DOMINIK SEUSS

EXPLOITING DOMAIN-SPECIFIC KNOWLEDGE FOR CLASSIFIER LEARNING — AU-BASED FACIAL EXPRESSION ANALYSIS AND EMOTION RECOGNITION

EXPLOITING DOMAIN-SPECIFIC KNOWLEDGE FOR CLASSIFIER LEARNING — AU-BASED FACIAL EXPRESSION ANALYSIS AND EMOTION RECOGNITION

DOMINIK SEUSS

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:
PD Dr.-Ing. habil. Thomas Wittenberg
Friedrich-Alexander Universität Elangen-Nürnberg

Further Committee Members:
Prof. Dr. Guido Wirtz

Submitted: January 29, 2021
Day of Defense: May 7th, 2021

EXPLOITING DOMAIN-SPECIFIC KNOWLEDGE FOR CLASSIFIER LEARNING — AU-BASED FACIAL EXPRESSION ANALYSIS AND EMOTION RECOGNITION

DOMINIK SEUSS

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

URN: urn:nbn:de:bvb:473-irb-499326

DOI: https://doi.org/10.20378/irb-49932

Dominik Seuß: *Exploiting Domain-specific Knowledge for Classifier Learning — AU-based Facial Expression Analysis and Emotion Recognition*, Submitted in partial fulfilment of the requirements for the degree of Doctor of Natural Sciences (Dr. rer. nat.) of the University of Bamberg, © January 29, 2021

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 May 26, 2021.

ABSTRACT

Facial expressions are one of the most important channels of human non-verbal communication. They allow us to draw conclusions about our mental state and are considered universal across ethnicities. The Facial Action Coding System (FACS) is used in psychology to describe these facial expressions. The FACS assigns identifiers, called Action Units (AUs), to the various possible facial muscle movements. The AUs are used to analyse the semantics of facial expressions. This dissertation addresses several challenges, in particular the lack of training data, the comprehensibility of system decisions to humans in emotion and pain research, and the incorporation of domain-specific knowledge in the context of automatic facial expression recognition systems.

Training data is usually required for the development of systems for automatic detection of facial expressions. There are already public databases for training such systems, but they all have shortcomings. One of the most important factors is the quality of the annotations, since these are crucial for the later performance of the system. For robust recognition of facial expressions, even with slightly rotated faces, people need to be captured from different angles. This robustness is important because in many domains it cannot be guaranteed that people always face the camera frontally. The image material must be of high quality so that even subtle changes in the face can be detected. As part of this dissertation, a new database called the Actor Study Dataset was curated, evaluated, and published. The database includes annotations of the facial expressions shown, from appraisal dimensions and emotions to AUs, in addition to high quality recordings. For this purpose, 21 actors were filmed from five perspectives using synchronised industry and high-speed cameras, and the resulting footage was annotated by FACS experts. Two current AU detection systems were used to produce benchmark results for the different camera angles. Both the footage and the annotations were curated, processed, and made available for non-commercial use.

Many experts in artificial intelligence (AI), respectively machine learning and computer vision have developed systems that can automatically recognise facial expressions. However, there are still many challenges: few approaches use AUs as a way to trace system decisions and instead use entirely data-driven approaches whose validation is often difficult from the perspective of domain experts. In this thesis, two two-stage approaches for classifying facial expressions based on AUs are presented. Expert knowledge from the domains of emotion research and pain research have been incorporated into the develop-

ment of both approaches to improve recognition performance and reduce complexity. A two-step approach enables interpretability and validation of the decisions and results of both systems and it provides an important contribution to the research field of explainability of AI decisions.

There are few approaches that classify facial expressions into so-called appraisal dimensions. These dimensions are continuous and important for classifying subtle facial expressions and for making inferences about mental states that cannot be assigned to basic emotions. Existing systems to classifying in the appraisal dimensions use at most two dimensions and therefore represent only a relatively small spectrum. This dissertation presents a two-step approach to classify facial expressions into the three appraisal dimensions: valence, control, and novelty. In this context, a specially developed approach for the automatic detection of AU intensities is introduced as a first step. This approach takes into account the temporal relationships of AUs and can easily be extended by new information sources without having to retrain the system, thus differing significantly from existing systems. Domain-specific knowledge from emotion research allowed the development to focus on the recognition of 22 relevant AUs. Based on the detected AU intensities, the second step is to classify them into appraisal dimensions using an ordinary least squares (OLS) regression. The two-step approach enables human comprehension of the system decision, since the weights of the regression allow direct conclusions about the contribution of each AU and thus the facial muscles used. This makes an additional validation of the system by experts possible.

Existing approaches to pain recognition usually attempt to learn pain directly from images or often use statistical features of various important points on the face. These systems often achieve good classification performance, but usually do not allow for conclusions about decision making. The second approach presented in this thesis deals with the recognition of pain based on a set of rules. This set of rules, called grammar, is inferred from AU sequences of a training dataset. Using the extracted rules, new sequences can be classified into "pain" and "non-pain". If a new sequence can be generated by deriving different rules, it is a pain sequence. Domain-specific knowledge from pain research was used in the development of the approach to allow optimisation of the rule extraction procedure. The advantages of the chosen approach are the traceability of the system decision for humans by tracking the used rules of the grammar and the possibility of validation by experts.

The conclusion of this work is a call to action for stronger joint research on approaches to the interpretability of AI systems by combining the research branches "Explainable AI" and "Quantification of Uncertainty in AI Decisions".

ZUSAMMENFASSUNG

Gesichtsausdrücke sind einer der wichtigsten Kanäle menschlicher nonverbaler Kommunikation. Sie lassen Rückschlüsse auf unseren mentalen Zustand zu und gelten als universell verständlich über Ethnien hinweg. Das Facial Action Coding System (FACS) wird in der Psychologie verwendet, um diese Mimiken beschreiben zu können. Dabei werden den verschiedenen möglichen Gesichtsmuskelbewegungen Bezeichner zugeordnet, sogenannte Action Units (AUs). Mit Hilfe der AUs wird die Analyse der Semantik von Gesichtsausdrücken durchgeführt. In dieser Dissertation werden mehrere Herausforderungen, insbesondere der Mangel an Trainingsdaten, die Nachvollziehbarkeit von Systementscheidungen für den Menschen in der Emotions- und Schmerzforschung und die Einbeziehung von domänespezifischen Wissen im Kontext von Systemen zur automatischen Erkennung von Gesichtsausdrücken adressiert.

Für die Entwicklung von Systemen zur automatischen Mimikerkennung werden meist Trainingsdaten benötigt. Es gibt bereits öffentliche Datenbanken zum Training von solchen Systemen, die aber alle Defizite aufweisen. Einer der wichtigsten Faktoren ist die Qualität der Annotationen, da diese maßgebend für die spätere Leistung des Systems sind. Für eine robuste Erkennung von Gesichtsausdrücken, auch bei leicht rotierten Gesichtern, müssen Personen aus verschiedenen Winkeln aufgenommen werden. Diese Robustheit ist wichtig, da in vielen Domänen nicht garantiert werden kann, dass Menschen immer frontal in die Kamera blicken. Das Bildmaterial muss hochwertig sein, damit auch subtile Änderungen im Gesicht erkennbar sind. Im Rahmen dieser Doktorarbeit wurde eine neue Datenbank, das sogenannte Actor Study Dataset, kuratiert, evaluiert und veröffentlicht. Die Datenbank beinhaltet neben qualitativ hochwertigem Bildmaterial Annotationen der dargestellten Gesichtsausdrücke, von Appraisal-Dimensionen und Emotionen bis hin zu AUs. Hierfür wurden 21 Schauspieler aus fünf Perspektiven mit synchronisierten Industrie und high-speed Kameras gefilmt und das daraus gewonnene Filmmaterial durch FACS-Experten annotiert. Zwei aktuelle Systeme zur AU-Detektion wurden zur Erstellung von Benchmark Ergebnissen für die verschiedenen Kamerawinkel genutzt. Sowohl die Aufnahmen als auch die Annotationen wurden kuratiert, aufbereitet und für die nicht-kommerzielle Nutzung zur Verfügung gestellt.

Viele Experten im Bereich der künstlichen Intelligenz (KI), respektive des maschinellen Lernens und der Computer Vision haben Systeme entwickelt, die automatisiert Mimiken erkennen können. Es bestehen aber immer noch viele Herausforderungen: Wenige Ansätze nutzen AUs als Möglichkeit zur Nachvollziehbarkeit von Systementscheidun-

gen und verwenden stattdessen rein datengetriebene Ansätze, deren Validierung aus Sicht von Domäneexperten oftmals schwierig ist. In dieser Arbeit werden zwei zweistufige Ansätze zur Klassifikation von Gesichtsausdrücken auf Basis von AUs vorgestellt. In die Entwicklung beider Ansätze ist Expertenwissen aus den Domänen der Emotionsforschung und der Schmerzforschung eingeflossen, um die Erkennungsleistung zu verbessern und die Komplexität zu verringern. Ein zweistufiges Verfahren ermöglicht Nachvollziehbarkeit und Validierung der Entscheidungen und Ergebnisse beider Systeme und liefert einen wichtigen Beitrag zum Forschungsfeld Erklärbarkeit von KI Entscheidungen.

Es gibt wenige Ansätze, die Gesichtsausdrücke in sogenannte Appraisal-Dimensionen einordnen. Diese Dimensionen sind kontinuierlich und wichtig für die Klassifikation von subtilen Gesichtsausdrücken und für Rückschlüsse auf mentale Zustände, die keiner Basisemotionen zugeordnet werden können. Bestehende Ansätze zur Klassifikation in die Appraisal-Dimensionen verwenden höchstens zwei Dimensionen und bilden daher nur ein relativ kleines Spektrum ab. Zur Klassifikation von Gesichtsausdrücken in die drei Appraisal-Dimensionen Valence, Control und Novelty wird in dieser Doktorarbeit ein zweistufiger Ansatz vorgestellt. In diesem Rahmen wird als erster Schritt ein eigens entwickelter Ansatz zur automatischen Detektion von AU-Intensitäten eingeführt. Dieser berücksichtigt die zeitlichen Zusammenhänge von AUs und kann einfach um neue Informationsquellen erweitert werden, ohne ein Neutraining des Systems durchführen zu müssen und unterscheidet sich damit maßgeblich von bereits bestehenden Systemen. Durch domänespezifisches Wissen aus der Emotionsforschung konnte die Entwicklung auf die Erkennung von 22 relevante AUs konzentriert werden. Aufbauend auf den erkannten AU-Intensitäten, wird im zweiten Schritt die Einordnung in die Appraisal Dimensionen mit Hilfe einer Kleinst-Quadrate-Regression vorgenommen. Der zweistufige Ansatz ermöglicht eine Nachvollziehbarkeit der Systementscheidung für den Menschen, da die Gewichtungen der Regression direkte Rückschlüsse auf den Beitrag jeder einzelnen AU und damit der verwendeten Gesichtsmuskeln erlauben Dies ermöglicht eine zusätzliche Validierung des Systems durch Experten.

Vorhandene Ansätze zur Schmerzerkennung versuchen meist Schmerz direkt aus Bildern zu lernen oder verwenden oftmals statistische Merkmale verschiedener wichtiger Punkte im Gesicht. Diese Systeme erreichen oft eine gute Klassifikationsleistung, lassen aber meist keine Rückschlüsse auf die Entscheidungsfindung zu. Der zweite vorgestellte Ansatz befasst sich mit der Erkennung von Schmerz auf Basis eines Regelwerks. Dieses Regelwerk, eine sogenannte Grammatik, wird aus AU-Sequenzen eines Trainingsdatensatzes inferiert. Mit Hilfe der extrahierten Regeln können neue Sequenzen in „Schmerz“ und „nicht Schmerz“ klassifiziert werden. Wenn eine neue Sequenz durch

Ableitung verschiedener Regeln generierbar ist, handelt es sich um eine Schmerzsequenz. Bei der Entwicklung des Ansatzes wurde domänespezifisches Wissen aus der Schmerzforschung verwendet, um eine Optimierung des Regelextraktionsverfahrens zu ermöglichen. Die Vorteile des gewählten Ansatzes sind die Nachvollziehbarkeit der Systementscheidung für den Menschen durch die Nachverfolgung der verwendeten Regeln der Grammatik und der Möglichkeit der Validierung durch Experten.

Den Abschluss dieser Arbeit bildet ein Aufruf zur verstärkten gemeinsamen Forschung an Ansätzen zur Nachvollziehbarkeit von KI-Systemen durch Kombination der Forschungszweige „Erklärbare KI“ und „Quantifizierung von Unsicherheit in KI Entscheidungen“.

PUBLICATIONS

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

- **Dominik Seuss**, Teena Hassan, Anja Dieckmann, Matthias Unfried, Klaus R. Scherer, Marcello Mortillaro, and Jens Garbas. "Automatic Estimation of Action Unit Intensities and Inference of Emotional Appraisals." in *IEEE Transactions on Affective Computing (2021)*, doi: 10.1109/TAFFC.2021.3077590.
  - My Contributions:
    * Conception of the paper
    * Comparison of Action Unit (AU) detection approach with a state-of-the-art approach
    * Overview and usage of Databases
    * Overview of AU intensity estimation and significant contribution to the approach
    * Generalisability of AU detection approach
    * Contribution to the comparison of methods for appraisal inference
  - Contributing to Section 4.1
- **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 Contributions:
    * Curation and publication of the dataset
    * Description of recording setup
    * Comparison of available datasets and their deficits
    * Description and Application of one state-of-the-art System, namely OpenFace
    * Conceptualisation and generation of multi-view benchmark results
  - Contributing to Section 3.1

- 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 Contributions:
    * Review of literature on automatic pain detection from facial expressions that were published during the period 2006 – 2018.
    * Categorisation of the reviewed papers with respect to the learning tasks, the features, and the machine learning methods used
  - Contributing to Section 2.2.1
- 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 Contributions:
    * Contributor to the AU detection framework, in particular in the following significant aspects
      · Conversion and utilisation of all face models
      · Training and integration of the two 68 point facial landmark detection models
      · Supervision of two theses supporting this approach
    * System overview (Section 3) and visualisation of results (Figure 4)
  - Contributing to Section 4.1.2
- Michael Siebers, Ute Schmid, **Dominik Seuß**, Miriam Kunz, and Stephan 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.
  - My Contributions:

        * Application of Alignment-based learning on the dataset
        * Extraction of the grammar
        * Enhancement of the used approach by implementing a frequency distribution for the extracted rules as the basis for the probability computations
      - Contributing to Section 4.2
  - **Dominik Seuss**. "Bridging the Gap Between Explainable AI and Uncertainty Quantification to Enhance Trustability.". 2021. arXiv: 2105.11828 [cs.AI].
    - My Contributions:
      * Entire Idea, conception and writing
      * Overview of approaches for Explainable AI and Uncertainty Quantification
      * Reasons for the need of joint research and suggestions for joint research
    - Contributing to Section 5.3
  - 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 Contributions:
      * Description of technical steps involving SHORE and preprocessing of facial images
      * Proofreading and correcting the manuscript
    - Contributing to Chapter 1 and Section 2.2.1

II. The following patent also contains some of the ideas 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. 13, 2021. [Online]. Available: USPTO PatFT Databases.
    - My Contributions:
      * Contributor to the AU detection framework in significant aspects from supervision of two theses to the further development
    - Contributing to Section 4.1.2

"The right person in the wrong place
can make all the difference in the world."

— The G-Man[1]

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance. I would like to thank my wife Miriam and my parents Henrike and Werner for listening to my ideas or concerns during pursuing my PhD. Thanks also to Rebecca, Katharina, Christian and Anita for their interest and understanding for many tired evenings.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Ute Schmid for accepting me as an external doctoral candidate, her guidance and always helpful advices. I would like to thank my colleagues and friends from Fraunhofer IIS, especially Jens Garbas for all his support, feedback, suggestions and so much more. Many thanks to Andreas Ernst and Sebastian Hettenkofer for all these stimulating discussions and help during my time at Fraunhofer. I would also like to thank Teena Hassan for our discussions, her valuable feedback, and for the sleepless nights we worked together before deadlines. I thank the Nuremberg Institute for Market Decisions, especially Dr. Matthias Unfried and Prof. Anja Dieckmann (Aalen University), for the very good discussions, their feedback and all the helpful advises during the project and beyond.

1 I changed the original quote by replacing "man" by "person", D.S.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ABL	Alignment-based Learning
AI	Artificial Intelligence
AU	Action Unit
AUC	Area Under ROC Curve
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
FACS	Facial Action Coding System
FPR	False Positive Rate
LSTM	Long Short-Term Memory
OLS	Ordinary Least Squares
ROC	Receiver Operator Characteristic
SVM	Support Vector Machine
SVR	Support Vector Regression
TPR	True Positive Rate

Part I

SYNOPSIS OF THESIS

1 INTRODUCTION

The human face is extremely expressive and facial expressions are one very important aspect of human communication. With facial expressions we are able to convey numerous emotions without having to speak. Unlike other forms of non-verbal communication, facial expressions are considered universal [22, 71] across cultures. The true value of this type of non-verbal communication becomes obvious when people are not able to interpret it. People who suffer from autism spectrum disorders, for example, often find it difficult to interpret the face of their counterpart. This can be met with incomprehension and makes social bonding more difficult if, for example, a reaction appears to be socially inappropriate.

