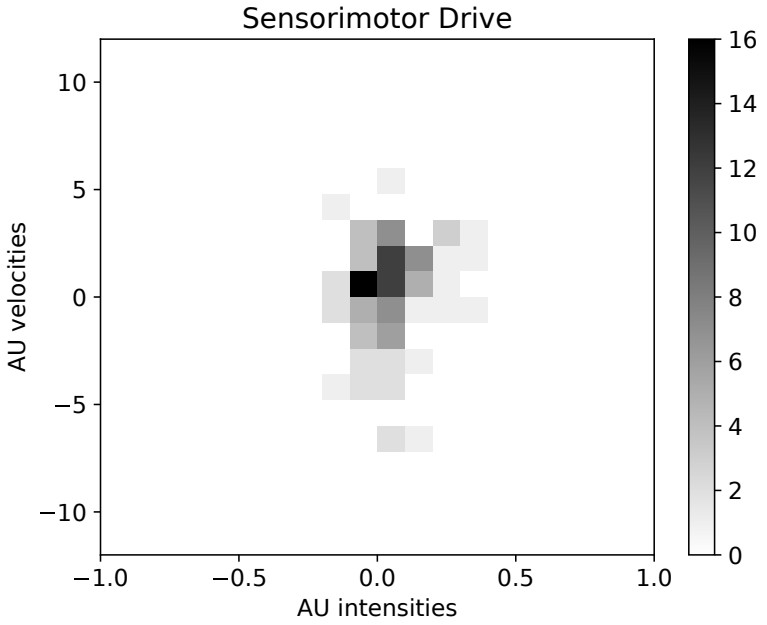
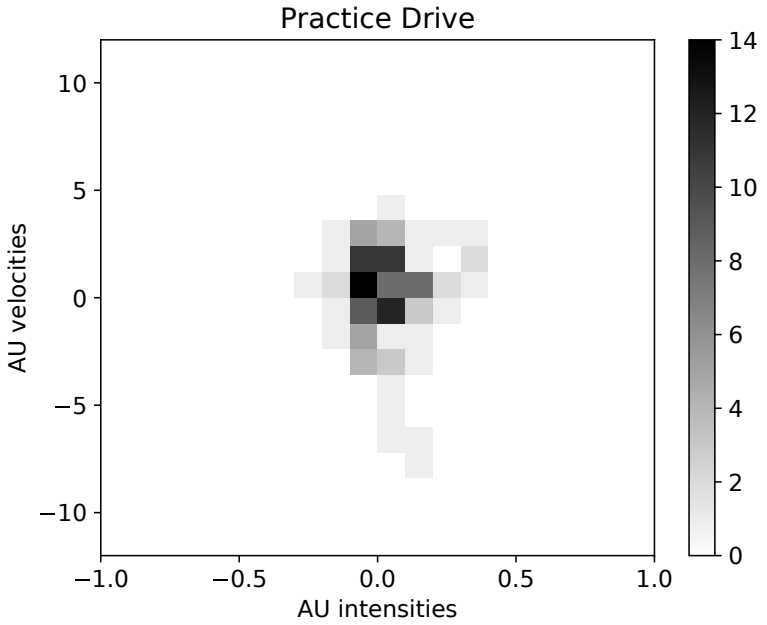
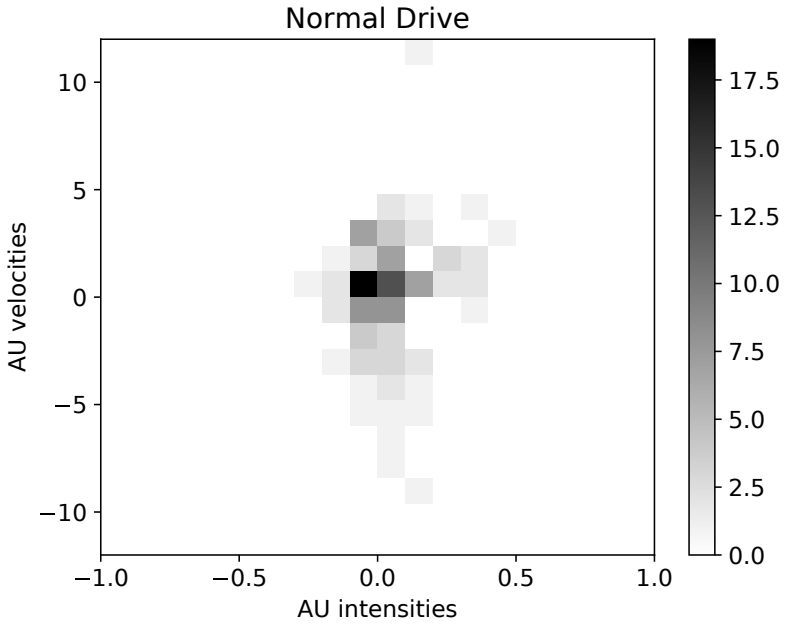


Figure A.1: 2D histogram showing the distribution of sequence-level facial activity while driving under the influence of sensorimotor stressors (dataset: [177]). Facial activity is represented by pairs of AU intensities and AU velocities. AU intensities (on the horizontal axis) are unitless. AU velocities (on the vertical axis) have the units: seconds^{-1}.

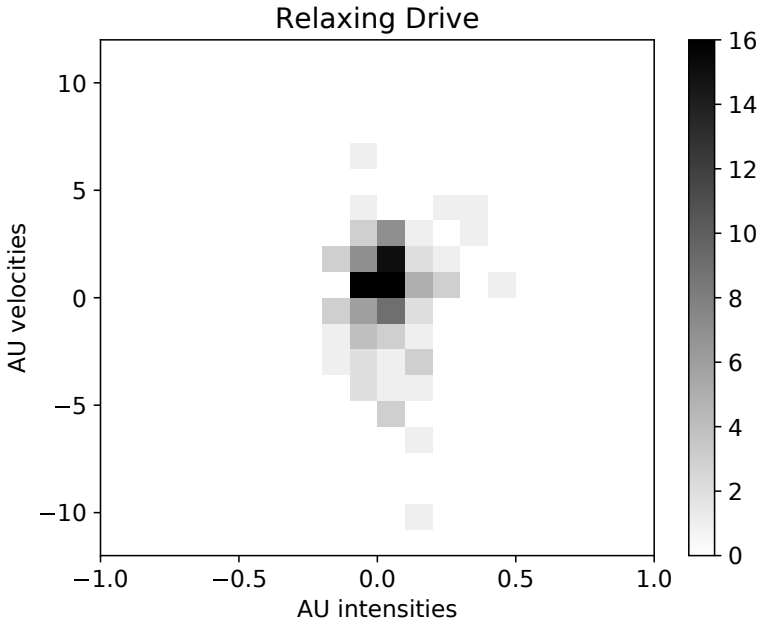


(a) Facial activity during practice drive

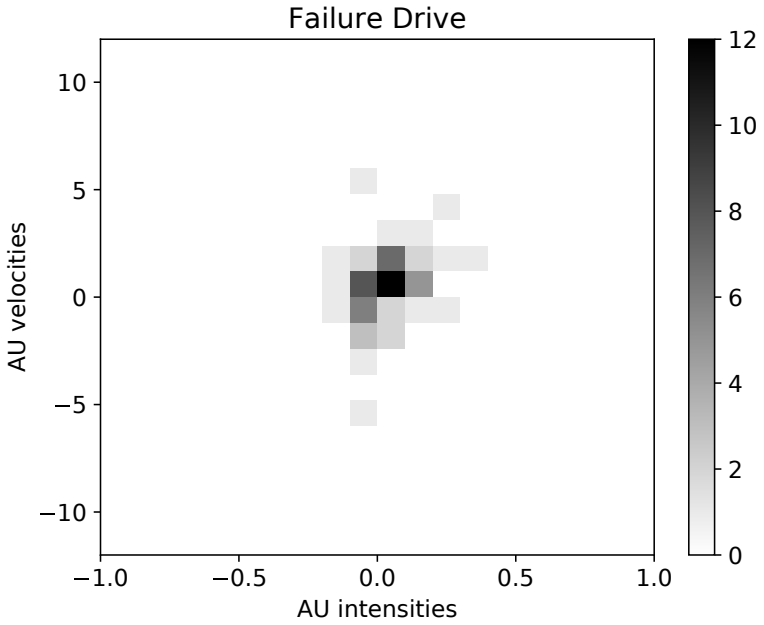


(b) Facial activity during normal drive

Figure A.2: 2D histograms showing the distribution of sequence-level facial activity during practice and normal drives (dataset: [177]). Facial activity is represented by pairs of AU intensities and AU velocities. AU intensities (on the horizontal axis) are unitless. AU velocities (on the vertical axis) have the units: seconds^{-1}.



(a) Facial activity during relaxed drive



(b) Facial activity during system failure drive

Figure A.3: 2D histograms showing the distribution of sequence-level facial activity during relaxed drive and drive with unexpected system failure (dataset: [177]). Facial activity is represented by pairs of AU intensities and AU velocities. AU intensities (on the horizontal axis) are unitless. AU velocities (on the vertical axis) have the units: seconds^{-1}.

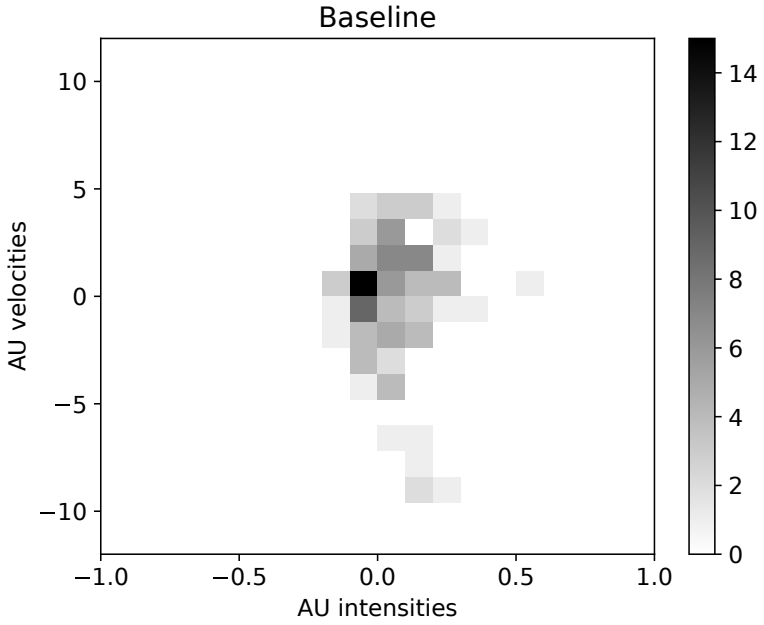


Figure A.4: 2D histogram showing the distribution of sequence-level facial activity while not driving (dataset: [177]). Facial activity is represented by pairs of AU intensities and AU velocities. AU intensities (on the horizontal axis) are unitless. AU velocities (on the vertical axis) have the units: seconds^{-1}.

B

PUBLICATIONS

B.1 MENTAL STATE ANALYSIS: AUTOMATIC PAIN DETECTION

B.1.1 *Hassan et al. "Automatic Detection of Pain from Facial Expressions: A Survey." In: IEEE TPAMI 2019*

Full Reference of Paper

Teena Hassan, Dominik Seuß, Johannes Wollenberg, Katharina Weitz, Miriam Kunz, Stefan Lautenbacher, Jens-Uwe Garbas, and Ute Schmid. "Automatic Detection of Pain from Facial Expressions: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–17. DOI: 10.1109/TPAMI.2019.2958341.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Defined the review methodology and framework for preparing the survey.
- Collected, reviewed, and structured literature on automatic pain detection from facial expressions that were published during the period 2006 – 2018.
- Categorised the pain detection methods into one-step and two-step methods depending on whether or not an intermediate step of AU detection was involved.
- Categorised the pain detection methods based on the learning tasks, the features, and the machine learning methods used.
- Identified deficits in the technical methods used to automatically detect pain from facial expressions.
- Outlined future research directions to address these deficits. Two of the key research directions proposed include the use of weakly supervised, semi-supervised, and unsupervised methods for automatic pain detection, and the development of integrative approaches combining multiple tasks, algorithms, and context-relevant information.

Written Contents Contributed by Myself

I wrote nearly 85% of the contents of the paper. I contributed to all sections in the paper, except Section 3 on facial expressions of pain. I also did not contribute the contents from the field of psychology that are included in the introduction (Section 1) and discussion (Section 6). I also did not contribute the recommendations for future datasets, listed at the end of Section 4. I also did not formulate the parts of discussion related to interpretable machine learning based on grammar inference and inductive logic programming, and their utility to medical staff.

B.1.2 *Kunz et al. "Problems of video-based pain detection in patients with dementia: a road map to an interdisciplinary solution." In: BMC Geriatrics 2017*

Full Reference of Paper

Miriam Kunz, Dominik Seuss, Teena Hassan, Jens U. Garbas, Michael Siebers, Ute Schmid, Michael Schöberl, and Stefan Lautenbacher. "Problems of video-based pain detection in patients with dementia: a road map to an interdisciplinary solution." In: *BMC Geriatrics* 17.33 (2017). DOI: 10.1186/s12877-017-0427-2.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which were published in this paper:

- Categorised the state of the art methods in automatic pain detection from facial expressions into four categories based on their learning goals (referred to in the paper as "aims of assessment"): (i) pain versus no pain; (ii) genuine versus faked pain; (iii) pain versus other emotions; and (iv) continuous-valued or discrete-valued pain intensities.
- Categorised the state of the art methods in automatic pain detection from facial expressions into single-level and two-level methods, based on whether or not an intermediate step of AU detection was involved.

Written Contents Contributed by Myself

In this paper, I wrote a brief overview of the existing methods for automatic pain detection from facial expressions. This included a categorisation of the methods as mentioned in the subsection above, and a brief comment on the different types of visual inputs used and the different types of features (shape, appearance, temporal) extracted from the visual input.

B.2 AUTOMATIC FACIAL ACTION ESTIMATION

B.2.1 *Hassan et al. "A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework." In: ECAI 2016*

Full Reference of Paper

Teena Hassan, Dominik Seuss, Johannes Wollenberg, Jens Garbas, and Ute Schmid. "A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework." In: *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August - 2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. Frontiers in Artificial Intelligence and Applications. The Hague, The Netherlands: IOS Press, 2016, pp. 1812–1817. DOI: 10.3233/978-1- 61499-672-9-1812.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Proposed and implemented a novel and practical approach to integrate categorical probability outputs from data-driven classifiers such as SVMs within a continuous state estimation framework that uses a Gaussian noise model. The proposed method uses the marginalisation technique as well as the variance of the Bernoulli distribution defined by the binary class probabilities.
- Applied the proposed method to fuse facial shape and appearance information within an extended Kalman fitler based framework for AU intensity estimation. Performance evaluation on three upper face AUs showed that fusion of shape and appearance information using the proposed method improved AU recognition performance, measured in terms of Area Under ROC Curve (AUC).

Written Contents Contributed by Myself

I wrote nearly 70% of the paper. This includes the abstract, conclusion, and the fusion approach (Section 4) entirely, as well as most part of the introduction, related work, and evaluation. Within the introduction (Section 1), I wrote the contents except those related to SHORE™. Within the related work (Section 2), I prepared (i) the summary of the state-of-the-art methods for facial expression analysis that are based on state estimation methods such as Kalman filter, particle filter, HMM, and DBN; (ii) the summary of the existing methods for fusing

probability outputs from SVMs within a discrete state estimation framework (HMM); and (iii) the problem formulation. Within the evaluation section (Section 5.3), I prepared the entire content, except Figure 4.

B.2.2 *Hassan et al. "A Kalman Filter with State Constraints for Model-based Dynamic Facial Action Unit Estimation." In: Forum Bildverarbeitung 2018*

Full Reference of Paper

Teena Hassan, Dominik Seuß, Andreas Ernst, and Jens Garbas. "A Kalman Filter with State Constraints for Model-based Dynamic Facial Action Unit Estimation." In: *Forum Bildverarbeitung 2018*. Ed. by Thomas Längle, Fernando Puente León, and Michael Heizmann. KIT Scientific Publishing, 2018. DOI: 10.5445/KSP/1000085290.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Extended the state constraints on AU intensities to the process models based on driven mass-spring-damper systems. This required additional constraints to be defined for the driving force acting on each mass-spring-damper system. The values for the constraint parameters associated with the driving force were determined empirically, and are provided in Table 1.1 in the paper.
- Validated the impact of these state constraints on the range of estimated AU intensities using histogram analysis. It was found that 99.95% of the AU intensities estimated using driven mass-spring damper process models were within the range $[-0.2, 1.2)$.
- The operation of these state constraints, with regard to the effects of the constraint coefficient and the constraint offset, was re-interpreted and visualised. This is illustrated in Figure 1.3 and explained as part of Section 3.1 in the paper.

It is to be noted that the results for the Regularized Landmark Mean-Shift (RLMS) [158] method and Constant Velocity (CVel) process model based method were generated as part of my master's thesis [69].

Written Contents Contributed by Myself

I wrote the entire contents of the paper, including figures and tables. However, Figures 1.1, 1.2 and 1.4 in the paper were originally created as part of my master's thesis [69].

B.2.3 *Seuss et al. "Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array." In: ACII 2019*

Full Reference of Paper

Dominik Seuss, Anja Dieckmann, Teena Hassan, Jens-Uwe Garbas, Johann Heinrich Ellgring, Marcello Mortillaro, and Klaus Scherer. "Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array." In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 35–41. DOI: 10.1109/ACII.2019.8925458.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contribution, which was published in this paper:

- Evaluated the performance of a state-of-the-art AU recognition system developed by Dapogny et al. [30] on the center view images in the Actor Study Database [168]. The performance was measured in terms of AUC, and the AU-wise results are presented in the column labelled 'ISIR' in Table IV in the paper.

Written Contents Contributed by Myself

I wrote the description of the state-of-the-art AU recognition system developed by Dapogny et al. [30]. This description is included in Section IV A of the paper, under the heading: 'AU detection system from ISIR'. In addition to this, I also proofread and corrected the manuscript.

B.3 ADDRESSING OPEN CHALLENGES IN AUTOMATIC MENTAL STATE ANALYSIS

B.3.1 *Mahesh et al. "Requirements for a Reference Dataset for Multimodal Human Stress Detection." In: PerCom Workshops 2019*

Full Reference of Paper

Bhargavi Mahesh, Teena Hassan, Erwin Prassler, and Jens-Uwe Garbas. "Requirements for a Reference Dataset for Multimodal Human Stress Detection." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 492–498. DOI: 10.1109/PERCOMW.2019.8730884.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Formulated the research questions to guide the requirements analysis for building a multimodal reference dataset for human stress detection. This is part of Section III of the paper.
- Conceptualised the technical requirements related to sensor calibration, sensor noise quantification, and time-synchronised recording of multiple modalities (REQ-5 in the paper).
- Conceptualised the annotation requirement (REQ-4 in the paper) related to the use of multiple self-reports to gather information about internal and external factors that could influence the response to stress stimuli.
- Conceptualised the requirement to include multiple modalities, especially the reliable stress modalities of heart and skin activity (REQ-3 n the paper).
- Formulated the evaluation criteria for evaluating the publicly available stress datasets based on the results of the requirements analysis. This is part of Section IV in the paper.
- Co-supervised this research work along with Prof. Dr. Erwin Prassler, Bonn-Rhein-Sieg University of Applied Sciences.

Written Contents Contributed by Myself

I reorganised the contents of the paper, with a focus on the scientific structure of arguments and flow of content. Apart from this, I wrote the research questions in Section III of the paper, and contributed written text to Section III E. (REQ-5) and Section IV (excluding Table II).

B.3.2 *Genc and Hassan. "Analysis of Personality Dependent Differences in Pupillary Response and its Relation to Stress Recovery Ability." In: PerCom Workshops 2019*

Full Reference of Paper

Pelin Genc and Teena Hassan. "Analysis of Personality Dependent Differences in Pupillary Response and its Relation to Stress Recovery Ability." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 505–510. DOI: 10.1109/PERCOMW.2019.8730779.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Conceptualised the personality-wise analysis of pupillary responses to arousal stimulus.
- Conceptualised the protocol for data collection (except the visual stimulus and recording setup).
- Conceptualised the analysis of personality-specific response characterisitcs based on the variances in pupil diameter.
- Conceptualised a visualisation of variances in pupil diameter for reporting the results (Fig. 5 in the paper).
- Supervised this research work.

Written Contents Contributed by Myself

I reorganised and rephrased the contents of the paper, with a focus on the scientific structure of arguments and flow of content. Apart from this, I formulated the research questions in Section II.

B.3.3 *Weitz et al. "Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods." In: tm-Technisches Messen 2019*

Full Reference of Paper

Katharina Weitz, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas. "Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods." In: *tm-Technisches Messen* 86.7-8 (2019), pp. 404–412. DOI: 10.1515/teme-2019-0024.

My Scientific Contributions

Within the scope of this dissertation, I made the following scientific contributions, which are published in this paper:

- Defined the research questions for investigating the predictive performance, interpretability, and explainability of a deep CNN model to distinguish pain from disgust and happiness.
- Defined the materials to be used, the data preparation steps, and the overall procedure to be followed for training, evaluating, and interpreting the deep CNN model.
- Co-supervised this research work along with Prof. Dr. Ute Schmid, University of Bamberg.

Written Contents Contributed by Myself

I reviewed the contents of the paper and suggested corrections with a focus on the flow of content and the scientific correctness of arguments, explanations, and conclusions. Apart from this, I contributed to the framing of the research questions in Section 3.

C

PREPRINTS

C.1 MENTAL STATE ANALYSIS: AUTOMATIC DISTRACTION DETECTION

C.1.1 *Gjoreski et al. "Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals."*

Reference of Preprint

Martin Gjoreski, Matjaž Gams, Mitja Luštrek, Pelin Genc, Jens-Uwe Garbas, and Teena Hassan. "Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals." Preprint. (Update 02.08.2020: Published in IEEE Access in April 2020. See full reference below.)

Full Reference of Paper

Martin Gjoreski, Matjaž Gams, Mitja Luštrek, Pelin Genc, Jens-Uwe Garbas, and Teena Hassan. "Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals." In: *IEEE Access* 8 (2020), pp. 70590–70603. DOI: 10.1109/ACCESS.2020.2986810.

My Scientific Contributions

Within the scope of this dissertation, I made the following (scientific) contributions, which are included in this preprint (Update 02.08.2020: This is now a published paper. See full reference above.):

- Conceptualised, supervised, and coordinated this collaborative research on driver distraction detection using facial and physiological signals.
- Extracted AU intensities from the facial videos in the dataset by applying the AU intensity estimation system that was developed as part of this doctoral work. This system is referred to as AUReader in the paper.
- Downsampled the extracted AU intensities to 1 Hz by computing the average of all valid intensities in each successive 1-second interval.

- Acquired funding via Fraunhofer Society's *Young Research Class 2016 programme on 'Cognitive Machines'*, and DAAD's *PPP Programmes for Project-Related Personal Exchange* with Slovenia.

Written Contents Contributed by Myself

I wrote the third paragraph of the Introduction section that defines driver distraction. In Section III (Data Description), I wrote the last paragraph on facial expression/AU analysis of the facial videos available in the dataset. I also reviewed and edited the contents of the paper.

C.2 AUTOMATIC FACIAL ACTION ESTIMATION AND PAIN DETECTION

C.2.1 *Hassan et al. "Automatic Facial Action Unit Detection for Psychological Research: Comparison between a Data-Driven and a Probabilistic Approach."*

Reference of Preprint

Teena Hassan, Dominik Seuss, Jens Garbas, and Ute Schmid. "Automatic Facial Action Unit Detection for Psychological Research: Comparison between a Data-Driven and a Probabilistic Approach." Preprint.

My Scientific Contributions

Within the scope of this dissertation, I made the following (scientific) contributions, which are included in this preprint:

- Conceptualised and carried out the performance evaluations of the proposed AU intensity estimation method and the state-of-the-art method for AU recognition from Dapogny et al. [30].
- Compared the qualitative and quantitative performance of the two methods, and drew conclusions about the pros and cons of each of them.
- Constructed and evaluated the different rules for detecting pain from AU intensities based on psychological evidence [97, 146].
- Illustrated some possible forms of verbal explanations for pain detected using the above-mentioned AU-based rules.

Written Contents Contributed by Myself

I prepared the entire content–text as well as visualisations–of this preprint.

Automatic Facial Action Unit Detection for Psychological Research: Comparison between a Data-Driven and a Probabilistic Approach

Teena Hassan
Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

Dominik Seuss
Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

Ute Schmid
University of Bamberg, Bamberg, Germany

Jens Garbas
Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

ABSTRACT

Psychologists use the Facial Action Coding System (FACS) to analyse facial expressions recorded as part of behavioural and psychological experiments. FACS defines basic facial muscle movements, known as Action Units (AUs). Trained coders annotate the video recordings of facial expressions in terms of these AUs. These annotated AUs are then used for statistical analyses of the facial behaviour or facial responses. However, manual coding of facial videos in terms of AUs is time-consuming, costly, and prone to errors due to subjective biases. Automatic AU detection systems can support and accelerate behavioural and psychological research by automating this process of facial expression coding. This paper compares the qualitative and quantitative performance of two automatic AU detection systems. One of them uses static features and data-driven machine learning methods, while the other uses a probabilistic framework to fuse dynamic as well as static information about AUs. The performance is evaluated on three different datasets that contain AU annotations: the Actor Study Database, the UNBC-McMaster Shoulder Pain Expression Archive Database, and a proprietary market research database. AU recognition as well as AU intensity estimation performance of the two systems are compared. The strengths and weaknesses of the two AU detection systems are then identified and the directions for future research are drawn.

KEYWORDS: facial expression analysis, facial action units, machine learning, probabilistic models, evaluation

1 INTRODUCTION

Facial expressions are an important channel used by humans to communicate non-verbal cues [22, 35, 40], including emotions [12] or other

psychophysiological states such as pain [32]. Hence, analysis of facial expressions is crucial not only for understanding human behaviour, but also for building affective and empathic human-computer interfaces. Psychologists study facial expressions by either analysing the 'message' conveyed by the expressions (e.g. emotions, attitude) or by decomposing the expressions into a set of objective 'signs' (e.g. facial muscle movements) [6]. The former approach is referred to as *message judgment* and the latter is referred to as *sign judgment* [6]. Sign judgment allows a finer analysis of facial expressions, which is necessary to investigate person-specific and context-specific variations in facial expressions (cf. [31]). Although the sign judgment approach allows a deeper analysis of facial responses, it involves the tedious process of annotating facial muscle movements according to the Facial Action Coding System (FACS) [16, 17]. FACS defines the fundamental, visually distinguishable facial muscle movements known as Action Units (AUs), and also specifies five coding levels (A–E) for the intensities of AUs. Trained FACS coders code facial expression videos in terms of AUs and their intensities. FACS based analysis of facial expressions has been performed extensively in pain research [32]. Furthermore, some of the appraisal theories of emotions [45] suggest the use of AUs as signs/cues to infer a person's emotional appraisal of an event [46]. The use of an automatic AU detection system could support human coders and has the potential to make the process of annotating facial expressions less tedious. Systems that allow a human coder to correct wrong AU predictions or AU intensity estimates would prove to be a useful human-computer collaborative tool for promoting behavioural and psychological research.

Several different approaches have been developed for automatic AU detection over the past two decades [34, 44]. Several systems (e.g. FaceReader,[1] Affectiva's AFFDEX[2]) are also available in the commercial market. However, a comprehensive evaluation of the quality of outputs produced by these approaches and a comparison of their predictive performance are often not performed. Therefore, in this study, we compare two recently developed out-of-the-box systems for automatic AU analysis. One of them, developed by Dapogny et al. [11] at ISIR, Sorbonne University, Paris, recognises the presence of 12 AUs in facial images. The other, developed by Hassan et al. [24] at Fraunhofer IIS, Erlangen, estimates continuous-valued intensities of 22 AUs. For ease of reference, the former system is referred to as ISIR-AU and the latter is referred to as AUReader. The two systems differ not only in the AU analysis tasks, but also in the features and methods used. In this study, we perform qualitative and quantitative analyses of the outputs of the two systems in order to identify

1 https://www.noldus.com/facereader
2 https://imotions.com/affectiva-requestdemo/

their strengths and weaknesses with respect to their potential for application in behavioural and psychological research.

In Section 2, an overview of the different approaches for automatic AU detection is presented. Following this, in Section 3, the approaches used in ISIR-AU and AUReader are described. The three datasets used for evaluation are described in Section 4, and the performance metrics used for quantitative evaluation are described in Section 5. The qualitative as well as quantitative results obtained on the three datasets are presented in Section 6. The general insights gained from the evaluation are discussed in Section 7. Section 8 concludes the paper.

2 RELATED WORK

As mentioned in the survey [34], computer vision research has focused on several different tasks within the realm of automatic AU detection. These mainly include the automatic recognition of whether an AU is present in a facial image or video, the automatic prediction of the discrete or continuous-valued intensities of AUs, and the automatic recognition of the temporal phases of AUs (onset, apex, offset) [34]. Most of the approaches developed for automatic AU detection are based on data-driven machine learning methods. Initially, such approaches used hand-designed features that described the facial shape and appearance or their changes over time. Facial shape was usually described in terms of positions of facial feature points, or distances and angles between them (e.g. [28, 51]). Facial appearance was described by image texture descriptors such as Gabor filters [13, 19, 20], Local Binary Patterns (LBP) [38, 39] and Histogram of Oriented Gradients (HOG) [10]. Such facial shape and appearance features extracted from images are usually referred to as *spatial* features, and the features that describe the changes in spatial features over time are usually referred to as *spatiotemporal* features [44]. These features were fed as input to machine learning methods such as Support Vector Machines (SVMs), its variants, and boosting algorithms, for making predictions related to the corresponding AU detection tasks [34]. Examples of approaches that used spatial features include [4, 5, 28, 47], and those that included spatiotemporal features include [3, 27]. More recently, deep architectures of Convolutional Neural Networks (CNNs) are increasingly being used to predict AUs (e.g. [21, 26]). These methods perform end-to-end learning by taking an image or image sequence as input, i.e. they extract facial image features on their own and do not require hand-designed features as input.

Another category of approaches for facial AU detection used parameterised, deformable facial shape or appearance models. Model-fitting methods such as Active Appearance Models (AAM) [7, 8] and Constrained Local Models (CLM) [9] were used to fit such deformable

models to the faces in images and image sequences [5]. Such model-fitting methods have been normally used for tracking the positions of facial landmarks (e.g. [5, 41]). However, by combining dynamic, probabilistic state estimation methods and AU-based deformable facial shape models such as CANDIDE-3 [2], approaches have been developed for AU intensity estimation [14, 15, 25]. In these approaches, state estimation methods such as Kalman filter, particle filter, or their variants are used to model the dynamics of the face model parameters, and to combine the dynamic model predictions with the observed image properties. Such probabilistic state estimation approaches inherently provide a quantification of uncertainty in the AU intensity estimates, which make them suitable for practical applications such as human-robot interaction.

While data-driven machine learning methods can learn robust features to deal with large variance in the data, the deformable face model based approaches support interpretability of the mathematical models involved. By combining the high predictive performance of the data-driven machine learning methods with the better interpretability of the state estimation and deformable face model based approaches, it would be feasible to build strong learning systems, as defined by Michie [36]. AUReader [23, 24] is an effort in this direction, whereas ISIR-AU [11] is a data-driven approach based primarily on random forest classifiers. The next section describes these systems in detail.

3 EVALUATED SYSTEMS

This section describes the methods used by the two AU detection systems evaluated in this paper, namely AUReader [23, 24] and ISIR-AU [11].

3.1 *AUReader*

The AUReader [24] uses a probabilistic framework for estimating continuous-valued intensities of 22 AUs (see Table 1). It is based on a Gaussian state estimation method that models the dynamics of facial muscle movements and integrates the predictions of these models with the observed facial shape and appearance. Facial shape is represented by the positions of 68 facial landmarks and is detected using a face alignment module [29] that uses a Random Forest (RF). Facial apperance is represented indirectly by a set of SVM classifiers trained on texture descriptors such as LBP and HOG. The Gaussian state estimation method fuses these shape and appearance information using the method described in [24]. A mass-spring damper system is used to model the dynamics of the viscoelastic motion of facial muscles. AUReader produces an AU intensity estimate for each image frame in a video. Apart from this, the probabilistic framework used by

AUReader also provides a measure of the Gaussian noise associated with the AU intensity estimates. This Gaussian noise is estimated on the basis of the noise in the modelling of facial muscle dynamics and the noise in the detections of facial shape and appearance.

The AUReader uses a deformable, parameterised model of facial shape that includes facial shape deformation vectors representing shape changes caused by facial muscle movements or AUs [23]. These AU deformation vectors were sampled from high-poly wireframe models designed by psychologists [30, 43]. The strength of facial shape deformation along each vector represents the AU intensity, which is estimated by AUReader. The AU deformation vectors were designed to correspond to the maximum possible shape deformation that could be caused when an AU is displayed. Therefore, the AU intensity estimates belong to a valid range of $[0, 1]$. The modelling of these range constraints is described in [23]. There are also a set of deformation vectors that represent the facial shape variations between individuals. Along with AU intensities, AUReader also estimates the person-dependent shape of the face. Several of the AU deformation vectors, especially those corresponding to AUs caused by neighboring muscles, show geometric similarities. Correlation coefficients computed between pairs of AU deformation vectors are used to adapt the Gaussian noise covariances used in the state estimation framework.

Using the features and parameters associated with AUReader, different configurations can be created. For the purpose of performance evaluation, a somewhat basic configuration of AUReader was used, where some of the features were disabled. Precisely speaking, anomaly detection and muscle-specific models were deactivated, since all the parameters associated with these models had not yet been empirically tuned. Consequently, all the mass-spring-damper models in AUReader used the same natural frequency (ω_0) of 3 Hz and damping ratio (ζ) of 1.2. In lieu of the anomaly detection feature that is based on the computation of normalised innovation squared [37] values, a simpler approach for detecting divergent AU intensity estimates was activated. This approach checked whether the estimated AU intensities diverged outside the interval $(-1, 2)$, and if so, marked the output for that image frame as invalid. In addition to these simple divergence checks, another quality assurance feature was also integrated in the AUReader. This feature marks AU intensity estimates as invalid, if no facial landmarks could be detected for three consecutive frames. Like anomaly detection and muscle-specific models, texture calibration was also disabled, since it was difficult to guarantee that all image sequences used in this evaluation began with a neutral face. However, facial shape calibration was performed online, simultaneously with AU intensity estimation. Furthermore, after preliminary trials, the correlation coefficients between certain pairs of AUs were disabled (set to zero) in order to improve the AU recognition performance. These AU pairs

were (AU01, AU02), (AU01, AU04), (AU02, AU04), (AU01, AU09) and (AU02, AU09). The SVM classifiers that detect these AUs compensate indirectly for this loss of information by providing appearance-based evidence to resolve ambiguities during the process of AU intensity estimation.

Table 1: List of the 22 AUs, whose intensities are estimated by AUReader. These AUs are defined in FACS [16, 17]. ISIR-AU detects the presence of 12 of these 22 AUs.

AU Code	**AU Name**
01	Inner Brow Raiser
02	Outer Brow Raiser
04	Brow Lowerer
05	Upper Lid Raiser
06	Cheek Raiser
07	Lid Tightener
09	Nose Wrinkler
10	Upper Lip Raiser
11	Nasolabial Deepener
12	Lip Corner Puller
13	Sharp Lip Puller
14	Dimpler
15	Lip Corner Depressor
16	Lower Lip Depressor
17	Chin Raiser
20	Lip Stretcher
23	Lip Tightener
24	Lip Pressor
25	Lips Part
26	Jaw Drop
27	Mouth Stretch
43	Eyes Closed

3.2 *ISIR-AU*

Unlike AUReader, ISIR-AU [11] uses data-driven machine learning methods for detecting the presence of AUs in an image. It uses a Random Forest (RF), in which each tree is trained on selected features from a randomly chosen local region of the face. The feature candi-

dates are appearance features based on HOG and geometric features such as distances and angles between facial landmarks. Features are thresholded at every non-leaf node of the trees to split training instances or to choose a search path. The leaf nodes of the trees predict a facial expression category.[3] The output of each tree is a probability vector, with one entry for each facial expression category. The probability vectors from all trees are concatenated to obtain a second set of candidate features from which a set of second-layer RFs are learned, one RF for each AU. An autoencoder is used to predict confidence levels for each local region. These confidence measures are intended to provide robustness against partial occlusions. A weighted average of these local confidence measures is computed, with the weights determined by the proportion of root-level decisions contributed by each local region towards AU classification.

It is to be noted that ISIR-AU system was trained on spatial features and performs AU classification. Each AU-specific RF provides a probability for the presence of that AU in the input image, along with a measure of confidence in the prediction. The ISIR-AU system used in this paper provided predictions for 12 AUs, namely AU01, AU02, AU04, AU05, AU06, AU09, AU12, AU15, AU17, AU20, AU25, and AU26. In contrast to ISIR-AU, AUReader uses spatial and temporal information, and estimates intensities of 22 AUs (see Table 1 for the list of 22 AUs). This set of 22 AUs include several AUs that closely resemble each other, such as AU06 and AU07, AU12 and AU13, and AU23 and AU24. It is also to be noted that while ISIR-AU has a separate classification model for each AU, the AUReader estimates intensities for all 22 AUs simultaneously. The features of AUReader and ISIR-AU are summarised in Table 2.

4 DATASETS

The *Actor Study Database* [48] has facial expression sequences from 21 actors recorded from five different views. The actors performed four different facial expression tasks. From these, 777 center (frontal) view recordings at 24 frames per second were used in this paper. These sequences are from Tasks 1 and 2, and contain acted/posed expressions of AUs and AU combinations. The recordings from Task 2 contain acted AU combinations associated with five basic emotions–sadness, happiness, anger, disgust and fear. The sequences from the first 11 actors were used for training the SVM classifiers and for tuning the parameters of other components of the AUReader. Therefore, only the remaining 370 sequences from Actors 12 to 21 were used for quantitative evaluation. Each sequence has frame-level annotations of FACS AUs. Annotations are available for all 22 AUs listed in Table 1.

3 Here, a facial expression category refers to the neutral expression or the facial expression of happiness, sadness, anger, disgust, surprise or fear.

Table 2: Characteristics of the two AU detection systems–AUReader and ISIR-AU–that are evaluated in this paper.

Characteristic	**AUReader**	**ISIR-AU**
Task	AU intensity estimation	AU recognition
Feature type	Spatial (shape, appearance), temporal	Spatial (shape, appearance)
Methods used	Continuous-discrete extended Kalman filter, SVM, RF	RF, autoencoder
Output	Continuous-valued AU intensities	Probabilities of presence of AUs
Confidence measure	Gaussian error variances and covariances between AU intensities	Scores based on probabilities of occlusions in input image
Paradigm	Estimates AU intensities jointly	Uses independent AU-specific models
No. of AUs	Twenty two	Twelve

The *UNBC-McMaster Shoulder Pain Expression Archive Database* [33] has 200 facial expression sequences recorded from 25 shoulder pain patients, while they performed range of motion tests on left and right arms. The database provides frame-wise annotations of 12 AUs, their intensities on a 5-point scale as defined in FACS, positions of 66 facial landmarks, and the Prkachin-Solomon Pain Intensity (PSPI) score. Sequence-level annotations of observer-reported and self-reported pain intensities are also available. In this study, only the frame-wise AU annotations and PSPI scores are used.

A *proprietary market research database* (introduced in [24]) consisting of recordings of spontaneous facial responses of 155 subjects watching commercial advertisement videos, is also used for quantitative evaluation. This database was collected by the Nuremberg Institute for Market Decisions (formerly GfK Verein), and includes 408 facial expression sequences containing a rich variety of spontaneous facial expressions. Frame-wise AU annotations are available for each sequence, and the annotations include all the 22 AUs listed in Table 1. However, this database is not publicly available.

Table 3 lists the number of annotated frames available for each AU in each of the above-mentioned databases. The facial expressions selected from the Actor Study Database [48] are deliberately displayed expressions, and the faces are nearly frontal. In contrast, the facial expressions in the UNBC-McMaster Shoulder Pain Expression Archive Database [33] and the proprietary market research database [24] are spontaneous expressions, displayed in response to specific stimuli.

These spontaneous facial expressions are occassionally accompanied by out-of-plane head movements. Therefore, the latter two databases pose a greater challenge for automatic AU recognition and automatic AU intensity estimation.

5 PERFORMANCE METRICS

AUReader produces continuous-valued outputs. To evaluate its performance, two strategies were adopted. The first strategy used each AU intensity estimate as a decision threshold to detect the presence or absence of an AU. The second strategy compared the AU intensity estimates directly with the annotated discrete intensity levels, after these annotations were converted into numerical values within the range $[0, 1]$. The former strategy reframed the intensity estimation problem into a classification task, and the latter retained it as a regression task. The performance evaluation method was chosen according to the strategy adopted for evaluation. Receiver Operating Characteristic (ROC) curves suited the first strategy, and computation of absolute errors suited the second strategy. The following paragraphs describe briefly how these evaluation methods were applied to quantify the performance of the AU intensity estimation method.