Systems that can recognise emotions can be helpful here. Even more, systems able to show the person why they recognised the particular emotion can in turn be used as tutor systems; thus showing the respective person by which characteristics one can recognise the emotion of the counterpart. Often, however, it is more helpful not to rely on the basic emotions but to make a classification in the emotional appraisal dimensions [77], such as the valence and control dimension. In a German Federal Ministry for Education and Research (BMBF) funded project, in which the author is involved, the aim is to create and test a new form of therapy with the help of a robotic system to support children suffering from autism spectrum disorders in the development of socio-emotional communication skills. In this context, it would be more helpful to recognise whether something is pleasant or unpleasant for the other person (positive or negative valence dimension) or whether the situation is currently overwhelming (lack of control dimension) than being limited to basic emotions which cannot cover these feelings. Existing approaches, however, do not classify in these continuous emotional appraisal dimensions but mostly use the basic emotions and often use models that are not comprehensible for humans. The development of an approach for automatic emotional appraisal inference by using domain-specific knowledge while maintaining human-interpretability is one of the main contributions of this thesis.

Facial expressions can also be taken as a form of prototypes. Prototypes are defined as mental knowledge structures that contain information about a particular object or concept in an abstract, generalised form [103]. They are not entities in memory, but illustrations of how learned knowledge can be used in information processing. Validation of prototypes is possible experimentally. In one experiment, avatars

with prototypical facial features for pain faces were used to validate the significant features for pain facial expressions [4]. In another experiment, a different effect was shown: The shift of an internal prototype. The experiment showed that a change of the internal prototype for the aesthetic perception of forks [99] was achieved. It was possible to move the aesthetic perception for such forks in a positive direction by constant exposure to rather distinctive fork shapes.

Such an approach will not lead to a change in the internal prototype for, for instance, a happy face. Raising the corners of the mouth will always be seen as smiling and associated with a happy mental state. What can change, however, is the perception of the intensity of a facial expression. Studies have shown [32] that medical staff who work in pain wards for an extended period of time are able to detect pain in a patient's face, but assess the intensity lower and lower over time. This can have serious consequences, for example, for the provision of pain medication. Kunz et al. [51] present a roadmap for the development of solutions for automatic pain detection based on facial expressions through joint interdisciplinary research. For human comprehensibility and the use in the care sector, it would be advantageous to have a rule-based system that can make the decision. However, most approaches do not classify their input based on such a set of rules. The development of a rule extraction system that can recognise pain in a way that can be interpreted by humans by exploiting domain-specific knowledge is another major contribution of this thesis.

The approaches mentioned above are both in a research direction that has recently received even greater attention: Explainable Artificial Intelligence (AI) and interpretability of AI approaches. This is about the need for decisions made by AI systems to be comprehensible and transparent to humans. The author is involved at the ADA Lovelace Center for Analytics, Data and Applications[1] and is coordinating one of the main research focus areas, namely "Explainable Learning", which deals with this research direction. In both cases described above, namely detection of emotional appraisals and recognition of pain, interpretability of the approaches is realised by combining an intermediate representation that is comprehensible to humans with methods that show the influence of these individual entities on the final decision. This intermediate representation is provided by the Facial Action Coding System (FACS) [19, 21]. With FACS it is possible to describe nearly any anatomically possible facial expression, deconstructing it into the specific Action Units (AUs) and their temporal segments that produce the expression. They are independent of any interpretation, which makes them usable for any higher order decision making process including emotional appraisal and pain detection. Furthermore they are part of the domain specific knowledge used for the development of these approaches. The manual annotation of

1 https://www.scs.fraunhofer.de/en/focus-projects/ada-center.html

AUs is a very time-consuming process that can only be performed by professionally trained coders and is therefore not applicable for fully automated systems. Existing approaches for automatic AU detection are static with respect to their expected inputs and usually do not incorporate temporal information into their models. Therefore, as a further contribution of this thesis, an automatic AU-recognition system is proposed based on a Gaussian state estimation approach that is easily extensible with additional information sources.

Training systems to automatically recognise AUs or affective states such as emotional appraisals have a major dependency: they require well-annotated data. None of the publicly available datasets provide annotated emotional appraisal dimensions and most databases do not provide recorded faces from multiple angles. As a main contribution of this thesis and to address this lack of data, a new high quality, multi-view dataset was curated and published. It contains both AU annotations as well as emotion annotations. These annotations have been created through an elaborate coding process and are available for academic use providing an important resource for the scientific research community.

Many systems and databases for automatic detection of facial expressions already exist. But many challenges still remain, such as explainability of AI approaches, the incorporation of domain-specific knowledge in classifiers and the lack of suitable and high-quality training data. For these reasons, this dissertation addresses these issues.

STRUCTURE OF THESIS

The thesis is organised as follows: Chapters 2 to 4 provide a synopsis of the research performed in this doctoral work. Chapter 2 summarises the state-of-the-art on automatic pain detection and automatic facial expression analyses and points out their deficits. Chapter 3 describes the datasets used in this thesis, while Section 3.1 addresses the research contribution towards the challenge of lack of data for automatic expression analyses by providing a new, high-quality and penta-view database, namely the Actor Study Dataset. Chapter 4 addresses the lack of human comprehensibility in systems for the automatic detection of mental states. Two detection systems, namely the emotional appraisal inference system and the pain detection approach are proposed. Both systems are two-step approaches using AUs as intermediate representation and are designed in such a way that their decisions are transparent to humans. Section 4.1 describes the approach for pursuing the detection of emotional appraisals and the incorporation of domain-specific knowledge in detail. As part of this approach, Section 4.1.2 addresses the lack of extensibility and use of temporal information of existing AU detection systems. Moreover, it describes the approach for automatic AU intensity estimation proposed in this

thesis. A system for the recognition of pain based on a rule extraction approach with incorporated domain-specific knowledge is described in Section 4.2. Chapter 5 concludes this thesis with a proposal for future research. Furthermore, it presents another contribution, namely a discussion of Explainable AI and Uncertainty Quantification, which are both research directions in the context of human-interpretability of AI approaches.

2 AUTOMATIC FACIAL EXPRESSION ANALYSES

For a better understanding of the performance of approaches and how they are embedded in their domains, definitions for measuring their success and background knowledge of domain-specific theories is required.

2.1 DEFINITIONS

2.1.1 *Performance Metrics*

- True Positive Rate
  The True Positive Rate (TPR), also known as sensitivity or recall, indicates the proportion of positive instances predicted as positive and is defined as:

  $$TPR = \frac{TP}{TP+FN}$$

- False Positive Rate
  The False Positive Rate (FPR), also known as the inversion of specificity, indicates the proportion of negative instances predicted as positive and is defined as:

  $$FPR = \frac{FP}{FP+TN}$$

- Precision
  The precision indicates the proportion of relevant instances among the retrieved instances and is defined as:

  $$PRECISION = \frac{TP}{TP+FP}$$

- Accuracy
  The accuracy indicates the proportion of right predictions and is defined as:

  $$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN}$$

- F1-score
  The F1-score is the harmonic mean of the precision and recall and is defined as:

  $$F1 - SCORE = 2 * \frac{precision*recall}{precision+recall}$$

  It outputs values between 0 and 1, with 1 indicating best result.

2.1.1.1 *Receiver Operating Characteristic*

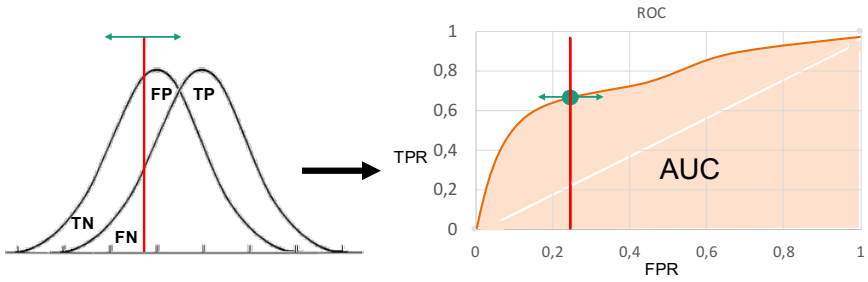

Figure 2.1: The Area Under the ROC curve. Schematic illustration about the impact to the true positive rate and the false positive rate caused by the varying threshold (left side) and the resulting area under curve plot (right side)

A Receiver Operator Characteristic (ROC) curve [24] evaluates the binary classification performance by measuring how well positive instances of a class can be distinguished from negative instances. First the output values from a binary classifier are compared to a decision threshold to predict whether an input counts to a positive or negative instance of the target class. This process is repeated for all instances in the test set. Second, the TPR and FPR are computed by comparing the predictions with the annotations. Then the decision threshold is varied to compute new sets of (TPR, FPR) pairs. A ROC curve is a two dimensional plot of these (TPR, FPR) pairs, TPR plotted along y-axis and FPR along x-axis (for a schematic overview, see Figure 2.1, left box).

2.1.1.2 *Area Under ROC Curve*

The Area Under ROC Curve (AUC) takes the results of a ROC and attempts to convert its statement into a numerical value. This is useful when a visual comparison of different ROC curves is too costly or ambiguous and a performance measurement needs to be broken down to a value. For computing the AUC, the area under the ROC curve is calculated. Good AUC values tend towards 1.0, whereas values towards 0.5 represent a random result. For a schematic overview of the combination of ROC and AUC, see Figure 2.1.

2.1.2 *The Facial Action Coding System*

The Facial Action Coding System (FACS, [19, 21]) assigns an AU to almost every visible movement of the facial muscles. These are units that summarise a single or several muscle movements. With this classification it is possible to translate facial expressions in writing -comparable to the notation of verbal expressions with written language.

Table 2.1: List of the annotated 27 AUs provided by the Actor Study Dataset. The first column contains the numerical code and the second column contains the actual name of the AU

AU Code	**AU Name**
AU01	Inner Brow Raiser
AU02	Outer Brow Raiser
AU04	Brow Lowerer
AU05	Upper Lid Raiser
AU06	Cheek Raiser
AU07	Lid Tightener
AU09	Nose Wrinkler
AU10	Upper Lip Raiser
AU11	Nasolabial Furrow Deepener
AU12	Lip Corner Puller
AU13	Cheek Puffer
AU14	Dimpler
AU15	Lip corner Depressor
AU16	Lower Lip Depressor
AU17	Chin Raiser
AU18	Lip Puckerer
AU20	Lip Stretcher
AU22	Lip Funneler
AU23	Lip Tightener
AU24	Lip Pressor
AU25	Lips Part
AU26	Jaw Drop
AU27	Mouth Stretch
AU38	Nostril Dilator
AU43	Eyes Closed
AU45	Squint
AU46	Wink

Figure 2.2: Some Examples of AUs and their combination. Images are part of the Actor Study Dataset [96]

There are 44 AUs, 12 in the upper face region and 32 in the lower face region. The AUs in the lower face are subdivided with respect to the direction of the movements. Thus, horizontal, vertical, oblique, circular, and mixed actions can be distinguished. The combinations of such AUs can be assigned to specific emotions or other mental states. Figure 2.2 shows some exemplary AUs and their combination. For this thesis, only a subset of the original 44 AUs is important. Table 2.1 shows all 27 AUs and their actual names available in the Actor Study Dataset –as defined in FACS [19, 21]. The approaches presented in Sections 4.1.2 and 4.2 use only a domain-specific subset of these 27 AUs.

2.1.3 *Emotional Appraisals*

In accordance with appraisal theories, emotions arise from our conscious or unconscious evaluations (appraisals) of an event [23]. According to the Component Process Model of Emotions (CPM, [89, 90]), emotions are determined by a cumulative sequence of evaluations (appraisals) about the relevance of events to an organism, for example:

- was the event expected or not?
- did the event promote or hinder goal attainment?
- how good could the organism deal with the event and potential consequences?

It is predicted, that the results of these appraisals will directly trigger specific psychophysiological responses as the preparation for potential behavioural reactions to this event. This is also true for the facial muscles, which are used to produce the expression of the internal state in the face. Experimental studies (e.g. [28, 92]) showed that facial muscle movements are directly based on the results of specific appraisal processes, which makes them recognisable for automatic AU detection systems as proposed in Section 4.1.2.

These observations classified in appraisal dimensions can also result in the inference of a basic emotion, e.g.

$$\text{expectancy} \downarrow + \text{lack of control} \uparrow + \text{valence} \downarrow = \text{(probably) fear}$$

Inference of the basic appraisal dimensions underlying emotion experience allows for the recognition of more nuanced emotions, including very subtle and/or volatile emotions. Therefore, the automatic recognition approach proposed in Section 4.1 focuses on the detection of emotional appraisals to detect even the subtle emotions that occur in an exemplary real-world scenario, namely market-research recordings.

2.2 STATE-OF-THE-ART

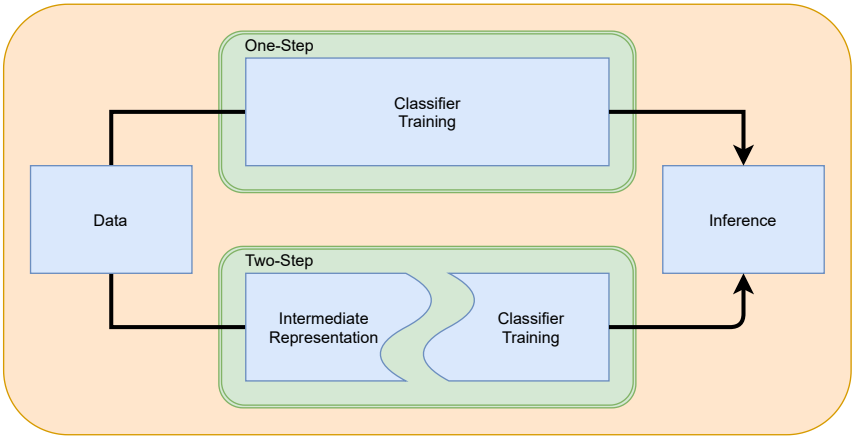

Figure 2.3: Processing pipeline for Computer Vision in Machine Learning. Starting from available data two approaches could be distinguished: (1) Two-step approaches consisting of a feature extraction and classifier training stage and (2) End-to-end approaches which pursue a classifier training directly on the images. Intermediate representation can be basic features like edges up to more sophisticated representations such as AUs. The classifier performs the inference depending on the particular task

All approaches presented in the following sections as state-of-the-art can be classified according to the schematic representation in Figure 2.3. In addition to the general categorisation, whether it is a one or two stage approach, a distinction is also made according to the datasets used and the classification objectives.

2.2.1 *Pain Detection*

Automatic detection of pain is a goal researchers have been pursuing for decades. One major contribution of this thesis is an exhaustive

research on the state-of-the-art for automatic pain detection (see [38]). The state-of-the-art can be divided into three categories as shown schematically in Figure 2.3. The pain databases used, whether it is a one-stage or two-stage approach, and the goal of the detection, for example whether it is distinct pain classes or a pain intensity estimation.

There are publicly available datasets for the training of automatic pain detection approaches (e.g. [5, 6, 66, 72, 108, 123]). The variation in the data related to the demographics of the subjects, the granularity of the annotations, and the type of annotation range from: the youngest subjects were between 18–72 hours old, the oldest subjects were 65 years old. Some databases provide annotations on frame level, some on sequence or segment level. The provided pain annotations range from binary (pain/no pain) to pain intensity levels. More datasets exist, but are partly not publicly available until today [38].

As shown in Figure 2.3 approaches could be categorised into one-step and two-step. One-step approaches predict pain or pain intensities directly from the input image or video based on geometric or textural features (see Figure 2.3, top path). Two-step approaches use an intermediate representation in terms of AUs or AU intensities before providing the actual pain estimate (see Figure 2.3, bottom path). Some approaches compute other features in addition to AUs, such as distances between facial feature points (e.g. [62]). Two-stage approaches are inspired by the way in which human coders would rate pain [30], namely based on specific facial expression elements [14].

Approaches using the intermediate representation for classifier training could make use of

- temporal information of AUs (e.g. [3, 30, 57, 58, 62, 101, 102])
- AU information of a single time-step (e.g. [64, 65, 121, 122])

This intermediate information is used in the second stage to finally train the classifier. Approaches use different machine learning methods which can be categorised depending on the prediction task:

- for classification tasks, many approaches use a Support Vector Machine (SVM) (e.g. [3, 30, 57, 58]). But also other methods, such as logistical linear regression (e.g. [64, 65]) were used.
- for regression tasks, Support Vector Regression (SVR) (e.g. [101]), linear regression (e.g. [102]) and multi-task Deep Neural Networks (DNNs) (e.g. [62]) were used.