ROC curves [18] evaluate binary classification performance by measuring how well positive instances of a class can be discriminated from the negative instances. For this, firstly, the output scores from a binary classifier are compared to a decision threshold to predict whether an input is a positive or a negative instance of the target class. This process is repeated on all instances in the test set. Subsequently, the True Positive Rate (TPR) and False Positive Rate (FPR) are computed by comparing the predictions with the annotations. TPR, also known as *sensitivity*, indicates the proportion of positive instances predicted as positive, and FPR, also known as inverted *specificity*, indicates the proportion of negative instances predicted as positive. TPR and FPR together represent the discriminative power of a classifier. By varying the decision threshold, a set of (TPR, FPR) pairs can be obtained. ROC curves are two dimensional plots of these (TPR, FPR) pairs, with TPR plotted along y-axis and FPR along x-axis. A scalar value, Area Under ROC Curve (AUC), is used to succinctly represent the information contained in an ROC curve. Like TPR and FPR, AUC also belongs to the range $[0, 1]$. AUC above 0.5 is preferred, since it represents performance better than chance. The higher the AUC above 0.5, the better is the performance. If the AUC is below 0.5, then the decisions of the classifier are inverted. ROC curves can also be used to find optimal decision thresholds for classifiers. The point on the ROC curve that maximises *F1-score* is usually chosen as the *operating point* for the classifier. F1-score is the harmonic mean of *precision* and *sensitivity*, where precision denotes the proportion of true positives

Table 3: Overview of the three datasets used for evaluation: The number of subjects, the number of sequences, the total number of frames, and the number of annotated frames per AU that are available in each dataset are listed. In the case of Actor Study Database [48], only sequences belonging to Actors 12 to 21 have been considered. For brevity, the UNBC-McMaster Shoulder Pain Expression Archive Database [33] is referred to as UNBC Pain in the table. Similarly, the proprietary market research database [24] is abbreviated as Market Research.

	Actor Study [48]	UNBC Pain [33]	Market Research [24]
Subjects	10	25	155
Sequences	370	200	408
Frames	47730	48398	80600
AU01	5609	0	4874
AU02	4013	0	5422
AU04	5941	1074	11200
AU05	2601	0	1009
AU06	3414	5557	16090
AU07	7048	3366	18182
AU09	1862	423	872
AU10	1524	525	1478
AU11	612	0	1790
AU12	4248	6887	26729
AU13	831	0	1205
AU14	1186	0	4592
AU15	1204	6	2987
AU16	494	0	644
AU17	3604	0	3552
AU20	2227	706	1566
AU23	783	0	1055
AU24	1550	0	1953
AU25	6489	2407	8381
AU26	3951	2093	4251
AU27	695	18	68
AU43	2136	2434	2456

among all positive predictions. An ROC curve is generated for each AU by using the AU intensity estimates produced on the test set as decision thresholds. AUC for each ROC curve is then computed and used as the measure of AU recognition performance. To generate the ROC curves, to compute AUCs, and to determine the operating points, an existing software library that was developed by Fraunhofer IIS was used.

In contrast to ROC curve based evaluation, the computation of absolute errors is used to measure performance in regression tasks. Absolute error is the absolute value of the difference between an annotated value and an estimated value. In this study, absolute error is computed between annotated AU intensities and estimated AU intensities. FACS based annotation of AU intensities is done on a 5-point ascending scale A–E, with 'A' denoting "traces" of an AU [17]. In this study, the discrete AU intensity annotations are converted into numerical values in two steps. The symbolic scale A–E is first mapped to the numerical scale 1–5[4], and then mulitplied by 0.2, in order to map the annotated intensity levels to the set $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. When an AU is not annotated, the intensity level is taken as 0. Mean and standard deviation of the absolute errors are computed for each AU over all frames in all sequences in the test set, and these are used as the measure of AU-wise intensity estimation performance. Mean Absolute Error is abbreviated henceforth as MAE.

6 RESULTS

In this section, the results of the qualitative and quantitative evaluation of the performance of AUReader and ISIR-AU are described. Table 4 provides an overview of the evaluations performed.

6.1 *Actor Study Database*

As mentioned in Section 4, the sequences belonging to Actors 12 to 21 were used for the evaluations on the Actor Study Database [48]. First of all, the AU intensity estimates produced by AUReader and the AU detection probability scores provided by ISIR-AU were compared qualitatively. Figures 1a and 1b illustrate the outputs for six AUs, namely, AU01, AU04, AU06, AU12, AU09 and AU25, for selected sequences from the Actor Study Database [48]. Figure 2 illustrates the outputs for a simultaneous display of AU01 and AU04. The ground truth curves in all cases were plotted by setting 1 to each frame that was annotated with the corresponding AU, and 0 to every other frame.

A visual inspection of the plots show that:

4 In the UNBC-McMaster Shoulder Pain Expression Archive Database [33], the AU intensities are already provided in the range 1–5.

Table 4: Overview of the evaluations of AUReader and ISIR-AU conducted using different datasets.

Evaluation Type	**AUReader**	**ISIR-AU**
Actor Study Database [48]		
Quantitative	AU recognition	AU recognition
Qualitative	AU intensity estimates	AU probability scores
UNBC-McMaster Shoulder Pain Expression Archive Database [33]		
Quantitative	AU recognition, AU intensity estimation	AU recognition, AU intensity estimation
Qualitative	AU intensity error variance	AU confidence score
Proprietary market research database [24]		
Quantitative	AU recognition	–
Qualitative	AU intensity estimates	–

- The intensities estimated by AUReader are temporally smoother than the probability scores from ISIR-AU. This is not surprising, since ISIR-AU does not use temporal information. As a consequence, the different phases of intensity variations are more clearly observable in the estimates from AUReader than in the scores from ISIR-AU (see the plots for AU01 in Figure 1a and AU04 in Figure 2). This makes the estimates from AUReader more suitable for a finer analysis of facial activity.

- The scores from ISIR-AU tend towards higher values, causing subtle displays to be completely missed (see AU09 in Figure 1b). This also shows that AUReader's estimates might be more suitable for detecting subtle displays of AUs.

- In general, ISIR-AU does not show delay in detecting an AU. In contrast, AUReader often requires some time to pick up the AU. This delay is mainly due to the constraints acting on the state estimates, and the effect of process and measurement noise models. In Figures 1a and 1b, this is especially visible in the intensity estimates for subtle AUs such as AU06 and AU09, for which a full correlation between noise in facial landmark detections was applied.

In sum, the delayed onsets and lower intensity estimates show that the estimates of AUReader are more conservative, which in turn could

potentially reduce sensitivity but improve specificity and precision. In order to examine this more closely, the qualitative analysis was followed up with a quantitative analysis based on ROC curves.

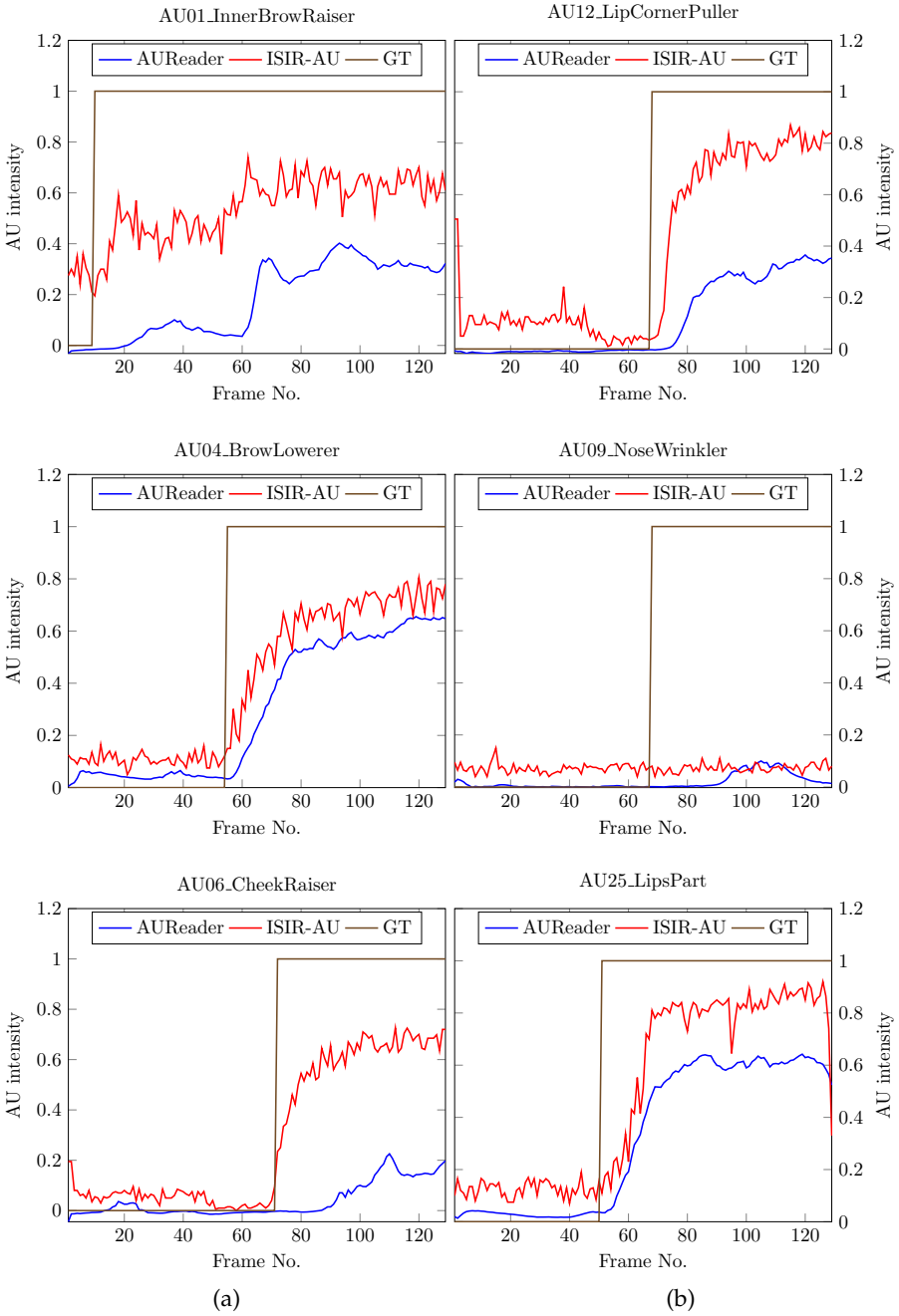


Figure 1: AUReader [24] versus ISIR-AU [11]: Qualitative results for selected single AU displays from Actor Study Database [48].

For each system, AU-specific ROC curves were generated. In the case of AUReader, the AU intensity estimates from all sequences were used as the decision thresholds to generate the ROC curves. In the

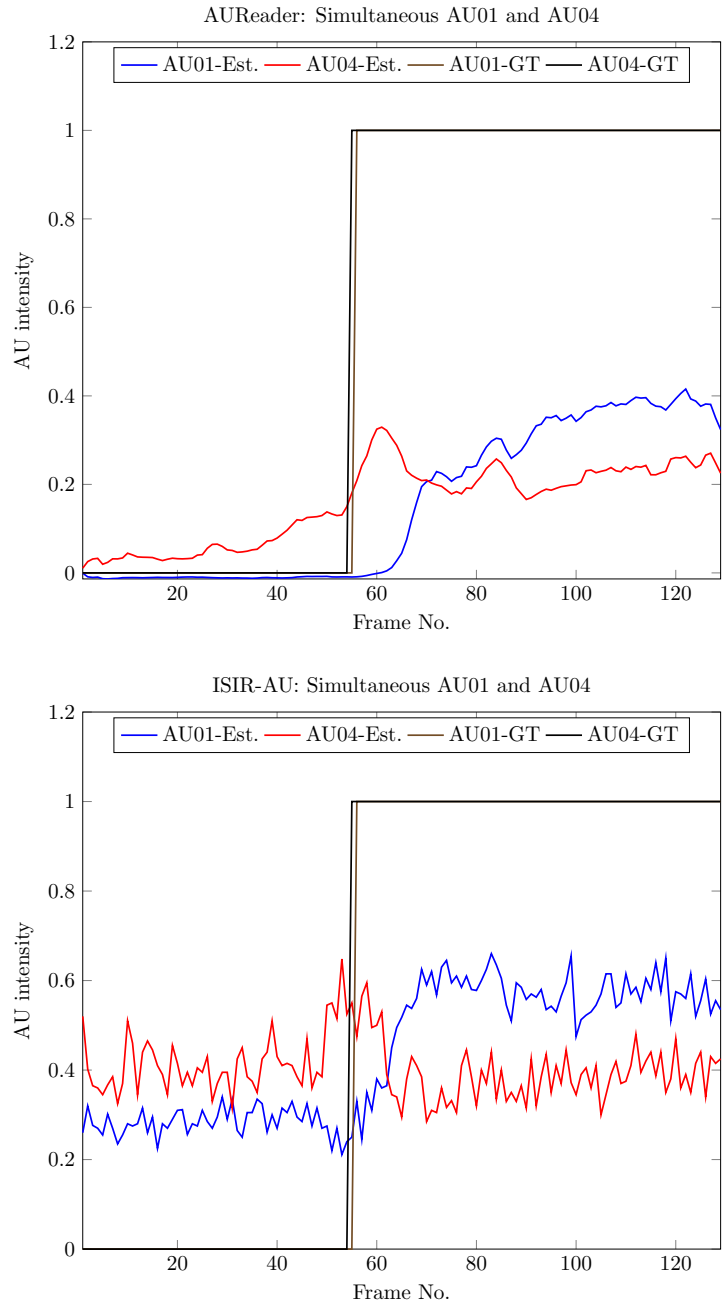


Figure 2: AUReader [24] versus ISIR-AU [11]: Qualitative results for a selected simultaneous display of AU01 and AU04 from Actor Study Database [48].

case of ISIR-AU, the probability scores from all sequences were used as the decision thresholds. AUC values were then computed for each AU and for each system. These are shown in Figure 3. The following observations can be made:

- The AUC values varied widely between AUs. In the case of AUReader, 20 AUs scored AUC values above 0.6. Among these,

15 AUs scored AUC values above 0.7, 11 AUs scored AUC values above 0.8, and 2 AUs touched or crossed 0.9. The AUC values for AU20 (LipStretcher) and AU24 (LipPressor) were quite low, touching 0.52 and 0.45, respectively. This could be due to their high resemblance with AU12 (LipCornerPuller) and AU23 (LipTightener), respectively. Another reason for this low performance could be the errors associated with the detections of landmarks on the innner boundaries of the lips. Due to the lack of edge features along the lips, the noise in the x-coordinates of these landmarks tend to be relatively high. However, including appearance-based evidence for AU20 and AU24 might help to improve the performance in the future.

- For the 12 AUs that can be detected using both AUReader and ISIR-AU, the AUC values for AUReader were either better or comparable to the AUC values for ISIR-AU, except for AU15 (LipCorrnerDepressor) and AU20 (LipStretcher). Together with the previous observation, this shows that AUReader is robust and can discriminate reasonably well between geometrically similar (non-orthogonal) AUs within the set of 22 AUs, whose intensities it simultaneously estimates.

- In the case of the subtle AU06 and AU09, the performance of AUReader and ISIR-AU were comparable (difference $\approx \pm 0.03$). Therefore, despite being conservative (due to delayed onsets and low intensity estimates), AUReader could achieve a comparably good sensitivity-specificity balance for AU06 and AU09.

- Overall, ISIR-AU was able to analyze more frames than AUReader. This difference in the number and index of analysed frames is caused by the differences in the facial landmark detection methods[5] utilised by the two systems, and by the quality assurance methods adopted by AUReader.

The sequences from Actors 1 to 11 in the Actor Study Database [48] were used for training and tuning the components of AUReader, whereas ISIR-AU was not trained on any image from this database. Therefore, for a fairer comparison of the generalisation performance, a database that was not used for training either AUReader or ISIR-AU should be used for evaluation. For this purpose, the UNBC-McMaster Shoulder Pain Expression Archive Database [33] was selected, due to the availability of annotations of both AU and pain intensities. The next subsection presents the results.

5 AUReader utilises facial landmark detection based on [29] and ISIR-AU uses the method described in [50]

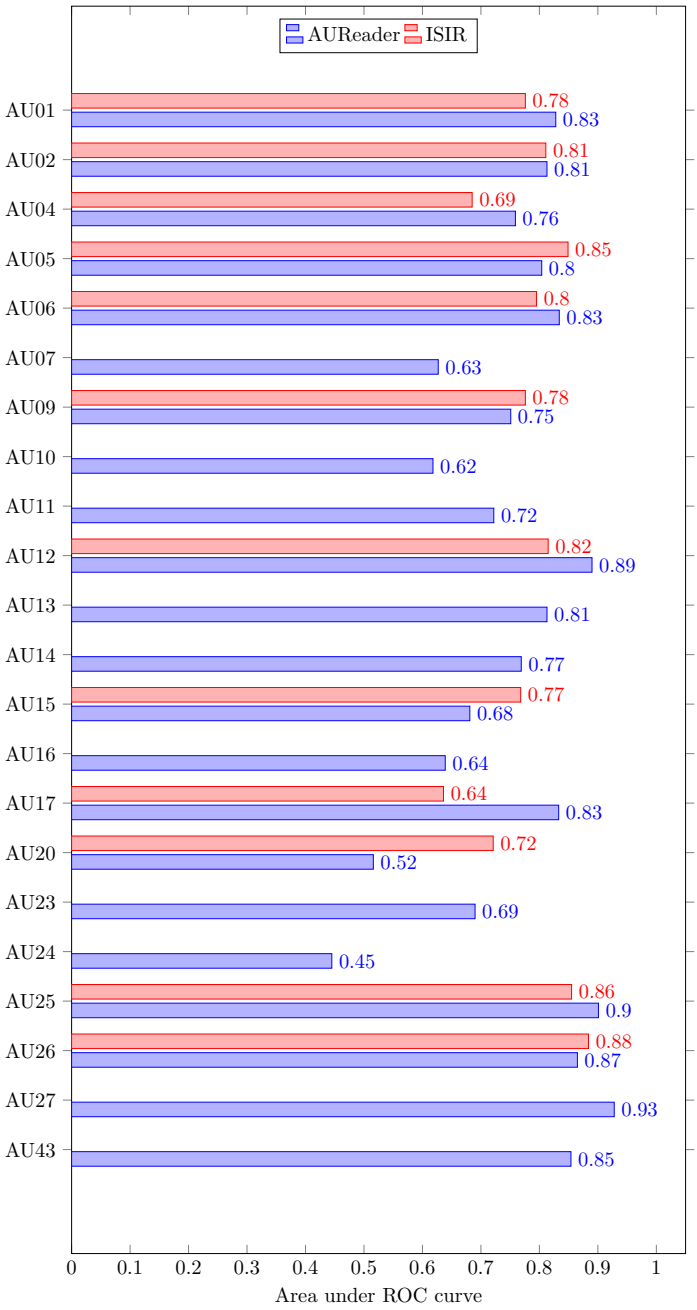


Figure 3: AUReader [24] versus ISIR-AU [11]: Areas under ROC curves obtained on the Actor Study Database [48] (Actors 12 to 21). AUReader analysed 42954 frames and ISIR-AU analysed 46956 frames.

6.2 *UNBC-McMaster Shoulder Pain Expression Archive Database*

All 200 sequences in the UNBC-McMaster Shoulder Pain Expression Archive Database [33] were used for quantitative evaluation. AUC and MAE values were computed for comparing the AU recognition and AU intensity estimation performance of AUReader and ISIR-AU.

A comparison of the AUC values is presented in Figure 4. The following observations can be made:

- Among the eight AUs that could be detected by both AUReader and ISIR-AU, the latter performed better on five AUs. For the remaining three AUs, the performances were almost identical. Therefore, ISIR-AU showed better overall generalisation performance for the task of AU recognition than AUReader.
- AUReader scored AUC values above 0.6 for ten AUs, and above 0.7 for six AUs. None of the AUC values were below 0.5.
- AUReader analyses more AUs than ISIR-AU, and this includes AUs that resemble each other closely in the facial shape deformations that they cause. Among the AUs annotated in this database, there are three AU groups (AU06 and AU07; AU09 and AU10; AUs 25, 26, and 27) that possess such geometric similarities. A complete comparison of performance between AUReader and the available ISIR-AU system was not possible, since the ISIR-AU system did not provide scores for AU07, AU10, and AU27. However, in the case of AU25 and AU26, AUReader performed slightly better than or comparable to ISIR-AU.
- Just like on Actor Study Database, AUReader analysed fewer frames than ISIR-AU. This could be due to the differences between the facial landmark detection methods used by the two systems, and the use of quality assurance methods such as tests for divergence and for successive missing observations in AUReader.

It can be concluded that overall ISIR-AU generalises better in terms of AU recognition performance. However, it is to be noted that AUReader analyses more AUs jointly, thereby dealing with greater possibilities for misclassification (due to either false attribution or shared attribution).[6] AUReader is conservative in its estimates due to the influence of the constraints and noise models. This could lead to lower sensitivity, which results in lower AUC values. As mentioned in [24], AUReader uses probability scores from SVM classifiers as appearance-based observations. Therefore, replacing the SVM models with the ISIR-AU system might improve the generalisation performance of AUReader on spontaneous facial expression databases.

In order to compare the AU intensity estimation performance of AUReader and ISIR-AU, the AU-wise MAEs were computed for the AU intensity estimates from AUReader and the AU probability scores from ISIR-AU (see Table 5). As explained in Section 5, in order to compute MAE, the five FACS AU intensity level codes $A < B <$

6 False attribution refers to the attribution of an observed facial deformation to a different but closely resembling AU. Shared attribution refers to the sharing/splitting of a facial deformation between multiple AUs.