All listed machine learning approaches are trained in a supervised way. For a categorisation of weakly-supervised or unsupervised approaches, see [38].

Most approaches in the review [38] implement a one-stage approach and don't make use of an intermediate representation. Many

approaches extract features like facial landmark positions (e.g. [74]), Gabor filters (e.g. [86]), combinations of spatial information (e.g. [36]) or spatiotemporal features like deep learned spatiotemporal features (e.g. [125]) in a first step. In the second step machine learning algorithms depending on the prediction task are for training the actual classifier:

- for classification tasks, again, most approaches use a SVM (e.g. [8, 66, 78, 112]).
- for regression tasks, SVR (e.g. [26]) or relevance vector regression (e.g. [18]) or its variants are used

More recently, deep learning methods, are increasingly being used for end-to-end learning of pain intensities. These approaches use single images (e.g. [109]) or image sequences (e.g. [84, 125]) as input.

While some approaches already use methods that make the classifier decision more interpretable to humans, such as regressions, the majority of the approaches presented are intransparent in their decision making. Kunz et al. [51] propose a roadmap to an interdisciplinary solution for pain detection for patients with dementia, suggesting a detection system based on AUs. As a major contribution of this thesis, an approach to pain recognition was developed that uses AUs as an intermediate representation to infer a grammar capable to classify pain in a human-interpretable way (see Section 4.2).

2.2.2 *Facial Expression Recognition*

Within the rapid development of AI, automatic recognition of facial expressions has been intensively studied in recent years. Systems like SHORE [50] offer the possibility to automatically recognise emotions in faces. As one building block of facial expressions, AUs, as described in Section 2.1.2, are an important concept. They are the language for interpreting a facial expression, such as of pain and emotions. However, annotating AUs is slow and tedious work that can only be done by trained coders. One example is the annotation of the Actor Study Dataset, presented in Section 3.1. As a rule of thumb, one minute of video material requires about one hour of annotation work. This is a major limitation especially for large studies and for practical systems that build on the information from AUs, therefore automatic AU detection became popular in recent years. There is a large overlap in the procedures and approaches for emotion and AU detection.

There are several models for emotions: One example is the valence-arousal space, which can be viewed as a 2D coordinate system for continuous emotion recognition. One axis encodes the positive and negative valence, the second axis encodes the degree of arousal. Sadness, as an example, has the classification medium negative valence

and medium low arousal. The well-known basic emotions, namely happiness, surprise, sadness, anger, disgust, and fear are another model [20]. Compound emotions are a rather newer model that encodes the combinations of two basic emotions each. In [17], 22 emotions were thus introduced, the 7 basic emotions, 12 combinations of basic emotions, and 3 additional emotions. Several survey articles have been published recently [40, 54], but common to all approaches is the categorisation of the approaches as shown schematically in Figure 2.3: the available datasets and whether it is a one-stage or two-stage approach. It is noteworthy that although there are several models for emotions, mostly the basic emotions are used for automatic classification of emotions.

There are publicly available datasets for emotion recognition, such as: JAFFE [68], CK+ [63], CE [17], DISFA [73], MMI [82], BU-3DFE [117], BP4D-Spontaneous [123], MPI [47], Multi-PIE [33], Oulu-CASIA [124], GEMEP-FERA [106], AFEW [15], RAF-DB [55], RAF-ML [53], GENKI-4K [113], UNBC [66]. Many of them, such as MMI [82] and BP4D-Spontaneous [123] also provide AU annotations. for a deeper comparison of AU databases please refer to [96].

The databases include between 10 and 337 subjects with posed or spontaneous induced emotions. On average, six basic emotions are annotated, up to 55 expressions [47]. None of the databases provide all three appraisal ratings, i.e. valence, novelty and control as presented in the Actor Study Dataset (see Section 3.1).

As shown in Figure 2.3 approaches could again be categorised into one-step and two-step. Here, one-step approaches predict emotions and AUs directly from the input image or video, based on geometric or textural features (see 2.3, top path). Two-step approaches use extracted features for training the final emotion classifier (see 2.3, bottom path).

Two step approaches use manually extracted features as input for the final classifier training. The approaches differ in the choice of the features they extract. The most used feature extraction methods are:

- Gabor Feature (e.g. [68, 118])
- Local Binary Pattern (e.g. [25, 35])
- Optical Flow (e.g. [9, 115])
- Active Shape Model (e.g. [10])
- Haar-like Feature (e.g. [116])
- Feature Point Tracking (e.g. [61])

While the mentioned approaches use for example the feature point coordinates from the feature point tracking for emotion detection, AU detection approaches like [16, 42] use specific models like CANDIDE-3 [1] to track AUs directly. Similar methods such as Feature Point

Tracking or the use of an AU-integrated face model were used for the approach to automatic estimation of AUs presented in this thesis (see Section 4.1.2), although they were used in modified forms.

These features are used in the second stage to finally train the classifier. Approaches use different machine learning methods to detect emotions, mostly:

- k-Nearest Neighbours (e.g. [104, 110])
- SVM (e.g. [75, 118])
- AdaBoost (e.g. [56, 111])

In addition, methods such as SVR (e.g. [44]) and relevance vector regression (e.g. [70]) were used for AU intensity estimation.

One-stage approaches are trained in an end-to-end fashion using mostly one of the following methods:

- Convolutional Neural Network (CNN) (e.g. [59, 76])
- Deep Belief Network (e.g. [60, 67])
- Long Short-Term Memory (LSTM) (e.g. [49])

While the aforementioned approaches are used for emotion recognition, these methods have also been used for AU recognition: Approaches that use CNNs (e.g. [34]) were used as well as LSTMs (e.g. [43]).

Almost all approaches of the state-of-the-art for automatic emotion recognition do not use an interpretable intermediate representation such as AUs and their recognition is often limited to basic emotions only. Although there is a system [7] which uses AUs to estimate the valence-arousal intensity in literature, it only classifies according to the valence-arousal model, has no additional dimensions integrated and its decisions are not comprehensible for humans. A continuous estimation of the emotion into a multidimensional classification while maintaining human-interpretability as presented in this thesis (see Section 4.1) is not performed by any approach. For AU detection, most approaches do not take temporal information into account or incorporate a model which encodes the knowledge about AUs, like anatomically possible motions. In addition, most systems are fixed and do not allow the integration of additional information without retraining the entire system. In this work, a new approach for automatic AU detection is proposed to address these challenges.

3 DATASETS

3.1 THE ACTOR STUDY DATASET

This section describes the contributions of this work to the challenges of low availability of suitable, high-quality and multi-view training data for facial expression recognition.

The Actor Study Dataset was created for automatic facial expression analyses namely action unit, emotion and emotional appraisal detection. For the Actor Study Dataset (see [96]), professional actors portrayed facial expressions of individual AUs, AU combinations, and staged facial expressions corresponding to different emotions and emotional appraisal scenarios:

- All video recordings contain annotations of AUs and their onset and offset by certified FACS coders
- Individual AUs and their relevant combinations which are important in facial appraisal and often difficult to detect were elaborated
- Appraisal scenarios were recorded intended to elicit relatively subtle emotion expressions
- The subjects faces were recorded from different angles simultaneously, which creates a penta-view of the person's face
- Two state-of-the-art action unit recognition systems were used for producing benchmark results on all five views

The dataset consists of 21 subjects, 11 women and 10 men, with an age range of 26-68 years and an average age of 42 years. It aimed for applications where the target person faces a frontal stimulus or observer, for example while interacting with a robot or in response to an advertisement in market-research use cases. The facial angles in these cases are in an expected range of -30 to 30 degrees. To tackle the challenge that AUs do not necessarily occur symmetrically, the recordings contain videos from 30° left and 30° right views.

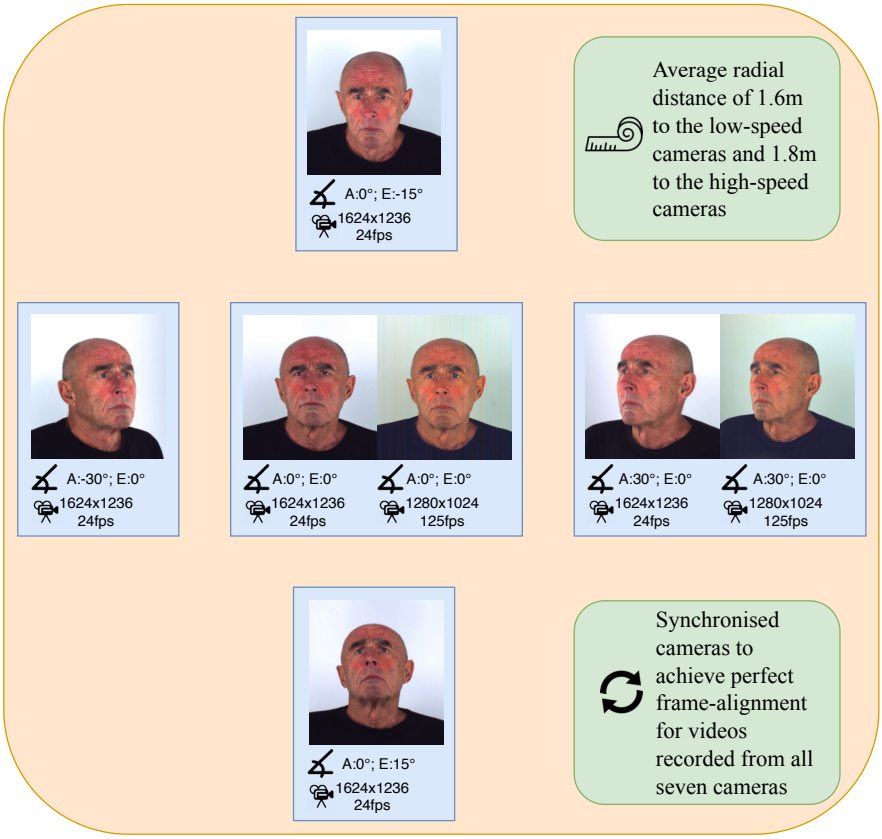

Figure 3.1: Overview of the Actor Study Dataset's recording setup with examples for each recording angle. Images are part of the Actor Study Dataset [96]

3.1.1 *Technical Setup*

Professional equipment was used to achieve best recording results:

- Five JAI CB-200 GE industrial cameras (24 frames per second; 1624x1236 px)
- Two high-speed Optronics CL300/2m cameras (125 frames per second; 1280x1024 px)
- The cameras were synchronised to achieve perfect frame-alignment for videos recorded from all seven cameras.
- Six high power MultiLED softbox lights were used to achieve uniform lighting over the actors' faces and for best visibility of AUs.

The total of 7 cameras were arranged as follows: The five low-speed cameras were positioned at five different angles, namely top, center, 30° left, 30° right and bottom. The high-speed cameras were positioned at 30° right and center positions. The average radial distance to the

low-speed cameras was 1.6m and that to the high-speed cameras was 1.8m. The setup provides real, multi-view data, which is captured using frame-synchronised cameras, in contrast to other databases, such as BP4D [123] dataset, which is synthesised through perspective distortion. For an overview of the setup with examples for each angle see Figure 3.1.

In summary, the database has 68 minutes (1,503,495 frames) of total length of recordings. 1002330 frames of these were from the high-speed cameras, and 501165 frames from low-speed cameras. The frames of the low-speed frontal camera were annotated by FACS coders, and the annotations were interpolated for the high-speed camera frames.

3.1.2 *Annotations*

The Actor Study Dataset consists of three parts: (1) Display of 32 single AUs and AU combinations, (2) Reactions to 8 appraisal scenarios and (3) Enactment of 13 emotion scenarios. Section 2.1.2 introduced the 27 AUs annotated in this database. Based on these, the relevant AU combinations for facial appraisal detection are:

- AU01+AU02
- AU01+AU04
- AU01+AU02+AU04
- AU06+AU12
- AU12+AU25

Much effort was spent to provide high-quality and frame-wise FACS annotations for this multi-view dataset. Sixteen certified FACS coders annotated 27 AUs and 5 AU combinations and 15 certified FACS coders annotated the appraisal and emotion recordings. For assessing the quality of the annotations, Cronbach's Alpha [11] was used to compute the inter-coder reliability. The actual values had a range from 0.64 to 0.72 (average = 0.69) for (1) and 0.65 to 0.87 (average=0.75) for (2) and (3), which indicates a good agreement of the coders and therefore a high quality of the annotations.

3.1.3 *Benchmarks*

For benchmarking, two state-of-the-art systems were used:

1. OpenFace [2]
   OpenFace is an open source software which is used among other things for facial landmark detection and head pose estimation. Furthermore, it is able to detect 18 AUs. For evaluation, only AU output was considered. The system uses appearance and

geometric features for AU detection in videos or single images. When a video as input is provided, OpenFace extracts frames and performs a calibration by estimating the person's expression at rest, before providing AU detection results. The system is not limited to full frontal faces.

2. AU detection system from ISIR [13]
   This system provides probability and confidence scores for six basic facial expression plus neutral expression and 12 AUs. For evaluation, only the AU probability scores were considered. The system uses a random forest where each tree uses features extracted from a randomly selected local facial region for predicting the prototypical expressions. The tree's local predictions were combined to get the global prototypical expression probability. The local predictions are furthermore used as inputs to another random forest for AU prediction. An autoencoder network is used to estimate confidence measures, used to weight the emotion predictions to achieve robustness against partial facial occlusions.

Table 3.1: Benchmark results of both systems for the center view

	OpenFace	**ISIR**
AU01	0.65	0.76
AU02	0.60	0.83
AU04	0.80	0.73
AU05	0.58	0.83
AU06	0.93	0.81
AU07	0.85	-
AU09	0.62	0.83
AU10	0.88	-
AU12	0.94	0.83
AU14	0.78	-
AU15	0.63	0.75
AU17	0.70	0.70
AU20	0.56	0.69
AU23	0.53	-
AU25	0.76	0.89
AU26	0.73	0.88
AU45	0.81	-

For the evaluation of both systems, first the ROC curve for each AU based on the output scores was computed. Second, the AUC was computed based on the ROC curves. This is a standard metric for performance evaluation in classification tasks.

OpenFace was evaluated using all five available views of the Actor Study Dataset recordings. The system from ISIR was evaluated only on the center view videos, since it was only trained on frontal faces. A comparison of the results from both approaches is shown in Table 3.1.

OpenFace analysed 100233 frames and 1290 more frames than the system from ISIR. A comparison of the mean AUC value show a slighty higher result for ISIR's system (0.79) compared to OpenFace (0.73). One explanation could be an increased chance for inter-AU confusion, since OpenFace recognises more AUs.

Table 3.2: Benchmark results for the five views, listed in the order: bottom; center; top; right; left

	F1-score	**AUC**
AU01	0.34;0.33;0.31;0.25;0.27	0.64;0.65;0.64;0.61;0.62
AU02	0.36;0.37;0.35;0.29;0.28	0.60;0.60;0.59;0.57;0.57
AU04	0.46;0.42;0.37;0.25;0.35	0.79;0.80;0.82;0.72;0.75
AU05	0.12;0.14;0.19;0.13;0.15	0.57;0.58;0.57;0.55;0.57
AU06	0.45;0.47;0.38;0.37;0.32	0.93;0.93;0.89;0.82;0.85
AU07	0.29;0.33;0.31;0.40;0.32	0.82;0.85;0.79;0.69;0.71
AU09	0.16;0.21;0.20;0.16;0.14	0.62;0.62;0.62;0.60;0.62
AU10	0.17;0.19;0.20;0.11;0.07	0.82;0.88;0.85;0.77;0.77
AU12	0.54;0.54;0.50;0.54;0.27	0.94;0.94;0.90;0.89;0.89
AU14	0.16;0.12;0.09;0.16;0.14	0.79;0.78;0.74;0.68;0.77
AU15	0.08;0.06;0.07;0.06;0.08	0.61;0.63;0.62;0.59;0.58
AU17	0.19;0.21;0.17;0.16;0.14	0.64;0.70;0.65;0.61;0.57
AU20	0.06;0.07;0.03;0.04;0.07	0.56;0.56;0.56;0.56;0.55
AU23	0.05;0.04;0.03;0.04;0.04	0.52;0.53;0.53;0.54;0.53
AU25	0.46;0.44;0.38;0.31;0.40	0.75;0.76;0.73;0.70;0.74
AU26	0.36;0.37;0.27;0.28;0.30	0.72;0.73;0.69;0.66;0.68
AU28	0.04;0.03;0.04;0.00;0.02	-
AU45	0.21;0.21;0.23;0.20;0.19	0.80;0.81;0.80;0.76;0.77

OpenFace was further evaluated using the four additional views provided by the Actor Study Dataset, since the system is able to detect AUs in non-frontal faces. These results are especially important for other researchers who want to compare their approaches and work in fields where a full frontal face as input cannot be guaranteed. For this

evaluation, F1-score was used as an additional standard metric. The F1-score could be computed since OpenFace also provides a binary decision wether an AU is present or not. Table 3.2 shows the F1 and AUC scores, obtained with OpenFace for all five views.