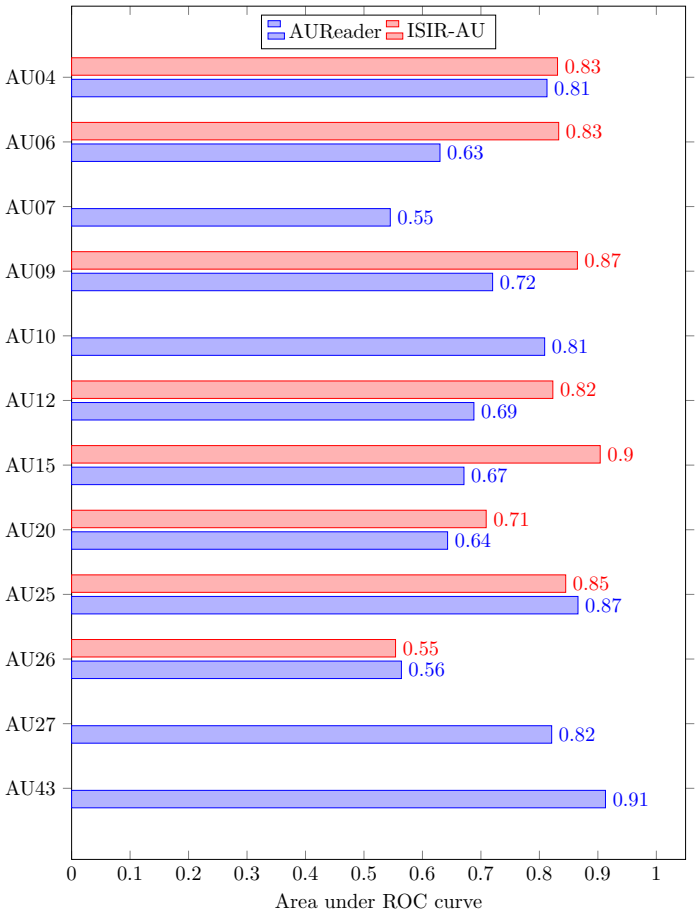


Figure 4: AUReader [24] versus ISIR-AU [11]: Areas under ROC curves obtained on the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. AUReader analysed 41288 frames and ISIR-AU analysed 47859 frames.

$C < D < E$ were converted into numerical codes with $A = 0.2$, and each successive level given an increment of 0.2. This gives the numerical intensity level codes $0.2 < 0.4 < 0.6 < 0.8 < 1.0$. These were used as AU intensity ground truth. The outputs from AUReader are already AU intensity estimates. The AU probability scores provided by ISIR-AU were interpreted as AU intensities for the purpose of evaluation. MAE values were computed under two conditions. In the first condition, all frames, for which AU intensity estimates or AU probability scores were available, were included in the MAE computation. In the second condition, only those frames, in which the corresponding AU was annotated, were considered for MAE computation.

The following observations can be made from the results presented in Table 5:

- When all frames are considered, the MAE values for AUReader are much lower than those for ISIR-AU. Since the number of

annotations are at least one order of magnitude smaller than the number of non-annotated frames (see Table 3), the MAE values are predominantly influenced by the intensity estimates for the non-annotated frames. Therefore, the MAE values reveal that the probability scores from ISIR-AU, in general, tend to indicate AU intensities at level *A* or *B*, even when no AU is annotated. In contrast, the AU intensity estimates from AUReader are, in general, very close to zero.

- However, when only annotated frames are considered, the probability scores from ISIR-AU tend to have lower MAE values that generally do not exceed two intensity levels on average (i.e. < 0.4). In comparison, the MAE values for AUReader tend to be numerically higher, although there are exceptions such as AU27, AU25, AU10, and AU04. This reinstates that AUReader estimates are conservative, i.e. they do not touch very high values rapidly.

- AUReader appears to have difficulty particularly in estimating intensities of AU15 (LipCornerDepressor), as indicated by the high MAE value of 0.59. However, there were only 6 frames that were annotated with AU15, which makes it difficult to draw a reliable conclusion about AUReader's performance. Similarly, the high MAE value of 0.55 for AU43 could be due to the binary annotation of AU43 in the database (intensity level: 0 or 1).

This evaluation has a key limitation. Due to the constraints, the noise models, and the damping and restoration forces acting on the mass-spring-damper models, the higher range of intensity values are not easy to attain for AUReader. Moreover, the intensity increments between consecutive AU intensity levels cannot be considered to be constant or linear, due to the nonlinearities in the models used for AU intensity estimation. In addition, for AUReader, an AU intensity of 1 represents the maximum possible anatomical facial shape deformation. This would not always correspond to the person-specific maximum intensity of expression of an AU. Due to these reasons, an annotated intensity level of *E* need not necessarily correspond to an AU intensity of 1 (one) in the deformable face model used by AUReader. Due to these reasons, the conversion of the symbolic FACS AU intensity codes into numeric values should be performed based on the distribution of AU intensities generated by an AU detection system.

Since ISIR-AU was not developed for intensity estimation, this comparison of AU intensity estimation performance may not be entirely justifiable. However, continuous-valued, FACS-conform, dynamic AU intensity estimation systems are scarce, which could likely be due to the lack of sufficient datasets with reliable AU intensity annotations.

Figure 5 illustrates an example of how the confidence associated with the AU intensity estimates from AUReader vary for a sample sequence. The confidence intervals shown in the figure are the 2-σ

Table 5: AUReader [24] versus ISIR-AU [11]: Mean and standard deviation of absolute errors on UNBC-McMaster Shoulder Pain Expression Archive Database [33]. AUReader analysed 41288 frames, of which 23501 had AU annotations. ISIR-AU analysed 47859 frames, of which 19153 had AU annotations. In the table, NAA stands for No Annotations Available and DND stands for Does Not Detect.

AU	**All Frames**		**Only Annotated Frames**	
	AUReader	**ISIR**	**AUReader**	**ISIR**
AU01	**0.14 ± 0.23**	0.40 ± 0.16	NAA	NAA
AU02	**0.08 ± 0.11**	0.36 ± 0.17	NAA	NAA
AU04	**0.14 ± 0.12**	0.36 ± 0.16	0.23 ± 0.17	0.23 ± 0.13
AU05	**0.05 ± 0.08**	0.37 ± 0.15	NAA	NAA
AU06	0.10 ± 0.15	0.12 ± 0.13	0.37 ± 0.21	**0.17 ± 0.15**
AU07	0.08 ± 0.13	DND	0.35 ± 0.22	DND
AU09	**0.04 ± 0.07**	0.23 ± 0.19	0.42 ± 0.25	**0.18 ± 0.15**
AU10	0.08 ± 0.12	DND	0.23 ± 0.20	DND
AU11	0.25 ± 0.13	DND	NAA	DND
AU12	0.10 ± 0.15	0.10 ± 0.12	0.34 ± 0.20	**0.24 ± 0.16**
AU13	0.16 ± 0.11	DND	NAA	DND
AU14	0.26 ± 0.17	DND	NAA	DND
AU15	**0.05 ± 0.07**	0.36 ± 0.11	0.59 ± 0.01	**0.08 ± 0.03**
AU16	0.07 ± 0.08	DND	NAA	DND
AU17	**0.21 ± 0.18**	0.44 ± 0.09	NAA	NAA
AU20	**0.06 ± 0.08**	0.32 ± 0.11	0.28 ± 0.17	**0.15 ± 0.11**
AU23	0.04 ± 0.06	DND	NAA	DND
AU24	0.04 ± 0.02	DND	NAA	DND
AU25	**0.06 ± 0.10**	0.20 ± 0.19	**0.26 ± 0.17**	0.31 ± 0.20
AU26	**0.08 ± 0.11**	0.27 ± 0.15	0.42 ± 0.23	**0.29 ± 0.21**
AU27	0.05 ± 0.05	DND	0.12 ± 0.01	DND
AU43	0.12 ± 0.16	DND	0.55 ± 0.20	DND

values computed from the state estimation error variances provided by AUReader. Figure 6 shows the relative standard deviations for the same sequence. The following key observations can be made:

- The confidence in the initial intensity estimates is low (larger variance, larger relative standard deviation), and gets better over time.

- The confidence increases (or, equivalently, variance decreases), when the AU intensity estimates change monotonously (same direction and rate). This is evident in the short interval between Frames 97 and 102, where a narrowing of the confidence interval (in Figure 5) and a decrease in the relative standard deviation (in Figure 6) are visible.
- The confidence decreases (or, equivalently, variance increases), whenever the rate or direction of update of intensity estimate changes. For example, the interval between Frame 120 and Frame 132 shows a change in direction and rate of the AU intensity update, due to which the confidence interval in Figure 5 enlarges and the relative standard deviation in Figure 6 increases.
- Based on the above two observations, it can be said that AUReader grows more confident of its estimate, as subsequent evidence confirms a pattern of update. This is as expected, since AUReader is based on a state estimation method.

ISIR-AU also provides a confidence measure, which is a probability score that is indicative of occlusions in different facial regions. Therefore, lower confidence values indicate more occlusion or poorer quality of input image. Figure 7 illustrates how the confidence associated with the AU detection probabilities given by ISIR-AU vary. It can be seen that the confidence is relatively low throughout, and drops further towards the end of the sequence, when the subject made an out-of-plane head rotation that caused some regions of the face to be occluded. It is to be noted that the confidence computation is based on image properties and the pre-learned RF models. Therefore, the confidence measure reflects the confidence in the input features used for AU detection. This measure could therefore be indicative of aleatoric uncertainty. Since ISIR-AU is trained on static features and examines only the current image frame, the confidence measure does not adapt over time.

The UNBC-McMaster Shoulder Pain Expression Archive Database [33] has spontaneous facial expressions. However, annotations are not available for all the AUs that AUReader can predict. Therefore, another spontaneous facial expression database with annotations for all 22 AUs was used for quantitatively evaluating the generalisation performance of AUReader. The results are presented in the next subsection.

6.3 *Proprietary Market Research Database*

The performance of AUReader was evaluated quantitatively on all 408 videos in the proprietary market research database. The AUC values are presented in Figure 8.

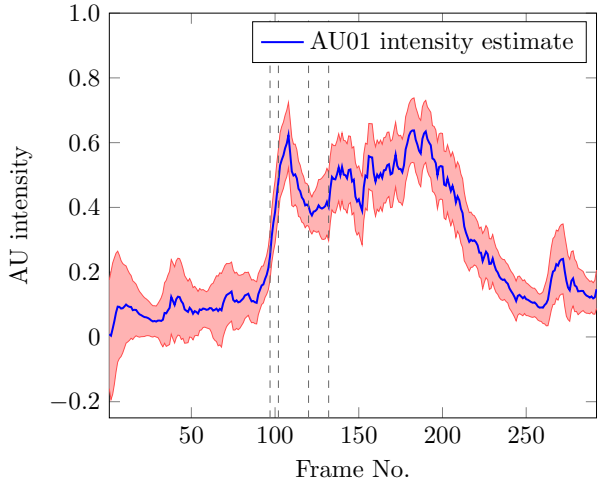


Figure 5: Confidence bounds for intensities estimated by AUReader for AU01 for a sequence in the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. For ease of identification of changes, the 2-σ confidence intervals are plotted here. Frames 97, 102, 120, and 132 have been marked using dashed, gray lines. Successive repetition of the same pattern of updates increases the confidence and reduces the variance (e.g. between Frames 97 and 102). When the direction and rate of update changes, the confidence temporarily decreases and variance increases (e.g. between Frames 120 and 132).

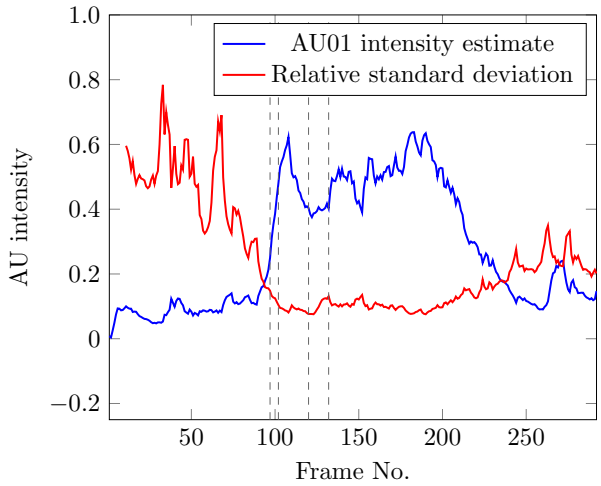


Figure 6: Illustration of the relative standard deviation for AU intensity estimates provided by AUReader for the same sequence as shown in Figure 5. It can be seen that the relative standard deviation (or, equivalently, uncertainty) drops as the intensity updates follow the same pattern over time (e.g. between Frames 97 and 102), and it increases temporarily, when the pattern changes (e.g. between Frames120, and 132). Frames 97, 102, 120, and 132 have been marked using dashed, gray lines.

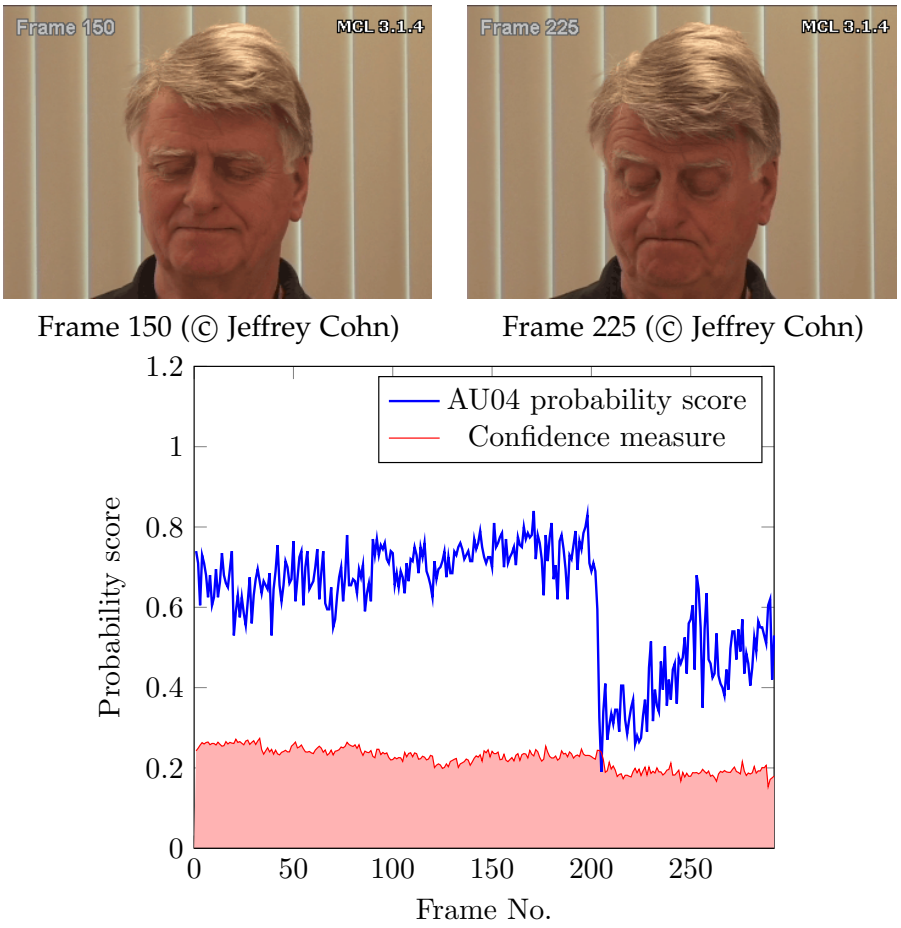


Figure 7: The plot at the bottom shows the confidence measures for AU04 probability scores provided by ISIR-AU for a sequence in the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. An out-of-plane head rotation made by the subject after Frame 200 led to occlusions on the forehead, which in turn caused a drop in confidence. This change in the head pose is visible in the images provided at the top. Frame 150 and Frame 225 illustrate the head pose before and after the out-of-plane head rotation. A sudden drop in the AU04 probability scores can also be observed, as a result of this head rotation.

The following key observations can be made:

- The AUC values vary considerably between AUs.
- Except for four AUs, the AUC values are well above 0.5. AU11 (NasolabialDeepener) and AU24 (LipPressor) scored nearly 0.5, while AU20 (LipStretcher) and AU15 (LipCornerDepressor) scored below 0.5. Recall that AU20 and AU24 also scored low AUC values on the Actor Study Database, indicating that these might

be particularly hard for AUReader to discriminate from other similar AUs.

- In the case of AU02, AU04, AU07, AU09, AU10, AU13, AU16, AU23 and AU27, the AUC values are comparable to those obtained on the Actor Study Database (difference in $[-0.05, 0.05]$).

Figures 9a and 9b present qualitative plots of intensity estimates produced by AUReader for a selection of six AUs. As in the illustrations on Actor Study Database (see Figures 1a, 1b, and 2), here too, similar observations can be made. These are listed below:

- The AU intensity estimates from AUReader follow a temporally smooth course, with visually discernible onset, apex, and offset phases.
- Delays in the onset of AUs are visible.
- The AU intensity estimates in these examples hardly came close to 0.8. As discussed in the earlier subsections, different aspects of the AUReader such as constraints, noise models, and the properties of the mass-spring-damper motion models oppose large and sudden changes in intensity estimates, and make it harder for the AUReader's intensity estimates to touch higher ranges of values. This makes the AUReader predictions robust and precise, however, at a potential cost to sensitivity or recall.

7 DISCUSSION

In this paper, we evaluated the qualitative and quantitative performance of two automatic AU detection systems, namely AUReader [23, 24] and ISIR-AU [11]. Three different databases were used for performance evaluation. The Actor Study Database [48] contains acted displays of AUs, the UNBC-McMaster Shoulder Pain Expression Archive Database [33] contains spontaneous facial expressions of pain, and the proprietary market research database contains spontaneous facial responses to commercial advertisements. The qualitative evaluation examined any delays in the detected onsets of AUs, and the range and temporal smoothness of the AU intensity estimates. The quantitative evaluation looked at the AU recognition and AU intensity estimation performance. The following paragraphs highlight the main strengths and weaknesses observed in AUReader and ISIR-AU.

The qualitative investigation of the outputs from AUReader and ISIR-AU was performed on selected sequences from the Actor Study Database [48] and the proprietary market research database. As expected, the outputs from AUReader were temporally smoother than ISIR-AU due to the integration of temporal information in the state estimation process. This made it easier to visually detect the different

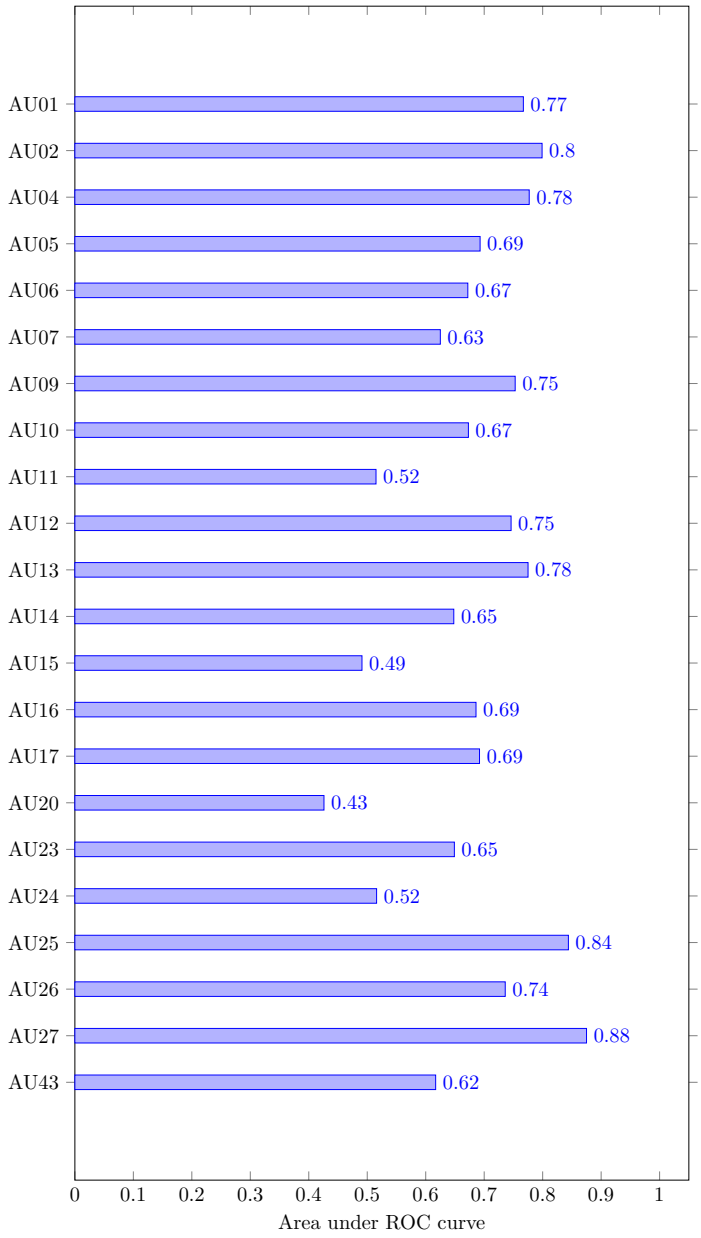


Figure 8: AUReader [24]: Areas under ROC curves obtained on the proprietary market research database. AUReader analysed 80,228 frames.

phases of AU displays (onset, apex, offset) as well as the subtle variations in the intensities of the displays. AUReader was also found to do well in recognising subtle (low intensity) displays of AUs (e.g. AU09 in Figure 1b). This makes AUReader more suitable for finer analysis of facial expressions. However, due to the constraints acting on the AU intensity estimates, the noise associated with the observations, and the viscoelastic resistance of mass-spring-damper models, the intensity estimates from AUReader showed delayed onset and tended more towards the lower range of values. This shows that the AU intensity

estimates from AUReader are more conservative, which makes the estimates suitable for real world applications such as pain monitoring.

The performance of AUReader and ISIR-AU was evaluated quantitatively for two different tasks, namely AU recognition and AU intensity estimation. The performance in the AU recognition task was measured in terms of AUC. It was found that, in general, ISIR-AU generalised

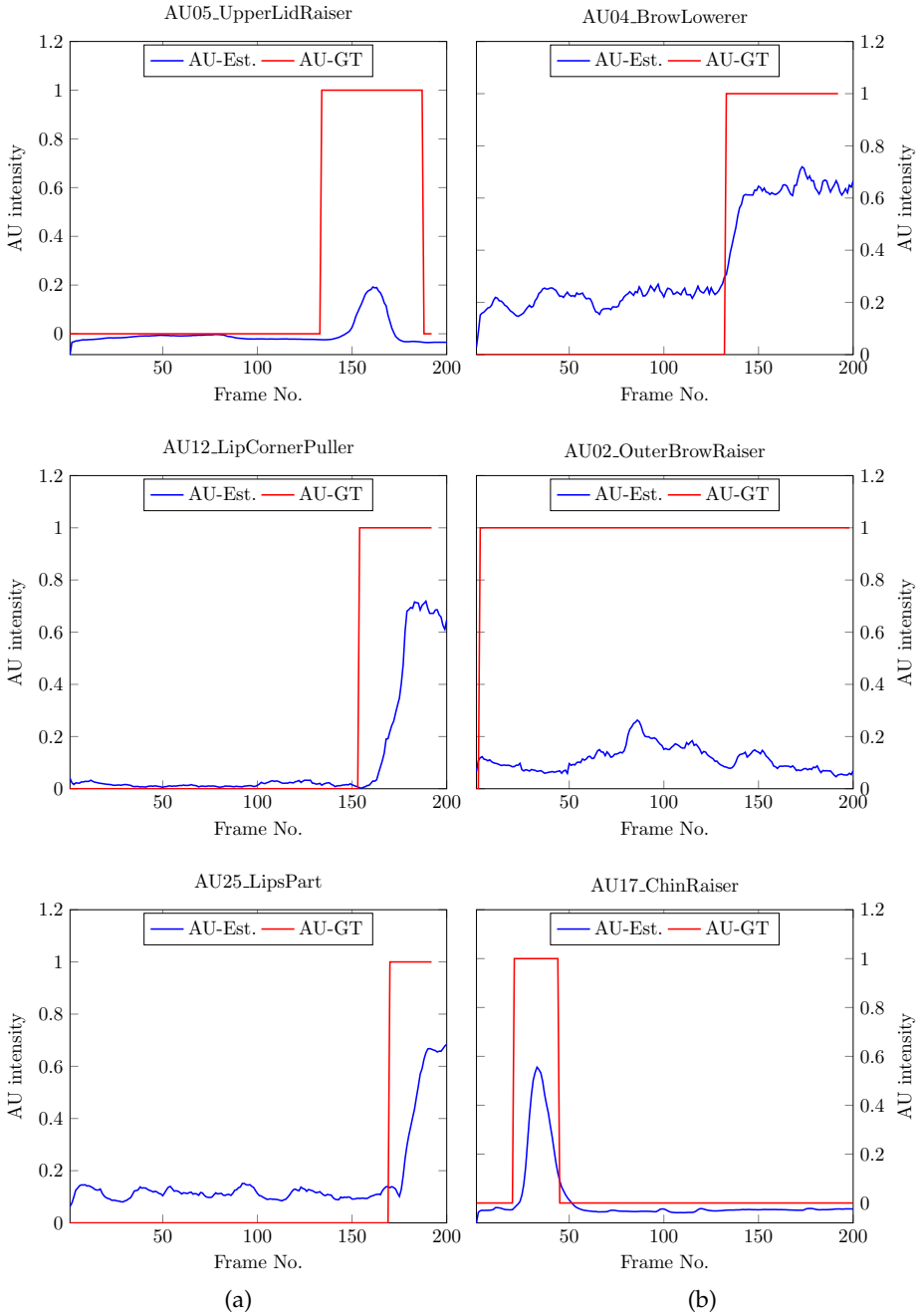


Figure 9: AUReader: Qualitative results on the proprietary market research database for a selection of six AUs.

better than AUReader to unseen subjects and spontaneous facial expressions. However, ISIR-AU did not include closely resembling AUs (except AU25 and AU26, which differ in appearance), whereas AUReader estimates several closely resembling AUs (e.g. AU06 and AU07; AU09 and AU10; AU12 and AU13; AU23 and AU24; AU26 and AU27). The AUC values obtained on all three datasets show that AUReader performs reasonably well in discriminating between geometrically similar AUs, which is a more difficult task, especially when the resembling AUs are subtly expressed. However, AU20 and AU24 appear to be difficult for AUReader to recognise well. This could be circumvented in the future by incorporating appearance-based evidence. Since ISIR-AU generalises well to unseen data, its AU predictions could be integrated within the probabilistic framework of AUReader using the method proposed in [24]. This might help in achieving better performance in real world settings by combining the strengths of both the systems.

The performance of AUReader and ISIR-AU in estimating AU intensities was evaluated on the UNBC-McMaster Shoulder Pain Expression Archive Database [33] by computing MAEs (see Table 5). To perform this evaluation, the annotated discrete intensity labels were rescaled to belong to the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. As already indicated by the qualitative results, the computed MAEs showed that AUReader's intensity estimates tended to the lower range of values, whereas those of ISIR-AU tended towards the higher range of values. This is visible in the non-zero offsets in ISIR-AU's probability scores (see Figures 1a, 1b and 2). This is also visible in the differences in the results between the two conditions under which the AU intensity estimation performance was evaluated. When the non-annotated frames were included, AUReader clearly showed less error. But, when only frames with annotated AU intensities were used, AUReader showed, in general, higher error than ISIR-AU. Therefore, it can be said that while ISIR-AU had difficulties in estimating the lower range of AU intensities, AUReader had difficulties in reaching the higher range of AU intensities. This difference is due to the differences in the methods used in the two systems. ISIR-AU is based on static, data-driven machine learning models trained to recognise AUs, and it does not use a deformable facial shape model that describes the semantics of AU intensities. But, AUReader uses such a semantic model of AU intensities. Furthermore, the probabilistic state estimation method used in AUReader allows the specification of initial values for AU intensities and performs a continous integration of information over time. Therefore, a combination of AUReader and ISIR-AU might help in increasing the coverage of both the lower and higher ranges of AU intensities, and thereby improve the AU intensity estimation performance.

Both AUReader and ISIR-AU provide estimates of uncertainties associated with their AU predictions. The uncertainty estimates produced by AUReader pertain directly to the AU intensity estimates,

and are created as part of the state estimation process. The AUReader models the uncertainty as a zero-mean Gaussian noise, which considers the errors associated with the image-based observations (aleatoric uncertainty) as well as with the dynamic process models (epistemic uncertainty). State estimation based approaches (like AUReader) that use empirically determined aleatoric uncertainty measures have the drawback that they could trust incorrect measurements to the same extent as correct measurements. This could produce incorrect AU intensities that gain trust over time, unless measures for detecting and handling such anomalies are introduced. In contrast, ISIR-AU estimates an uncertainty score based on the quality of the facial image (aleatoric uncertainty), and therefore, only indirectly measures the uncertainties in the AU probability scores. In the future, aleatoric uncertainty quantification methods that combine direct empirical evidence from past instances with an indirect measure of confidence based on the quality of the present instance could be developed or explored.

The modular structure of the state estimation framework used in AUReader [24] makes it possible to incorporate new methods for extracting facial shape and appearance information, either as additional evidence of facial expression or as a replacement for the methods currently used. Such changes do not require changes to other existing modules. This flexibility is however difficult to achieve in data-driven machine learning methods such as those used by ISIR-AU. Changes to any stage in the pipeline could require a retraining of several (subsequent) stages. The state estimation framework of AUReader also facilitates the integration of interdisciplinary knowledge from diverse fields. For example, evidences from histological or biomechanical studies of facial muscles could help in adapting the parameters of the mass-spring-damper models; methods developed in the field of navigation and tracking could be applied to enhance the robustness of the state estimation framework; control system theories could help in analysing the stability of the system; evidences from experimental psychology could be used to define higher-order semantic relationships between AUs. Furthermore, the state estimation framework used by AUReader has the potential to support on-the-fly user feedback in the form of control inputs. Future work could explore this possibility to develop AUReader into a collaborative tool for facial expression analysis in psychological and behavioural research.

8 CONCLUSION

The comparison of AUReader and ISIR-AU, presented in this paper, exemplarily elucidated the strengths and weaknesses of data-driven machine learning methods and deformable face model based methods for automatic AU analysis. It is hypothesised that a combination of the

two approaches would help in building systems that generalise and perform better in estimating AU intensities. Several future research directions have been outlined. We hope that this paper would encourage an interdisciplinary discussion on the merits and applicability of different techonological solutions for automatically detecting AUs in human-centered research.

9 APPENDIX

As mentioned in the Introduction, AU analysis is widely pursued in pain research [32]. Appendix 9.1 illustrates how the AU intensities estimated by AUReader could be used to automatically detect pain by applying evidences from experimental psychology. Appendix 9.2 illustrates how verbal explanations of AU-based pain detections could be automatically generated to assist humans in understanding the decisions made by the system.

9.1 *Automatic Detection of Pain using AU Intensities*

This section investigates the performance of the AU intensity estimates from AUReader for the message judgment task of pain detection. A set of simple rules were defined for computing pain intensities based on the continuous-valued AU intensities from AUReader. The rules are based on findings from psychological research on facial expressions of pain. Precisely speaking, the rules are based on the PSPI scale [42] and on the four clusters of facial expressions of pain (except the stoic cluster) that were identified by the psychologists Miriam Kunz and Stefan Lautenbacher at University of Bamberg (UB) [31]. Table 6 describes in detail, the rules that have been defined and applied for pain detection in this study.

As given in Table 6, the pain rules have been categorised into continuous, discrete and continuous-discrete rules, depending on the type of AU intensity values used in the rule. Continuous rules used the continuous-valued AU intensity estimates directly in the pain rule. Discrete rules binarised the AU intensity estimates before using them in the pain rule. The binarisation was performed on the basis of decision thresholds corresponding to the operating points with maximum F1-score on the ROC curves obtained on the Actor Study Database [48]. These AU recognition thresholds are listed in Table 7. Continuous-discrete rules used continuous AU intensity values for all AU terms except AU43, for which a binary value was used. Continuous and continuous-discrete rules produce continuous-valued pain intensities, and discrete rules produce discrete-valued pain intensities. Table 6 also shows that the pain rules have been categorised into max-rule and sum-rule, depending on how the intensities of certain AU groups are included in the rule. Max-rules consider only the maximum AU

intensities (continuous or discrete) within pre-defined groups of AUs. Sum-rules compute the sum of AU intensities (continuous or discrete) within each predefined group of AUs.

Table 6: Definitions of AU-based rules for pain detection. In the rules with discrete terms, the AU intensities were discretised using the operating point thresholds chosen from the ROC curves obtained on the Actor Study Database [48] (see Table 7).

Rule ID	**Category**	**Rule Definition**
PSPI-I	continuous, max-rule	AU04 + max(AU06, AU07) + max(AU09, AU10) + AU43
PSPI-II	continuous, sum-rule	AU04 + AU06 + AU07 + AU09 + AU10 + AU43
PSPI-III	continuous-discrete, max-rule	AU04 + max(AU06, AU07) + max(AU09, AU10) + bool_AU43
PSPI-IV	continuous-discrete, sum-rule	AU04 + AU06 + AU07 + AU09 + AU10 + bool_AU43
PSPI-V	discrete, max-rule	bool_AU04 + max(bool_AU06, bool_AU07) + max(bool_AU09, bool_AU10) + bool_AU43
PSPI-VI	discrete, sum-rule	bool_AU04 + bool_AU06 + bool_AU07 + bool_AU09 + bool_AU10 + bool_AU43
UB-C-I	continuous, max-rule	AU04 + max(AU06, AU07) + max(AU09, AU10)
UB-C-II	continuous, max-rule	max(AU06, AU07) + max(AU25, AU26, AU27)
UB-C-III	continuous, max-rule	max(AU01, AU02)
UB-C-IV	continuous, max-rule	AU04 + max(AU06, AU07)

The pain rules were evaluated on the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. The PSPI scores were used as the ground truth. All frames with PSPI score greater than zero were considered as positive instances of pain. Table 8 presents the AUC values for all pain rules. The following key observations can be made:

- It is not surprising that the PSPI-based rules performed better than the UB rules, since the ground truth are labeled using PSPI scale. Among the PSPI-based rules, the ones that used

Table 7: Decision thresholds used to binarise intensities of AU that are part of the pain rules. The thresholds were determined from the ROC curves computed on Actor Study Database [48]. The decision threshold corresponding to the operating point with maximum F1-score was chosen for each AU. The thresholds have been rounded to two decimal places.

AU	AU01	AU02	AU04	AU06	AU07	AU09
Decision Threshold	0.09	0.13	0.14	0.18	0.09	0.09
AU	AU10	AU25	AU26	AU27	AU43	
Decision Threshold	0.14	0.39	0.24	0.4	0.15	

continuous intensity values for some or all of the AUs performed better than those that used only discrete (binary) AU intensities.

- It can also be seen that the pain rule UB-C-II performed comparably to the PSPI-based rules. UB-C-II combines the maximum of AU06 and AU07 (narrowing of eyes) and the maximum of AU25, AU26, and AU27 (opening of mouth) [31]. The pain rules UB-C-I and UB-C-IV also include 'narrowing of eyes', but perform poorly in comparison to UB-C-II. From this, it can be concluded that the opening of mouth often accompanies the pain expressions annotated in the UNBC-McMaster Shoulder Pain Expression Archive Database [33], even though it is not part of the PSPI scale used for annotation.

- The poor performance of UB-C-III indicates that raised eyebrows did not often accompany the pain expressions annotated in the dataset.

The quantitative analysis has been followed by a qualitative analysis of pain intensities at sequence and frame levels. Among the PSPI-based rules, PSPI-III was chosen for qualitative analysis, since it performed best in quantitative analysis and resembles the original PSPI scale the most. Figures 10 and 11 show the pain intensities estimated by different pain rules for two sequences selected from the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. It can be seen that, in general, PSPI-III follows the ground truth more closely. PSPI-III and UB-C-I differ only in a single term, namely AU43 detection. While PSPI-III inlcudes AU43, UB-C-I excludes it. Therefore, the difference between the two becomes marked, when AU43 is detected in the pain sequence. In all other cases, they give the same pain intensity estimates. UB-C-IV is also a subset of PSPI-III, and hence follows the overall pattern of PSPI-III, but is insensitive to opening and closing of eyes (AU43) and to the wrinkling of nose (AU09, AU10). In contrast to the

Table 8: Performance of different pain rules on the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. The pain rules used AU intensity estimates from AUReader [24].

Rule ID	**Category**	**Performance (AUC)**
PSPI-I	continuous, max-rule	**0.65**
PSPI-II	continuous, sum-rule	0.64
PSPI-III	continuous-discrete, max-rule	**0.65**
PSPI-IV	continuous-discrete, sum-rule	**0.65**
PSPI-V	discrete, max-rule	0.62
PSPI-VI	discrete, sum-rule	0.62
UB-C-I	continuous, max-rule	0.57
UB-C-II	continuous, max-rule	0.62
UB-C-III	continuous, max-rule	0.39
UB-C-IV	continuous, max-rule	0.56

general observation from quantitative analysis, in the examples shown in Figures 10 and 11, raised eyebrows (AU01, AU02) were detected during the pain episodes, as indicated by the UB-C-III scores. Opening of mouth (AU25, AU26, AU27) was detected during the pain episode in Figure 11, as indicated by the changes in UB-C-II scores.

Figure 12 shows pain intensity scores for Frame 149 from the pain sequence analysed in Figure 10. Intensity estimates for the AUs constituent in the pain rules are also provided. The application of the decision thresholds in Table 7 would indicate that AU01, AU04, AU09 and AU43 are active in Frame 149. Pain rules PSPI-III, UB-C-I, UB-C-III and UB-C-IV, give pain intensities greater than the decision thresholds of any constituent AU. Since these rules had at least one active AU, each of these rules are considered to indicate the presence of pain in the analysed image. This follows the logic that was used in the UNBC-McMaster Shoulder Pain Expression Archive Database [33] to annotate pain intensities based on the PSPI scale. In [33], if any of the AUs in the PSPI scale were active, a non-zero pain intensity was assigned.

By empirically determining decision thresholds for all five FACS AU intensity levels (A–E), the AU intensity estimates from AUReader can be converted into FACS intensity labels. For example, let us suppose that the estimated AU intensities are mapped to FACS intensity levels as follows:

1. AU intensity estimates ≤ 0.1 represent absence of AU;
2. AU intensity estimates in $(0.1, 0.25]$ represent intensity level A;

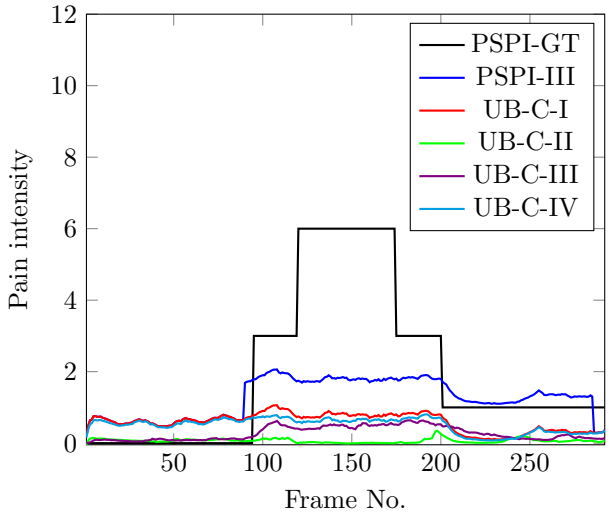


Figure 10: Rule-based pain intensity estimates computed for a sequence selected from the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. The pain rules used AU intensity estimates provided by AUReader. PSPI-GT refers to the ground truth or annotated pain score.

3. AU intensity estimates in $(0.25, 0.45]$ represent intensity level B;
4. AU intensity estimates in $(0.45, 0.65]$ represent intensity level C;
5. AU intensity estimates in $(0.65, 0.85]$ represent intensity level D;
6. AU intensity estimates > 0.85 represent intensity level E.

By applying this mapping of AU intensity estimates to FACS intensity levels, the intensity of activation of AU43, AU01, AU04, and AU09 in Frame 149 can be determined (see Figure 12). The estimated intensity of AU01 is mapped to the FACS intensity level B, that of AU04 is mapped to level C, and that of AU09 is mapped to level A. Following the method in [33], this results in a PSPI score of 5, which indicates a mild level of pain.[7] This estimate is close enough to the annotated PSPI score of 6. However, some mismatches were observed between the detected and annotated AUs and their intensities (see the description of Figure 12). Before a conclusive comparison of the pain predictions and pain annotations can be done, the thresholds for mapping the continuous AU intensities from AUReader into discrete intensity levels should be empirically validated and the ground truth AU intensity annotations should be verified by experts.

Although Boolean or discrete terms would be better for generating explanations, the use of continuous-valued AU intensities in the estimation of pain enables a fine-granular analysis of changes in pain

7 This PSPI score was computed using the formula AU04 + max(AU06, AU07) + max(AU09, AU10) + bool_AU43 [42], with discrete and symbolic intensity levels A–E mapped to the discrete and numeric intensity levels 1–5, as done in [33].