Both systems provide promising results. However, automatic AU extraction sometimes fails even on this high quality data. This indicates that there are still challenges in AU detection and hopefully, the Actor Study Dataset as a new dataset will help to improve AU recognition.

This chapter described the contribution of this dissertation in relation to the lack of training data by introducing a new multi-view database with 1,503,495 annotated frames and benchmark results of two state-of-the-art approaches. The included benchmark results give other researchers the opportunity to compare their approaches and probably find some hints for improvements like robustness in change of camera angles.

This dataset is used in Section 4.1.2 to train the SVMs for AU classification and for benchmarking the AU detection approach.

OTHER METHODS TO OVERCOME DATA SCARCITY Many methods, especially Deep Learning (DL)-based methods, require a large number of training data. In addition to the provision of a new, high-quality dataset, other approaches, namely the synthetic production of data was also investigated. Here, the aim is to generate new, similar data from already existing and preferably annotated data. There are approaches for generating synthetic data (e.g. [31, 46]) and approaches for expression transfer (e.g. [105, 126]). The latter are about transferring facial expressions from one class of training data to another. This task was pursued in the form of two supervised master's theses[1][2]. These theses focused on generating additional samples for an underrepresented class by means of facial expression transfer using the overrepresented class. Both theses showed promising results for the transfer of the facial expressions and the evaluation of the quality of transferred images, although they could not yet show a significant improvement for classifier training.

3.2 THIRD PARTY DATASETS

3.2.1 *Proprietary Market-Research Database*

This undisclosed dataset was generated for applying appraisal inference in a real-world scenario. It contains images of natural and spontaneous emotion displays. The dataset covers the use case of con-

1 Maximilian Iftner: "Facial expression transfer on unpaired image sets by using deep neural networks", Master's Thesis

2 Stefan Saloman: "Defining a new Metric for Generative Adversarial Networks in context of Facial Expression Transfer", Master's Thesis

sumer research studies and consists of 408 video sequences, extracted from longer face recordings of 155 subjects, exposed to a series of TV commercials. The sequences have full HD resolution and an average length of 8 seconds [91]. In a verification step, psychology students made sure that an identifiable facial movement was shown and its apex (the point of maximum muscle contraction) was within the sequence. For FACS annotations, the 408 sequences were distributed to certified FACS coders, who generated the AU annotations including three levels of intensity. These three are a derivative of the original five levels, which are as follows:

- A $\widehat{=}$ a and b in FACS manual (small action)
- B $\widehat{=}$ c in FACS manual (moderate to strong action,)
- C $\widehat{=}$ d and e in FACS manual (estimated maximum action)

This dataset is used in Section 4.1 for the assessment of the generalisation performance of the AU detection approach and the training and evaluation of the emotional appraisal inference system.

3.2.2 *Extended Cohn-Kanade Dataset*

The Extended Cohn-Kanade (CK+) dataset [63] is a well-known database for facial expression recognition. It contains 593 sequences of subjects showing mostly posed facial expressions. The subjects are between 18 and 50 years old while 69% are female and 31% are male. Different ethnicities are represented: 81% Euro-American, 13% Afro-American and 6% other groups. The sequences contain AU annotations according to FACS and additionally, annotations of basic emotions and the coordinates of a 68-point facial mesh. The image resolution is mostly 640x480 pixels. An example image is given in Figure 3.2.

Figure 3.2: Example picture from the CK+ Database [63]

This dataset is used in Section 4.1.2 to empirically determine the noise in the face alignment method and to train the SVMs for AU classification.

3.2.3 *Pain Sequence Dataset*

This undisclosed dataset is a derivate of the work from Kunz et al. [52]. The original data was obtained in a psychological study of facial expression of pain in patients with dementia. In this study, demented patients and healthy persons were exposed to pain of various intensities resulting in 347 sequences of 86 subjects with a maximum length of 17 AUs and a mean of 4.03 AUs. These pain episodes were converted into a sequence of AUs, while maintaining properties like chronological order and simultaneous occurrence. Following these definition two entities can be distinguished:

1. Singular AUs, such as AU06 and AU07, where the string "AU06 AU07" denotes a sequential event of AU07 follows after AU06 was present.

2. AU compound, such as AU06-07, denoting the simultaneous presence of AU06 and AU07 and therefore coding a single event.

Converting the original data into this format results in a total of 76 distinct AU and AU compounds with a maximum of 13 and a mean of 3.54 AUs and AU compounds per sequence.

This dataset is used in section 4.2 to extract the set of rules for classifying pain and for its performance evaluation.

4 AUTOMATIC FACIAL EXPRESSION RECOGNITION

4.1 AUTOMATIC APPRAISAL INFERENCE

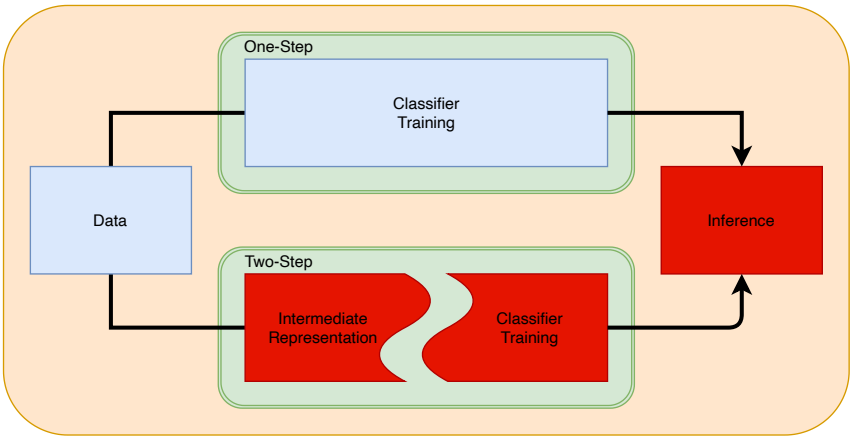

Figure 4.1: Processing pipeline for Computer Vision in Machine Learning. The focus is on the use of an intermediate representation (AUs) for training a classifier for appraisal detection (red boxes)

Classifying facial expressions into emotional appraisal dimensions instead of basic emotions while maintaining human-interpretability is one of the challenges addressed as another main contribution in this thesis.

An approach for automatic emotional appraisal inference combined with an approach for automatic AU intensity estimation (see [98]) is proposed. It is implemented as a two-stage approach (see Figure 4.1, bottom path) and is inspired by the way how humans are thought to recognise internal states of others. Therefore, the detection of facial expressions is separated from the inference of the hidden mental state.

As another contribution of this thesis, an approach for AU intensity estimation was developed (see Section 4.1.2 and is used in the first step. In the second step, the AU intensity observations are used in an Ordinary Least Squares (OLS) regression model to infer the emotional appraisal dimension, namely valence, control and novelty. For a schematic overview see Figure 4.2.

Using a two-stage approach has two important advantages:

- Interpretability
  Using AUs as an intermediate representation in combination with a regression model allows the tracing of the system's decision. The different weights of a regression represent the contribution

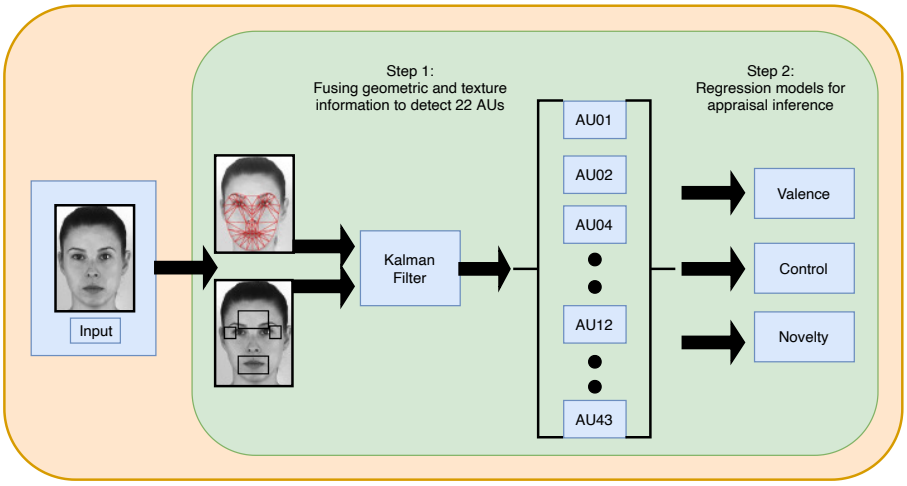

Figure 4.2: Overview of two-step approach for appraisal inference. Images are part of the Actor Study Dataset [96]

of each AU to the detected appraisal, while each AU is defined via FACS, which makes the entire system's decision process understandable to humans. The regression weights could be further used for a comparison with theories of emotion, comparing data-driven approaches with psychological theorems.

- Flexibility
  The generic nature of the approach allows the exchange of approaches in the individual steps. For instance, detected AUs could be used as input for other types of inferences, for example for pain detection (see Section 4.2). In the same way, additional input like vocal expressions or physiological parameters [94] could be added to infer emotional states.

4.1.1 *Domain Knowledge*

Scherer et al. [93] conducted studies using synthesised avatar expressions in a similar way as mentioned in the introduction of this thesis to empirically determine which AUs contribute to the different appraisals. For the three important appraisal dimensions, namely valence, control and novelty, following AUs are important:

- Positive valence:
  AU05, AU06, AU12, AU13, AU14, AU23, AU25, AU27, AU43
- Negative valence:
  AU04, AU09, AU10, AU11, AU14, AU15, AU17, AU20, AU23, AU24

- Lack of control/confusion:
  AU01, AU04, AU16, AU25, AU26
- Novelty:
  AU01, AU02, AU04, AU05, AU07, AU26

Exploiting this domain-specific knowledge, only a subset of the AUs, introduced by the Actor Study Dataset (see Section 3.1) was used for the development of the AU intensity estimation approach. For appraisal inference, these 22 AUs (see Table 4.1) are important and integrated into the system.

Table 4.1: List of the annotated 22 AUs, integrated in the AU detection approach. The first column contains the numerical code and the second column contains the actual name of the AU

AU Code	**AU Name**
AU01	Inner Brow Raiser
AU02	Outer Brow Raiser
AU04	Brow Lowerer
AU05	Upper Lid Raiser
AU06	Cheek Raiser
AU07	Lid Tightener
AU09	Nose Wrinkler
AU10	Upper Lip Raiser
AU11	Nasolabial Furrow Deepener
AU12	Lip Corner Puller
AU13	Cheek Puffer
AU14	Dimpler
AU15	Lip corner Depressor
AU16	Lower Lip Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU23	Lip Tightener
AU24	Lip Pressor
AU25	Lips Part
AU26	Jaw Drop
AU27	Mouth Stretch
AU43	Eyes Closed

4.1.2 *Automatic Action Unit Detection*

This section describes the contributions of this thesis to address the challenge of integrating temporal information into an automatic AU detection system without being limited to a fixed number of information sources.

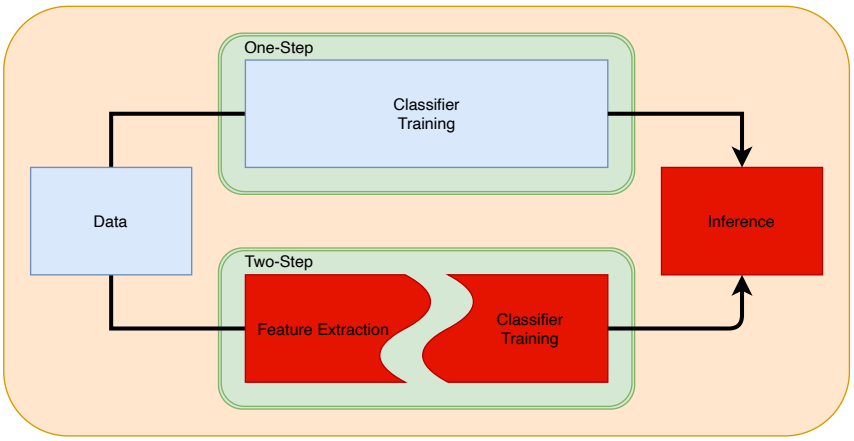

Figure 4.3: Processing pipeline for Computer Vision in Machine Learning. The focus is on the use of extracted features for training a classifier for AU detection (red boxes)

The AU detection approach realises the first stage of the two-stage emotional appraisal inference approach and is another contribution of this thesis. Its development started as a supervised masters' thesis[1], was further developed [39] and finally patented (see Appendix). The AU intensity estimation is implemented as a two-step approach (see [39, 98]) visualised in Figure 4.3. It uses extracted features, namely shape and texture information for inferencing AU intensity values.

4.1.2.1 *Face Models*

The shape information is calculated using two face models, describing a 68 point mesh of a face. This domain-knowledge in comparison with the actual measurement of the corresponding facial feature point locations are crucial and incorporated for the final AU intensity estimation.

FACIAL LANDMARK DETECTION A detection model to identify landmark positions in the current face region is essential for further processing. The detected landmarks cover the same points as introduced in the CK+ [63] database (see Figure 4.4 for example meshes). The first approach for training such a model was a former state-of-

1 Teena Hassan: "Dynamic Facial Expression Estimation by means of Model Fitting", Master's Thesis

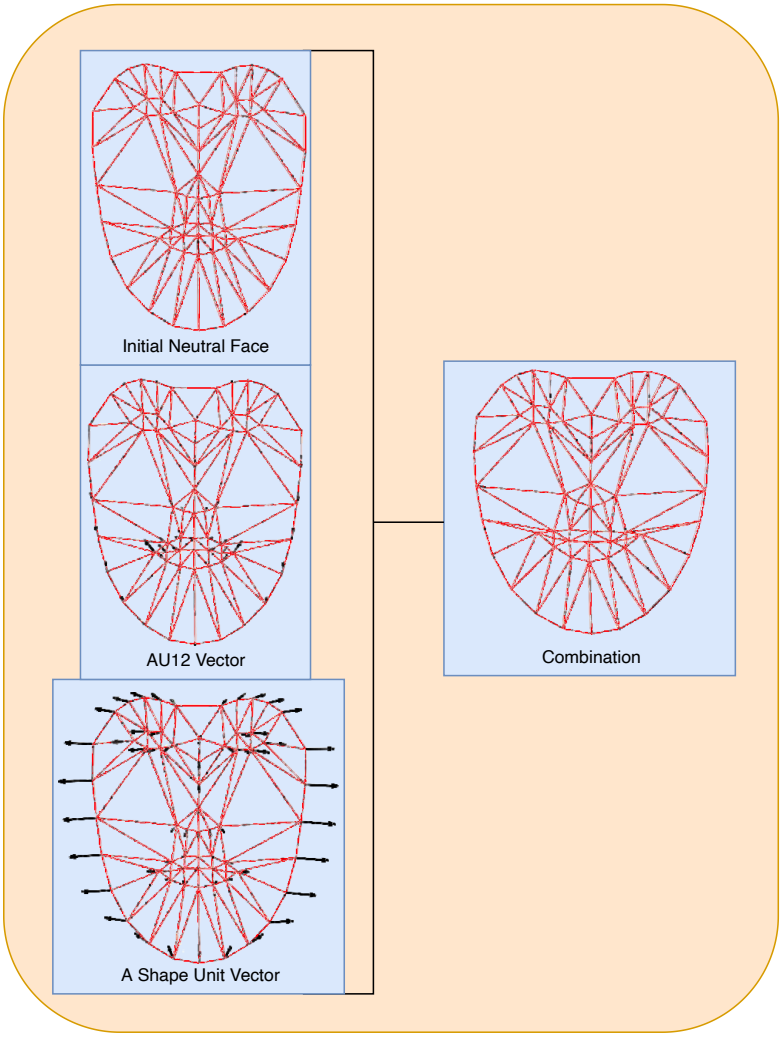

Figure 4.4: Visualisation for the combination of the different face models. The neutral face (top left) shows the start for the AU detection. The fitting to the detected facial feature points in the detected face is done via the combination of the AU models (here, for example AU12, mid left) and the shape unit models (here, an exemplary shape unit, bottom left) to get the final facial shape (mid right)

the-art constraint local model algorithm [88]. In a supervised thesis[2], the parameter of the algorithm with respect to processing time and accuracy were evaluated. While the results for the detected facial landmarks showed a good accuracy, the processing time was quite high. This is due to the fact that the approach must perform several optimisation steps for retrieving the final positions of the points. In this Bachelor's thesis, the possibility of reducing the number of optimisation steps was investigated. However, this reduction had a huge impact on accuracy and was not feasible.