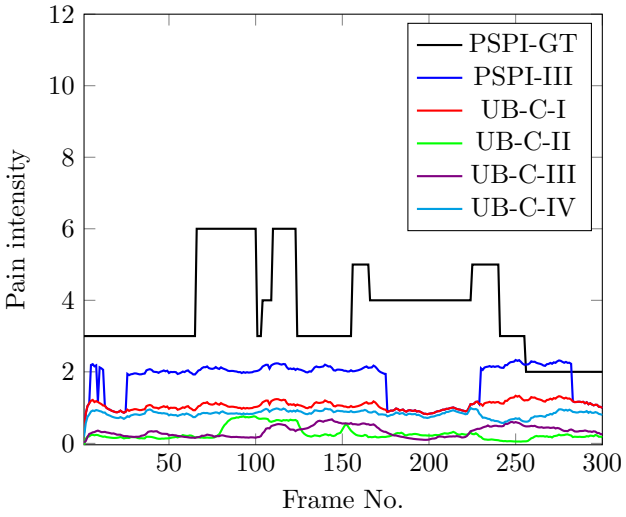


Figure 11: Rule-based pain intensity estimates computed for yet another sequence selected from the UNBC-McMaster Shoulder Pain Expression Archive Database [33]. The pain rules used AU intensity estimates provided by AUReader. PSPI-GT refers to the ground truth or annotated pain score.

intensity. This is indicated by the better quantitative performance of continuous rules (see Table 8). The continous-valued pain intensities could assist in analysing pain experience over longer periods of time, and in adapting the dosage of pain medication.

9.2 *Explaining Pain Detections: An Illustration*

Pain rules, such as those defined in Table 6, enable automatic generation of explanations for pain detections. Explanations could be generated at frame-level or sequence-level. In this illustration, frame-level explanations for pain detection are explored. These pain detections can be explained as follows, in terms of the codes of activated AUs and the facial actions they represent:

- PSPI-III detects pain in the image due to the activation of AU04, AU09 and AU43. In other words, pain is detected due to the lowering of eyebrows, wrinkling of nose, and closing of eyes.
- UB-C-I detects pain in the image due to the activation of AU04 and AU09. In other words, pain is detected due to the lowering of eyebrows and wrinkling of nose.
- UB-C-III detects pain in the image due to the activation of AU01. In other words, pain is detected due to the raising of inner corners of eyebrows.
- UB-C-IV detects pain in the image due to the activation of AU04. In other words, pain is detected due to the lowering of eyebrows.

- UB-C-II does not detect pain in the image because none of AU06, AU07, AU25, AU26 and AU27 is activated. In other words, pain is not detected because neither the narrowing of eyes nor the opening of mouth is detected.

Using the FACS intensity labels and their verbal descriptions as provided in FACS [17], more detailed explanations could be generated for the pain detections made by the pain rules, as given below:

- PSPI-III detects pain in the image due to "pronounced" lowering of eyebrows (AU04, level *C*), "traces of" nose wrinkling (AU09, level *A*), and closure of eyes (AU43).
- UB-C-I detects pain in the image due to "pronounced" lowering of eyebrows (AU04, level *C*) and "traces of" nose wrinkling (AU09, level *A*).
- UB-C-III detects pain in the image due to "slight" raising of inner corners of eyebrows (AU01, level *B*).
- UB-C-IV detects pain in the image due to the "pronounced" lowering of eyebrows (AU04, level *C*).
- UB-C-II does not detect pain in the image because neither the narrowing of eyes nor the opening of mouth is detected.

The explanations presented so far interpreted both '+' and 'max' operators in the pain rules as logical 'OR'. However, in the case of UB pain clusters [31], the '+' and 'max' operators represent logical 'AND' and logical 'OR', respectively. Going by this definition, only UB-C-III detects pain, due to the "slight" raising of inner corners of eyebrows (AU01, level *B*).

Even though the use of pain rules based on psychological evidence is helpful in generating explanations for pain detections, the automatic facial expression systems often suffer from missed detections or incorrect detections. Therefore, the question arises about how trustworthy the pain detections made by an automatic system are. Empirical evidence is not sufficient to guarantee that a detected AU is truly present. Therefore, robust mechanisms should be built-in into the automatic pain detection methods to improve the reliability of the predictions at run-time. These mechanisms could include information about the dynamics of facial expressions and the history of detected facial expressions. For example, if AU01-InnerBrowRaiser is detected to be active for several minutes, then it can be deduced that it is highly likely to be a false detection. It would also be useful to explore the use of uncertainty measures that combine aleatoric and epistemic uncertainty, while generating explanations for pain detections. In addition to these, multi-resolution explanations for pain detections might increase the robustness of the generated explanations, and help in identifying any

© Jeffrey Cohn

PSPI-III	1.76
UB-C-I	0.76
UB-C-II	-0.005
UB-C-III	0.42
UB-C-IV	0.59

AU01	0.42 ± 0.05	Level *B*
AU02	0.08 ± 0.04	Absent
AU04	0.58 ± 0.05	Level *C*
AU06	0 ± 0.02	Absent
AU07	0.01 ± 0.02	Absent
AU09	0.17 ± 0.04	Level *A*
AU10	0.03 ± 0.02	Absent
AU25	-0.02 ± 0.01	Absent
AU26	-0.02 ± 0.01	Absent
AU27	-0.03 ± 0.01	Absent
AU43	0.28 ± 0.03	Active

Figure 12: Pain intensities estimated by five different pain rules for Frame 149 from the pain sequence analysed in Figure 10 (Dataset: The UNBC-McMaster Shoulder Pain Expression Archive Database [33]) . The AU intensities from AUReader that contributed towards these estimates are listed below the image.

hidden errors. As an initial step in this direction, a combination of image-based explanations (e.g. [49]) and AU-based explanations could be explored.

Before generating explanations, it is essential to define the objective of the explanation (see Figure 5 in [1]). Depending on whether the objective is to "justify", "improve", "control" or "discover" [1], different types of explanation generation strategies might be necessary. In any case, the generation of explanations for detections of pain is a very challenging task, that should be looked at from different perspectives. This calls for intense interdisciplinary collaborations to identify and develop innovative solutions.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

- *Teena Hassan*: Conceptualisation of this comparative study, Investigation of systems' performance, Formal analysis of results, Visualisation of results, Writing – Original draft preparation
- *Dominik Seuss*: Data curation, Project administration
- *Ute Schmid*: Supervision
- *Jens Garbas*: Supervision, Writing – Review & Editing

ACKNOWLEDGEMENTS

We would like to thank Anja Dieckmann and Matthias Unfried from the Nuremberg Institute for Market Decisions, for their support during the development of AUReader, as well as for their consent to use the proprietary market research database for this work. We also thank Miriam Kunz from the University of Augsburg and Stefan Lautenbacher from the University of Bamberg for providing insights and clarifying doubts regarding facial expressions of pain and FACS.

REFERENCES

[1] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." In: *IEEE Access* 6 (2018), pp. 52138–52160. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2870052.

[2] J. Ahlberg. *CANDIDE-3 – an updated parameterized face*. Tech. rep. LiTH-ISY-R-2326. Sweden: Department of Electrical Engineering, Linköping University, 2001.

[3] T. R. Almaev and M. F. Valstar. "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition." In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 356–361. DOI: 10.1109/ACII.2013.65.

[4] Jixu Chen, Xiaoming Liu, Peter Tu, and Amy Aragones. "Learning person-specific models for facial expression and action unit recognition." In: *Pattern Recognition Letters* 34.15 (2013). Smart Approaches for Human Action Recognition, pp. 1964 –1970. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2013.02.002. URL: http://www.sciencedirect.com/science/article/pii/S0167865513000469.

[5] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. "In the Pursuit of Effective Affective Computing: The Relationship Between Features and Registration." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 1006–1016. ISSN: 1941-0492. DOI: 10.1109/TSMCB.2012.2194485.

[6] Jeffrey F. Cohn, Zara Ambadar, and Paul Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System." In: *Series in affective science. Handbook of emotion elicitation and assessment*. Ed. by J. A. Coan and J. J. B. Allen. Oxford University Press, 2007, pp. 203 –221.

[7] T.F. Cootes, G.J. Edwards, and C.J. Taylor. "Active appearance models." In: *Computer Vision - ECCV'98*. Ed. by Hans Burkhardt and Bernd Neumann. Vol. 1407. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, pp. 484–498. ISBN: 978-3-540-64613-6. DOI: 10.1007/BFb0054760. URL: http://dx.doi.org/10.1007/BFb0054760.

[8] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. "Active appearance models." In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6 (2001), pp. 681–685.

[9] David Cristinacce and Timothy F. Cootes. "Feature Detection and Tracking with Constrained Local Models." In: *British Machine Vision Conference (BMVC'06)*. 2006, pp. 929–938.

[10] Navneet Dalal and Bill Triggs. "Histograms of Oriented Gradients for Human Detection." In: *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*. Ed. by Cordelia Schmid, Stefano Soatto, and Carlo Tomasi. Vol. 1. San Diego, United States: IEEE Computer Society, June 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[11] A. Dapogny, K. Bailly, and S. Dubuisson. "Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection." In: *International Journal of Computer Vision* 126.2 (2018), pp. 255–271. ISSN: 1573-1405. DOI: 10.1007/s11263-017-1010-1. URL: https://doi.org/10.1007/s11263-017-1010-1.

[12] Charles Darwin. *The expression of the emotions in man and animals*. Original work published 1872. New York: Oxford University Press, 1998.

[13] John G. Daugman. "Image Analysis And Compact Coding By Oriented 2D Gabor Primitives." In: *Image Understanding and the Man-Machine Interface*. Ed. by Eamon B. Barrett and James J. Pearson. Vol. 0758. International Society for Optics and Photonics. SPIE, 1987, pp. 19 –30. DOI: 10.1117/12.940063. URL: https://doi.org/10.1117/12.940063.

[14] Yanchao Dong, Zhencheng Hu, Yufeng Zhou, K. Uchimura, and N. Murayama. "A robust and efficient face tracker for driver inattention monitoring system." In: *Intelligent Control and Automation (WCICA), 2011 9th World Congress on*. 2011, pp. 1212–1217. DOI: 10.1109/WCICA.2011.5970709.

[15] F. Dornaika and F. Davoine. "Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion." In: *International Journal of Computer Vision* 76.3 (2008), pp. 257–281.

[16] P. Ekman, W. V. Friesen, and J. C. Hager. *The Facial Action Coding System*. 2nd ed. Salt Lake City, UT: Research Nexus eBook, 2002.

[17] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[18] Tom Fawcett. "An introduction to ROC analysis." In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861 –874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: http://www.sciencedirect.com/science/article/pii/S016786550500303X.

[19] I. Fogel and D. Sagi. "Gabor filters as texture discriminator." In: *Biological Cybernetics* 61.2 (1989), pp. 103 –113. DOI: 10.1007/BF00204594. URL: https://doi.org/10.1007/BF00204594.

[20] D. Gabor. "Theory of communication. Part 1: the analysis of information." English. In: *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* 93 (26 1946), pp. 429–441. ISSN: 0367-7540.

[21] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. "Deep learning based FACS Action Unit occurrence and intensity estimation." In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 06. 2015, pp. 1–5. DOI: 10.1109/FG.2015.7284873.

[22] Robert G. Harper, Arthur N. Wiens, and Joseph D. Matarazzo. *Nonverbal communication: The state of the art*. Oxford, England: John Wiley & Sons, 1978.

[23] T. Hassan, D. Seuß, A. Ernst, and J. Garbas. "A Kalman filter with state constraints for model-based dynamic facial action unit estimation." In: *Forum Bildverarbeitung 2018*. Ed. by Thomas Längle, Fernando Puente León, and Michael Heizmann. KIT Scientific Publishing, 2018. DOI: 10.5445/KSP/1000085290.

[24] Teena Hassan, Dominik Seuss, Johannes Wollenberg, Jens Garbas, and Ute Schmid. "A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework." In: *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August - 2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. Frontiers in Artificial Intelligence and Applications. The Hague, The Netherlands: IOS Press, 2016, pp. 1812–1817. ISBN: 978-1-61499-671-2, 978-1-61499-672-9. DOI: 10.3233/978-1-61499-672-9-1812. URL: http://ebooks.iospress.nl/volumearticle/45031.

[25] Nils Ingemars. "A feature based face tracker using extended Kalman filtering." Bachelor's Thesis. Institutionen för Systemteknik, Department of Electrical Engineering, Linköping University, 2007.

[26] S. Jaiswal and M. Valstar. "Deep learning the dynamic appearance and shape of facial action units." In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp. 1–8. DOI: 10.1109/WACV.2016.7477625.

[27] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling." In: *IEEE Transactions on Cybernetics* 44.2 (2014), pp. 161–174. ISSN: 2168-2275. DOI: 10.1109/TCYB.2013.2249063.

[28] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. "Continuous Pain Intensity Estimation from Facial Expressions." In: *Advances in Visual Computing*. Ed. by George Bebis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 368–377. ISBN: 978-3-642-33191-6.

[29] Vahid Kazemi and Josephine Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[30] E. G. Krumhuber, L. Tamarit, E. B. Roesch, and K. R. Scherer. "FACSGen 2.0 animation software: Generating three-dimensional FACS-valid facial expressions for emotion research." In: *Emotion* 12.2 (2012), pp. 351–363.

[31] M. Kunz and S. Lautenbacher. "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain." In: *European Journal of Pain* 18.6 (2014), pp. 813–823.

[32] M. Kunz, D. Meixner, and S. Lautenbacher. "Facial muscle movements encoding pain–a systematic review." In: *Pain* 160.3 (2019), pp. 535–549.

[33] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. "Painful data: the UNBC-McMaster shoulder pain expression archive database." In: *Face and Gesture 2011*. 2011, pp. 57–64.

[34] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. "Automatic Analysis of Facial Actions: A Survey." In: *IEEE Transactions on Affective Computing* 10.3 (2019), pp. 325–347. ISSN: 2371-9850. DOI: 10.1109/TAFFC.2017.2731763.

[35] Albert Mehrabian and Susan R. Ferris. "Inference of attitudes from nonverbal communication in two channels." In: *Journal of Consulting Psychology* (1967), pp. 248 –252.

[36] Donald Michie. "Machine Learning in the Next Five Years." In: *Proceedings of the 3rd European Conference on European Working Session on Learning*. EWSL'88. Glasgow, UK: Pitman Publishing, Inc., 1988, pp. 107–122. ISBN: 0-273-08800-9. URL: http://dl.acm.org/citation.cfm?id=3108771.3108781.

[37] R. Niu, P. K. Varshney, M. Alford, A. Bubalo, E. Jones, and M. Scalzo. "Curvature nonlinearity measure and filter divergence detector for nonlinear tracking problems." In: *2008 11th International Conference on Information Fusion*. 2008, pp. 1–8.

[38] T. Ojala, M. Pietikäinen, and T. Mäenpää. "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns." In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 24.7 (July 2002), pp. 971–987. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2002.1017623.

[39] Timo Ojala, Matti Pietikäinen, and David Harwood. "A comparative study of texture measures with classification based on featured distributions." In: *Pattern Recognition* 29.1 (1996), pp. 51–59. ISSN: 0031-3203. DOI: https://doi.org/10.1016/0031-3203(95)00067-4.

[40] Brian Parkinson. "Do Facial Movements Express Emotions or Communicate Motives?" In: *Personality and Social Psychology Review* 9.4 (2005). PMID: 16223353, pp. 278–311. DOI: 10.1207/s15327957pspr0904_1. URL: https://doi.org/10.1207/s15327957pspr0904_1.

[41] Utsav Prabhu, Keshav Seshadri, and Marios Savvides. "Automatic Facial Landmark Tracking in Video Sequences Using Kalman Filter Assisted Active Shape Models." In: *Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I*. Ed. by Kiriakos N. Kutulakos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 86–99. ISBN: 978-3-642-35749-7.

[42] Kenneth M. Prkachin and Patricia E. Solomon. "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain." In: *PAIN* 139.2 (2008), pp. 267–274. ISSN: 0304-3959. DOI: https://doi.org/10.1016/j.pain.2008.04.010.

[43] Etienne B. Roesch, Lucas Tamarit, Lionel Reveret, Didier Grandjean, David Sander, and Klaus R. Scherer. "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units." In: *Journal of Nonverbal Behavior* 35.1 (2011), pp. 1 –16.

[44] E. Sariyanidi, H. Gunes, and A. Cavallaro. "Automatic Analysis of Facial Affect: a Survey of Registration, Representation, and Recognition." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6 (2015), pp. 1113–1133. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2366127.

[45] Klaus. R. Scherer. "Emotion in action, interaction, music, and speech." In: *Language, music, and the brain: A mysterious relationship*. Ed. by Michael A. Arbib. MIT Press, 2013, pp. 107 –140. DOI: 10.7551/mitpress/9780262018104.003.0005.

[46] Klaus R. Scherer, Marcello Mortillaro, Irene Rotondi, Ilaria Sergi, and Stéphanie Trznadel. "Appraisal-driven facial actions as building blocks for emotion inference." In: *Journal of Personality and Social Psychology* 114.3 (2018), pp. 358 –379.

[47] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 993–1005. ISSN: 1941-0492. DOI: 10.1109/TSMCB.2012.2193567.

[48] D. Seuss, A. Dieckmann, T. Hassan, J. Garbas, J. H. Ellgring, M. Mortillaro, and K. Scherer. "Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array." In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 35–41. DOI: 10.1109/ACII.2019.8925458.

[49] Katharina Weitz, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas. "Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods." In: *tm-Technisches Messen* 86.7-8 (2019), pp. 404–412.

[50] Xuehan Xiong and Fernando De la Torre. "Supervised Descent Method and Its Applications to Face Alignment." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

[51] Z. Zafar and N. A. Khan. "Pain Intensity Evaluation through Facial Action Units." In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 4696–4701.

D

PATENTS

D.1 AUTOMATIC FACIAL ACTION ESTIMATION

D.1.1 *Patent US 10,275,640 B2. "Determining Facial Parameters."*

Full Reference of Patent

Determining Facial Parameters, by Dominik Seuss, Teena Chakkalayil Hassan, Johannes Wollenberg, Andreas Ernst, and Jens-Uwe Garbas. (2019, Apr. 30). *Patent* US 10,275,640 B2. Accessed on: Jan. 26, 2020. [Online]. Available: USPTO PatFT Databases.

Patent Search URL

`http://patft.uspto.gov/netahtml/PTO/search-bool.html`

Innovations Contributed by Myself

Among the innovations/contributions included in this patent, the following were created as part of my doctoral research:

- The application of driven mass-spring-damper models as transition models for AU parameters/intensities within a continuous-discrete Gaussian state estimation framework.
- Fusion of noisy facial landmark positions and noisy probability scores from SVM AU classifiers within a Gaussian state estimation framework.
- Modelling of AU correlation coefficients on the basis of AU shape deformation vectors, and the application of these correlation coefficients in the Gaussian noise covariance matrices of the state estimation framework.

Written Contents Contributed by Myself

Written contents pertaining to the above-mentioned innovations were contributed originally by me. These have been included in Sections 4 and 5 in the patent, after mild revisions by the patent lawyers. It is to be noted that these sections also contain material from my master's thesis that preceded this doctoral work. Among the figures listed in Section 3, Fig. 8 and Fig. 9 were created by me as part of this doctoral work.

BIBLIOGRAPHY

[1] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] J. Ahlberg. *CANDIDE-3 – an updated parameterized face*. Tech. rep. LiTH-ISY-R-2326. Sweden: Department of Electrical Engineering, Linköping University, 2001.

[3] J. Ahlberg. *CANDIDE- a parameterized face*. http://www.icg.isy.liu.se/candide/. [Accessed 27-December-2019].

[4] T. R. Almaev and M. F. Valstar. "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition." In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 356–361. DOI: 10.1109/ACII.2013.65.

[5] A. B. Ashraf, S. Lucey, J. F. Cohn, Tsuhan Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. "The painful face – Pain expression recognition using active appearance models." In: *Image and Vision Computing* 27.12 (2009), pp. 1788–1796. DOI: 10.1016/j.imavis.2009.05.007.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. "Robust Discriminative Response Map Fitting with Constrained Local Models." In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. 2013, pp. 3444–3451. DOI: 10.1109/CVPR.2013.442.

[7] "Chapter 13 - Social cognition." In: *Fundamentals of Cognitive Neuroscience*. Ed. by B. J. Baars and N. M. Gage. San Diego: Academic Press, 2013, pp. 357–382. DOI: 10.1016/B978-0-12-415805-4.00013-8.

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." In: *PLOS ONE* 10.7 (July 2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140.

[9] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. "Automatic recognition of facial actions in spontaneous expressions." In: *Journal of Multimedia* 1.6 (2006), pp. 22–35.

[10] M. Bartlett, G. Littlewort, M. Frank, and K. Lee. "Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions." In: *Current Biology* 24.7 (2014), pp. 738–743. DOI: 10.1016/j.cub.2014.02.009.

[11] P. A. Beach, J. T. Huck, M. M. Miranda, K. T. Foley, and A. C. Bozoki. "Effects of Alzheimer Disease on the Facial Expression of Pain." In: *The Clinical Journal of Pain* 32.6 (2016), pp. 478–487. DOI: 10.1097/AJP.0000000000000302.

[12] S. Brahnam, L. Nanni, and S. Randall. "Neonatal Facial Pain Detection Using NNSOA and LSVM." In: *The 2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'08), Las Vegas*. 2008.

[13] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack. "SVM Classification of Neonatal Facial Images of Pain." In: *Fuzzy Logic and Applications*. Ed. by I. Bloch, A. Petrosino, and A. G. B. Tettamanzi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 121–128. DOI: 10.1007/11676935_15.

[14] R. G. Brown and P. Y. C. Hwang. *Introduction to random signals and applied Kalman filtering*. Third. New York: John Wiley & Sons, Inc., 1997.

[15] Y. S. Can, B. Arnrich, and C. Ersoy. "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey." In: *Journal of Biomedical Informatics* 92 (2019). DOI: 10.1016/j.jbi.2019.103139.

[16] X. Cao, Y. Wei, F. Wen, and J. Sun. "Face Alignment by Explicit Shape Regression." In: *International Journal of Computer Vision* 107.2 (2014), pp. 177–190. DOI: 10.1007/s11263-013-0667-3.

[17] C.-C. Chang and C.-J. Lin. "LIBSVM: A library for support vector machines." In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, accessed: 15-February-2020, 27:1–27:27.

[18] J. Chen, X. Liu, P. Tu, and A. Aragones. "Learning person-specific models for facial expression and action unit recognition." In: *Pattern Recognition Letters* 34.15 (2013). Smart Approaches for Human Action Recognition, pp. 1964–1970. DOI: 10.1016/j.patrec.2013.02.002.

[19] J. Chen, Z. Chen, Z. Chi, and H. Fu. "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning." In: *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI '14. Istanbul, Turkey: ACM, 2014, pp. 508–513. DOI: 10.1145/2663204.2666277.

[20] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan. "In the Pursuit of Effective Affective Computing: The Relationship Between Features and Registration." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 1006–1016. DOI: 10.1109/TSMCB.2012.2194485.

[21] J. F. Cohn. "Foundations of Human Computing: Facial Expression and Emotion." In: *Proceedings of the 8th International Conference on Multimodal Interfaces*. ICMI '06. Banff, Alberta, Canada: ACM, 2006, pp. 233–238. DOI: 10.1145/1180995.1181043.

[22] J. F. Cohn, Z. Ambadar, and P. Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System." In: *Series in affective science. Handbook of emotion elicitation and assessment*. Ed. by J. A. Coan and J. J. B. Allen. Oxford University Press, 2007, pp. 203–221.

[23] T. Cootes, G. Edwards, and C. Taylor. "Active appearance models." In: *Computer Vision - ECCV'98*. Ed. by H. Burkhardt and B. Neumann. Vol. 1407. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, pp. 484–498. DOI: 10.1007/BFb0054760.

[24] T. Cootes and C. Taylor. "Active Shape Models - 'Smart Snakes'." English. In: *BMVC92*. Ed. by D. Hogg and R. Boyle. Springer London, 1992, pp. 266–275. DOI: 10.1007/978-1-4471-3201-1_28.

[25] T. Cootes, C. Taylor, D. Cooper, and J. Graham. "Training Models of Shape from Sets of Examples." English. In: *BMVC92*. Ed. by D. Hogg and R. Boyle. Springer London, 1992, pp. 9–18. DOI: 10.1007/978-1-4471-3201-1_2.

[26] T. F. Cootes, G. J. Edwards, and C. J. Taylor. "Active appearance models." In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6 (2001), pp. 681–685. DOI: 10.1109/34.927467.

[27] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil. "Universals and cultural variations in 22 emotional expressions across five cultures." In: *Emotion* 18.1 (2018), pp. 75–93. DOI: 10.1037/emo0000302.

[28] D. Cristinacce and T. F. Cootes. "Feature Detection and Tracking with Constrained Local Models." In: *British Machine Vision Conference (BMVC'06)*. 2006, pp. 929–938. DOI: 10.5244/C.20.95.

[29] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection." In: *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*. Ed. by C. Schmid, S. Soatto, and C. Tomasi. Vol. 1. San Diego, United States: IEEE Computer Society, June 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.

[30] A. Dapogny, K. Bailly, and S. Dubuisson. "Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection." In: *International Journal of Computer Vision* 126.2 (2018), pp. 255–271. DOI: 10.1007/s11263-017-1010-1.

[31] C. Darwin. *The expression of the emotions in man and animals*. Original work published 1872. New York: Oxford University Press, 1998.

[32] J. G. Daugman. "Image Analysis And Compact Coding By Oriented 2D Gabor Primitives." In: *Image Understanding and the Man-Machine Interface*. Ed. by E. B. Barrett and J. J. Pearson. Vol. 0758. International Society for Optics and Photonics. SPIE, 1987, pp. 19–30. DOI: 10.1117/12.940063.

[33] H. Dehghani, H. Tavangar, and A. Ghandehari. "Validity and reliability of behavioral pain scale in patients with low level of consciousness due to head trauma hospitalized in intensive care unit." In: *Archives of Trauma Research* 3.1 (2014). DOI: 10.5812/atr.18608.

[34] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. "EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction." In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI '18. Boulder, CO, USA: ACM, 2018, pp. 653–656. DOI: 10.1145/3242969.3264993.

[35] H. Ding, S. K. Zhou, and R. Chellappa. "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition." In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 2017, pp. 118–126. DOI: 10.1109/FG.2017.23.

[36] Distraction. *Cambridge Dictionary*. Accessed 9-November-2019. 2019.

[37] Y. Dong, Z. Hu, Y. Zhou, K. Uchimura, and N. Murayama. "A robust and efficient face tracker for driver inattention monitoring system." In: *Intelligent Control and Automation (WCICA), 2011 9th World Congress on*. 2011, pp. 1212–1217. DOI: 10.1109/WCICA.2011.5970709.

[38] F. Dornaika and F. Davoine. "Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion." In: *International Journal of Computer Vision* 76.3 (2008), pp. 257–281. DOI: 10.1007/s11263-007-0059-7.

[39] P. N. Druzhkov and V. D. Kustikova. "A survey of deep learning methods and software tools for image classification and object detection." In: *Pattern Recognition and Image Analysis* 26.1 (2016), pp. 9–15. DOI: 10.1134/S1054661816010065.

[40] J. Egede, M. Valstar, and B. Martinez. "Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation." In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 2017, pp. 689–696. DOI: 10.1109/FG.2017.87.

[41] L. Egghe and L. Leydesdorff. "The relation between Pearson's correlation coefficient r and Salton's cosine measure." In: *Journal of the American Society for Information Science and Technology* 60.5 (2009), pp. 1027–1036. DOI: 10.1002/asi.21009.

[42] P. Ekman, W. V. Friesen, and J. C. Hager. *The Facial Action Coding System*. 2nd ed. Salt Lake City, UT: Research Nexus eBook, 2002.

[43] P. Ekman. "An argument for basic emotions." In: *Cognition and Emotion* 6.3–4 (1992), pp. 169–200. DOI: 10/bh2cq3.

[44] P. Ekman and D. Cordaro. "What is Meant by Calling Emotions Basic." In: *Emotion Review* 3.4 (2011), pp. 364–370. DOI: 10.1177/1754073911410740.

[45] P. Ekman and W. V. Friesen. "Constants across cultures in the face and emotion." In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129. DOI: 10.1037/h0030377.

[46] P. Ekman and W. V. Friesen. *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[47] P. Ekman and W. Friesen. "EMFACS-7: Emotional Facial Action Coding System." In: *Unpublished manual* (1983).

[48] I. Fogel and D. Sagi. "Gabor filters as texture discriminator." In: *Biological Cybernetics* 61.2 (1989), pp. 103–113. DOI: 10.1007/BF00204594.

[49] P. Foley and C. Kirschbaum. "Human hypothalamus–pituitary–adrenal axis responses to acute psychosocial stress in laboratory settings." In: *Neuroscience & Biobehavioral Reviews* 35.1 (2010). Psychophysiological Biomarkers of Health, pp. 91–96. DOI: 10.1016/j.neubiorev.2010.01.010.

[50] G. Freilinger, W. Happak, G. Burggasser, and H. Gruber. "Histochemical mapping and fiber size analysis of mimic muscles." In: *Plastic and reconstructive surgery* 86.3 (1990), pp. 422–428. DOI: 10.1097/00006534-199009000-00005.

[51] R. W. Frick. "Communicating emotion: The role of prosodic features." In: *Psychological Bulletin* 97.3 (1985), pp. 412–429. DOI: 10.1037/0033-2909.97.3.412.

[52] D. Gabor. "Theory of communication. Part 1: the analysis of information." English. In: *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering* 93 (26 1946), pp. 429–441. DOI: 10.1049/ji-3-2.1946.0074.

[53] J. Garbas, T. Ruf, M. Unfried, and A. Dieckmann. "Towards Robust Real-Time Valence Recognition from Facial Expressions for Market Research Applications." In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 570–575. DOI: 10.1109/ACII.2013.100.

[54] P. Genc and T. Hassan. "Analysis of Personality Dependent Differences in Pupillary Response and its Relation to Stress Recovery Ability." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 505–510. DOI: 10.1109/PERCOMW.2019.8730779.

[55] B. Gholami, W. M. Haddad, and A. R. Tannenbaum. "Relevance Vector Machine Learning for Neonate Pain Intensity Assessment Using Digital Imaging." In: *IEEE Transactions on Biomedical Engineering* 57.6 (2010), pp. 1457–1466. DOI: 10.1109/TBME.2009.2039214.

[56] M. Gjoreski, M. Gams, M. Luštrek, P. Genc, J. Garbas, and T. Hassan. "Machine Learning and End-to-End Deep Learning for Monitoring Driver Distractions From Physiological and Visual Signals." In: *IEEE Access* 8 (2020), pp. 70590–70603.

[57] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski. "Monitoring stress with a wrist device using context." In: *Journal of Biomedical Informatics* 73 (2017), pp. 159–170. DOI: 10.1016/j.jbi.2017.08.006.

[58] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. "Deep learning based FACS Action Unit occurrence and intensity estimation." In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 06. 2015, pp. 1–5. DOI: 10.1109/FG.2015.7284873.

[59] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. "A Survey of Methods for Explaining Black Box Models." In: *ACM Comput. Surv.* 51.5 (Aug. 2018). DOI: 10.1145/3236009.

[60] N. Hamelin, O. E. Moujahid, and P. Thaichon. "Emotion and advertising effectiveness: A novel facial expression analysis approach." In: *Journal of Retailing and Consumer Services* 36 (2017), pp. 103–111. DOI: 10.1016/j.jretconser.2017.01.001.

[61] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. "Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders." In: *Journal of Neuroscience Methods* 200.2 (2011), pp. 237–256. DOI: 10.1016/j.jneumeth.2011.06.023.

[62] Z. Hammal and M. Kunz. "Pain monitoring: a dynamic and context-sensitive system." In: *Pattern Recognition* 45.4 (2012), pp. 1265–1280. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.09.014.

[63] Hao Wu, Xiaoming Liu, and G. Doretto. "Face alignment via boosted ranking model." In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587753.

[64] W. Happak, G. Burggasser, and H. Gruber. "Histochemical characteristics of human mimic muscles." In: *Journal of the Neurological Sciences* 83.1 (1988), pp. 25–35. ISSN: 0022-510X. DOI: 10.1016/0022-510X(88)90017-2.

[65] J. L. Harbluk, Y. I. Noy, P. L. Trbovich, and M. Eizenman. "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance." In: *Accident Analysis and Prevention* 39.2 (2007), pp. 372–379. ISSN: 0001-4575. DOI: 10.1016/j.aap.2006.08.013.

[66] R. G. Harper, A. N. Wiens, and J. D. Matarazzo. *Nonverbal communication: The state of the art*. Oxford, England: John Wiley & Sons, 1978.

[67] T. Hassan, D. Seuß, A. Ernst, and J. Garbas. "A Kalman filter with state constraints for model-based dynamic facial action unit estimation." In: *Forum Bildverarbeitung 2018*. Ed. by T. Längle, F. Puente León, and M. Heizmann. KIT Scientific Publishing, 2018. DOI: 10.5445/KSP/1000085290.

[68] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J. Garbas, and U. Schmid. "Automatic Detection of Pain from Facial Expressions: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–17. DOI: 10.1109/TPAMI.2019.2958341.

[69] T. Hassan. "Dynamic Facial Expression Estimation by means of Model Fitting." Unpublished. Master's Thesis. Bonn-Rhein-Sieg University of Applied Sciences and Fraunhofer Institute for Integrated Circuits (IIS), 2014, 2014.

[70] T. Hassan, D. Seuss, J. Wollenberg, J. Garbas, and U. Schmid. "A Practical Approach to Fuse Shape and Appearance Information in a Gaussian Facial Action Estimation Framework." In: *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August - 2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*. Frontiers in Artificial Intelligence and Applications. The Hague, The Netherlands: IOS Press, 2016, pp. 1812–1817. DOI: 10.3233/978-1-61499-672-9-1812.

[71] C. Herrmann, D. Willersinn, and J. Beyerer. "Low-resolution Convolutional Neural Networks for video face recognition." In: *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2016, pp. 221–227. DOI: 10.1109/AVSS.2016.7738017.

[72] A. V. Hill. *First and last experiments in muscle mechanics*. Cambridge: Cambridge University Press, 1970.

[73] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard. "The effect of mental stress on heart rate variability and blood pressure during computer work." In: *European Journal of Applied Physiology* 92.1 (2004), pp. 84–89. ISSN: 1439-6327. DOI: 10.1007/s00421-004-1055-z.

[74] R. Ierusalimschy. *Programming in Lua*. 2nd ed. Rio de Janeiro: Lua.org, 2006.

[75] S. I. Inc. *FaceGen - 3D Human Faces*. http://www.facegen.com/. [Accessed 22-December-2019].

[76] N. Ingemars. "A feature based face tracker using extended Kalman filtering." Bachelor's Thesis. Institutionen för Systemteknik, Department of Electrical Engineering, Linköping University, 2007.

[77] R. Irani, K. Nasrollahi, and T. B. Moeslund. "Pain recognition using spatiotemporal oriented energy of facial muscles." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, pp. 80–87. DOI: 10.1109/CVPRW.2015.7301340.

[78] M. Itoh. "Individual differences in effects of secondary cognitive activity during driving on temperature at the nose tip." In: *2009 International Conference on Mechatronics and Automation*. 2009, pp. 7–11. DOI: 10.1109/ICMA.2009.5246188.

[79] L. G. Ixaru and G. Vanden Berghe. "Runge-Kutta Solvers for Ordinary Differential Equations." In: *Exponential Fitting*. Dordrecht: Springer Netherlands, 2004, pp. 223–304. DOI: 10.1007/978-1-4020-2100-8_6.

[80] S. Jaiswal and M. Valstar. "Deep learning the dynamic appearance and shape of facial action units." In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp. 1–8. DOI: 10.1109/WACV.2016.7477625.

[81] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970.

[82] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre. "Continuous AU intensity estimation using localized, sparse facial feature space." In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, pp. 1–7. DOI: 10.1109/FG.2013.6553808.

[83] B. Jiang, M. F. Valstar, and M. Pantic. "Action unit detection using sparse appearance descriptors in space-time video volumes." In: *Face and Gesture 2011*. 2011, pp. 314–321. DOI: 10.1109/FG.2011.5771416.

[84] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling." In: *IEEE Transactions on Cybernetics* 44.2 (2014), pp. 161–174. DOI: 10.1109/TCYB.2013.2249063.

[85] R. Kalman. "A new approach to linear filtering and prediction problems." In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45. DOI: 10.1115/1.3662552.

[86] R. Kalman and R. Bucy. "New results in linear filtering and prediction theory." In: *Journal of Basic Engineering* 83.1 (1961), pp. 95–108. DOI: 10.1115/1.3658902.

[87] S. Kaltwang, O. Rudovic, and M. Pantic. "Continuous Pain Intensity Estimation from Facial Expressions." In: *Advances in Visual Computing*. Ed. by G. Bebis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 368–377. DOI: 10.1007/978-3-642-33191-6_36.

[88] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt. "Driver Behavior Analysis for Safe Driving: A Survey." In: *IEEE Transactions on Intelligent Transportation Systems* 16.6 (2015). DOI: 10.1109/TITS.2015.2462084.

[89] S. R. Kaufman and L. A. Abel. "The Effects of Distraction on Smooth Pursuit in Normal Subjects." In: *Acta Oto-Laryngologica* 102.1-2 (1986), pp. 57–64. DOI: 10.3109/00016488609108647.

[90] V. Kazemi and J. Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[91] F. J. Keefe and R. W. Pryor. "Assessment of Pain Behaviors." In: *Encyclopedia of Pain*. Ed. by R. F. Schmidt and W. D. Willis. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 136–138. ISBN: 978-3-540-29805-2. DOI: 10.1007/978-3-540-29805-2_302.

[92] R. Kharghanian, A. Peiravi, and F. Moradi. "Pain detection from facial images using unsupervised feature learning approach." In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, pp. 419–422. DOI: 10.1109/EMBC.2016.7590729.

[93] J. Korfage, J. Koolstra, G. Langenbach, and T. van Eijden. "Fiber-type Composition of the Human Jaw Muscles –Part 2) Role of Hybrid Fibers and Factors Responsible for Inter-individual Variation." In: *Journal of Dental Research* 84.9 (2005). PMID: 16109985, pp. 784–793. DOI: 10.1177/154405910508400902.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105.

[95] S. E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth. "Speech recognition with support vector machines in a hybrid system." In: *INTERSPEECH-2005*. Lisbon, Portugal, 2005, pp. 993–996.

[96] E. G. Krumhuber, L. Tamarit, E. B. Roesch, and K. R. Scherer. "FACSGen 2.0 animation software: Generating three-dimensional FACS-valid facial expressions for emotion research." In: *Emotion* 12.2 (2012), pp. 351–363. DOI: 10.1037/a0026632.

[97] M. Kunz and S. Lautenbacher. "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain." In: *European Journal of Pain* 18.6 (2014), pp. 813–823. DOI: 10.1002/j.1532-2149.2013.00421.x.

[98] M. Kunz, D. Meixner, and S. Lautenbacher. "Facial muscle movements encoding pain–a systematic review." In: *Pain* 160.3 (2019), pp. 535–549. DOI: 10.1097/j.pain.0000000000001424.

[99] M. Kunz, V. Mylius, S. Scharmann, K. Schepelman, and S. Lautenbacher. "Influence of dementia on multiple components of pain." In: *European Journal of Pain* 13.3 (2009), pp. 317–325. DOI: 10.1016/j.ejpain.2008.05.001.

[100] M. Kunz, D. Seuss, T. Hassan, J. Garbas, M. Siebers, U. Schmid, M. Schöberl, and S. Lautenbacher. "Problems of video-based pain detection in patients with dementia: a road map to an interdisciplinary solution." In: *BMC Geriatrics* 17.33 (2017). DOI: 10.1186/s12877-017-0427-2.

[101] M. Kunz, S. Scharmann, U. Hemmeter, K. Schepelmann, and S. Lautenbacher. "The facial expression of pain in patients with dementia." In: *PAIN* 133.1 (2007), pp. 221–228. DOI: 10.1016/j.pain.2007.09.007.

[102] J. Lawrence, D. Alcock, P. McGrath, J. Kay, S. MacMurray, and C. Dulberg. "The development of a tool to assess neonatal pain." In: *Neonatal Network : NN* 12.6 (1993), pp. 59–66.

[103] Y. Li, S. M. Mavadati, M. H. Mahoor, Y. Zhao, and Q. Ji. "Measuring the intensity of spontaneous facial action units with dynamic Bayesian network." In: *Pattern Recognition* 48.11 (2015), pp. 3417–3427. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2015.04.022.

[104] Y. Liang and J. D. Lee. "Combining cognitive and visual distraction: Less than the sum of its parts." In: *Accident Analysis and Prevention* 42.3 (2010). Assessing Safety with Driving Simulators, pp. 881–890. ISSN: 0001-4575. DOI: 10.1016/j.aap.2009.05.001.

[105] Y. Liao, G. Li, S. E. Li, B. Cheng, and P. Green. "Understanding driver response patterns to mental workload increase in typical driving scenarios." In: *IEEE Access* 6 (2018), pp. 35890–35900. DOI: 10.1109/ACCESS.2018.2851309.

[106] H.-P. Lin, H.-Y. Lin, W.-L. Lin, and A. C.-W. Huang. "Effects of stress, depression, and their interaction on heart rate, skin conductance, finger temperature, and respiratory rate: sympathetic-parasympathetic hypothesis of stress and depression." In: *Journal of Clinical Psychology* 67.10 (2011), pp. 1080–1091. DOI: 10.1002/jclp.20833.

[107] H.-T. Lin, C.-J. Lin, and R. C. Weng. "A note on Platt's probabilistic outputs for support vector machines." In: *Machine Learning* 68.3 (2007), pp. 267–276. DOI: 10.1007/s10994-007-5018-6.

[108] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. "Dynamics of facial expression extracted automatically from video." In: *2004 Conference on Computer Vision and Pattern Recognition Workshop*. 2004. DOI: 10.1109/CVPR.2004.327.

[109] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. "The computer expression recognition toolbox (CERT)." In: *Face and Gesture 2011*. 2011, pp. 298–305. DOI: 10.1109/FG.2011.5771414.