2 Jan Dörntlein: "Weiterentwicklung eines Constrained Local Model Fitting Verfahrens", Bachelor's Thesis

Since the system was developed to be real-time capable another, more recent state-of-the-art approach for face alignment was investigated and implemented [48], which uses regression trees to estimate the facial landmark coordinates. Due to the structure of the new model, the processing time is much shorter, while higher accuracy is achieved than with the previous approach. Different combinations of parameters like width and depth of the tree structure were evaluated. The final facial landmark detection model was trained in such a way that its detection model size (around 100MB) and processing time (20-50 millisecond range on a regular system) is suitable for real-time applications. Changing the parameters to achieve higher model capacity resulted in only minor improvements in accuracy, but made the model significantly larger and longer in execution.

ACTION UNIT MODELS A 3D, parameterized and deformable model of human facial shape (similar to CANDIDE model [1, 87]) is used to describe the facial shape deformations corresponding to the 22 AUs, shown in Table 4.1. The deformation information is stored as a vector which encodes how the shape of the face is changed when an AU is displayed (see Figure 4.4, mid left for an example). In this system, each AU is a 3D 68-point displacement vector and could be seen as the encoding of a prototypical movement of points for the respective AU. The actual values for the displacement can be interpreted as the intensities of the corresponding AU, since the vectors were designed to show the maximum anatomically possible deformation when the value is set to maximum. Those 22 models were obtained from FACSGen [85] by the Swiss Center for Affective Sciences. These deformation information are vectors for a high dimensional face mesh and needed to be scaled down to a 68 point mesh that corresponds to the landmark detection output.

SHAPE UNIT MODELS A similar model of human facial shape is used to describe the person-specific facial shape. This facial shape information is crucial for the AU intensity estimation approach, since it incorporates information about the person-specific facial shape, which is seen as relatively unvarying. This information can be encoded as shape units in contrast to AUs, enabling calibration to a persons's neutral face (see Figure 4.4, bottom left for an example). One can imagine a person whose corners of the mouth point downward even when showing a neutral face. In this case, however, it is not the activation of one or more AUs but an individual facial feature. Shape units and AUs can be treated as antagonists, all information covered by the shape units is not encoded by the AUs, reducing the risk of false intensity estimates and false positive AU detections. These 61 shape units were obtained in a similar fashion as the AU deformation vectors by reducing high dimensional models from Singular Inversion's FaceGen

software [41] to a 68 point mesh that corresponds to the landmark detection output.

4.1.2.2 *Description of the System*

The system consists of three steps: (1) Preprocessing, (2) Extraction of shape and appearance information and (3) dynamic state (AU intensity) estimation (see [98]). Steps 1 and 2 are performed only on every single frame, therefore this information is referred as static information.

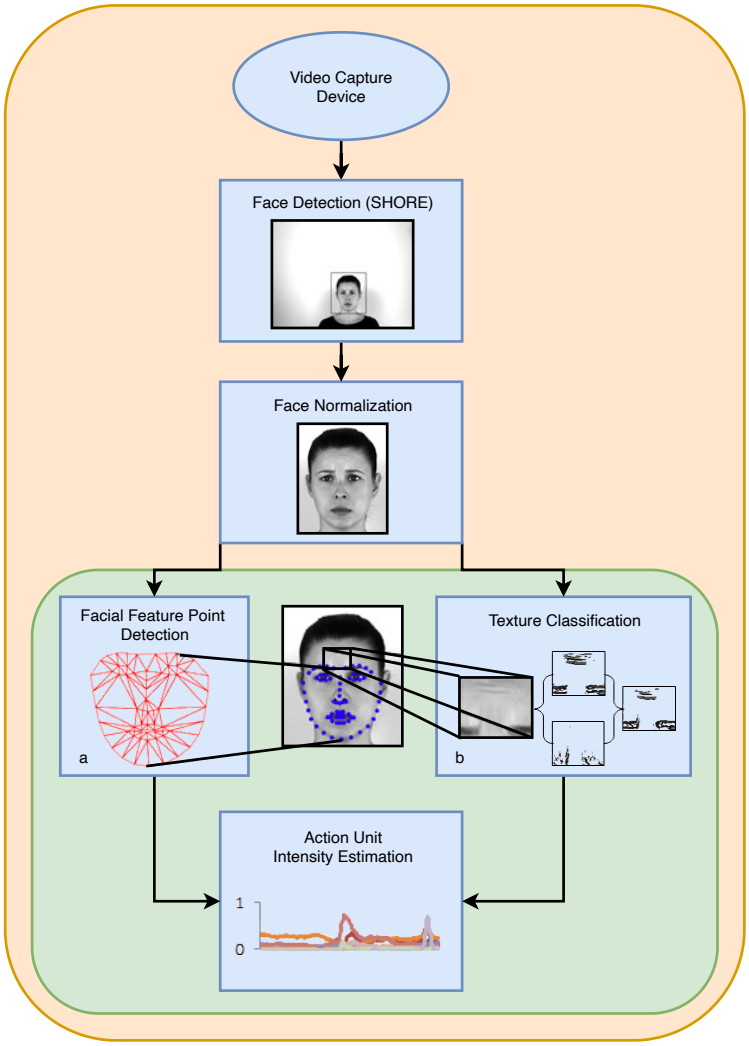

Figure 4.5: Flow chart of the AU intensity estimation approach. Extracted face mesh of the detected face is shown in (a). Two texture features, namely horizontal and vertical edges, as an example are shown in (b). Images are part of the Actor Study Dataset [96]

1. Preprocessing:
   In this step, the video stream provided as input is processed. Each frame is analysed with the software SHORE [50] to detect

a face in the image. If more than one face is detected, the largest face based on the number of pixels it contains is selected.

2. Extraction of shape and appearance information
   The face region, obtained in the previous step is processed further for extracting information about geometric deformations in the facial shape and characteristic textural patterns in the facial appearance.

   The detection model (introduced in Section 4.1.2.1) to locate the 68 landmark positions in the current face is crucial for extracting geometric deformations. The detected positions (see Figure 4.5 a) contain information about the facial shape, facial muscle movement and are very person-specific. The combination of this measurement and the two face models described in Section 4.1.2.1 transform the problem of estimating the intensities of AUs into the problem of estimating the parameters of the face models. The face alignment results were compared with the annotations of the CK+ Dataset (presented in Section 3.2.2). This comparison allows to empirically determine the noise associated with the measurement, which is important for the dynamic state estimation step.

   The appearance information contain information about specific patterns, like wrinkles in the face (see Figure 4.5 b). The location of those wrinkles is often very AU-specific and limited to a particular region of the face. AU04, appearing as a single AU, produces vertical wrinkles between the eye brows. AU01, if appearing as single AU, produces horizontal wrinkles on the forehead. Both AUs in combination produce a different specific pattern, namely a rectangular wrinkle pattern. Detecting these appearance information was done following the same steps visualised in Figure 4.3: Textural features are first extracted using appearance-based features, such as Histograms of Oriented Gradients [12] and Local Binary Patterns [79] on the CK+ dataset (see Section 3.2.2) and parts of the Actor Study Dataset (see Section 3.1, namely actors 1-11). These features are then classified using a SVM, extended by pairwise-coupling [114] to obtain probability estimates which are necessary for the dynamic state estimation step. In this system, appearance information for nine AUs was used, namely AU01, AU02, AU04, AU06, AU09, AU12, AU15, AU17 and AU25.

3. Dynamic State Estimation:
   The last step in the framework is the dynamic state estimation for AU intensity prediction. In this step, AU intensities are predicted using the estimates from the previous frame and dynamic models that incorporate the face and head motion. The prediction is updated using the static information obtained in the previous step. This is possible, since the Kalman filter [45] allows to fuse

measurements from multiple sources as long as each source has a noise model during the update step. An advantage of using a Kalman filter is the possibility to easily integrate new information sources for AUs in the future, since these only require a noise model without having to completely retrain the system as would be the case with other classifiers. The fusion is performed with the aim to minimise the uncertainty of the predictions on the basis of the noise of the measurements: While sources with less noise contribute more to the state update, sources with greater noise contribute less. This results in the filtered AU intensity estimates, where predictions for each AU are based more on geometric information in cases where it's noise is low and based more on appearance information in cases where it has less noise.

The dynamic state estimation, which is handled as a Gaussian state estimation, is implemented as an extended Kalman filter to incorporate the non-linearities in the models. The filters' state vector has to cover all properties of a detected face:

- Head pose parameters:
  For predicting the x, y, z rotation and x, y, z translation of the head, a constant velocity model was used for modelling head motion and head pose changes.
- AU intensity estimates:
  For predicting the AU intensities, a driven mass-spring-damper model is used for modelling the dynamics of AUs. Soft constraints were applied [37] to restrict the values to the range [0,1], as per definition of an AU, where each AU has only one direction of muscular movement.
- Facial shape information:
  For predicting the facial shape information, a constant position model is used for the person-specific facial shape. No constraints were applied, since the movement of the points is allowed in both directions, proving adaptability to for example wider or narrower faces.

For more information about the different models, see patent in the appendix.

4.1.2.3 *Results*

For benchmarking the proposed AU intensity estimation approach, results are compared with OpenFace (introduced in Section 3.1). Table 4.2 provides an overview of the classification performance using AUC as evaluation metric on parts of the Actor Study Dataset not used for training the system.

A comparison of the mean AUC values shows, that both approaches seem to be equal in their performance measurement. In order to

Table 4.2: Comparison of AUC values between proposed AU intensity estimation approach and OpenFace based on common detected AUs for Actor Study Dataset

	Proposed Approach	**OpenFace**
AU01	.82	.63
AU02	.70	.61
AU04	.58	.75
AU05	.85	.59
AU06	.78	.93
AU07	.63	.83
AU09	.67	.62
AU10	.60	.81
AU12	.88	.92
AU14	.79	.77
AU15	.65	.62
AU17	.81	.69
AU20	.58	.56
AU23	.70	.51
AU25	.89	.73
AU26	.89	.74
Average	**.74**	**.71**

implement this approach in the big picture of automatic emotion and appraisal inference a closer look at the requirements is necessary. In Section 4.1 the important AUs for appraisal recognition are pointed out. OpenFace provides only a few of the important AUs, determined both theoretically and data-driven. Especially in the important dimensions for use cases like market research, namely positive valence and novelty, the proposed approach outperforms OpenFace.

- For positive valence (important AUs: AU05, AU06, AU13, AU14, AU25, AU27, AU43), the proposed approach outperforms OpenFace in 3 of 4 AUs, providing two AUs more.
- For novelty (important AUs: AU01, AU02, AU24, AU25), this approach outperforms OpenFace in 3 of 3 AUs, providing one AU more.

The proposed approach performs an AU calibration online, i.e. while processing a video stream. This allows use in real-time scenarios such as market research applications and, in perspective, pain assessment (presented in Section 4.2). OpenFace, in contrast, performs a calibration

after processing the input video by estimating the person's neutral expression, before providing AU detection results.

In this section, a two-stage approach was proposed for automatic detection of AU intensities for 22 AUs based on domain-specific knowledge about emotional appraisals. A Gaussian state estimation approach was used to incorporate the temporal dynamics of AUs and can be extended with additional information sources without re-training the system.

ACTION UNIT DETECTION USING DEEP LEARNING METHODS In a supervised thesis[3], end-to-end training of a classifier for AU detection was evaluated. In contrast to the approach presented in this section, the classifier did not use an intermediate representation but learned directly on provided input images. This corresponds to the top path, shown in Figure 4.3. The results were promising, but only a subset of the provided AUs by the system described in this section could be learned, showing the disadvantage of most DL methods: the need of a large amount of labeled training data. In other DL approaches for AU detection, the lack of training data was tried to be mitigated. This was done by combining datasets where possible and enhancing algorithms for a better handling in case of imbalanced classes. These approaches also used the Actor Study Dataset and showed promising results [80, 81, 83]. However, a drawback of these approaches is that additional sources of information cannot be easily integrated because these trained models are static with respect to the expected input.

4.1.3 *Appraisal Inference*

The inference of emotional appraisals is the second stage of the proposed two-stage approach.

4.1.3.1 *Generalisation of the Action Unit Intensity Estimation Approach*

The AU intensity estimation approach proposed in Section 4.1.2 was benchmarked on the Proprietary Market-Research Database (see Section 3.2.1) before its' output was used to train the appraisal inference regression. This was done to obtain an impression of the generalisability and the performance in respect to spontaneous facial expressions in an applied market-research setting. The results are shown in Table 4.3.

The results are promising although the most AUC values are lower than the ones in Table 4.2. This was expected, since the Proprietary Market-Research Database consists of webcam videos with different lighting conditions and different distances between the subjects and

3 Jeanette Lobentanzer: "Performance of Deep Neural Networks for Facial Action Unit Detection in Cross-Dataset Evaluation", Master's Thesis

Table 4.3: AUC values and correlations between detected AU intensities and annotations for Proprietary Market-Research Database

	AUC	**Correlations**
AU01	.73	.39
AU02	.77	.32
AU04	.76	.43
AU05	.71	.04
AU06	.68	.27
AU07	.59	.10
AU09	.79	.07
AU10	.79	.05
AU11	.64	.07
AU12	.74	.39
AU13	.68	.10
AU14	.58	.06
AU15	.57	.32
AU16	.57	.01
AU17	.76	.26
AU20	.52	-.03
AU23	.54	-.04
AU24	.46	.18
AU25	.88	.65
AU26	.73	.33
AU27	.95	.13
AU43	.64	.08
Average	**.69**	**.19**

the camera. In contrast to the Actor Study Dataset, the facial expressions are spontaneous rather than posed, which results in far more subtle facial expressions. Correlations were introduced as an additional performance metric. While computing the AUC is a common metric for classification tasks, correlations are often used for regression tasks. The more similar the detected AU intensities and the annotated temporal AU courses, the higher are the correlation values.

4.1.3.2 *Comparison of Approaches for Appraisal Inference*

In a further step the Proprietary Market-Research Database was used to train the actual classifier, i.e. the mapping between detected AUs and the annotated appraisal dimensions. Choosing the most appropriate classifier is not a simple matter, as it is usually a trade-off between several objectives. Methods that produce a decision that can be interpreted by humans often perform worse than approaches that have no such constraint. For the training of the classifier different methods, namely OLS regression, SVR and Multilayer Perceptron were compared. This procedure was chosen because, although the focus of this work is on interpretable models (e.g. OLS regression), their performance should be compared to other approaches that do not have this limitation of human-interpretability and are used in literature (e.g. SVR, Multilayer Perceptron).

Optimising all approaches would be extremely complex and does not correspond to the actual goals of this thesis. The parameters of the SVR and the Multilayer Perceptron were tested for intervals of common values for that type of problem category[4][5].

Table 4.4: Comparison of different methods for appraisal inference: OLS regression was compared to SVR and Multilayer Perceptron with correlation as performance metric

	OLS regression	**SVR**	**Multilayer Perceptron**
Valence	.73	.64	.7
Control	.42	.42	.25
Novelty	.26	.37	.11

Table 4.4 shows the results of each method for each emotional appraisal dimension. All approaches yield similar results for the valence dimension with at most marginal differences. For the control dimension, OLS regression and SVR were on par and outperformed the Multilayer Perceptron. Only for the novelty dimension does the SVR yield an around .1 better result compared to the OLS regression. This performance difference is accepted here for the advantage of interpretability and therefore OLS regression was also chosen for the novelty dimension.

4 SVR: (RBF-kernel, gamma ($[10^{-6}; 10^{-2}]$) and C ($[10^{0}; 10^{2}]$)

5 Multilayer Perceptron: Sigmoid Activation Function, 2 layers with size of 12 and 6 were used (Different number of layers and different layers sizes did not change the results significantly)

4.1.3.3 *Ordinary Least Squares Regression Model*

Different OLS model specifications were compared to find the best configuration of AUs for predicting each appraisal dimension (see [98]). As performance criterion, the model with the highest correlation of predicted appraisal intensities using AU data and the actual annotated appraisal intensity values was chosen. For better comparison of the OLS model with findings from literature afterwards, the valence dimension was divided into positive and negative valence.

The definition of the OLS model used for appraisal inference is:

$$App_{i,t} = \alpha + \beta_{i,j} sAU^{T}_{i,j,t} + \delta_{i,l} Val^{T}_{i,l,t} + \epsilon_{i,t}$$

$sAU_{i,j,t}$ is a vector and contains j single AUs which were identified to contribute to the appraisal dimension i. $Val_{i,l,t}$ is the positive and negative valence classification from SHORE enhanced with a valence classification module [27]. For each appraisal dimension i the model was optimised separately. Training of the final OLS regression model is done in two steps:

1. Variables were selected using a step-wise regression approach with forward adding and backward elimination of variables. The Akaike Information Criterion was used to determine the optimal set of variables. It is a common selection criterion in economic research, where low values encode the best.

2. A ten-fold cross validation was used to evaluate the different models per appraisal dimension, using unique assignment of respondents to a certain subset. The correlation values between predicted appraisal value and the ground truth was computed for each of the ten iterations and averaged across the ten folds.

Finally, the model with the highest average correlation was selected. Table 4.5 shows the coefficient values for the final model, trained on the AU predictions from the approach presented in Section 4.1.2.

4.1.3.4 *Evaluation*

For the evaluation of the automatic appraisal inference approach, the OLS regression was applied two times:

1. using the annotated AU ground truth annotation to predict the appraisal ground truth ratings

2. using the actual output of the automatic AU intensity estimation approach to predict the appraisal ground truth ratings

The first evaluation can be seen as a benchmark for the second one, since it provides the theoretically best result possible. Table 4.6 shows the results when applying the OLS regression model shown in Table 4.5 to both cases. Since the positive and negative valence predictions were inferred separately, their predictions needed to be combined

Table 4.5: Overview of regression coefficients for the final appraisal inference model. Val_{pos} and Val_{neg} denote the output of the valence detection model of SHORE. Please note that these values are up to 100 times higher than the AU intensity estimates

	Positive Valence	**Negative Valence**	**Lack of Control**	**Novelty**
Intercept	.0649	.0249	.018	.0121
Val_{pos}	.0015	-.0002	-.0004	-.0003
Val_{neg}	-.0001	.0007	.0004	
AU01	-.0947	.1553	.0525	.4242
AU02	-.0942		.1924	.3792
AU04	-.0384	.2267	.4144	
AU05	.1451			
AU06	.1522	-.049	-.0555	
AU07	-.1303	.364	.1504	-.1425
AU09		-.4251	-.4083	
AU10		.0878	.1007	.0857
AU11				.0696
AU12	.0571	-.0345		
AU13	.3838	.0576	.0804	
AU14	.1065	-.0463	-.0316	
AU15	-.0625	.1297	.0757	
AU16	-.4074	.238	.1288	.0911
AU17		.0835		
AU20	.1737		-.1132	
AU23		-.0627	-.0623	
AU24		-.5234	-.6012	.356
AU25	.1877	-.1017	-.0555	.1069
AU26	.3712	-.1459		-.2587
AU27	.2766	.1198		
AU43	.1344	.0397	.0528	-.0447

for the final results. This was done by choosing the maximum value from either the positive or the negative valence prediction. Only few cases were found for positive control and negative novelty. To reduce the complexity of the prediction problem, control and novelty were therefore transformed to unipolar dimension, i.e. only negative control (lack of control) and positive novelty were considered.

Table 4.6: Correlation values for the different appraisal dimensions. Appraisal inference was done using (1) annotated AU values (second column) and (2) the estimated AU intensity values obtained by the automatic AU intensity estimation approach (third column) as input

Appraisal dimension	**Correlations between appraisal inference and ground truth using annotated AU values**	**Correlations between appraisal inference an ground truth using automatic AU intensity estimation approach**
Valence	.72	.72
Control	.59	.56
Novelty	.51	.39

The correlation values are equal for both models. This is noteworthy, since classifier outputs from the appearance-based detection system [27] were integrated into the regression model. The correlation coefficients presented in Table 4.5 show that the detected AUs still contribute and enhance this system's classifiers. Only for novelty, predictions from the AU intensity estimation approach are behind those from ground truth. A possible reason could be that responses in the novelty dimension are often subtle and short (see [29]), which can be a challenging for detection systems.

As the last part in this section, the data-driven determined important AUs per dimension (shown in Table 4.5) are compared with the findings in literature [93]. Note that to maintain readability, only coefficients greater than 0.1 were considered (see [98]).

Table 4.7 shows a great overlap between the AUs from the regression model and the theoretically determined AUs for the positive valence dimension. However, this overlap is smaller for the other dimension, namely negative valence, lack of control and novelty. One explanation could be that the respondents in the Proprietary Market-Research Database were recorded while passively watching commercials. This situation is more likely to expose only a limited range of observable

Table 4.7: Comparison of the data-driven determined AUs with the findings in literature. For better clarity, common AUs are bolded

Appraisal Dimension	Data-driven determined AUs	AUs from literature
Positive valence	**AU05**, **AU06**, **AU13**, **AU14**, AU20, **AU25**, AU26, **AU27**, **AU43**	**AU05**, **AU06**, AU12, **AU13**, **AU14**, AU23, **AU25**, **AU27**, **AU43**
Negative valence	AU01, **AU04**, AU07, **AU15**, AU16, AU27	**AU04**, AU09, AU10, AU11, AU14, **AU15**, AU17, AU20, AU23, AU24
Lack of control/ confusion	AU02, **AU04**, AU07, AU10, **AU16**	AU01, **AU04**, **AU16**, AU25, AU26
Novelty	**AU01**, **AU02**, AU24, AU25	**AU01**, **AU02**, AU04, AU05, AU07, AU26

expressions. Particularly strong responses like being overwhelmed, angry and fearful are missing. For this dimensions, more overlap in "signature" AUs is found:

- Negative valence: AU04 (lowered brows) and AU15 (depressed lip corners)
- Lack of control: AU04 (lowered brows) and AU16 (depressed lower lip)
- Novelty: AU01 (inner brows raised) and AU02 (outer brows raised)

Albeit this example is very limited, no evidence, apart from the positive valence dimension, for the contribution of the more subtle AUs from literature could be found.

This section described the contributions of this thesis to address the challenge of human-interpretable emotional appraisal inference by proposing a new two-stage approach. Domain-specific knowledge in the form of important AUs and their contribution to each appraisal dimension was integrated into the classification in both steps of the approach. The use of AUs as an intermediate representation and OLS regression as the final classifier allowed humans to understand the decisions of this system.

4.2 AUTOMATIC PAIN RECOGNITION

The challenge of training a system that incorporates knowledge about facial pain expressions and produces a set of rules to recognise pain in a human comprehensible way is addressed in this section.

For the recognition of pain a modified grammar inference approach (see [100], [95]) is used on the pain sequence dataset from Section 3.2.3. These AU sequences represent an intermediate representation and are used for creating a classifier, in this case a grammar. This procedure follows the workflow shown in Figure 4.6 (red boxes).

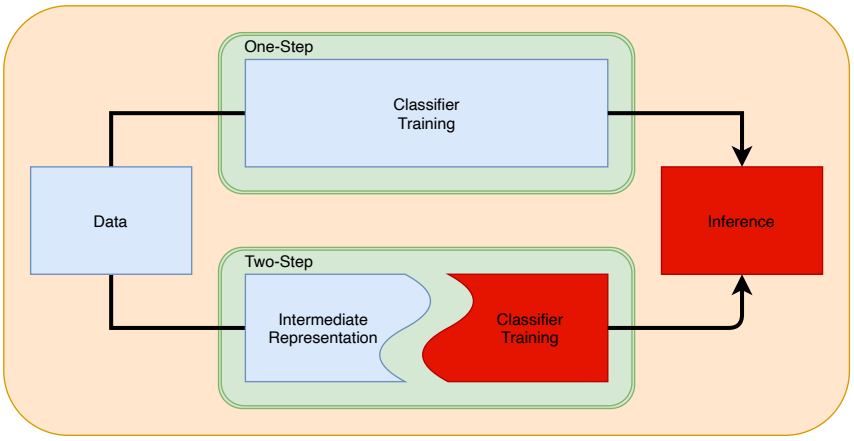

Figure 4.6: Processing pipeline for Computer Vision in Machine Learning. The focus is on the use of an intermediate representation (AUs) for training a classifier (red boxes)

4.2.1 *Definition of Grammar*

A grammar, from a formal language point of view, is described as the production of words by rules. For understanding the formal definition of the rule set, the most important definitions of the buildings blocks are given below:

- A word is a concatenation of symbols
- An alphabet predefines all symbols which are allowed to appear in words. Those symbols are called terminals
- Non-terminals can be seen as variables in the production rules and describe how words over the alphabet can be generated
- Production rules are defined as a finite set
- A grammar starts with a distinguished non-terminal called start symbol

Every production rule has the form: $A \rightarrow \alpha$ where A is a single non-terminal symbol, and α is a string of terminals and/or non-terminals (additionally, α can be empty). A formal grammar is called "context free" when its production rules can be applied no matter of the context of a non-terminal. The single non-terminal on the left hand side can always be replaced by the right hand side, regardless of which symbols surround it (in contrast to context-sensitive grammars). In the following we will focus on context free grammars. Usually grammars are used to characterise classes of languages, deriving for example rule sets for accepting or generating those languages. In this machine learning context the inverse procedure is pursued: a grammar should be infered from sample words.

4.2.2 *Grammar Extraction from Sequences*

For the extraction of a grammar for pain sequences, a method called Alignment-based Learning (ABL) is used [107, 119, 127]. This approach is an unsupervised grammar inference algorithm creating a context-free grammar from input words. ABL extracts a set of rules for the given input in two phases: "alignment phase" and "selection learning phase".

1. Alignment Phase:
   The input words (actually the AU sequences) are pair-wisely compared to align identical symbols between the words. The differing symbols are seen as candidates for possible interchanging. Consider following two AU sequences as an example:

   AU01 AU02 AU04 AU06
   AU01 AU02 AU12 AU06

   Both words are identical besides the AU04 in the first word and AU12 in the second. These symbols are therefore recognised as to create a potential constituent. This results in an annotation for both words:

   AU01 AU02 $(\text{AU04})_1$ AU06
   AU01 AU02 $(\text{AU12})_1$ AU06

   This process is repeated with the original words and the newly annotated words. It is important to also consider the newly annotated words, since this allows for nesting of the constituents. Consider the case that the second word is compared to a new word like, for example "AU01 AU02 AU10", new constituents are created:

   AU01 AU02 $((\text{AU04})_1$ AU06$)_2$
   AU01 AU02 $((\text{AU12})_1$ AU06$)_2$
   AU01 AU02 $(\text{AU10})_2$

2. Selection Learning Phase:
   In this phase, the results from the predecessing step are reduced before the transfer from the constituents to the final grammar is applied. This is necessary, because the proposed alignments might not be unique, i.e. overlapping candidates for constituents might exist. To perform such a reduction, frequency information is used by computing the possibility of each possible constituent. Finally, for each word the constituent combination with the highest probability is kept.

For the final inference of the grammar, each constituent receives an unused non-terminal and for each possible replacement in the constituent a production rule ist created. Applied to the example above, the following set of *non-terminal→ production rule* is extracted:

- nt1→ AU04
- nt1→ AU12
- nt2→ nt1 AU06
- nt2→ AU10

4.2.3 *Performance Optimisation through Domain Knowledge*

ABL applied on the pain sequence dataset produces a large amount of rules and seems to incorporate individual specific information. Following the goal to extract typical facial expressions with a good generalisation performance, changes and extensions were made to both: (1) the dataset and (2) the algorithm.

1. The AU sequences in the dataset have been transferred into a more compressed representation. Pursuing the goal of pain recognition, only the pain-relevant [52] AUs, namely AU04, AU06, AU07, AU09 and AU10 were kept. The other, non pain-relevant AUs were replaced by a wildcard character, namely "I". Exploiting this domain-specific knowledge lead to a change in the distribution of the AUs and AU compounds (see [100]): The maximal number of distinct AUs and AU compounds per sequence decreased from 13 to 9 and the mean number of distinct AUs and AU compounds decreased from 3.54 to 2.69. The aggregation of different entities to a more generic representation in a similar way is a well-known method used for example in natural language processing [69].

2. ABL was enhanced to perform a heuristic rule extraction. The selection learning step was extended to obtain a frequency distribution of the applied production rules. Based on this, the probability of each production rule is the relative frequency of

the replacement within the constituent. For detailed explanation on how to handle special cases like production rules with a probability of 1, please refer to [100]. This procedure allows the introduction of a threshold, which controls the amount of production rules of the grammar. Each production rule used in the derivation is kept as long as a word can be derived from the start symbol with a sequence of rules with a probability higher than the threshold.

Applied on the example above, the rules nt1→ AU04 and nt1→ AU12 get the probability .5 assigned, since each of them occurs once in the first constituent. The rule nt2→ nt1 AU06 has the probability of .66 since the replacement occurs twice compared to nt2→ AU10 which occurs only once and gets a probability of .33 assigned.

The extension of the selection learning phase, the extraction of the grammar up to this step, and the application to the pain sequence dataset are the main contributions of this work. In the last section the results of the evaluation of this approach is discussed.

4.2.4 *Results*

Evaluating the output of an unsupervised empirical grammar inference is different [120] compared to the evaluation metrics used in Sections 3.1, 4.1 and 4.1.2. To benchmark the performance of the proposed method, the ability to classify a given input whether it belongs to the language of the induced grammar or not is tested (as proposed by [120]). Given words from the same language, the induced grammar should be able to derive all of them. Words not included in the language should, in contrast, not be derivable. For performance evaluation, a ten-fold cross validation (like in Section 4.1) was conducted. For generating the necessary negative examples (i.e. words that are not part of the language), for each positive example a random negative example was created. Properties like equal distribution of word length and terminal frequency were ensured and applied to each fold.

For final performance measurement the mean of precision, recall and accuracy over all ten folds was calculated. These metrics are calculated in the usual way but with following definitions:

- True positives are words from the positive example set, which can be derived by the induced grammar
- False negatives are words from the positive example set, which cannot be derived by the induced grammar
- False positives are words from the negative example set, which can be derived by the induced grammar
- True negatives are words from the negative example set, which cannot be derived by the induced grammar

After applying the modifications introduced in Section 4.2.3 the benchmark values on the pain sequence dataset are as follows. The mean precision is .50 with a deviation of .02, which is an increase by over 25% compared to the results with the full alphabet. The mean recall is .91 with a deviation of .04 and decreased by just 2.5%. The mean accuracy is .51 with a deviation of .03 and decreased by just 3% compared to the results using the full alphabet. For all evaluation results and an excerpt of the extracted grammar please refer to [100]. Despite the set having an amount of 897 production rules, the actual derivation for each example is tractable and understandable by humans. Therefore, inferencing a grammar using domain-specific knowledge, namely AUs and their pain-relevant subset provides an advantage compared to black box methods especially in such a sensitive field like pain recognition.

This section described the contribution of this thesis to address the challenge of creating human-interpretable pain classifiers using a rule extraction approach. It was shown that the incorporated domain-specific knowledge in the form of AUs and their occurrence in facial pain expressions improves the grammar inference method. The applied rules that have led to a decision of the system are traceable and comprehensible for humans and provide explanations for experts and medical staff.

PAIN RECOGNITION USING DEEP LEARNING METHODS Section 2.2.1 provided a review of methods for automatic pain detection. Many methods use end-to-end learning to classify images of facial expressions of pain, which is illustrated as the top path in Figure 4.6. Those approaches do not use an intermediate representation before the actual classifier is trained and are mostly treated as black boxes. In a co-supervised thesis[6], a DNN was trained to classify sequences of images of facial expressions of pain. Although the results seemed very promising, the lack of available data for training such a classifier, as well as the non-explainability of the approach, made it more suitable for other, not so sensitive, research areas.

6 Johannes Rabold: "Exploring Deep Learning for Image Data — Effort and Performance For Learning a Classifier for Facial Expressions of Pain", Bachelor's Thesis

5 CONCLUSION AND OUTLOOK

5.1 EMOTIONAL APPRAISAL INFERENCE

As an important contribution of this thesis, a new high quality database, namely the Actor Study Dataset was curated and published and thus provides an important resource for the scientific research community. It includes frame-synchronised penta-views of actors enacting different AUs, AU combinations, emotion and appraisal scenarios. All images have a high resolution and are FACS annotated by professional FACS coders. Benchmark results of state-of-the-art approaches for the frontal and for all views are provided. This makes the Actor Study Dataset an important contribution to scientific community: It mitigates the problem of data scarcity by providing training data for the development of AU detection as well as emotion recognition approaches. Additionally, the Actor Study Dataset is a valuable database for benchmarking new as well as existing approaches for automatic AU recognition, and for emotional phenomena research.

In this thesis, a novel, hybrid and adaptable approach for the validated inference of emotional appraisals was presented as another important contribution. Implemented in this framework, a method for automatic action unit detection incorporating domain-specific knowledge was proposed and a comparison with a state-of-the-art approach provided. In contrast to other static approaches, the AU detection method uses a modified Kalman filter to model the temporal characteristics of AUs. Furthermore, it allows adding new information sources without re-training the whole system. This makes it more flexible in dynamic multimodal environments than other static approaches. The detected AU intensities provide the input for a regression approach to finally infer the three appraisal dimensions, namely valence, control and novelty. A comparison between different methods showed that the chosen regression model is on par and at the same time provides interpretability of the model's decision in the way that a human observer can see the influence of the different detected AU intensities. Evaluation shows that this entire approach for automatic emotional appraisal inference, implementing domain-specific knowledge in form of selected AUs as an intermediate representation performs well while maintaining a human-interpretable decision model. The findings of this data-driven approach could further be validated to expert knowledge and findings in literature.

The use of the three appraisal dimensions make the classifications of the system more meaningful compared to the use of basic emotions.