[110] G. C. Littlewort, M. S. Bartlett, and K. Lee. "Faces of Pain: automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain." In: *Proceedings of the 9th International Conference on Multimodal Interfaces*. ICMI '07. Nagoya, Aichi, Japan: ACM, 2007, pp. 15–21. DOI: 10.1145/1322192.1322198.

[111] G. C. Littlewort, M. S. Bartlett, and K. Lee. "Automatic coding of facial expressions displayed during posed and genuine pain." In: *Image and Vision Computing* 27.12 (2009), pp. 1797–1803. DOI: 10.1016/j.imavis.2008.12.010.

[112] C. Liu, P. Rani, and N. Sarkar. "Affective state recognition and adaptation in human-robot interaction: a design approach." In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2006, pp. 3099–3106. DOI: 10.1109/IROS.2006.282328.

[113] D. Liu and E. S. Ebbini. "Viscoelastic property measurement in thin tissue constructs using ultrasound." In: *Ultrasonics, Ferroelectrics, and Frequency Control, IEEE Transactions on* 55.2 (2008), pp. 368–383. DOI: 10.1109/TUFFC.2008.655.

[114] D. Liu, F. Peng, A. Shea, O. Rudovic, and R. W. Picard. "DeepFaceLIFT: Interpretable Personalized Models for Automatic Estimation of Self-Reported Pain." In: *Proceedings of Machine Learning Research*. Vol. 66. 2017, pp. 1–16.

[115] L. Lo Presti and M. La Cascia. "Using Hankel matrices for dynamics-based facial emotion recognition and pain detection." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2015, pp. 26–33.

[116] R. R. Looser, P. Metzenthin, S. Helfricht, B. M. Kudielka, A. Loerbroks, J. F. Thayer, and J. E. Fischer. "Cortisol Is Significantly Correlated With Cardiovascular Responses During High Levels of Stress in Critical Care Personnel." In: *Psychosomatic Medicine* 72.3 (2010), pp. 281–289. DOI: 10.1097/PSY.0b013e3181d35065.

[117] D. Lopez-Martinez and R. Picard. "Multi-task neural networks for personalized pain recognition from physiological signals." In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2017, pp. 181–184. DOI: 10.1109/ACIIW.2017.8272611.

[118] D. Lopez-Martinez, O. Rudovic, and R. Picard. "Personalized Automatic Estimation of Self-Reported Pain Intensity from Facial Expressions." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 2318–2327. DOI: 10.1109/CVPRW.2017.286.

[119] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M. Prkachin. "Automatically detecting pain using facial actions." In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 2009, pp. 1–8. DOI: 10.1109/ACII.2009.5349321.

[120] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression." In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[121] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. "Painful data: the UNBC-McMaster shoulder pain expression archive database." In: *Face and Gesture 2011*. 2011, pp. 57–64. DOI: 10.1109/FG.2011.5771462.

[122] S. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. Schramek. "The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition." In: *Brain and Cognition* 65.3 (2007), pp. 209–237. DOI: 10.1016/j.bandc.2007.02.007.

[123] B. Mahesh, T. Hassan, E. Prassler, and J. Garbas. "Requirements for a Reference Dataset for Multimodal Human Stress Detection." In: *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 2019, pp. 492–498. DOI: 10.1109/PERCOMW.2019.8730884.

[124] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. "Automatic Analysis of Facial Actions: A Survey." In: *IEEE Transactions on Affective Computing* 10.3 (2019), pp. 325–347. DOI: 10.1109/TAFFC.2017.2731763.

[125] F. Meawad, S.-Y. Yang, and F. L. Loy. "Automatic Detection of Pain from Spontaneous Facial Expressions." In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ICMI '17. Glasgow, UK: ACM, 2017, pp. 397–401. DOI: 10.1145/3136755.3136794.

[126] A. Mehrabian and S. R. Ferris. "Inference of attitudes from nonverbal communication in two channels." In: *Journal of Consulting Psychology* (1967), pp. 248–252. DOI: 10.1037/h0024648.

[127] M. Mehu, M. Mortillaro, T. Bänziger, and K. R. Scherer. "Reliable facial muscle activation enhances recognizability and credibility of emotional expression." In: *Emotion* 12.4 (2012), pp. 701–715. DOI: doi.org/10.1037/a0026717.

[128] H. Merskey and N. Bogduk, eds. *PART III pain terms, a current list with definitions and notes on usage*. IASP Press, 2012, pp. 209–214.

[129] D. Michie. "Machine Learning in the Next Five Years." In: *Proceedings of the 3rd European Conference on European Working Session on Learning*. EWSL'88. Glasgow, UK: Pitman Publishing, Inc., 1988, pp. 107–122.

[130] M. Miyaji, H. Kawanaka, and K. Oguri. "Driver's cognitive distraction detection using physiological features by the adaboost." In: *2009 12th International IEEE Conference on Intelligent Transportation Systems*. 2009, pp. 1–6. DOI: 10.1109/ITSC.2009.5309881.

[131] A. Moors. "Integration of Two Skeptical Emotion Theories: Dimensional Appraisal Theory and Russell's Psychological Construction Theory." In: *Psychological Inquiry* 28.1 (2017), pp. 1–19. DOI: 10.1080/1047840X.2017.1235900.

[132] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda. "Appraisal Theories of Emotion: State of the Art and Future Development." In: *Emotion Review* 5.2 (2013), pp. 119–124. DOI: 10.1177/1754073912468165.

[133] S. H. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. Besold. "Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP." In: *Machine Learning* 107.7 (2018), pp. 1119–1140. DOI: 10.1007/s10994-018-5707-3.

[134] D. J. Murdoch and E. D. Chow. "A Graphical Display of Large Correlation Matrices." In: *The American Statistician* 50.2 (1996), pp. 178–180.

[135] R. Niese, A. Al-Hamadi, A. Panning, D. Brammen, U. Ebmeyer, and B. Michaelis. "Towards Pain Recognition in Post-Operative Phases Using 3D-based Features From Video and Support Vector Machines." In: *International Journal of Digital Content Technology and its Applications* 3.4 (Dec. 2009). DOI: 10.4156/jdcta.vol3.issue4.2.

[136] R. Niu, P. K. Varshney, M. Alford, A. Bubalo, E. Jones, and M. Scalzo. "Curvature nonlinearity measure and filter divergence detector for nonlinear tracking problems." In: *2008 11th International Conference on Information Fusion*. 2008, pp. 1–8.

[137] T. Ojala, M. Pietikäinen, and T. Mäenpää. "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns." In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 24.7 (July 2002), pp. 971–987. DOI: 10.1109/TPAMI.2002.1017623.

[138] T. Ojala, M. Pietikäinen, and D. Harwood. "A comparative study of texture measures with classification based on featured distributions." In: *Pattern Recognition* 29.1 (1996), pp. 51–59. DOI: 10.1016/0031-3203(95)00067-4.

[139] V. Ojansivu and J. Heikkilä. "Blur Insensitive Texture Classification Using Local Phase Quantization." In: *Image and Signal Processing*. Ed. by A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 236–243. DOI: 10.1007/978-3-540-69905-7_27.

[140] R. L. Olson, R. J. Hanowski, J. S. Hickman, and J. Bocanegra. *Driver Distraction in Commercial Vehicle Operations*. Tech. rep. FMCSA-RRR-09-042. Blacksburg, VA 24061: Center for Truck and Bus Safety, Virginia Tech Transportation Institute, 2009.

[141] O. M. Parkhi, A. Vedaldi, and A. Zisserman. "Deep Face Recognition." In: *British Machine Vision Conference*. 2015.

[142] B. Parkinson. "Do Facial Movements Express Emotions or Communicate Motives?" In: *Personality and Social Psychology Review* 9.4 (2005). PMID: 16223353, pp. 278–311. DOI: 10.1207/s15327957pspr0904_1.

[143] J. C. Platt. "Probabilistic outputs for support vector machines and comparison to regularized like-lihood methods." In: ed. by A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans. Cambridge, MA: MIT Press, 2000.

[144] U. Prabhu, K. Seshadri, and M. Savvides. "Automatic Facial Landmark Tracking in Video Sequences Using Kalman Filter Assisted Active Shape Models." In: *Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I*. Ed. by K. N. Kutulakos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 86–99. DOI: 10.1007/978-3-642-35749-7_7.

[145] L. L. Presti and M. L. Cascia. "Boosting Hankel matrices for face emotion recognition and pain detection." In: *Computer Vision and Image Understanding* 156 (2017). Image and Video Understanding in Big Data, pp. 19–33. DOI: 10.1016/j.cviu.2016.10.007.

[146] K. M. Prkachin and P. E. Solomon. "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain." In: *PAIN* 139.2 (2008), pp. 267–274. DOI: 10.1016/j.pain.2008.04.010.

[147] C. Qu, C. Herrmann, E. Monari, T. Schuchert, and J. Beyerer. "Robust 3D Patch-Based Face Hallucination." In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 1105–1114. DOI: 10.1109/WACV.2017.128.

[148] M. A. Regan, C. Hallett, and C. P. Gordon. "Driver distraction and driver inattention: Definition, relationship and taxonomy." In: *Accident Analysis and Prevention* 43.5 (2011), pp. 1771–1781. DOI: 10.1016/j.aap.2011.04.008.

[149] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[150] F. Ringeval et al. "AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition." In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. AVEC'18. Seoul, Republic of Korea: ACM, 2018, pp. 3–13. DOI: 10.1145/3266302.3266316.

[151] F. Ringeval et al. "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition." In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC '19. Nice, France: ACM, 2019, pp. 3–12. DOI: 10.1145/3347320.3357688.

[152] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. "Deep Pain: exploiting Long Short-Term Memory Networks for Facial Expression Classification." In: *IEEE Transactions on Cybernetics* (2017), pp. 1–11. DOI: 10.1109/TCYB.2017.2662199.

[153] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Scherer. "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units." In: *Journal of Nonverbal Behavior* 35.1 (2011), pp. 1–16. DOI: 10.1007/s10919-010-0095-9.

[154] A. Rowlerson, G. Raoul, Y. Daniel, J. Close, C.-A. Maurage, J. Ferri, and J. J. Sciote. "Fiber-type differences in masseter muscle associated with different facial morphologies." In: *American Journal of Orthodontics and Dentofacial Orthopedics* 127.1 (2005), pp. 37–46. DOI: 10.1016/j.ajodo.2004.03.025.

[155] J. A. Russell. "A circumplex model of affect." In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[156] SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Norm. J3016. 2018.

[157] D. Sander, D. Grandjean, S. Kaiser, T. Wehrle, and K. R. Scherer. "Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion." In: *European Journal of Cognitive Psychology* 19.3 (2007), pp. 470–480. DOI: 10.1080/09541440600757426.

[158] J. M. Saragih, S. Lucey, and J. F. Cohn. "Deformable model fitting by regularized landmark mean-shift." In: *International Journal of Computer Vision* 91.2 (2011), pp. 200–215. DOI: 10.1007/s11263-010-0380-4.

[159] E. Sariyanidi, H. Gunes, and A. Cavallaro. "Automatic Analysis of Facial Affect: a Survey of Registration, Representation, and Recognition." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.6 (2015), pp. 1113–1133. DOI: 10.1109/TPAMI.2014.2366127.

[160] S. W. Savage, D. D. Potter, and B. W. Tatler. "Does preoccupation impair hazard perception? A simultaneous EEG and Eye Tracking study." In: *Transportation Research Part F: Traffic Psychology and Behaviour* 17 (2013), pp. 52–62. DOI: https://doi.org/10.1016/j.trf.2012.10.002.

[161] A. Savran, B. Sankur, and M. T. Bilge. "Regression-based intensity estimation of facial action units." In: *Image and Vision Computing* 30.10 (2012). 3D Facial Behaviour Analysis and Understanding, pp. 774–784. DOI: 10.1016/j.imavis.2011.11.008.

[162] K. R. Scherer. "The dynamic architecture of emotion: Evidence for the component process model." In: *Cognition and Emotion* 23 (2009), pp. 1307–1351. DOI: 10.1080/02699930902928969.

[163] K. R. Scherer, M. Mortillaro, I. Rotondi, I. Sergi, and S. Trznadel. "Appraisal-driven facial actions as building blocks for emotion inference." In: *Journal of Personality and Social Psychology* 114.3 (2018), pp. 358–379. DOI: 10.1037/pspa0000107.

[164] U. Schmid, M. Siebers, D. Seuss, M. Kunz, and S. Lautenbacher. "Applying Grammar Inference to Identify Generalized Patterns of Facial Expressions of Pain." In: *Heinz, Jeffrey; de la Higuera, Colin; Oates, Tim (Hrsg.): Proceedings of the Eleventh International Conference on Grammatical Inference, PMLR*. Vol. 21. Heidelberg: Springer, 2012, pp. 183–188.

[165] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection." In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI '18. New York, NY, USA: ACM, 2018, pp. 400–408. DOI: 10.1145/3242969.3242985.

[166] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven. "Wearable-Based Affect Recognition-A Review." In: *Sensors (Basel, Switzerland)* 19.19 (2019). DOI: 10.3390/s19194079.

[167] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 993–1005. DOI: 10.1109/TSMCB.2012.2193567.

[168] D. Seuss, A. Dieckmann, T. Hassan, J. Garbas, J. H. Ellgring, M. Mortillaro, and K. Scherer. "Emotion Expression from Different Angles: A Video Database for Facial Expressions of Actors Shot by a Camera Array." In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 35–41. DOI: 10.1109/ACII.2019.8925458.

[169] R. Shadmehr and S. P. Wise. "A simple muscle model." In: *Supplementary documents for Computational Neurobiology of Reaching and Pointing*. 2005.

[170] N. Sharma and T. Gedeon. "Objective measures, sensors and computational techniques for stress recognition and classification: A survey." In: *Computer Methods and Programs in Biomedicine* 108.3 (2012), pp. 1287–1301. DOI: 10.1016/j.cmpb.2012.07.003.

[171] E. Sheu, J. Versloot, R. Nader, D. Kerr, and C. K.D. "Pain in the elderly: validity of facial expression components of observational measures." In: *The Clinical Journal of Pain* 27.7 (2011), pp. 593–601. DOI: 10.1097/AJP.0b013e31820f52e1.

[172] M. Siebers, U. Schmid, D. Seuß, M. Kunz, and S. Lautenbacher. "Characterizing facial expressions by grammars of action unit sequences - a first investigation using ABL." In: *Information Sciences* 329 (2016). Special issue on Discovery Science, pp. 866–875. DOI: 10.1016/j.ins.2015.10.007.

[173] K. Sikka, A. Dhall, and M. Bartlett. "Weakly supervised pain localization using multiple instance learning." In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553762.

[174] D. Simon. "Kalman filtering with state constraints: a survey of linear and nonlinear algorithms." In: *Control Theory Applications, IET* 4.8 (2010), pp. 1303–318. DOI: 10.1049/iet-cta.2009.0032.

[175] D. Simon. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. 1st ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[176] O. Soliemanifar, A. Soleymanifar, and R. Afrisham. "Relationship between Personality and Biological Reactivity to Stress: A Review." In: *Psychiatry Investigation* 15.12 (2018), pp. 1100–1114. DOI: 10.30773/pi.2018.10.14.2.

[177] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis. "A multimodal dataset for various forms of distracted driving." In: *Scientific Data* (2017). DOI: 10.1038/sdata.2017.110.

[178] D. I. Tamir, M. A. Thornton, J. M. Contreras, and J. P. Mitchell. "Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence." In: *Proceedings of the National Academy of Sciences* 113.1 (2016), pp. 194–199. DOI: 10.1073/pnas.1511905112.

[179] M. Tavakolian and A. Hadid. "Deep Spatiotemporal Representation of the Face for Automatic Pain Intensity Estimation." In: *2018 24th International Conference on Pattern Recognition (ICPR)*. 2018, pp. 350–354. DOI: 10.1109/ICPR.2018.8545324.

[180] D. Terzopoulos and K. Waters. "Physically-based facial modelling, analysis, and animation." In: *The Journal of Visualization and Computer Animation* 1.2 (1990), pp. 73–80. DOI: 10.1002/vis.4340010208.

[181] Y. Tong, W. Liao, and Q. Ji. "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007), pp. 1683–1699. DOI: 10.1109/TPAMI.2007.1094.

[182] F. Tsai, Y. Hsu, W. Chen, Y. Weng, C. Ng, and C. Lee. "Toward Development and Evaluation of Pain Level-Rating Scale for Emergency Triage based on Vocal Characteristics and Facial Expressions." In: *Interspeech 2016*. 2016, pp. 92–96. DOI: 10.21437/Interspeech.2016-408.

[183] M. F. Valstar and M. Pantic. "Fully Automatic Recognition of the Temporal Phases of Facial Actions." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.1 (2012), pp. 28–43. DOI: 10.1109/TSMCB.2011.2163710.

[184] M. F. Valstar and M. Pantic. "Fully Automatic Recognition of the Temporal Phases of Facial Actions." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.1 (2012), pp. 28–43. DOI: 10.1109/TSMCB.2011.2163710.

[185] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. "FERA 2017 - Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge." In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 2017, pp. 839–847. DOI: 10.1109/FG.2017.107.

[186] M. Valstar. "Automatic Behaviour Understanding in Medicine." In: *RFMIR '14: Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*. 2014, pp. 57–60. DOI: 10.1145/2666253.2666260.

[187] S. D. Vanderark and D. Ely. "Cortisol, Biochemical, and Galvanic Skin Responses to Music Stimuli of Different Preference Values by College Students in Biology and Music." In: *Perceptual and Motor Skills* 77.1 (1993). PMID: 8367245, pp. 227–234. DOI: 10.2466/pms.1993.77.1.227.

[188] A. Vinciarelli, M. Pantic, and H. Bourlard. "Social signal processing: Survey of an emerging domain." In: *Image and Vision Computing* 27.12 (2009). Visual and multimodal analysis of human spontaneous behaviour: pp. 1743–1759. DOI: 10.1016/j.imavis.2008.11.007.

[189] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features." In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

[190] S. Walter, S. Gruss, H. Ehleiter, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. Moreira da Silva. "The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system." In: *2013 IEEE International Conference on Cybernetics (CYBCO)*. 2013, pp. 128–131. DOI: 10.1109/CYBConf.2013.6617456.

[191] H. Wang and D. Yeung. "Towards Bayesian Deep Learning: A Framework and Some Existing Methods." In: *IEEE Transactions on Knowledge and Data Engineering* 28.12 (2016), pp. 3395–3408. DOI: 10.1109/TKDE.2016.2606428.

[192] V. Warden, A. Hurley, and L. Volicer. "Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale." In: *Journal of the American Medical Directors Association* 4.1 (2003), pp. 9–15. DOI: 10.1097/01.JAM.0000043422.31640.F7.

[193] K. Waters. "A Muscle Model for Animation Three- dimensional Facial Expression." In: *SIGGRAPH Comput. Graph.* 21.4 (Aug. 1987), pp. 17–24. DOI: 10.1145/37402.37405.

[194] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas. "Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods." In: *tm-Technisches Messen* 86.7-8 (2019), pp. 404–412. DOI: 10.1515/teme-2019-0024.

[195] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. "Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition." In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 1. 2005, 786–791 Vol. 1. DOI: 10.1109/ICCV.2005.147.

[196] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue. "Automatic Pain Recognition from Video and Biomedical Signals." In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 4582–4587. DOI: 10.1109/ICPR.2014.784.

[197] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue. "Automatic Pain Assessment with Facial Activity Descriptors." In: *IEEE Transactions on Affective Computing* 8.3 (2017), pp. 286–299. DOI: 10.1109/TAFFC.2016.2537327.

[198] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. Picard. "Automatic Recognition Methods Supporting Pain Assessment: A Survey." In: *IEEE Transactions on Affective Computing* (2019). DOI: 10.1109/TAFFC.2019.2946774.

[199] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. Traue. "Towards Pain Monitoring: facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges." In: *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.

[200] A. Wesley, D. Shastri, and I. Pavlidis. "A Novel Method to Monitor Driver's Distractions." In: *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '10. Atlanta, Georgia, USA: ACM, 2010, pp. 4273–4278. DOI: 10.1145/1753846.1754138.

[201] T.-F. Wu, C.-J. Lin, and R. C. Weng. "Probability Estimates for Multi-class Classification by Pairwise Coupling." In: *J. Mach. Learn. Res.* 5 (Dec. 2004), pp. 975–1005. ISSN: 1532-4435.

[202] Z. Zafar and N. A. Khan. "Pain Intensity Evaluation through Facial Action Units." In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 4696–4701. DOI: 10.1109/ICPR.2014.803.

[203] G. Zhao and M. Pietikainen. "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions." In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 29.6 (2007), pp. 915–928. DOI: 10.1109/TPAMI.2007.1110.

[204] S. Zhu, S. Liu, C. C. Loy, and X. Tang. "Deep Cascaded Bi-Network for Face Hallucination." In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Cham: Springer International Publishing, 2016, pp. 614–630. DOI: 10.1007/978-3-319-46454-1_37.