Furthermore, the fact that the decisions of the presented approach are comprehensible to humans sets it apart from other approaches and can make it a valuable contribution to emotion research.

5.2 PAIN RECOGNITION

In the context of this thesis, an approach for inferencing a context-free grammar and its application to AU sequences of pain was proposed. Grammar inference approaches extract an interpretable model out of its training data, although the number of rules might be high. To diminish this effect, the used approach was extended to calculate a frequency distribution of applied production rules. That information was used to compute the probability of each production rule wich allows the introduction of a threshold to control the amount of production rules adopted from the grammar. This grammar, modified by implementing domain-specific knowledge is able to differentiate pain from not pain sequences based on the pain sequence dataset. Thus, an approach is provided that generates a set of rules for pain recognition. Since this system can be validated by humans, it can make an important contribution to pain research and to medical staff.

5.3 FUTURE RESEARCH

Human-interpretability was always an aspect discussed in all approaches in this thesis. For AI research, frameworks like "Responsible AI" are now coming to the fore, postulating properties like fairness, interpretability, privacy and security of applications. In [97], the author did a call to action and discussed the necessity of joint research in the fields of Explainable AI and Uncertainty Quantification, to enhance trust and trustability in AI approaches. The more we use machine learning approaches to solve even sensitive tasks, the more important it is to know what the system is doing. This awareness of the system includes, in addition to the explanation for a decision, the certainty of the system about its output. The author pursues the combination of these two research fields as the coordinator for "Explainable Learning", one of the main research focus areas at the ADA Lovelace Center for Analytics, Data and Applications[1]. In this setting, the author brings those researchers working on explainability of AI approaches together with those working on Uncertainty Quantification. To spread awareness to this topic, the author was also responsible for a tutorial about Bayesian Deep learning, a research field for Uncertainty Quantification at KI2020[2].

1 https://www.scs.fraunhofer.de/en/focus-projects/ada-center.html
2 43rd German Conference on Artificial Intelligence, September 21–25, 2020: Bamberg, Germany

The methods proposed in this work each cover one aspect of the call to action mentioned above. The presented approach for automatic action unit detection in Section 4.1.2 deals with uncertainties internally, since it is based on a Kalman filter, although without explaining the source of the uncertainty to the user. The grammar inference approach, proposed in Section 4.2 produces an explanation of its decision regarding the presence of pain or not that is understandable to humans. This approach does not consider any uncertainty information of the input and handles each part of the input as entirely trustworthy.

In future research it is thinkable to combine both approaches and extend the grammar extraction by incorporating uncertainty information obtained by the Kalman filter which is part of the AU detection approach. An induced grammar could then incorporate the uncertainties of the AU detection, caused for example by a partially covered face, and could adapt its production rules accordingly, resulting in a more differentiated and probabilistic classification of the input data by maintaining its explanatory component. These modifications could lead to a more natural understanding of the power and weaknesses of future approaches.

Part II

APPENDIX

A

PUBLICATIONS

FULL REFERENCES OF PAPERS

- **Dominik Seuss**, Teena Hassan, Anja Dieckmann, Matthias Unfried, Klaus R. Scherer, Marcello Mortillaro, and Jens Garbas. "Automatic Estimation of Action Unit Intensities and Inference of Emotional Appraisals." in *IEEE Transactions on Affective Computing (2021)*, doi: 10.1109/TAFFC.2021.3077590.

- **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.

- 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.

- 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.

- Michael Siebers, Ute Schmid, **Dominik Seuß**, Miriam Kunz, and Stephan 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.

- 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.

B

PREPRINTS

FULL REFERENCES OF PAPERS

- **Dominik Seuss**. "Bridging the Gap Between Explainable AI and Uncertainty Quantification to Enhance Trustability.". 2021. arXiv: 2105.11828 [cs.AI].

BRIDGING THE GAP BETWEEN EXPLAINABLE AI AND UNCERTAINTY QUANTIFICATION TO ENHANCE TRUSTABILITY

A Call to Action

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

Abstract

After the tremendous advances of deep learning and other AI methods, more attention is flowing into other properties of modern approaches, such as interpretability, fairness, etc. combined in frameworks like Responsible AI. Two research directions, namely Explainable AI and Uncertainty Quantification are becoming more and more important, but have been so far never combined and jointly explored. In this paper, I show how both research areas provide potential for combination, why more research should be done in this direction and how this would lead to an increase in trustability in AI systems.

Keywords: Responsible AI, Trust, Explainable AI, Uncertainty Quantification

1 Introduction

Deep Learning methods have achieved tremendous success in almost all disciplines in the field of Artificial Intelligence and Machine Learning. Now, there are signs of a shift away from the goal of improving the recognition performance of the models alone. Frameworks like "Responsible AI" are now coming to the fore, postulating properties like fairness, interpretability, privacy and security of AI applications. In this paper, I will focus on the notion of interpretability in particular, expand interpretability to include "Trust" and show why more research should go into combining the methods of Explainable AI and Uncertainty Quantification. To this end, I will take up these two separate research branches, point out their methods, and show their relatedness in contributing towards trustability - the quality or state of being trustable. There are already research fields such as "Planning under Uncertainty" and "Decision making under Uncertainty" which combine known methods with uncertainty but these either

assume that uncertainties already exist or refer to methods for quantifying uncertainty [5, 7] without having properties such as trustability towards humans as a goal.

No one would question that explainability of approaches is conducive to building trust. But uncertainties and awareness of them also have an impact on people. A number of behavioral and electrophysiological studies indicate "[...] a strong relationship between uncertainty and a key component of cognitive control - outcome monitoring. In particular, it appears that highly uncertain environments tend to increase the recruitment of monitoring processes" [21]. So the more uncertain a situation is for us, the more we try to monitor the outcome. Applied to AI systems, this would mean a stronger need for explanation and predictability of the outcome. Both properties that can be achieved through a combination of Explainable AI and Uncertainty Quantification methods. As the coordinator for "Explainable Learning", one of the main research focus areas at the ADA Lovelace Center for Analytics, Data and Applications[1], I bring those researchers working on explainability of AI approaches together with those working on Uncertainty Quantification.

In the next section, both branches of research with their definitions and methods will be introduced. In the third section, possible combinations of the two research directions and the importance of their combination will be discussed. The outlook will be a suggestion on how to handle trust in the future.

2 Overview of Explainable AI and Uncertainty Quantification

2.1 Explainable AI

There is no clear definition of Explainable AI. The Defense Advanced Research Projects Agency (DARPA, [12]) formulates the goals as "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners". Another definition comes from the organisers of xML Challenge, FICO [9], who see XAI as "an innovation towards opening up the black-box of ML" and as "a challenge to create models and techniques that are both accurate and provide good trustworthy explanation that will satisfy customers' needs". Both definitions already contain the term "trust" without going further into it. An overview of the individual methods without concrete approaches can be seen in Figure 1 (for an in-depth survey please refer to [2]).

One way to group Explainable AI methods is to divide them into three categories

[1] https://www.scs.fraunhofer.de/en/focus-projects/ada-center.html

according to their chronological order of use: "Pre-modelling explainability", "Explainable Modelling" and "Post-modelling Explainability" [15]. We will leave out the "Pre-modelling explainability" category for this paper, as this deals with understanding the data itself before actually training a model.

- Explainable modelling
  For explainable modelling, a component for comprehensibility is already implemented in the model during training. Two types of approaches can be distinguished: "Representation Explaining" and "Explanation Producing Systems" (see lower block in Figure 1). An example for Representation Explaining are Concept Activation Vectors [16], which can be used to show for example how sensitive a prediction of "zebra" is to the presence of stripes. Approaches like Beta VAE [13], from the field of Explanation Producing Systems, try to introduce interpretable latent representations, which means that the latent representations contain human-interpretable information already.

- Post-Explainability
  In Post-Explainability, after the decision of the model, the processing is examined up to the decision (see upper block in Figure 1). The different approaches use different types of and representations for explanations. An example for a graphical representation is "Layerwise Relevance Propagation" [18] and is often used with images and shows the user which pixels of the input made what contribution to the decision. An example for a different type of representation is another approach named ANN-DT [22]. It extracts decision trees from neural networks to make the underlying decision processes more comprehensible for humans.

2.2 Uncertainty Quantification

One comprehensive description [20] of Uncertainty Quantification (UQ) from the field of applied mathematics is: "UQ involves the quantitative characterisation and management of uncertainty in a broad range of applications. It employs both computational models and observational data, together with theoretical analysis. UQ encompasses many different tasks, including uncertainty propagation, sensitivity analysis, statistical inference and model calibration, decision making under uncertainty, experimental design, and model validation. UQ thus draws upon many foundational ideas and techniques in applied mathematics and statistics (e.g., approximation theory, error estimation, stochastic modelling, and Monte Carlo methods) but focuses these techniques on

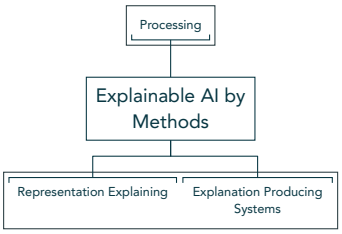

Figure 1: Overview of Explainable AI by Methods. All three categories (Processing, Representation Explaining, and Explanation Producing Systems) can be further divided into post-explainability (upper box) and explainable modelling (lower box).

complex models, for instance, of physical or socio-technical systems that are primarily accessible through computational simulation. [...]"

This definition also holds for the field of machine learning, since we do almost always deal with complex models. There are two main sources of uncertainty, aleatoric uncertainty and epistemic uncertainty [14]. The former is caused by noise in data or labels, and the latter is caused by data sparsity or out of data distribution. The scope of this paper is not to give a comprehensive overview of uncertainty quantification methods, but to provide a categorisation of selected approaches (see Figure 2) to show their combinability with Explainable AI methods. For a more extensive treatise on uncertainty quantification methods, see [1]. In AI research, Uncertainty Quantification can be used to pursue two goals. First, to calibrate neural networks so that the output confidence reflects the empirical accuracy, and second, for out-of-distribution detection, i.e., to detect whether a new classification example really corresponds to the distribution with which the network was trained.

Uncertainty quantification methods can be divided into two types, discriminative methods and generative methods. Generative methods are used, among other things, to perform out-of-distribution detection [11], or to reconstruct data, such as image reconstruction from noisy and incomplete images [6]. Discriminative methods - the focus of this work - are used to perform e.g. classification, i.e. to assign an input to one or more or no output classes. Here, one can distinguish between model-agnostic approaches, Bayesian and Non-Bayesian methods.

- Model-agnostic methods

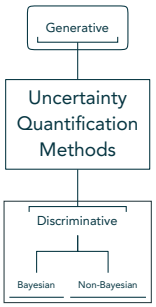

Figure 2: Overview of Uncertainty Quantification Methods in Machine Learning

This includes approaches that are independent of the method used. They generate e.g. through data augmentation, better calibrated DNNs (Mixup [23]) or alleviate the over-confidence of the model (CutMix [24]).

- Bayesian methods
  They use Bayesian Inference, in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Two well-known approaches of this type are Bayes by Backprop [4] and MC Dropout [10].

- Non-bayesian methods
  These include approaches that are not based on Bayesian inference. Reasons for this are, e.g., that Bayesian approaches require significant modifications to the training procedure and are computationally expensive compared to non-Bayesian neural networks. One well-known approach is to estimate uncertainty using deep ensembles [17]. In this context, the quantification of Uncertainty comes from sampling from several models, where each model is trained separately on the same data or subsets. The Consensus of the models gives confidence in overall model ensemble.

3 Discussion

In this section, I will address the question of why more research should be done in the direction of joint approaches for Explainable AI and Uncertainty Quantification and will also discuss the desired properties of such an approach. Then two short exemplary possible combinations of methods from Explainable AI and Quantification of Uncertainties will be shown. Finally, I will discuss the concept of trust and highlight why both research areas can jointly contribute to improving trustability.

3.1 Combination of Explainable AI and Uncertainty Quantification

When one speaks of Interpretability, one automatically thinks of Explainable AI methods. The research field of Uncertainty Quantification is almost not present here. In my opinion, however, current research needs to be more concerned with merging the two research directions, as in a sense they show two sides of the same coin. While the methods of Explainable AI try to show the way to the decision, the methods of Quantification of Uncertainty try to give a realistic evaluation regarding the reliability of the decision. Thus, through their respective statements, "I know why I made a decision and can show it to you" and "I know what I don't know or how certain I am about my decision", both research directions contribute to paint a more complete picture and thus should end up in their combination being more than the sum of their parts.

Sections 2.1 and 2.2 have shown different methods that differ in their complexity and in their integrability, but can form a basis for combinations. On a meta-level, the result of such a combination should be a model that combines both Explainable AI and Uncertainty Quantification, so that one gets back not only an explanation but also an uncertainty in the decision including an indication of where the uncertainty comes from. This must be done in a way that is appropriate for humans: The nature of the explanations must be both understandable to humans and delivered in an appropriate manner. A few possibilities for this are:

- Concrete explanations: The user is directly provided with the explanation in the form of decision $\rightarrow$ certainty $\rightarrow$ explanation. An example of this in an emotion classification task would be the feedback "The person is classified as sad to 30% due to a covering of the mouth."

- Counterfactual Explanations: The user is shown under which conditions the system would have decided differently and with what degree of certainty. Applied to the example just presented, the feedback would be "If the person's mouth

was not covered, I would classify the person as 80% happy". User-studies [3, 8] have shown, for example, that people prefer counterfactual explanations over case-based reasoning, i.e., solving new problems based on the solutions of similar past problems.

The type of representation also has to be understandable to humans. It is very input-specific and must be appropriate and expectable for the user depending on the problem. Some forms of representation that could be considered are:

- Textual
  The system provides an explanation in the form of a text in which the uncertainty of the decision is also verbalised. For example, a combination of approaches for generating interpretable latent representations, presented in section 2.1 and "Bayes by Backprop", presented in section 2.2, would be conceivable.

- Visual
  The system provides an explanation in the form of a visualisation with an appropriate representation of uncertainties. One possibility here would be, for example, to link the method LRP, presented in section 2.1, with an elliptic coding of uncertainties obtained by Monte Carlo dropout procedure, presented in section 2.2.

3.2 The Role of Trust

Early in the introduction, I talked about trust and its psychological role and necessity to humans. In other disciplines, especially in human-machine interaction, trust has played a major role from the very beginning. Parts of these definitions from other research branches can also be transferred to the requirements of Responsible AI.

In robotics, Lewis et al. [19] distinguish two properties of systems: System predictability and system intelligibility and transparency. The former is important because "[...] knowing that the automation may fail reduces the uncertainty and consequent risk associated with use of the automation. In other words, predictability may be as (or more) important as reliability." This in combination with "Systems that can explain their reasoning will be more likely to be trusted, since they would be more easily understood by their users [...]" describe in my opinion exactly this combination of Explainable AI ("Systems that can explain their reasoning [...]") and Uncertainty Quantification ("[...] predictability may be as (or more) important as reliability").

A system that can tell the user that it is uncertain in combination with a reasoning for it corresponds to my propositions from section 3.1 and should have as a goal an enhancement of trustability in AI systems.

4 Outlook

It is remarkable that both definitions of Explainable AI in section 2.1 address trust, even if it only seems more like trust is a pleasant side effect of explainability. At this point, I wonder if this has led to less attention being paid to linking it to quantification of uncertainty, on the one hand, and whether this classification is appropriate, on the other hand. It would also be conceivable to rename the "Interpretability" pillar in the Responsible AI framework to "Trust" and to downgrade Explainable AI to a method for increasing the trustability of an AI system. Even if the actual idea behind the point "Interpretability" is perhaps to check whether AI systems function as intended, this goal can also be subordinated to the term "Trust" or trust-building measure. Perhaps it will also come about in the near future that one will focus even more on the human component and human acceptance in this case and carry out such a renaming.

Acknowledgements

The author wants to thank Teena Hassan for the helpful discussions.

References

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges (2020)

[2] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)

[3] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, p. 1–14. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3173574.3173951. URL `https://doi.org/10.1145/3173574.3173951`

[4] Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: F. Bach, D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 1613–1622. PMLR, Lille, France (2015). URL `http://proceedings.mlr.press/v37/blundell15.html`

[5] Blythe, J.: An overview of planning under uncertainty. In: Artificial intelligence today, pp. 85–110. Springer (1999)

[6] Böhm, V., Lanusse, F., Seljak, U.: Uncertainty quantification with generative models. arXiv preprint arXiv:1910.10046 (2019)

[7] Dimitrakakis, C., Ortner, R.: Decision making under uncertainty and reinforcement learning (2018)

[8] Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285 (2019)

[9] FICO: Explainable Machine Learning Challenge (2018 (accessed December 28, 2020)). URL `https://community.fico.com/s/explainable-machine-learning-challenge`

[10] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of The 33rd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (2016). URL `http://proceedings.mlr.press/v48/gal16.html`

[11] Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one (2020)

[12] Gunning, D.: Darpa's explainable artificial intelligence (xai) program. pp. ii–ii (2019). DOI 10.1145/3301275.3308446

[13] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)

[14] Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods (2020)

[15] Khaleghi, B.: The How of Explainable AI: Pre-modelling Explainability (2019 (accessed December 28, 2020)). URL https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4

[16] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, pp. 2668–2677. PMLR (2018)

[17] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems, pp. 6402–6413 (2017)

[18] Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**, e0130140 (2015). DOI 10.1371/journal.pone.0130140

[19] Lewis, M., Sycara, K., Walker, P.: The role of trust in human-robot interaction. In: Foundations of trusted autonomy, pp. 135–159. Springer, Cham (2018)

[20] Marzouk, Y.M., Willcox, K.E.: Uncertainty Quantification, vol. II, chap. 34, pp. 131–134. Princeton University Press (2015)

[21] Mushtaq, F., Bland, A.R., Schaefer, A.: Uncertainty and cognitive control. Frontiers in psychology **2**, 249 (2011)

[22] Schmitz, G.P.J., Aldrich, C., Gouws, F.S.: Ann-dt: an algorithm for extraction of decision trees from artificial neural networks. IEEE Transactions on Neural Networks **10**(6), 1392–1401 (1999). DOI 10.1109/72.809084

[23] Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems **32**, 13888–13899 (2019)

[24] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6023–6032 (2019)

C

PATENTS

FULL REFERENCE

- 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. 13, 2021. [Online]. Available: USPTO PatFT Databases.

BIBLIOGRAPHY

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

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. "Openface: an open source facial behavior analysis toolkit." In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–10.

[3] 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.

[4] C. Blais, D. Fiset, H. Furumoto-Deshaies, M. Kunz, D. Seuss, and S. Cormier. "Facial features underlying the decoding of pain expressions." In: *The Journal of Pain* 20.6 (2019), pp. 728–738.

[5] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih. "Machine assessment of neonatal facial expressions of acute pain." In: *Decision Support Systems* 43.4 (2007), pp. 1242–1254.

[6] S. Brahnam, C.-F. Chuang, F. Y. Shih, and M. R. Slack. "Machine recognition and representation of neonatal facial displays of acute pain." In: *Artificial intelligence in medicine* 36.3 (2006), pp. 211–222.

[7] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien. "FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 17–25.

[8] J. Chen, Z. Chi, and H. Fu. "A new approach for pain event detection in video." In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, pp. 250–254.

[9] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. "Feature-point tracking by optical flow discriminates subtle differences in facial expression." In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE. 1998, pp. 396–401.

[10] D. Cristinacce, T. F. Cootes, and I. M. Scott. "A multi-stage approach to facial feature detection." In: *Bmvc*. Vol. 1. 2004, pp. 277–286.

[11] L. J. Cronbach. "Coefficient alpha and the internal structure of tests." In: *psychometrika* 16.3 (1951), pp. 297–334.

[12] 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.

[13] 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.

[14] 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.

[15] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. "Collecting Large, Richly Annotated Facial-Expression Databases from Movies." In: *IEEE MultiMedia* 19 (July 2012), p. 0034. DOI: 10.1109/MMUL.2012.26.

[16] 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.

[17] S. Du, Y. Tao, and A. M. Martinez. "Compound facial expressions of emotion." In: *Proceedings of the National Academy of Sciences* 111.15 (2014), E1454–E1462.

[18] J. O. Egede and M. Valstar. "Cumulative attributes for pain intensity estimation." In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 2017, pp. 146–153.

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

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

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

[22] H. A. Elfenbein and N. Ambady. "On the universality and cultural specificity of emotion recognition: a meta-analysis." In: *Psychological bulletin* 128.2 (2002), p. 203.

[23] P. C. Ellsworth and K. R. Scherer. *Appraisal processes in emotion.* Oxford University Press, 2003.

[24] T. Fawcett. "An introduction to ROC analysis." In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[25] X. Feng, M. Pietikäinen, and A. Hadid. "Facial expression recognition based on local binary patterns." In: *Pattern Recognition and Image Analysis* 17.4 (2007), pp. 592–598.

[26] C. Florea, L. Florea, R. Butnaru, A. Bandrabur, and C. Vertan. "Pain intensity estimation by a self-taught selection of histograms of topographical features." In: *Image and Vision Computing* 56 (2016), pp. 13–27.

[27] 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.

[28] K. Gentsch, D. Grandjean, and K. R. Scherer. "Appraisals generate specific configurations of facial muscle movements in a gambling task: Evidence for the component process model of emotion." In: *PloS one* 10.8 (2015), e0135837.

[29] K. Gentsch, D. Grandjean, and K. R. Scherer. "Temporal dynamics and potential neural sources of goal conduciveness, control, and power appraisal." In: *Biological Psychology* 112 (2015), pp. 77 –93. ISSN: 0301-0511. DOI: https://doi.org/10.1016/j.biopsycho.2015.10.001. URL: http://www.sciencedirect.com/science/article/pii/S0301051115300624.

[30] A. Ghasemi, X. Wei, P. Lucey, S. Sridharan, and C. Fookes. "Social signal processing for pain monitoring using a hidden conditional random field." In: *2014 IEEE Workshop on Statistical Signal Processing (SSP)*. IEEE. 2014, pp. 61–64.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks." In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[32] M Gregoire, M. Coll, M. Tremblay, K. Prkachin, and P. Jackson. "Repeated exposure to others pain reduces vicarious pain intensity estimation." In: *European Journal of Pain* 20.10 (2016), pp. 1644–1652.

[33] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. "Multi-pie." In: *Image and Vision Computing* 28.5 (2010), pp. 807–813.

[34] 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.

[35] Z. Guo, L. Zhang, and D. Zhang. "A completed modeling of local binary pattern operator for texture classification." In: *IEEE transactions on image processing* 19.6 (2010), pp. 1657–1663.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Y. Huang, F. Chen, S. Lv, and X. Wang. "Facial expression recognition: A survey." In: *Symmetry* 11.10 (2019), p. 1189.

[41] S. I. Inc. *FaceGen Modeller (Software)*. 2009. URL: http://www.facegen.com/ (visited on 01/03/2021).

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

[43] 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.

[44] 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.

[45] 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.

[46] T. Karras, S. Laine, and T. Aila. "A style-based generator architecture for generative adversarial networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[47] K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven. "The MPI facial expression database - a validated database of emotional and conversational facial expressions." In: *PloS one* 7.3 (2012), e32321.

[48] 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.

[49] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro. "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition." In: *IEEE Transactions on Affective Computing* 10.2 (2017), pp. 223–236.

[50] C. Küblbeck and A. Ernst. "Face detection and tracking in video sequences using the modified census transformation." In: *Image and Vision Computing* 24.6 (2006), pp. 564–572.

[51] 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.

[52] 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.

[53] S. Li and W. Deng. "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning." In: *International Journal of Computer Vision* 127.6-7 (2019), pp. 884–906.

[54] S. Li and W. Deng. "Deep facial expression recognition: A survey." In: *IEEE Transactions on Affective Computing* (2020).

[55] S. Li, W. Deng, and J. Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[56] C. F. Liew and T. Yairi. "Facial expression recognition and analysis: a comparison study of feature descriptors." In: *IPSJ transactions on computer vision and applications* 7 (2015), pp. 104–120.

[57] 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.

[58] 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.

[59] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. "Deeply learning deformable facial action parts model for dynamic expression analysis." In: *Asian conference on computer vision*. Springer. 2014, pp. 143–157.

[60] P. Liu, S. Han, Z. Meng, and Y. Tong. "Facial expression recognition via a boosted deep belief network." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1805–1812.

[61] Y. Liu, J. Wang, and P. Li. "A feature point tracking method based on the combination of SIFT algorithm and KLT matching algorithm." In: *Journal of Astronautics* 32.7 (2011), pp. 1618–1625.

[62] D. Lopez-Martinez, O. Rudovic, and R. Picard. "Physiological and behavioral profiling for nociceptive pain estimation using personalized multitask learning." In: *arXiv preprint arXiv:1711.04036* (2017).

[63] 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 Pattern Recognition*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[64] 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.

[65] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. "Automatically detecting pain in video through facial action units." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41.3 (2010), pp. 664–674.

[66] 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*. IEEE. 2011, pp. 57–64.

[67] Y. Lv, Z. Feng, and C. Xu. "Facial expression recognition via deep learning." In: *2014 International Conference on Smart Computing*. IEEE. 2014, pp. 303–308.

[68] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. "Coding facial expressions with gabor wavelets." In: *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE. 1998, pp. 200–205.

[69] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. "Building a Large Annotated Corpus of English: The Penn Treebank." In: *Computational Linguistics* 19.2 (1993), pp. 313–330. URL: https://www.aclweb.org/anthology/J93-2004.

[70] 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.

[71] D. Matsumoto. *The handbook of culture and psychology*. Oxford University Press, 2001.

[72] B. J. Matuszewski, W. Quan, and L.-K. Shark. "High-resolution comprehensive 3-D dynamic database for facial articulation analysis." In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE. 2011, pp. 2128–2135.

[73] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. "Disfa: A spontaneous facial action intensity database." In: *IEEE Transactions on Affective Computing* 4.2 (2013), pp. 151–160.

[74] H. Meng and N. Bianchi-Berthouze. "Affective state level recognition in naturalistic facial and vocal expressions." In: *IEEE Transactions on Cybernetics* 44.3 (2013), pp. 315–328.

[75] P. Michel and R. El Kaliouby. "Real time facial expression recognition in video using support vector machines." In: *Proceedings of the 5th international conference on Multimodal interfaces*. 2003, pp. 258–264.

[76] A. Mollahosseini, D. Chan, and M. H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." In: *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE. 2016, pp. 1–10.

[77] 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.

[78] N. Neshov and A. Manolova. "Pain detection from facial characteristics using supervised descent method." In: *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*. Vol. 1. IEEE. 2015, pp. 251–256.

[79] 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.

[80] J. Pahl, I. Rieger, and D. Seuss. "Multi-Label Class Balancing Algorithm for Action Unit Detection." In: (2020). arXiv: 2002.03238 [cs.CV].

[81] J. Pahl, I. Rieger, and D. Seuss. "Multi-label Learning with Missing Values using Combined Facial Action Unit Datasets." In: (2020). arXiv: 2008.07234 [cs.CV].

[82] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. "Web-based database for facial expression analysis." In: *2005 IEEE international conference on multimedia and Expo*. IEEE. 2005, 5–pp.

[83] I. Rieger, J. Pahl, and D. Seuss. "Unique Class Group Based Multi-Label Balancing Optimizer for Action Unit Detection." In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 2020, pp. 619–623. DOI: 10.1109/FG47880.2020.00101.

[84] 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.

[85] E. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. Scherer. "FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units." In: *Journal of Nonverbal Behavior* 35 (Mar. 2011), pp. 1–16. DOI: 10.1007/s10919-010-0095-9.

[86] S. D. Roy, M. K. Bhowmik, P. Saha, and A. K. Ghosh. "An approach for automatic pain detection through facial expression." In: *Procedia Computer Science* 84 (2016), pp. 99–106.

[87] M. Rydfalk. *CANDIDE, a parameterized face*. Tech. rep. LiTH-ISY-I-866. Sweden: Department of Electrical Engineering, Linköping University, 1987.

[88] 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.

[89] K. R. Scherer. "Appraisal considered as a process of multilevel sequential checking." In: *Appraisal processes in emotion: Theory, methods, research* 92.120 (2001), p. 57.

[90] 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.

[91] K. R. Scherer, A. Dieckmann, M. Unfried, H. Ellgring, and M. Mortillaro. "Investigating appraisal-driven facial expression and inference in emotion communication." In: *Emotion* (2019).

[92] K. R. Scherer and H. Ellgring. "Multimodal expression of emotion: Affect programs or componential appraisal patterns?" In: *Emotion* 7.1 (2007), p. 158.

[93] 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.

[94] K. Scherer. "Emotions are emergent processes: They require a dynamic computational architecture." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364 (Dec. 2009), pp. 3459–74. DOI: 10.1098/rstb.2009.0141.

[95] 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.

[96] 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.

[97] D. Seuss. *Bridging the Gap Between Explainable AI and Uncertainty Quantification to Enhance Trustability*. 2021. arXiv: 2105.11828 [cs.AI].

[98] D. Seuss, T. Hassan, A. Dieckmann, M. Unfried, K. R. R. Scherer, M. Mortillaro, and J.-U. Garbas. "Automatic Estimation of Action Unit Intensities and Inference of Emotional Appraisals." In: *IEEE Transactions on Affective Computing* (2021), pp. 1–1. DOI: 10.1109/TAFFC.2021.3077590.

[99] M. Siebers, J. Folger, S. Schineller, D. Seuß, S. Faerber, and U. Schmid. "Modelling Adaptation Effects as Similarity to Dynamic Prototypes." In: *Proceedings of the 11th Conference of the German Cognitive Science Society*. 2012.

[100] 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.

[101] K. Sikka. "Facial expression analysis for estimating pain in clinical settings." In: *Proceedings of the 16th International Conference on Multimodal Interaction*. 2014, pp. 349–353.

[102] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang. "Automated assessment of childrens postoperative pain using computer vision." In: *Pediatrics* 136.1 (2015), e124–e131.

[103] E. Smith and S. Queller. "Mental representations." In: *Blackwell handbook in social psychology* 1 (Jan. 2001), pp. 111–133.

[104] A. S. M. Sohail and P. Bhattacharya. "Classification of facial expressions using k-nearest neighbor classifier." In: *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*. Springer. 2007, pp. 555–566.

[105] J. Thies, M. Zollhoefer, M. Niessner, L. Valgaerts, M. Stamminger, and C. Theobalt. "Real-time expression transfer for facial reenactment." In: *ACM Trans. Graph.* 34.6 (2015), pp. 183–1.

[106] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. "Meta-analysis of the first facial expression recognition challenge." In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 966–979.

[107] M. M. Van Zaanen. "Bootstrapping structure into language: alignment-based learning." In: *arXiv preprint cs/0205025* (2002).

[108] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. 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)*. IEEE. 2013, pp. 128–131.

[109] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille. "Regularizing face verification nets for pain intensity regression." In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 1087–1091.

[110] X.-H. Wang, A. Liu, and S.-Q. Zhang. "New facial expression recognition based on FSVM and KNN." In: *Optik* 126.21 (2015), pp. 3132–3134.

[111] Y. Wang, H. Ai, B. Wu, and C. Huang. "Real time facial expression recognition with adaboost." In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. IEEE. 2004, pp. 926–929.

[112] 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.

[113] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. "Toward practical smile detection." In: *IEEE transactions on pattern analysis and machine intelligence* 31.11 (2009), pp. 2106–2111.

[114] 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.

[115] Y. Yacoob and L. S. Davis. "Recognizing human facial expressions from long image sequences using optical flow." In: *IEEE Transactions on pattern analysis and machine intelligence* 18.6 (1996), pp. 636–642.

[116] P. Yang, Q. Liu, and D. N. Metaxas. "Boosting encoded dynamic features for facial expression recognition." In: *Pattern Recognition Letters* 30.2 (2009), pp. 132–139.

[117] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. "A 3D facial expression database for facial behavior research." In: *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE. 2006, pp. 211–216.

[118] J. Yu and B. Bhanu. "Evolutionary feature synthesis for facial expression recognition." In: *Pattern Recognition Letters* 27.11 (2006), pp. 1289–1298.

[119] M. van Zaanen. "Implementing Alignment-Based Learning." In: *Grammatical Inference: Algorithms and Applications*. Ed. by P. Adriaans, H. Fernau, and M. van Zaanen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 312–314. ISBN: 978-3-540-45790-9.

[120] M. van Zaanen and J. Geertzen. "Problems with Evaluation of Unsupervised Empirical Grammatical Inference Systems." In: *Grammatical Inference: Algorithms and Applications*. Ed. by A. Clark, F. Coste, and L. Miclet. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 301–303. ISBN: 978-3-540-88009-7.

[121] 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.

[122] X. Zhang, L. Yin, and J. F. Cohn. "Three dimensional binary edge feature representation for pain expression analysis." In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 1. IEEE. 2015, pp. 1–7.

[123] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database." In: *Image and Vision Computing* 32.10 (2014), pp. 692–706.

[124] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. "Facial expression recognition from near-infrared videos." In: *Image and Vision Computing* 29.9 (2011), pp. 607–619.

[125] J. Zhou, X. Hong, F. Su, and G. Zhao. "Recurrent convolutional neural network regression for continuous pain intensity estimation in video." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 84–92.

[126] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[127] M. van Zaanen. "Theoretical and Practical Experiences with Alignment-Based Learning." English. In: *Proceedings of the Australasian Language Technology Workshop*. Ed. by A. Knott and D. Estival. The Australian Language Technology Association, 2003, pp. 25–32.